Recent Advances in Social Data and Artificial Intelligence 2019

Edited by
Hari Mohan Srivastava, Gautam Srivastava and Vijay Mago

Printed Edition of the Special Issue Published in *Symmetry*

www.mdpi.com/journal/symmetry

MDPI

Recent Advances in Social Data and Artificial Intelligence 2019

Recent Advances in Social Data and Artificial Intelligence 2019

Editors

Hari Mohan Srivastava
Gautam Srivastava
Vijay Mago

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editors

Hari Mohan Srivastava
University of Victoria
Canada

Gautam Srivastava
Brandon University
Canada

Vijay Mago
Lakehead University
Canada

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Symmetry* (ISSN 2073-8994) (available at: https://www.mdpi.com/journal/symmetry/special_issues/Social_Data_Artificial_Intelligence).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-4021-4 (Hbk)
ISBN 978-3-0365-4022-1 (PDF)

Contents

About the Editors

Hari Mohan Srivastava

Hari Mohan Srivastava is Professor Emeritus in the Department of Mathematics and Statistics at the University of Victoria in Canada. He currently holds several advisory, honorary, visiting and chair professorships at universities and research institutes around the world. Prof. Hari Mohan Srivastava's current research interests include (for example) Real and Complex Analysis, Fractional Calculus and Its Applications, Integral Equations and Integral Transformations, Higher Transcendental Functions and Their Applications, q-Series and q-Polynomials, Analytic and Geometric Inequalities, Probability and Statistics, and Inventory Modelling and Optimization. Further biographical and professional details about Professor Hari Mohan Srivastava are available at the following link: https://www.math.uvic.ca/~harimsri/.

Gautam Srivastava

Gautam Srivastava is a senior-level Associate Professor in the Department of Mathematics and Computer Science at Brandon University in Canada. Professor Gautam Srivastava's current research interests include (for example) Blockchain Technology, Cryptography, Big Dada, Data Mining, Social Networks, Security and Privacy, Anonymity, and Graphs. Further biographical and professional details about Professor Gautam Srivastava are available at the following link: https://people.brandonu.ca/srivastavag/.

Vijay Mago

Vijay Mago is a senior-level Associate Professor in the Department of Computer Science at Lakehead University in Canada. Professor Vijay Mago's current research interests include (for example) Social Data Analytics, Big Data, Health Informatics, and Mathematical and Computational Modelling. Further biographical and professional details about Professor Vijay Mago are available at the following link: https://www.lakeheadu.ca/users/M/vmago/node/25295.

Preface to "Recent Advances in Social Data and Artificial Intelligence 2019"

This volume consists of a collection of a total of 27 accepted submissions (including several invited feature articles) to the Special Issue of the MDPI's journal, Symmetry, on the general subject-area of "Social Data and Artificial Intelligence" from all over the world.

The importance and usefulness of subjects and topics involving social data and artificial intelligence are becoming widely recognized. In this Special Issue, we cordially invited and welcome review, expository, and original research articles dealing with the recent advances in the subjects of social data and artificial intelligence, and potentially their links to Cyberspace (that is, the seamless integration of physical, social, and mental spaces), which is an integral part of our society, ranging from learning and entertainment to business and cultural activities, and so on. However, there are a number of pressing challenges associated with cyberspace. For example, how do we strike a balance between the need for strong cybersecurity and preserving the privacy of ordinary citizens?

This Special Issue has emerged from the International Conference on Social Data and Artificial Intelligence (SDAI 2020) held in Toronto, Canada on 26–27 May 2020 and the IEEE Cyber Science and Technology Congress (CyberSciTech 2020) which will also be held in Canada (CyberSciTech 2020, Calgary, Canada, 22–26 June 2020).

To address the challenges described for both conferences, there is a need to establish new science and research portfolios that incorporate social data and artificial intelligence alone or in combination with cyber-physical, cyber-social, cyber-intelligent, and cyber-life technologies in a cohesive and efficient manner.

In this Special Issue, we invited and welcome review, expository and original research articles dealing with the recent state-of-the-art advances on the topics of integral transformations and operational calculus as well as their multidisciplinary applications, together with some relevance to the aspect of symmetry.

The suggested topics of interest for the call of papers for this Special Issue included, but were not limited to, the following keywords: Social data inadequacies and inconsistencies; Predictive models of social behaviors; Infrastructure and architecture for testing social theories; Data collection and analysis platforms; Relevance of IoT for social science theories; Building capacity to continuously collect data across a range of social media networks; Designing efficient parsers to deal with noisy social media data-sets for real-time tracking of health issues, diseases, and wellness; Designing tools to map and measure the effectiveness of health campaigns by healthcare organizations; Cross-validating the predictive models of social media data-sets with ground truth data; Developing frameworks and algorithms to perform real-time analysis of social media data-sets; Cyberspace theory and technology; Cyber social computing and networks; Cyber life and wellbeing; Cyber intelligence and cognitive science

Finally, it gives us great pleasure in thanking all of the participating authors, and the referees and the peer-reviewers, for their invaluable contributions toward the remarkable success of each of the above-mentioned Special Issues. We do also express our appreciation for the editorial and managerial help and assistance provided efficiently and generously by Mr. Philip Li and other colleagues and associates in the Editorial Office of Symmetry. The dedicated and wholehearted support and help of one and all are indeed greatly appreciated.

Hari Mohan Srivastava, Gautam Srivastava, and Vijay Mago

Editors

symmetry

MDPI

Article

Color Revolution: A Novel Operator for Imperialist Competitive Algorithm in Solving Cloud Computing Service Composition Problem

Amin Jula [1,*], Elankovan A. Sundararajan [2], Zalinda Othman [1] and Narjes Khatoon Naseri [2]

[1] Centre for Artificial Intelligent (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; zalinda@ukm.edu.my
[2] Centre for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; elan@ukm.edu.my (E.A.S.); narges.naseri85@gmail.com (N.K.N.)
* Correspondence: amin.jula@gmail.com

Abstract: In this paper, a novel high-performance and low-cost operator is proposed for the imperialist competitive algorithm (ICA). The operator, inspired by a sociopolitical movement called the color revolution that has recently arisen in some countries, is referred to as the color revolution operator (CRO). The improved ICA with CRO, denoted as ICACRO, is significantly more efficient than the ICA. On the other hand, cloud computing service composition is a high-dimensional optimization problem that has become more prominent in recent years due to the unprecedented increase in both the number of services in the service pool and the number of service providers. In this study, two different types of ICACRO, one that applies the CRO to all countries of the world (ICACRO-C) and one that applies the CRO solely to imperialist countries (ICACRO-I), were used for service time-cost optimization in cloud computing service composition. The ICACRO was evaluated using a large-scale dataset and five service time-cost optimization problems with different difficulty levels. Compared to the basic ICA and niching PSO, the experimental and statistical tests demonstrate that the ability of the ICACRO to approach an optimal solution is considerably higher and that the ICACRO can be considered an efficient and scalable approach. Furthermore, the ICACRO-C is stronger than the ICACRO-I in terms of the solution quality with respect to execution time. However, the differences are negligible when solving large-scale problems.

Keywords: cloud computing; color revolution operator; imperialist competitive algorithm; quality of service; service composition; service time-cost

Citation: Jula, A.; Sundararajan, E.A.; Othman, Z.; Naseri, N.K. Color Revolution: A Novel Operator for Imperialist Competitive Algorithm in Solving Cloud Computing Service Composition Problem. *Symmetry* **2021**, *13*, 177. https://doi.org/10.3390/sym13020177

Academic Editors: Hari M. Srivastava and José Carlos R. Alcantud
Received: 26 November 2020
Accepted: 18 January 2021
Published: 22 January 2021

1. Introduction

Providing enhanced processing facilities and appropriate on-demand self-service systems has always been the major concern of web service suppliers [1]. Recently, these concerns were largely resolved by the introduction and development of cloud computing [2,3]. Its appealing financial and technical characteristics [4] and the increasing complexity of the requested services have led to a growing trend among service providers and customers toward cloud computing.

Although clouds and their applications are growing rapidly, service providers are not able to prepare all the complex combinations of required services. Hence, clouds suppliers need a mechanism for composing the required composite services using the unique services already provided in the service pool (see Figure 1).

The dramatic increase in cloud computing customers has made the environment a lucrative commercial space that encourages facility owners to provide services. Hence, a considerable number of service providers are currently offering a large set of different unique services in the pool, allowing one to find a large number of instances of the same

service, each with different functional and quality of service (QoS) specifications [5–8]. In the current situation, cloud suppliers are confronted with a difficult optimization problem [4,9–12]: providing the optimal compositions of unique services that will satisfy composite service customers and persuade them to make their next requests. This is the problem that network function virtualization (NFV) [13] systems are also confronting [14,15]. Due to the undeniable importance of the cloud computing service composition (CCSC) problem, which is an NP-hard problem [9,16–18], an increasing amount of research on the development of cloud computing has been conducted, which will be studied in Section 1.1. Despite the research conducted and progress made thus far, considerable effort is still needed to close the gaps between the solutions obtained to date and the optimal CCSC solutions.

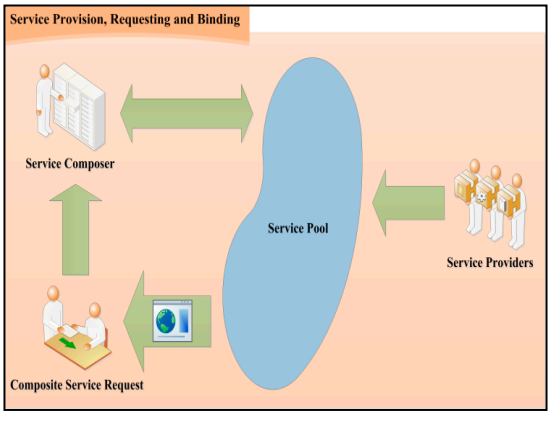


Figure 1. Service provision, requesting and binding in cloud computing.

These gaps, which have mostly arisen because of the very large problem search space, necessitate the customization of more efficient evolutionary algorithms (EAs) to solve the CCSC problem with higher accuracy and optimality. In addition, service time and service cost have been identified as the two most important QoS parameters in cloud service composition that should be further considered in optimizing the QoS [4]. Hence, in this paper, the researchers are strongly motivated to focus on service time and cost optimization for CCSC by improving the imperialist competitive algorithm (ICA) [18–20].

In this study, a new operator is designed based on the sociopolitical movements known as the color revolutions, which have largely been observed in newly independent states of the former Soviet Union and the Balkans and applied to the ICA. The operator is denoted as the color revolution operator (CRO) and attempts to replace the worst part of a country (solution) with a better option within an acceptable amount of time. The algorithm, obtained by applying the *CRO* to the ICA, is called the imperialist competitive algorithm with color revolution operator (ICACRO), two forms of which are used in the service time-cost optimization for CCSC.

To provide a more explicit description of the subject, a complete explanation of the research motivations is presented. The strategy employed by many EAs in finding the optimal solutions to a given optimization problem is to perform a random search in the specified search space and impose a variety of information to guide the search process such that it is carried out more efficiently. Accordingly, the applied algorithm is expected to reach the optimal solution of a given problem within a reasonable amount of time provided that the problem does not include a very large search space and a large number of dimensions.

When facing a problem with a vast search space and numerous dimensions, the number of space points that are actually potential solutions increases to the extent that, in practice, very few of them can be investigated using the algorithm.

Example 1. *In this paper, WSDream-QoSDataset2 [21] is used as a real-world dataset to evaluate the ability of the proposed algorithms to solve the service time-cost optimization problem in CCSC (STCOCCSC) (please refer to part 2.1 for the problem description). A total of 339 service providers exist in the dataset, each of which provides 5825 different simple services. On the one hand, suppose that 100 different simple services are required to represent a composite cloud service. Each of the required simple services can be provided by any of the 339 service providers. In this STCOCCSC problem, the algorithm faces a 100-dimensional problem in which each dimension involves 339 options. Hence, there are* 339^{100} *distinct potential solutions in the search space. On the other hand, assume that 1000 members exist in the first generation of solutions of the applied evolutionary algorithm and that the algorithm is executed 6000 times. Accordingly, in the most optimistic situation, 6,000,000 different potential solutions are examined by the algorithm provided that repeated investigated solutions are neglected. Hence,* $5.73e-245\%$ *of all available points in the search space are examined during the execution time. In other words, many potential solutions are neglected by the algorithm.*

Example 2. *A visual description of the issue is illustrated in Figure 2, in which Nh is the radius of the neighborhood circle of an investigated potential solution. As mentioned above, many potential solutions located in the neighborhood circle are neglected by the algorithm. Although ICA designers have made commendable efforts to prepare a more successful search using irregular but purposive movement of the solutions in the search space, the large search space of some problems such as the STCOCCSC greatly reduces the efficiency of the algorithm. In this study, a new operator is designed and proposed for the ICA to improve its efficiency in investigating large search spaces. The remainder of the paper is organized as follows. A motivation section is provided as the last part of the introduction. Related works are briefly studied in Section 1.1. Descriptions of the STCOCCSC and the ICA are presented in Section 2. Complete descriptions of the proposed operator (CRO) and algorithm (ICACRO) are provided in Section 3. The experimental design and test results obtained using two different types of the ICACRO as well as two other algorithms, i.e., the ICA and niching particle swarm optimization (PSO), including numerical and statistical investigations, are discussed in detail in Sections 4 and 5, respectively. Finally, our conclusions and potential topics for future research are presented in Section 6.*

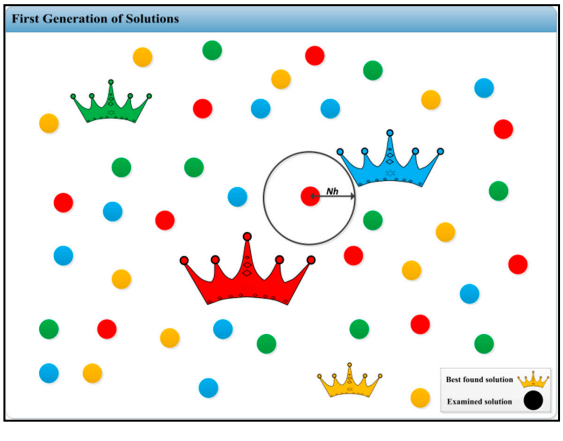


Figure 2. Countries of the world and their neighborhoods.

1.1. Literature Review

Different goals have been pursued by different researchers in solving the CCSC problem, but a factor common to most of these studies is the consideration of QoS parameters, particularly service time and service cost. The objectives pursued in the literature, the QoS parameters investigated, and their importance percentages were detailed by Jula et al. in [4]. The previous studies conducted to solve the CCSC as an optimization problem can be divided into two main categories.

The first category consists of studies that used different classic algorithms and designed customized structures or workflows. Koflet et al. mapped the problem onto the knapsack problem and solved it by applying a parallel form of the branch-and-bound method [22]. Backtracking and dynamic programming are other classic algorithms used to solve the CCSC problem [23,24]. Mixed integer programming (MIP) and linear programming (LP) were employed by Shangguang et al. and Zhu et al. [25–27] and Hossain et al. [28], respectively, to find appropriate unique services. CCSC is also often mapped onto graph structures to employ graph-based algorithms.

A weighted directed graph with an improved fast-EP algorithm was used in [29], and other examples of applying graph structures in the field include the directed acyclic graph (DAG) [30], discovery graph with Markov chain [31], and the graphical construction of the analytic hierarchy process (AHP) [32]. Petri nets are another non-heuristic approach that has been used to solve the CCSC problem [33].

Artificially intelligent, combinatorial, and evolutionary algorithms [34–39] constitute the second category of studies conducted to solve the CCSC problem. Although these algorithms have been used less frequently than the algorithms included in the first category, the importance of achieving closer-to-optimal solutions and shorter execution times and increases in the size of CCSC problems have forced researchers to prioritize the application of EAs [40,41]. Different types of artificial neural networks (ANNs) and fuzzy logic have been used in several studies [33,42–44]. A genetic algorithm (GA) including a roulette wheel selection was applied by Ye et al. in 2011 [45]. Two improved GAs were also proposed by Klein et al. and Ludwig [46,47].

The most successful proposed approaches are based on PSO [48,49], the chaos optimization algorithm (COA) [50], game theory [51], and hybrid algorithms. In addition, unlike almost all conducted studies focusing on applying different search approaches to find the best combination of single services, PROCLUS, as a high-dimensional clustering algorithm, was utilized in [52] to categorize service providers. The categorized search space enhanced the ability of the ICA to select the best possible services by optimizing the service time in CCSC. To accomplish a comprehensive review of the literature of the domain, including the proposed methods, used datasets and tools, and different research objectives, a systematic literature review was presented in [4].

2. Problem and Algorithm Description

2.1. Service Time-Cost Optimization in Cloud Computing Service Composition (STCOCCSC)

With the development of computer-based system procedures, process execution and the required services have become more complex. Due to the growing complexity and variety of systems, a simple independent service is incapable of satisfying functional prerequisites of various real-world requests. Hence, it is necessary to prepare a set of simple atomic services that can effectively work together to perform a complex service. Therefore, a cloud service composer system (CSC) must be incorporated into cloud computing. Increasing the number of service providers will increase the number of similar unique services. Because the similar services are found in different parts of the network and have absolute QoS values, the CSC should efficiently choose a unique service for each requirement among the many similar services provided by different service providers. Application of the most suitable method yields the highest QoS based on customers' requirements and priorities.

Due to fundamental changes in cloud environments, accessible services, and service customer needs, automatic functional capabilities are mandatory in designing the CSC. Therefore, choosing appropriate unique services for merging to provide an optimal combination that satisfies the functional and QoS requirements of a customer can be considered one of the foremost critical problems of service composition. This problem is called Cloud Computing Service Composition (*CCSC*) that is defined in detail and discussed in [4].

This paper assumes that within a cloud, each composite service (CS) consists of n unique services (*USs*), each of which has service time (*ST*) and service cost (*SC*) as its QoS parameters. A combination of *USs* must correspondingly act in an ordinal workflow (*wf*) to provide a required CS. However, if wf_k is the workflow of CS_k, then $ST(wf_k)$ and $SC(wf_k)$ will be defined as the *ST* and *SC* of the workflow k, where the *ST* and *SC* vectors of the workflow can be expressed as (1) and (2), respectively.

$$ST(wf_k) = (ST_1(wf_k), ST_2(wf_k), \ldots, ST_n(wf_k)) \quad (1)$$

$$SC(wf_k) = (SC_1(wf_k), SC_2(wf_k), \ldots, SC_n(wf_k)) \quad (2)$$

The merit value (*MV*) of wf_k is the sum of the total *STs* and total *SCs* of all elements of wf_k and is calculated using (3), where TW and CW are the weights of time and cost, which should be determined by the user from [0,1] such that $TW + CW = 1$. Hence, the optimal solution to *STCOCCSC* is the solution with the minimum *MV* value. Note that the *ST* and cost values should be normalized between 1 and 10 via min-max normalization [53] to be used in (3). Section 4 shows that the proposed operator uses the *MV* of solutions. Hence, different structures of the *STCOCCSC* have no effect on the application and efficiency of the operator and are therefore neglected here.

$$MV(wf_k) = \sum_{i=1}^{n} (TW \times ST_i(wf_k) + CW \times SC_i(wf_k)) \quad (3)$$

2.2. Imperialist Competitive Algorithm (ICA)

In evolutionary computation, the recently proposed algorithm, ICA, was formed based on social and political activities [20,54,55], in contrast to other EAs, which are based on the physical events or animals' natural behaviors. The ICA starts with an initial population created at random, in which members of the population are regarded as countries. A few of the most powerful countries are considered imperialist, whereas the rest represent colonies of the imperialists. Assuming that there are n dimensions in the given optimization problem, a country is treated as an $1 \times n$ array as in (4):

$$Country = [p_1, p_2, \ldots, p_n] \quad (4)$$

The power of country i is calculated using the objective function f, which is a function of the variables $(p_1, p_2, \ldots, p_n)$, yielding the following equation:

$$Power(Country_i) = f(Country_i) = f(p_{i1}, p_{i2}, \ldots, p_{in}) \quad (5)$$

The ICA commences with m countries, and n_{imp} of the most powerful ones are labeled as the imperialists. The other countries are called colonies each of which belongs to an empire. A simple procedure is used to disperse the non-imperialists among the imperialists. An imperialist is randomly selected for each non-imperialist country, which is dedicated to that empire. During the execution of the algorithm, every imperialist attracts their colonies according to the total power of the empire and the colonies. As Equation (6) states, the total power of each empire is the sum of the corresponding imperialist's power and average of power of all the colonies of the empire multiplied to a coefficient.

$$TP_n = c_n + (\alpha \times Average\{power(coloniesofempire_n)\}) \quad (6)$$

where TP_n is the total power of the nth empire, c_n is the power of the imperialist country and α is a positive decimal number between 0 and 1.

In the course of movement stage, a colony moves a x units toward its imperialist. As illustrated in Figure 3, the direction of the movement can be represented by a vector from the colony to its associated imperialist, where d is the distance between the colony and the imperialist, and x is a value taken at random that can be obtained following a uniform distribution like what is shown in (7):

$$x \approx Uniform(0, \beta \times d) \tag{7}$$

where β is greater than one and close to 2.

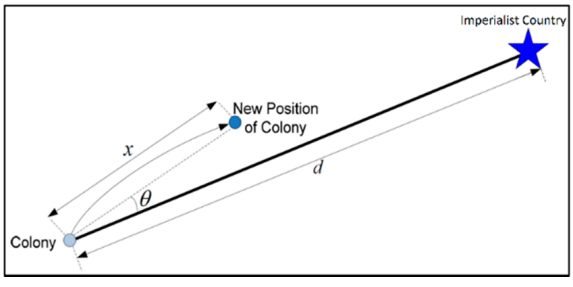


Figure 3. Country movement [20].

In ICA, the imperialistic competition plays an important role in the convergence of the algorithm. This operator decreases the power of the weakest empire by taking some of its colonies and assigning them to other empires. The process will be continued until the weakest empire is destroyed. Imperialistic competition can be applied in various forms based on the requirements of the problem. To facilitate the description of the proposed algorithm, a mapping description of the key terms of the *STCOCCSC*, ICA and optimization is presented in Table 1.

Table 1. Mapping description of key terms of *service time-cost optimization problem in CCSC (STCOCCSC)* and imperialist competitive algorithm (ICA).

STCOCCSC	ICA	Optimization
Composite Service (CS)	Country	Solution
Merit Value (MV)	Power of Country	Objective Function

3. Imperialist Competitive Algorithm with the Color Revolution Operator (ICACRO)

Manipulation of the routine process may enhance the ability to navigate the search space using heuristic algorithms if it is performed intelligently. Changes must be made such that the optimal solution can be reached without changing the nature of the algorithm. For this purpose, the CRO is applied to the ICA to obtain closer-to-optimal solutions for the *STCOCCSC*. As described in Section 4.1 and shown in Figure 4, by making a heuristic change in a selected part of a solution structure that does not yield an evolution of the solution structure, the CRO increases the likelihood of achieving more suitable solutions within a smaller execution time. The CRO can be imposed on either all existing countries of the ICA world or only selected ones. The functional details and use conditions of the CRO are discussed in the following sections.

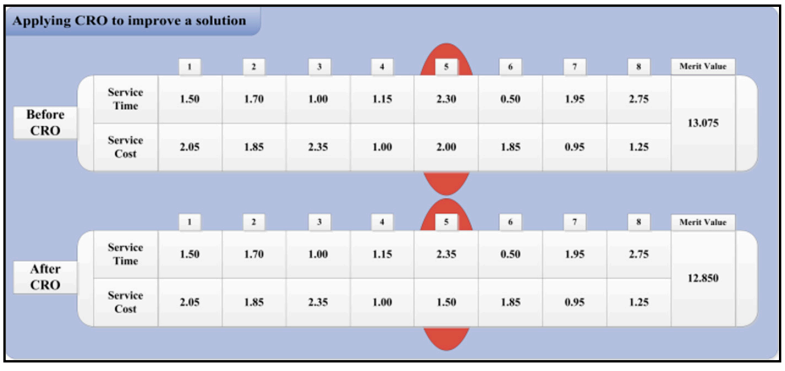


Figure 4. A simple example of applying the color revolution operator (CRO).

3.1. Color Revolution Operator (CRO)

New operators designed for an algorithm should ideally be similar to the algorithm in terms of nature and structure. Hence, the operator designed to improve the ICA is similar to the ICA and rooted in the sociopolitical attitudes of society. The main concept of the CRO is inspired by the color revolutions that occurred in some newly independent states of the former Soviet Union and Balkans during the early 21st century.

The color revolutions are movements in which activists protesting against the government or ruling party under a common symbol carry out civil disobedience and nonviolent resistance to exert strong pressure to enact change. The desired changes occur in the laws governing the country, with the structures remaining intact, increasing hope for social and political reform in a society. The quality of life is expected to be enhanced after a color revolution; however, the targeted improvements are uncertain [56,57].

Based on the characteristics of color revolutions described above, the CRO must impose changes in the selected countries without changing either the structure of the countries or the process through which the algorithm is executed. The operator should try to replace one of the less optimal features of the target solution with a more suitable feature with the ultimate goal of improving the solution. To reach this goal, in this study, the CRO examines the USs that compose the existing CS to identify the unique service with the greatest time-cost value (TCV), where the TCV is the sum of normalized ST and SC values for the service. To do this, the CRO should traverse the solution to find its maximum TCV.

If the number of service providers is m, then the process of finding the maximum value of the list using a linear time operation has a time complexity of $O(m)$. In the next step, the CRO randomly seeks a better alternative among service providers to replace a service with one that provides a shorter TCV. Replacing the service that requires the greatest TCV with an equivalent service that terminates for a smaller TCV will decrease the MV of the CS. An important factor of the performance and time complexity of the CRO is the number of service providers considered by the operator when searching for a smaller TCV. How many attempts should the CRO make to find an alternative service provider? This question has no single answer.

The service provider that provides a smaller TCV and is sought by the CRO could be located anywhere on the list of providers considered; it might be the first one considered or may not even be included in the list because there is no better provider. Therefore, it is best to choose the number of attempts randomly. To provide a reasonable likelihood of finding a better provider while preventing an excessive search time, the random number should be neither too small nor too large. Hence, one should determine a limited range from which the number of attempts should be chosen. In this paper, $\left[\log_2 m, \log_2^2 m\right]$ is applied to obtain the number of attempts needed to find a better service provider, where m indicates the number of service providers in the list. As a simple example, if the

number of service providers is 1024, then the number of attempts will be a random number between 10 and 100. Clearly, $\left(\log_2 m + \log_2^2 m\right)/2$ attempts will result in better choices being found. After the replacing process, the CRO can calculate the new MV of the solution simply and quickly by finding the difference between the old TCV and the new one and subtracting it from the current MV. Accordingly, the CRO time complexity, as shown in (8), will be $O\left(\log_2^2 m\right)$.

Figure 4 provides an example of applying the CRO to improve a solution that includes 8 USs, in which the worst selected service (located in the 5th field) is replaced with a new one, yielding a time-cost reduction in the solution. In this example, the weights of time and cost are 0.5. The pseudocode for the CRO is presented in Figure 5. Another factor to consider is the number of countries that would be affected by the color revolution. This factor has a considerable impact on the overall performance of the algorithm and will be discussed in part 6.

$$T(CRO) = O(\log_2 m) + O\left(\log_2^2 m\right) \Rightarrow T(CRO) = O\left(\log_2^2 m\right) \quad (8)$$

Algorithm 1. Function CRO
Input: Country ID i
Output: Updated Country ID i

```
BEGIN
m = Number of available service providers;
ReqServiceNo = Number of required unique services;
WorstServiceID = 1;
WorstProviderID = i.AssignedProvider[ 1 ];
SelectedProviderID = 1;
WorstServiceTimeCost = i.ServiceTimeCost[WorstServiceID];
for all assigned services j ∈ [2, ReqServiceNo] to country i do
Begin
    if ( i.ServiceTimeCost[ j ] > WorstServiceTimeCost ) then
        WorstServiceID = j;
        WorstServiceTimeCost = i.ServiceTimeCost[ j ] ;
        WorstProviderID = i.AssignedProvider[ j ];
    end if;
end;
Rnd = A random number in [log2 m, log2^2 m];
for k =1 to rnd do
Begin
    TmpServPro = A random number in [1, m];
    if ( Provider [TmpServPro, WorstServiceID].TimeCost <
        Provider[WorstProviderID, WorstServiceID].TimeCost )
    then
        i.AssignedProvider[WorstServiceID] = TmpServPro;
        SelectedProviderID = TmpServPro;
        WorstProviderID = SelectedProviderID;
    end if;
MeritDecreaseAmount =
    WorstServiceTimeCost –
    Provider[SelectedProviderID, WorstServiceID].TimeCost;
i.MeritValue = i.MeritValue – (MeritDecreaseAmount);
return i;
END;
```

Figure 5. Pseudocode of applied *CRO* in imperialist competitive algorithm with color revolution operator (ICACRO).

3.2. ICA with CRO (ICACRO)

Designing an efficient and simple structure for solutions is important for enhancing the implementation performance. Hence, the structure of a country in the ICACRO consists of two separate parts. The first part is an array of d elements, representing the index of service providers that are selected to provide the *USs* required, where d is the number of required *USs* that should be composed to provide the required *CS* and is thus identified as the size of the problem. The second part of the country structure is a decimal variable for storing the total *ST*, which is the *MV* of the solution.

The execution of the algorithm begins with the generation of a set of countries called the 'world'. The world consists of *CountryNo* countries, in which each country represents a solution to the intended problem. Each country is generated by randomly selecting a service provider for each of the required *USs*. The *MV* of each country is calculated immediately following its generation. The *MV* of country i is equal to the sum of the *STs* and *SCs* of all its constituent services and is calculated using (9).

$$Country\_TST(i) = -\frac{1}{2}\sum_{j \in Country_i} (ST(\,j\,) + SC(\,j\,)) \tag{9}$$

where *Country_TST(i)* is the *MV* of country i and *ST(j)* and *SC(j)* are, respectively, the *ST* and *SC* of the required service j obtained from the selected service provider; the *TW* and *CW* values are assumed to both equal 0.5. Because the ICA was originally designed for maximization and the objective of the *STCOCCSC* is to achieve the minimum total service time, the negative sign reflects a smaller *MV* and thus a larger quantity. The countries are sorted in ascending order at the end of the generation process, which helps the algorithm by placing the best solution at the top of the world list. In this phase, the first *imperialistNo* countries of the world list are selected as imperialist countries. The remaining countries are considered colonies. The number of colonies can be obtained using (10).

$$ColonyNo = CountryNo - imperialistNo \tag{10}$$

The policy applied by the ICACRO to distribute the colonies among the imperialist countries is an equal division. Hence, the number of colonies or all imperialist countries is equal in the first iteration of the algorithm and can be easily obtained using (11).

$$ColonialNo = \frac{ColonyNo}{imperialistNo} \tag{11}$$

Colony displacement is the next phase of the ICACRO, in which each colony finds and moves to a new location closer to its imperialist country. Displacements of various sizes are made in distinct dimensions such that the displacement in each dimension is the distance between the colony and imperialist country in the dimension multiplied by a random coefficient. The distance between colony i and its imperialist country (imp_k) and the displacement of colony i in dimension d can be calculated using (12) and (13), respectively.

$$dist^d(imp_k, colony_i) = imp_k.AssignedServer(d) - colony_i.AssignedServer(d) \tag{12}$$

$$disp^d(colony_i) = rnd \times dist^d(imp_k, colony_i) \tag{13}$$

where *AssignedServer(d)* is the dedicated service provider for the required service d and *rnd* is a random decimal in $(0, 1]$.

The obtained distance is a positive number provided that the index of the imperialist's assigned server is greater than the index of the server dedicated to the colony and vice versa. Similarly, the size of the displacement may also be positive or negative. Hence, the index of the next assigned service provider can be less or greater than the index of the current service provider. The index of the next assigned service provider of colony i for dimension d can be obtained using (14).

$$colony_i.AssignedServer^{t+1}(d) = colony_i.AssignedServer^{t}(d) + disp^{d}(colony_i) \quad (14)$$

The MV of each country is calculated after every displacement. The newly calculated MV of the colony may be greater than that of its imperialist country, in which case the power of the colony is greater than that of the imperialist country and their positions should be exchanged. After the position exchange, the previous colony and imperialist country are now an imperialist country and colony, respectively.

After the displacement and position exchange of the countries, the *CRO* is introduced to enhance the performance of the ICA. Excessive usage of the *CRO* may affect the convergence of the ICA and disturb its evolutionary process, although it is not permitted to make basic changes according to the color revolution concept. Excessive usage may also lead to an unacceptable increase in the execution time of the ICACRO. Hence, one must apply a *CRO* rate by which the effects of the *CRO* are controlled. For this purpose, a *CRO* rate of 0.05 is used in the ICACRO, i.e., 5 out of 100 iterations of the algorithm will execute the *CRO*. The MV of the colonies should be recalculated after the *CRO* is executed for every colony. Accordingly, the colony and its imperialist country should exchange their positions if the newly obtained MV of the colony is better than that of the imperialist country.

In the last phase of the ICACRO, the imperialist countries enter an imperialistic competition. In this competition, each imperialist country attempts to overcome the other imperialist countries by taking the weakest colony of the weakest empire. The victorious imperialist country separates the colony from its empire and adds it to its own colony set. An imperialist country with no colonies in its empire will itself be taken by the victorious country. To ensure that there is an adequate opportunity for each imperialist country to increase its number of colonies and thus strengthen its empire, the imperialist competition function is called at a rate of 0.05, i.e., the imperialist competition is executed in 5 out of every 100 iterations, as described for the *CRO*.

The ICACRO is designed such that except during the first initialization, it does not require all the countries to be sorted to find the optimal solution. Therefore, it is sufficient to sort the imperialist countries in descending order at the end of each iteration of the algorithm. Because there are fewer imperialist countries than non-imperialist countries, a clear reduction in the ICACRO execution time compared to that of the ICA occurs. The ICACRO can be terminated after a specified number of iterations, the achievement of a specific MV, or the disappearance of all but one empire in the world. If the termination criterion is not satisfied, the next iteration will start via displacement. The ICACRO flowchart and pseudocode are shown in Figures 6 and 7, respectively.

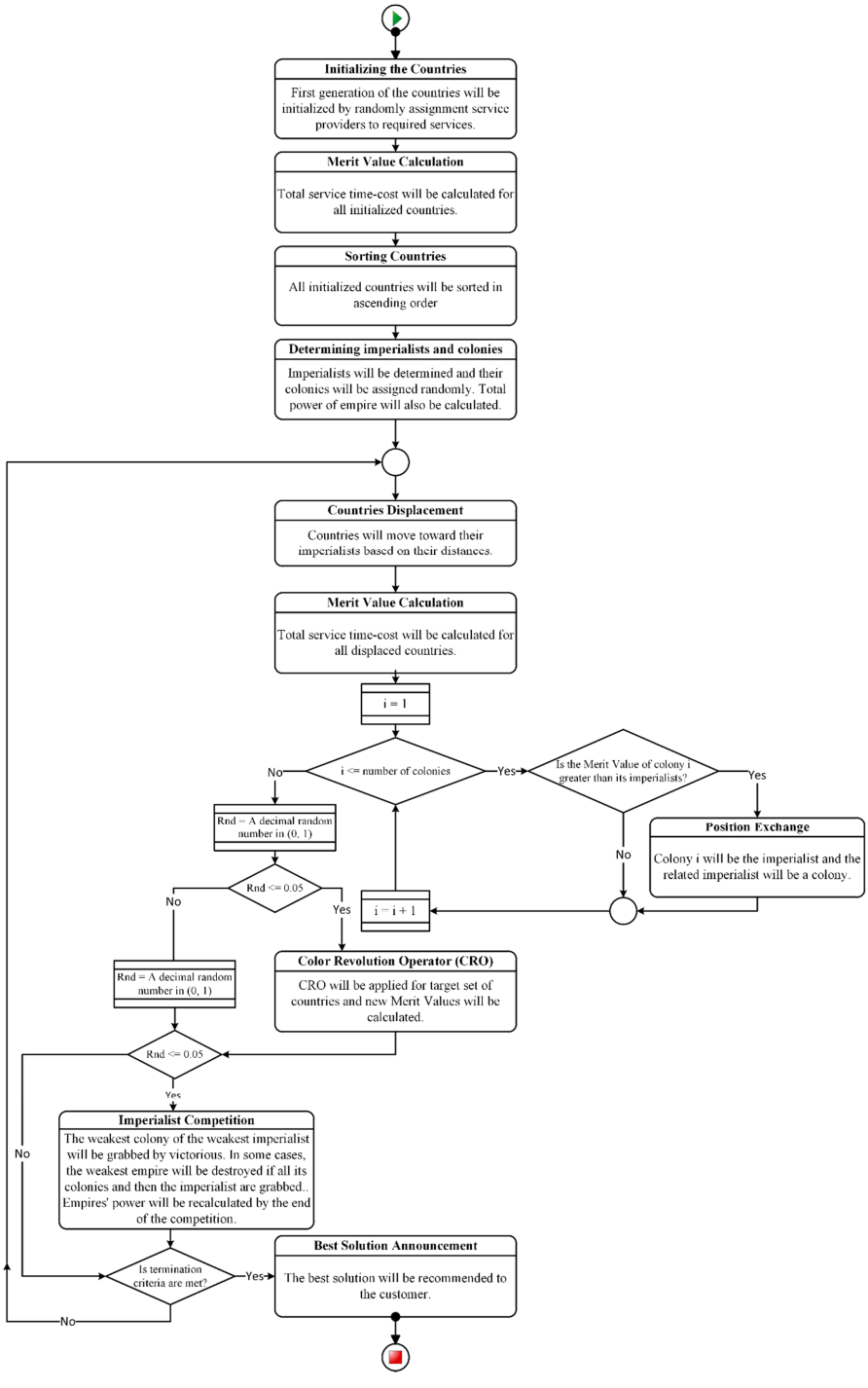


Figure 6. ICACRO flowchart.

Algorithm 2. Function ICACRO
Input: Termination criteria
Output: The best solution

```
1   BEGIN
2   Initialize the first generation of countries ;
3   Calculate total service time-cost (Merit Value) of all countries ;
4   Sort the countries ascendingly according to their Merit Value ;
5   Determine the Imperialists and assign their colonies randomly ;
6   Calculate total power of every empire ;
7   While the termination criteria are not met do
    begin
8       Move the countries toward their imperialist ;
9       Calculate Merit Value of the moved countries ;
10      for every colony i do
        begin
10          if Merit Value of i > Merit Value of imperialist of i
            then
11              i is the new imperialist of the empire and
                current imperialist will be a colony for i ;
            end if;
        end;
12      Rnd = a decimal random number in (0, 1) ;
13      if Rnd <= 0.5 then
14          Apply CRO operator for the target set ;
15          Calculate Merit Value for the CRO target set ;
        end if;
16      Rnd = a decimal random number in (0, 1) ;
17      if Rnd <= 0.5 then
18          Apply Imperialist Competition ;
19          Calculate total power of all the empires ;
        end if;
20  Sort the countries ascendingly according to their Merit Value ;
    end;
21  Return the best imperialist as the best solution ;
22  END;
```

Figure 7. Pseudocode of ICACRO.

4. Experimental Design

4.1. ICACRO Implementation and Execution

To apply the ICACRO to the *STCOCCSC*, the algorithm is implemented in Microsoft Visual Studio C#.NET 2012. Five *STCOCCSCs* of different sizes, denoted as problems A, B, C, D and E, are generated randomly based on the WSDream-QoSDataset2 [21,58], which is a large real-world QoS dataset that includes *STs* collected from 339 service providers

and 5825 *USs*. Problems A, B, C, D and E consist of 100, 200, 300, 400 and 500 unrepeated required *USs*, respectively. In other words, the 100 required *USs* in A should be combined to prepare a requested *CS*. With respect to the description of the *STCOCCSC*, the composition algorithm should select a service among 339 similar services for each required unique service such that the sum of the service time-costs of all selected services is minimized. The classic ICA and niching PSO [49,59], which is one of the most efficient methods proposed to date, are also implemented in the same environment for comparison with the ICACRO for the five described problems. The ICACRO, the classic ICA, and niching PSO are executed 40 times independently on a PC with an Intel Core i7—3.40 GHz processor and 8 GB of RAM under identical conditions. The average value of the results of each method is employed for comparison. Because WSDream-QoSDataset2 does not provide *SCs*, these values are generated randomly by the authors such that their statistical distribution and value range follow the *ST* distribution and value range, respectively.

For a more accurate evaluation of the ICACRO and more useful comparisons between the ICACRO and the ICA and niching PSO, the algorithms are compared for each problem at two fixed points of the execution process. The first point is at iteration 1500, and the second one is at iteration 6000, which is the intended endpoint of the run. The importance of these two points can be considered from the perspective of the abilities of the algorithms to reach better solutions within a limited number of iterations and their achievements after the termination of the evolutionary process. Because 1500 is 25% of 6000, the execution time of the algorithm after 1500 iterations should be 25% of that after 6000 iterations. In this study, the solutions obtained after 1500 and 6000 iterations are termed the first proper solution (*fps*) and final solution (*fs*), respectively.

4.2. Definitions

In this section, some new analysis factors are introduced and defined, with which a more meticulous comparison can be carried out among the different algorithms.

Definition 1. *The appropriate solution at the lowest time (APLT) policy can be applied by cloud suppliers to select the best obtained solution as the final solution after a limited number of iterations, although there is still a high probability that better solutions could be achieved. This policy is useful for suppliers who prefer to respond to requests as rapidly as possible. To satisfy the APLT policy, the fps should be considered by cloud suppliers.*

Definition 2. *The optimal solution at the appropriate time (OSAT) policy is an appropriate policy for cloud suppliers who prefer to wait a reasonable amount of time to achieve closer-to-optimal solutions. Application of this policy causes slightly more delay in responding to service requests than APLT but yields more QoS-satisfying solutions. Utilization of the fs will satisfy the OSAT policy.*

Definition 3. *To assess the optimal method of applying the CRO to empower the ICA for obtaining better solutions, the time-cost consumption check (TCC) can be defined, using (15), and used for the consumption check. The TCC allows the optimality of all algorithms to be evaluated in comparison to that of one of the algorithms under consideration. For this purpose, the weakest algorithm in the experimental test is chosen as the basis, and the other investigated algorithms are compared to the basis. In (15), Best(A) and Best(basis) are the best MVs obtained using algorithm A and the basis, respectively. A larger difference between Best(A) and Best(basis) indicates a higher quality of A.*

$$TCC(\%) = \left(1 - \frac{Best(A)}{Best(basis)}\right) \times 100 \tag{15}$$

In this study, the *TCC* is employed to compare the results obtained using niching PSO, the ICA, the ICACRO-I, and the ICACRO-C for problems A to E, where niching PSO is selected as the basis.

Definition 4. *The ratio of the MV of the final obtained solution to the execution time of an algorithm is called the time utilization (TU) and can be a suitable and reliable factor for comparing the efficiencies of algorithms in solving the same problem. A larger/smaller TU value indicates the greater efficiency of an algorithm compared to that of others when solving maximization/minimization problems.*

Definition 5. *Let us define TU-based optimality (TUO) as the ratio of the TU value of an algorithm ALG1 to the TU value of another algorithm ALG2. TUO can be used to show the efficiency of ALG1 compared to that of ALG2 by taking their execution times as effective parameters.*

Definition 6. *Merit value variation (MVV) can be studied from the 1500th iteration to 6000th iteration to identify the improvement of the best solution obtained by the algorithms between these two points of execution. The MVV value can easily be calculated by finding the difference between the best MVs obtained at the two points. The MVV values of the ICACRO-I and ICACRO-C are obtained using (16) and (17), respectively. In the same way, execution time variation (ETV) can also be defined as the difference between execution times of an algorithm in the 1500th iteration and 6000th iteration. According to this definition, the ETVs of the two algorithms are calculated using (18) and (19), respectively.*

$$\Delta MC = MVV(ICACRO - C) = Cfps - Cfs \tag{16}$$

$$\Delta MI = MVV(ICACRO - I) = Ifps - Ifs \tag{17}$$

$$\Delta TC = ETV(ICACRO - C) = CfsT - CfpsT \tag{18}$$

$$\Delta TI = ETV(ICACRO - I) = IfsT - IfpsT \tag{19}$$

Definition 7. *The merit value gradient (MVG), which is defined in (20) and (21) for the two algorithms, can be considered a meaningful factor that includes the MVV and ETV of an algorithm and is used to evaluate how well an algorithm can avoid premature convergence by looking for better solutions. For minimization problems, the greater the MVG, the more convincing the evidence for allowing the algorithm to more thoroughly search through a search space. The very low MVG for small problems may indicate that a solution very close to the optimal solution is obtained.*

$$MVG(ICACRO - C) = \frac{\Delta MC}{\Delta TC} \tag{20}$$

$$MVG(ICACRO - I) = \frac{\Delta MI}{\Delta TI} \tag{21}$$

5. Comparison of Results and Discussion

As previously stated, the number of countries and their positions in the ranking are expected to play a decisive role in the efficiency and effectiveness of the CRO. To assess this role, the CRO is applied in the algorithm in two different ways. The first is the ICACRO-C, in which the CRO is utilized for all countries. The second is the ICACRO-I, in which the CRO is applied only to the imperialist countries.

As also discussed earlier, five different-sized problems with different degrees of difficulty are generated and addressed using the ICACRO-C, ICACRO-I, ICA, and niching PSO. For the first three algorithms, the number of countries initialized in the first iteration is 500. Similarly, 500 particles are generated for niching PSO to provide an equivalent comparison for all five algorithms.

Part a of Figure 8 indicates that the solutions obtained using the ICACRO-C and ICACRO-I are significantly better than those generated using the ICA and niching PSO in terms of the APLT policy. Similarly, part *b* of the same figure indicates the superiority of the ICACRO-C and ICACRO-I over the ICA and niching PSO in terms of the OSAT policy. The execution results, obtained using the two different types of the ICACRO, and their trends shown in Figure 8 also reveal that the ICACRO-C achieved a better solution than that of the ICACRO-I for problem A. Hence, based on the first evaluation, the ICACRO-C is the most suitable algorithm for solving problem A due to the better solution achieved. Despite this conclusion, the following paragraphs will show that other factors can affect the decision.

The improvements achieved by applying the *CRO* to the ICA are demonstrated by comparing the quality of the solutions in parts a and b of Figures 9–12. The total service time-cost of the best solutions, obtained using the ICACRO-C and ICACRO-I, are consistently lower than the service time-cost of the solutions generated using the ICA and niching PSO for problems B, C, D and E. The ICACRO-C and ICACRO-I obtain solutions that are significantly closer to optimal considering that these two algorithms, when evaluating the CRO, achieved optimality for different problems in which different levels of difficulty arose because of the different numbers of *USs* required. Since the ICA demonstrated better results than those of niching PSO, and because the ICA and the ICACRO differ only in the application of the CRO, it can be concluded that the high-performance level achieved resulted from the CRO efficiency and its appropriate embedding in the ICA.

On the other hand, Figure 13 presents the optimality calculated for the five investigated problems in five separate categories, each in two parts reflecting the checkpoints after 1500 and 6000 iterations. Based on the TCCs calculated based on niching PSO for problems A to E, the minimum and maximum optimality values obtained by applying the *CRO* are 39.05% and 47.86% for the *fs* and 38.68% and 46.66% for the *fps*, respectively. The averages of the optimality values achieved using the ICACRO-C, ICACRO-I, and ICA are 42.89%, 40.51%, and 21.13% for the *fs* and 46.12%, 44.11%, and 22.88% for the *fps*, respectively. The optimality achieved using the CRO in the ICA compared to that achieved using the classic ICA can be calculated by considering the ICA as the basis. Figure 14 shows that the minimum and maximum optimality values obtained for problems A to E are 20.48% and 32.80% for the *fs* and 22.04% and 31.63% for the *fps*, respectively. Hence, the averages of the optimality values obtained using the ICACRO-C and ICACRO-I are 30.30% and 27.49% for the *fs* and 27.81% and 24.72% for the *fps*, respectively. The optimality values of the ICACRO-C based on the ICACRO-I for problems A to E are presented in Figure 15.

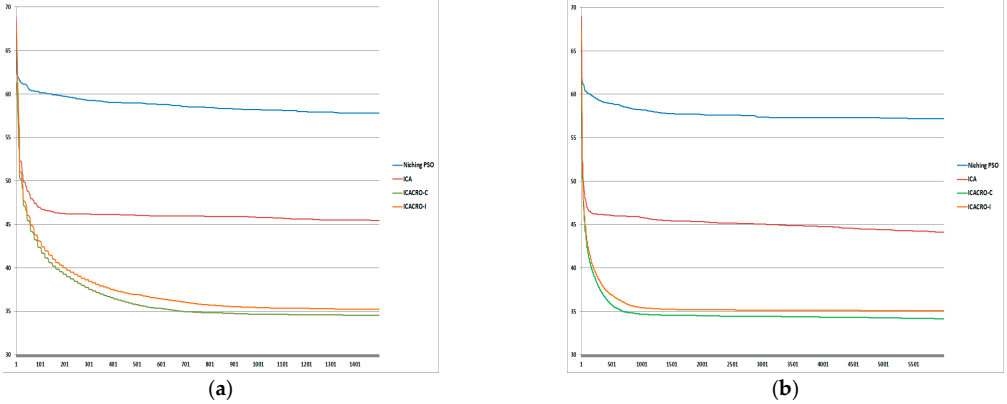
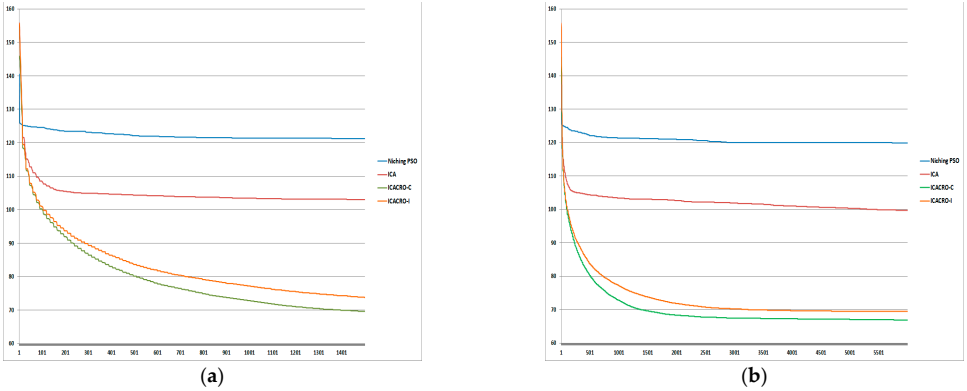
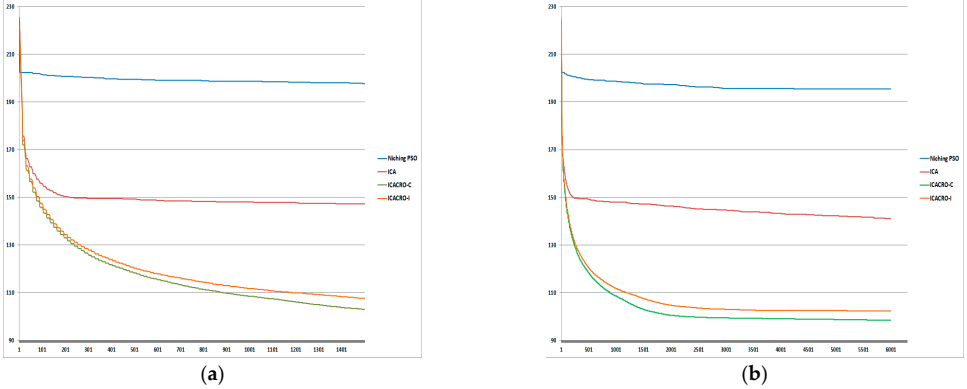


Figure 8. Comparison of the total service time-costs obtained using the four algorithms for problem A: (**a**) after 1500 iterations and (**b**) after 6000 iterations.



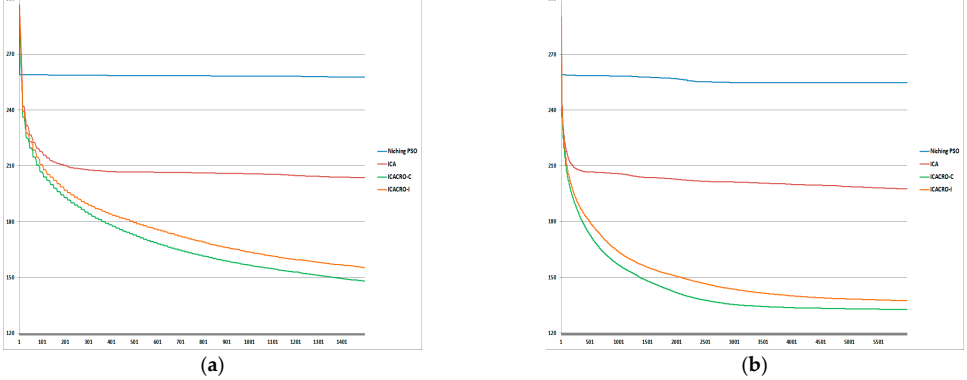
(**a**) (**b**)

Figure 9. Comparison of the total service time-costs obtained using the four algorithms for problem B: (**a**) after 1500 iterations and (**b**) after 6000 iterations.



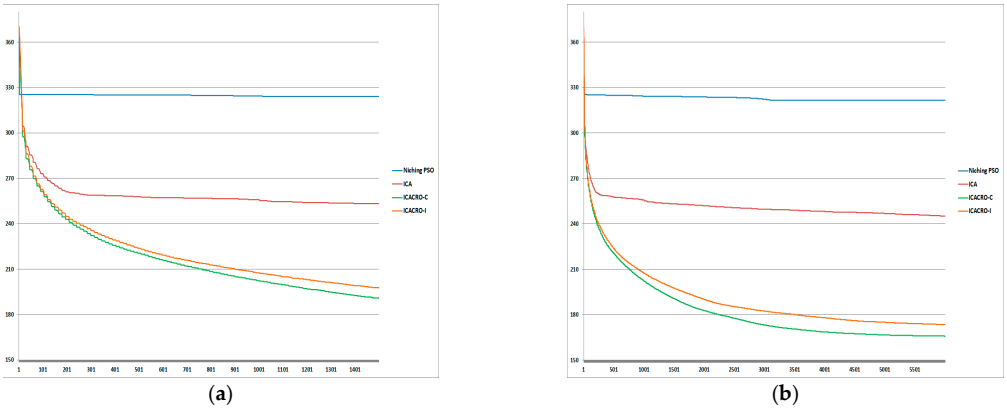
(**a**) (**b**)

Figure 10. Comparison of the total service time-costs obtained using the four algorithms for problem C: (**a**) after 1500 iterations and (**b**) after 6000 iterations.



(**a**) (**b**)

Figure 11. Comparison of the total service time-costs obtained using the four algorithms for problem D: (**a**) after 1500 iterations and (**b**) after 6000 iterations.



(**a**) (**b**)

Figure 12. Comparison of the total service time-costs obtained using the four algorithms for problem E: (**a**) after 1500 iterations and (**b**) after 6000 iterations.

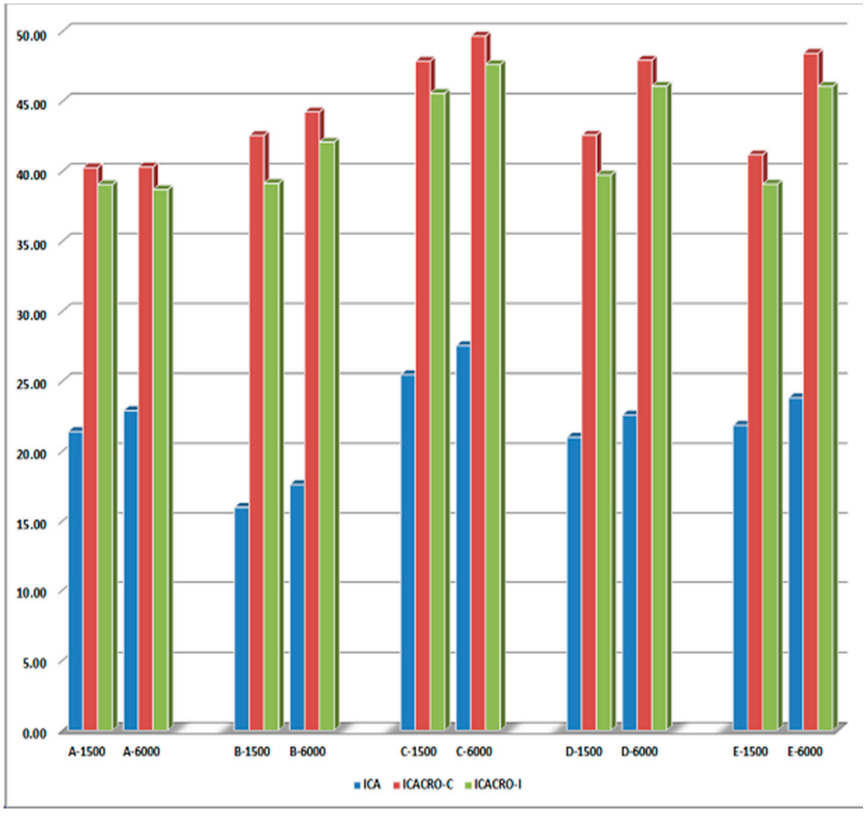


Figure 13. Optimality values of the four algorithms based on niching particle swarm optimization (PSO) for problems A–E.

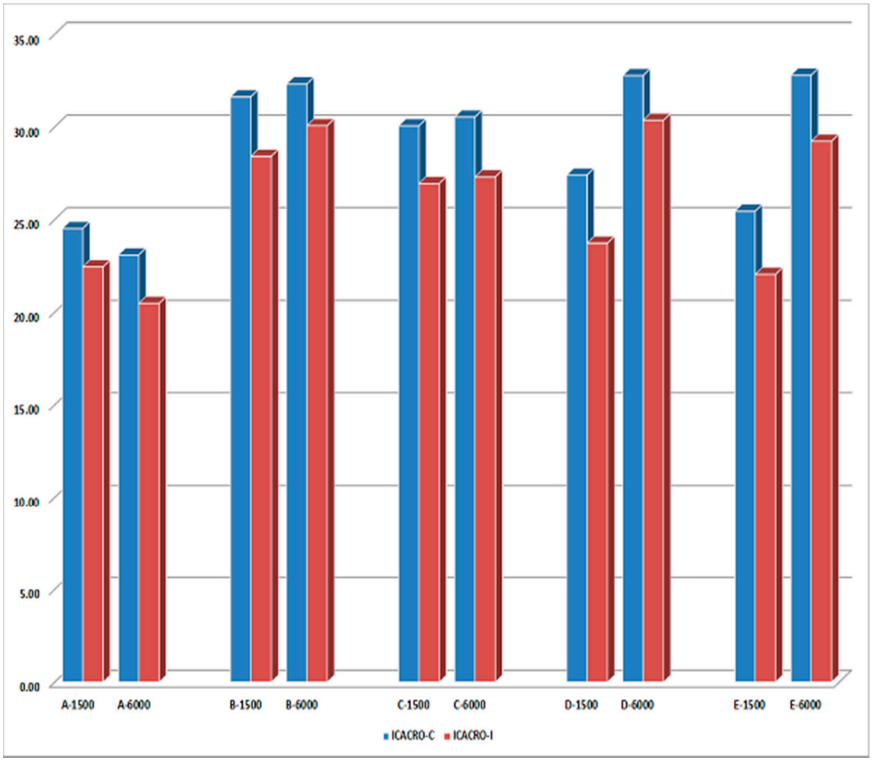


Figure 14. Optimality values of the ICACRO-C, ICACRO-I and ICA based on the ICA for problems A–E.

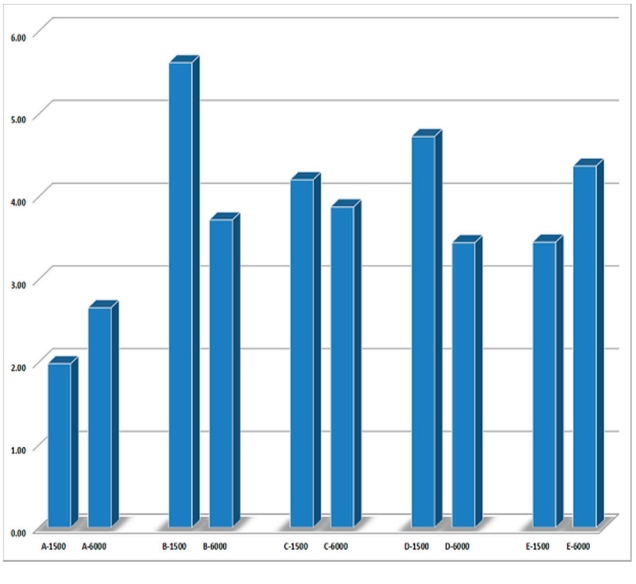


Figure 15. Optimality values of the ICACRO-C based on the ICACRO-I for problems A–E.

5.1. ICACRO-C or ICACRO-I?

The performances of the ICACRO-C and ICACRO-I are examined more closely to identify the differences between the two algorithms (see Table 2). Comparison of the best solutions obtained using the ICACRO-C and ICACRO-I in solving the different problems demonstrates that although the ICACRO-C performs better than the ICACRO-I for all investigated problems and for both APLT and OSAT policies, investigation of the execution times of the algorithms and the effect of the problem size can help us to study the performances of the two algorithms more precisely.

Table 2. *fs* and *fst* results for ICACRO-I and ICACRO-C.

	Problem A	**Problem B**	**Problem C**	**Problem D**	**Problem E**
ICACRO-C Best result after 1500 iterations (Cfps)	34.53	69.63	103.02	147.96	190.64
ICACRO-C Best result after 6000 iterations (Cfs)	34.15	66.94	98.49	132.70	165.87
ICACRO-C 1500-iteration execution time (CfpsT)	6.5 s	13.4 s	20.2 s	26.8 s	34.4 s
ICACRO-C 6000-iteration execution time (CfsT)	25.7 s	53.5 s	80.4 s	107 s	136.6 s
ICACRO-I Best result after 1500 iterations (Ifps)	35.23	73.78	107.54	155.29	197.45
ICACRO-I Best result after 6000 iterations (Ifs)	35.08	69.50	102.39	137.41	173.46
ICACRO-I 1500-iteration execution time (IfpsT)	5.9 s	12.2 s	18.7 s	24.5 s	31.1 s
ICACRO-I 6000-iteration execution time (IfsT)	23.5 s	49.1 s	74.6 s	97.5 s	123.8 s
Cfps—Cfs	0.38	2.69	4.53	15.26	24.77
Ifps—Ifs	0.15	4.28	5.15	17.88	23.99
CfsT—CfpsT	19.2 s	40.1 s	60.2 s	80.2 s	102.2 s
IfsT—IfpsT	17.6 s	36.9 s	55.9 s	73 s	92.7 s

To study the execution times of the ICACRO-C and ICACRO-I, denoted as *CfsT* and *IfsT*, respectively, the times, shown in Table 2, are recorded after execution of the algorithms for problems A to E. Figure 16 demonstrates that *CfsT* and *IfsT* increased in linear fashion according to the problem size. It also shows that the larger the problem size, the larger the difference between *CfsT* and *IfsT*. With the execution times and information regarding their trends, it is possible to compare the ICACRO-C and ICACRO-I results more precisely.

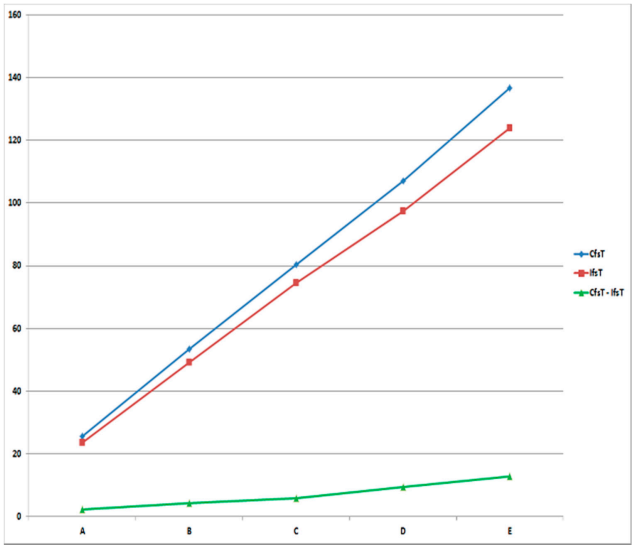


Figure 16. Execution times of the ICACRO-C and ICACRO-I for solving problems A–E.

In accordance with definitions 4 and 5, Figure 17 shows that the ICACRO-C achieved smaller TU values for all 5 problems than those achieved by the ICACRO-I. Hence, it can be concluded that with respect to the required execution time, the ICACRO-C outperforms the ICACRO-I and can achieve more proper solutions when solving the same *STCOCCSCs*. The TUO values of the five problems are also shown to be almost equal. Therefore, the same level of optimality can be achieved when solving problems of various sizes using the ICACRO-C instead of the ICACRO-I.

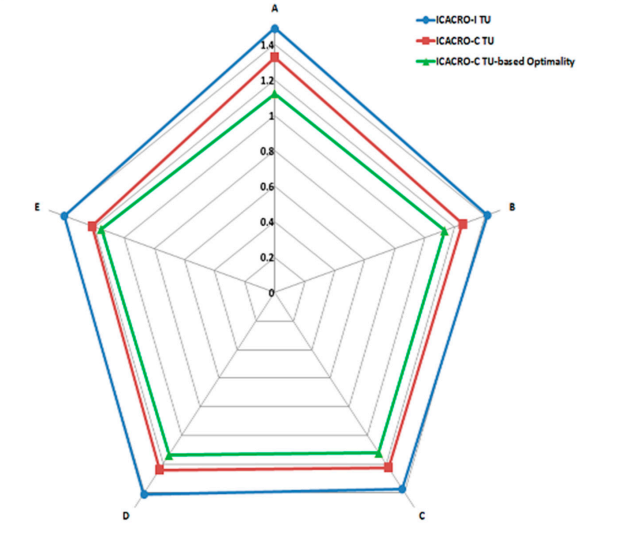


Figure 17. TU values of ICACRO-C and ICACRO-I and TU-based optimality of ICACRO-C for problems A–E.

5.2. Fps or Fs?

As mentioned in definitions 1 and 2, cloud suppliers can prepare a solution based on the APLT or OSAT policies. Table 2 shows the execution times and MVs of the ICACRO-I and ICACRO-C for problems A-E for both policies.

In accordance with definitions 6 and 7, Figure 18 demonstrates that although the MVG is very close to zero for simple problem A, it experiences an increasing trend for both investigated algorithms. Also, the ICACRO-I is shown to perform slightly better than the ICACRO-C based on the MVG investigation.

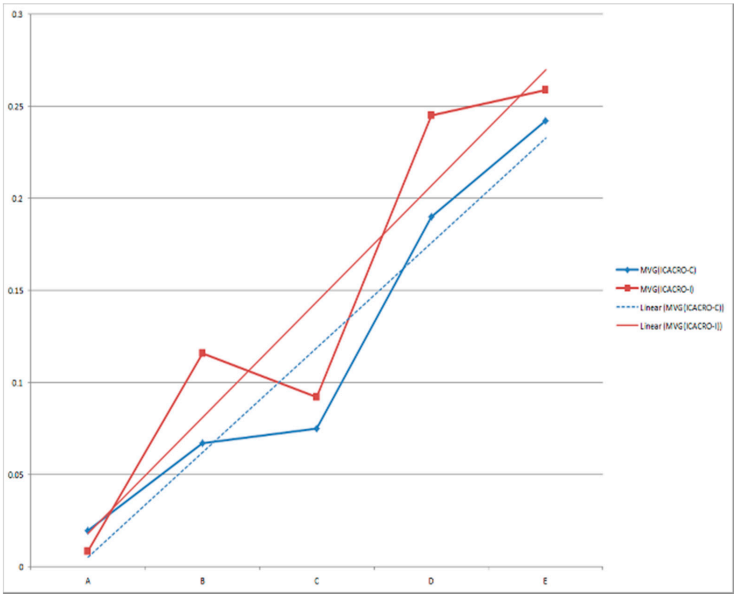


Figure 18. Merit value gradient (MVG) and trend of MVG of ICACRO-C and ICACRO-I for problems A–E.

Therefore, it can be concluded that for problems with a small number of required simple services, obtaining the fps seems to be sufficient and that execution of the algorithms to reach the fs will not lead to a significant improvement in the best obtained solution to the problem. Nevertheless, as the number of required simple services increases, the difference becomes more significant.

5.3. Performance Statistical Test

Statistically evaluating and comparing the results obtained using the previous algorithms can provide more information regarding the algorithm functionality and performance. To this end, different statistical tests are performed using IBM SPSS STATISTICS version 22.

A repeated measures analysis of variance, with the Greenhouse-Geisser correction [60], is conducted to consider the difference in mean total service time-cost obtained using the four algorithms. The result of the repeated measures analysis of variance indicates that the mean *STs* of the four investigated algorithms are statistically significantly different for all investigated problems (see Table 3). Pairwise comparisons with the Bonferroni correction [61–63] also reveal that the two different types of ICACRO obtained a significantly lower mean total service time-cost than that of the ICA and niching PSO for all four investigated problems.

Table 3. Results of repeated measures ANOVA.

		df	Mean Square	F	Sig.
Problem A	Between Groups	1.234	1654453.748	1064352.466	<0.001
	Error	7400.915	1.554		
Problem B	Between Groups	1.079	9596518.233	258544.375	<0.001
	Error	6473.570	37.117		
Problem C	Between Groups	1.082	31014381.517	441298.360	<0.001
	Error	6493.645	70.280		
Problem D	Between Groups	1.052	46227282.898	199913.106	<0.001
	Error	6309.354	231.237		
Problem E	Between Groups	1.040	72555514.778	195415.062	<0.001
	Error	6239.068	371.289		

Furthermore, the ICA results are also significantly better than the results of niching PSO according to the results shown in Table 4. To statistically analyze the behavior of the ICACRO-C and ICACRO-I, Table 4 is provided. The table shows that there is a significant difference between the results obtained using the two algorithms. Further inspection of the results suggests that the mean difference between these two algorithms is significant and increases as the problem size increases. Hence, based on the trends of the algorithm results, it can be concluded that the larger the problem size, the more efficient the performance of the ICACRO-C compared to that of the ICACRO-I.

Table 4. Results of Bonferroni pairwise comparisons.

	(I) Algorithm	(J) Algorithm	Mean Difference (J—I)	Std. Error	Sig.
Problem A	ICACRO-C	Niching PSO	−22.836	0.018	<0.001
	ICACRO-I	Niching PSO	−22.037	0.019	<0.001
	ICA	Niching PSO	−12.575	0.009	<0.001
	ICACRO-C	ICA	−10.261	0.015	<0.001
	ICACRO-I	ICA	−9.462	0.016	<0.001
	ICACRO-C	ICACRO-I	−0.798	0.002	<0.001
Problem B	ICACRO-C	Niching PSO	−50.116	0.089	<0.001
	ICACRO-I	Niching PSO	−47.144	0.088	<0.001
	ICA	Niching PSO	−18.518	0.026	<0.001
	ICACRO-C	ICA	−31.598	0.072	<0.001
	ICACRO-I	ICA	−28.626	0.071	<0.001
	ICACRO-C	ICACRO-I	−2.972	0.01	<0.001
Problem C	ICACRO-C	Niching PSO	−92.806	0.127	<0.001
	ICACRO-I	Niching PSO	−89.269	0.121	<0.001
	ICA	Niching PSO	−51.471	0.044	<0.001
	ICACRO-C	ICA	−41.335	0.098	<0.001
	ICACRO-I	ICA	−37.798	0.091	<0.001
	ICACRO-C	ICACRO-I	−3.537	0.009	<0.001
Problem D	ICACRO-C	Niching PSO	−111.665	0.221	<0.001
	ICACRO-I	Niching PSO	−104.956	0.218	<0.001
	ICA	Niching PSO	−53.594	0.061	<0.001
	ICACRO-C	ICA	−58.071	0.178	<0.001
	ICACRO-I	ICA	−51.362	0.174	<0.001
	ICACRO-C	ICACRO-I	−6.709	0.019	<0.001
Problem E	ICACRO-C	Niching PSO	−139.396	0.293	<0.001
	ICACRO-I	Niching PSO	−132.027	0.27	<0.001
	ICA	Niching PSO	−71.467	0.086	<0.001
	ICACRO-C	ICA	−67.928	0.225	<0.001
	ICACRO-I	ICA	−60.56	0.2	<0.001
	ICACRO-C	ICACRO-I	−7.369	0.026	<0.001

6. Conclusions and Directions for Future Research

Designing effective operators for EAs can enhance the searching capabilities of the algorithms provided that they are utilized appropriately. Furthermore, the designed operators should be developed based on the same origins used as inspiration in the design of the algorithms. Hence, in this paper, a new operator termed the CRO is designed and applied to the ICA based on the color revolutions, a sociopolitical movement that has occurred in some countries. Application of the CRO significantly increases the ability of the ICA to achieve closer-to-optimal solutions when considering five problems of different size. Two types of the proposed algorithm (ICACRO-C and ICACRO-I) considerably outperformed niching PSO and the ICA in terms of the APLT and OSAT policies. The ICACRO-C is experimentally and statistically evaluated to generally be more efficient than the ICACRO-I in terms of execution time and the obtained results.

The CRO can be improved by providing a dynamic probability rate and optimum number of target countries in future research. This can be done by evaluating the trends of changes in the MV and convergence speed of solutions. The CRO can also be made more effective provided that service providers are categorized precisely based on all possible QoS parameters.

Author Contributions: methodology, A.J.; software, A.J. and N.K.N.; validation, E.A.S. and Z.O.; formal analysis, A.J.; investigation, A.J. and N.K.N.; resources, A.J., N.K.N. and E.A.S.; Writing—Original draft preparation, A.J.; Writing—review and editing, A.J., N.K.N., E.A.S. and Z.O.; visualization, A.J. and N.K.N.; supervision, E.A.S. and Z.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Ministry of Higher Education Malaysia Grant: FRGS/1/2014/ICT07/UKM/02/1, the Research University Grant: DIP-2018-041 and Research Incentive Grant: GPP-2020-032.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The results of the experimental tests can be found at http://www.scidb.cn/en/s/pj6b2uq.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hamdaqa, M.; Tahvildari, L. Cloud Computing Uncovered: A Research Landscape. *Adv. Comput.* **2012**, *86*, 41–85. [CrossRef]
2. Mell, P.; Grance, T. *The NIST Definition of Cloud Computing*; National Institute of Standards and Technology, U.S. Department of Commerce: Gaithersburg, MA, USA, 2011; pp. 1–7.
3. Vaquero, L.M.; Rodero-Merino, L.; Caceres, J.; Lindner, M. A break in the clouds: Towards a cloud definition. *Comput. Commun. Rev.* **2008**, *39*, 50–55. [CrossRef]
4. Jula, A.; Sundararajan, E.; Othman, Z. Cloud computing service composition: A systematic literature review. *Expert Syst. Appl.* **2014**, *41*, 3809–3824. [CrossRef]
5. Ding, S.; Yang, S.; Zhang, Y.; Liang, C.; Xia, C. Combining QoS prediction and customer satisfaction estimation to solve cloud service trustworthiness evaluation problems. *Knowl. Based Syst.* **2014**, *56*, 216–225. [CrossRef]
6. Yousefipour, A.; Rahmani Amir, M.; Jahanshahi, M. Energy and cost-aware virtual machine consolidation in cloud computing. *Softw. Pract. Exp.* **2018**, *48*, 1758–1774. [CrossRef]
7. Yuan, Y.; Zhang, W.; Zhang, X.; Zhai, H. Dynamic Service Selection Based on Adaptive Global QoS Constraints Decomposition. *Symmetry* **2019**, *11*, 403. [CrossRef]
8. Lu, P.; Zhang, L.; Liu, X.; Yao, J.; Zhu, Z. Highly efficient data migration and backup for big data applications in elastic optical inter-data-center networks. *IEEE Netw.* **2015**, *29*, 36–42. [CrossRef]
9. Fei, T.; Yuanjun, L.; Lida, X.; Lin, Z. FC-PACO-RM: A Parallel Method for Service Composition Optimal-Selection in Cloud Manufacturing System. *IEEE Trans. Ind. Inform.* **2013**, *9*, 2023–2033. [CrossRef]
10. Yu, T.; Lin, K.-J. Service selection algorithms for composing complex services with multiple qos constraints. In Proceedings of the Third International Conference on Service-Oriented Computing, Amsterdam, The Netherlands, 12–15 December 2005; pp. 130–143.

11. Anselmi, J.; Ardagna, D.; Cremonesi, P. A QoS-based selection approach of autonomic grid services. In Proceedings of the 2007 Workshop on Service-Oriented Computing Performance: Aspects, Issues, and Approaches, Monterey, CA, USA, 25 June 2007; pp. 1–8.
12. Li, J.; Zheng, X.-L.; Chen, S.-T.; Song, W.-W.; Chen, D. An efficient and reliable approach for quality-of-service-aware service composition. *Inf. Sci.* **2014**, *269*, 238–254. [CrossRef]
13. Mijumbi, R.; Serrat, J.; Gorricho, J.; Bouten, N.; Turck, F.D.; Boutaba, R. Network Function Virtualization: State-of-the-Art and Research Challenges. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 236–262. [CrossRef]
14. Ocampo, A.F.; Gil-Herrera, J.; Isolani, P.H.; Neves, M.C.; Botero, J.F.; Latré, S.; Zambenedetti, L.; Barcellos, M.P.; Gaspary, L.P. Optimal Service Function Chain Composition in Network Functions Virtualization. In Proceedings of the 11th IFIP WG 6.6 International Conference on Autonomous Infrastructure, Management, and Security, Zurich, Switzerland, 10–13 July 2017; pp. 62–76.
15. Wang, M.; Cheng, B.; Li, B.; Chen, J. Service Function Chain Composition and Mapping in NFV-Enabled Networks. In Proceedings of the 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 8–13 July 2019; pp. 331–334.
16. Wada, H.; Suzuki, J.; Yamano, Y.; Oba, K. A Multiobjective Optimization Framework for SLA-Aware Service Composition. *IEEE Trans. Serv. Comput.* **2012**, *5*, 358–372. [CrossRef]
17. Hayyolalam, V.; Pourhaji Kazem, A.A. A systematic literature review on QoS-aware service composition and selection in cloud environment. *J. Netw. Comput. Appl.* **2018**, *110*, 52–74. [CrossRef]
18. Jula, A.; Othman, Z.; Sundararajan, E. A Hybrid Imperialist Competitive-Gravitational Attraction Search Algorithm to Optimize Cloud Service Composition. In Proceedings of the 2013 IEEE Workshop on Memetic Computing (MC), Singapore, 15–19 April 2013; pp. 37–43.
19. Wang, Z.-S.; Lee, J.; Song, C.G.; Kim, S.-J. Efficient Chaotic Imperialist Competitive Algorithm with Dropout Strategy for Global Optimization. *Symmetry* **2020**, *12*, 635. [CrossRef]
20. Atashpaz-Gargari, E.; Lucas, C. Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. In Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2007, Singapore, 25–28 September 2007; pp. 4661–4667.
21. Zibin, Z.; Yilei, Z.; Lyu, M.R. Distributed QoS Evaluation for Real-World Web Services. In Proceedings of the 8th IEEE International Conference on Web Services (ICWS 2010), Miami, FL, USA, 5–10 July 2010; pp. 83–90.
22. Kofler, K.; ul Haq, I.; Schikuta, E. A Parallel Branch and Bound Algorithm for Workflow QoS Optimization. In Proceedings of the ICPP 2009, International Conference on Parallel Processing, Vienna, Austria, 22–25 September 2009; pp. 478–485.
23. Moura, L.D.; Bjørner, N. Satisfiability modulo theories: Introduction and applications. *Commun. ACM* **2011**, *54*, 69–77. [CrossRef]
24. Worm, D.; Zivkovic, M.; van den Berg, H.; van der Mei, R. Revenue maximization with quality assurance for composite web services. In Proceedings of the 2012 5th IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2012), Taipei, Taiwan, 17–19 December 2012; pp. 1–9.
25. Shangguang, W.; Qibo, S.; Fangchun, Y. Towards Web Service selection based on QoS estimation. *Int. J. Web Grid Serv.* **2010**, *6*, 424–443. [CrossRef]
26. Zhu, Y.; Li, W.; Luo, J.; Zheng, X. A novel two-phase approach for QoS-aware service composition based on history records. In Proceedings of the 2012 5th IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2012), Taipei, Taiwan, 17–19 December 2012; pp. 1–8.
27. Qi, Y.; Bouguettaya, A. Efficient Service Skyline Computation for Composite Service Selection. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 776–789. [CrossRef]
28. Hossain, M.S.; Hassan, M.M.; Al Qurishi, M.; Alghamdi, A. *Resource Allocation for Service Composition in Cloud-Based Video Surveillance Platform*; IEEE: New York, NY, USA, 2012; pp. 408–412. [CrossRef]
29. Zeng, C.; Guo, X.A.; Ou, W.J.; Han, D. Cloud Computing Service Composition and Search Based on Semantic. In *Lecture Notes in Computer Science*; Jaatun, M.G., Zhao, G., Rong, C., Eds.; Springer: Berlin, Germany, 2009; Volume 5931, pp. 290–300.
30. Huang, J.; Liu, Y.; Yu, R.; Duan, Q.; Tanaka, Y. Modeling and Algorithms for QoS-Aware Service Composition in Virtualization-Based Cloud Computing. *IEICE Trans. Commun.* **2013**, *96*, 10–19. [CrossRef]
31. Zhou, X.; Mao, F. A Semantics Web Service Composition Approach Based on Cloud Computing. In Proceedings of the 2012 Fourth International Conference on Computational and Information Sciences (ICCIS 2012), Chongqing, China, 17–19 August 2012; pp. 807–810.
32. Karim, R.; Chen, D.; Miri, A. An End-to-End QoS Mapping Approach for Cloud Service Selection. In Proceedings of the 2013 IEEE Ninth World Congress on Services (SERVICES), Santa Clara, CA, USA, 28 June–3 July 2013; pp. 341–348.
33. Barzegar, S.; Davoudpour, M.; Meybodi, M.R.; Sadeghian, A.; Tirandazian, M. Formalized learning automata with adaptive fuzzy coloured Petri net; an application specific to managing traffic signals. *Sci. Iran.* **2011**, *18*, 554–565. [CrossRef]
34. Zhao, H.; Gao, W.; Deng, W.; Sun, M. Study on an Adaptive Co-Evolutionary ACO Algorithm for Complex Optimization Problems. *Symmetry* **2018**, *10*, 104. [CrossRef]
35. Jaddi, N.S.; Alvankarian, J.; Abdullah, S. Kidney-inspired algorithm for optimization problems. *Commun. Nonlinear Sci. Numer. Simul.* **2017**, *42*, 358–369. [CrossRef]
36. He, J.; Lin, G.M. Average Convergence Rate of Evolutionary Algorithms. *IEEE Trans. Evol. Comput.* **2016**, *20*, 316–321. [CrossRef]

37. Vesterstrom, J.; Thomsen, R. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2004, Portland, OR, USA, 19–23 June 2004; pp. 1980–1987.
38. Bäck, T.; Fogel, D.B.; Michalewicz, Z. *Handbook of Evolutionary Computation*; IOP Publishing Ltd.: Bristol, UK, 1997; p. 988.
39. Naseri, N.K.; Sundararajan, E.A.; Ayob, M.; Jula, A. Smart Root Search (SRS): A Novel Nature-Inspired Search Algorithm. *Symmetry* **2020**, *12*, 2025. [CrossRef]
40. Knuth, D.E. *The Art of Computer Programming, Volume 4A: Combinatorial Algorithms, Part 1*; Pearson Education: Tamil Nadu, India, 2011.
41. Abonyi, J.; Akerkar, R.; Alavi, A.H.; Arango, C.; Aydogdu, I.; Brest, J.; Cai, X.; Cordeiro, J.; Cortés, P.; Costa, K.A.P.; et al. List of Contributors. In *Swarm Intelligence and Bio-Inspired Computation*; Yang, X.-S., Cui, Z., Xiao, R., Gandomi, A.H., Karamanoglu, M., Eds.; Elsevier: Oxford, UK, 2013; pp. xv–xviii. [CrossRef]
42. Zhang, X.; Dou, W. Preference-Aware QoS Evaluation for Cloud Web Service Composition Based on Artificial Neural Networks. In *Web Information Systems and Mining*; Wang, F., Gong, Z., Luo, X., Lei, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6318, pp. 410–417.
43. Wu, Q.; Zhang, M.; Zheng, R.; Lou, Y.; Wei, W. A QoS-Satisfied Prediction Model for Cloud-Service Composition Based on a Hidden Markov Model. *Math. Probl. Eng.* **2013**, *2013*, 7. [CrossRef]
44. Lie, Q.; Yan, W.; Orgun, M.A. Cloud Service Selection Based on the Aggregation of User Feedback and Quantitative Performance Assessment. In Proceedings of the 2013 IEEE International Conference on Services Computing (SCC), Santa Clara, CA, USA, 28 June–3 July 2013; pp. 152–159.
45. Ye, Z.; Zhou, X.; Bouguettaya, A. Genetic Algorithm Based QoS-Aware Service Compositions in Cloud Computing. In *Database Systems for Advanced Applications*; Yu, J., Kim, M., Unland, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6588, pp. 321–334.
46. Klein, A.; Ishikawa, F.; Honiden, S. Towards network-aware service composition in the cloud. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 959–968.
47. Ludwig, S.A. Clonal selection based genetic algorithm for workflow service selection. In Proceedings of the 2012 IEEE Congress on Evolutionary Computation (CEC), Brisbane, QLD, Australia, 10–15 June 2012; pp. 1–7.
48. Wang, S.G.; Sun, Q.B.; Zou, H.; Yang, F.C. Particle Swarm Optimization with Skyline Operator for Fast Cloud-based Web Service Composition. *Mob. Netw. Appl.* **2013**, *18*, 116–121. [CrossRef]
49. Liao, J.X.; Liu, Y.; Wang, J.Y.; Zhu, X.M. Service Composition Based on Niching Particle Swarm Optimization in Service Overlay Networks. *KSII Trans. Internet Inf. Syst.* **2012**, *6*, 1106–1127. [CrossRef]
50. Wang, Y.W. *Application of Chaos Ant Colony Algorithm in Web Service Composition Based on QoS*; IEEE Computer Soc: Los Alamitos, CA, USA, 2009; pp. 225–227. [CrossRef]
51. Yang, Y.; Mi, Z.; Sun, J. Game theory based iaas services composition in cloud computing environment. *Adv. Inf. Sci. Serv. Sci.* **2012**, *4*, 238–246.
52. Jula, A.; Othman, Z.; Sundararajan, E. Imperialist competitive algorithm with PROCLUS classifier for service time optimization in cloud computing service composition. *Expert Syst. Appl.* **2015**, *42*, 135–145. [CrossRef]
53. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA, 2011; p. 696.
54. Bahrami, H.; Faez, K.; Abdechiri, M. Imperialist Competitive Algorithm Using Chaos Theory for Optimization (CICA). In Proceedings of the 2010 12th International Conference on Computer Modelling and Simulation (UKSim), Brisbane, Australia, 24–26 March 2010; pp. 98–103.
55. Zarandi, M.H.F.; Zarinbal, M.; Ghanbari, N.; Turksen, I.B. A new fuzzy functions model tuned by hybridizing imperialist competitive algorithm and simulated annealing. Application: Stock price prediction. *Inf. Sci.* **2013**, *222*, 213–228. [CrossRef]
56. Zherebkin, M. In search of a theoretical approach to the analysis of the 'Colour revolutions': Transition studies and discourse theory. *Communist Post-Communist Stud.* **2009**, *42*, 199–216. [CrossRef]
57. Marples, D.R. Color revolutions: The Belarus case. *Communist Post-Communist Stud.* **2006**, *39*, 351–364. [CrossRef]
58. Jula, A.; Nilsaz, H.; Sundararajan, E.; Othman, Z. A new dataset and benchmark for cloud computing service composition. In Proceedings of the 2014 5th International Conference on Intelligent Systems, Modelling and Simulation, Langkawi, Malaysia, 27–29 January 2014; pp. 83–86.
59. Liao, J.X.; Liu, Y.; Zhu, X.M.; Xu, T.; Wang, J.Y. Niching Particle Swarm Optimization Algorithm for Service Composition. In *2011 IEEE Global Telecommunications Conference*; IEEE: Houston, TX, USA, 2011.
60. Abdi, H. *Greenhouse-Geisser Correction. Encyclopedia of Research Design*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2010; pp. 545–549. [CrossRef]
61. Nakagawa, S. A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behav. Ecol.* **2004**, *15*, 1044–1045. [CrossRef]
62. Cabin, R.; Mitchell, R. To Bonferroni or not to Bonferroni: When and how are the questions. *Bull. Ecol. Soc. Am.* **2000**, *81*, 246–248. [CrossRef]
63. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70. [CrossRef]

symmetry

MDPI

Article

Overlapping Community Discovery Method Based on Two Expansions of Seeds

Yan Li [1], Jing He [1], Youxi Wu [2,*] and Rongjie Lv [1]

[1] School of Economics and Management, Hebei University of Technology, Tianjin 300401, China; lywuc@hebut.edu.cn (Y.L.); 2002016@hebut.edu.cn (J.H.); rjlv@hebut.edu.cn (R.L.)
[2] School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
* Correspondence: wuc@hebut.edu.cn; Tel.: +86-22-6043-5882

Abstract: The real world can be characterized as a complex network sto in symmetric matrix. Community discovery (or community detection) can effectively reveal the common features of network groups. The communities are overlapping since, in fact, one thing often belongs to multiple categories. Hence, overlapping community discovery has become a new research hotspot. Since the results of the existing community discovery algorithms are not robust enough, this paper proposes an effective algorithm, named Two Expansions of Seeds (TES). TES adopts the topological feature of network nodes to find the local maximum nodes as the seeds which are based on the gravitational degree, which makes the community discovery robust. Then, the seeds are expanded by the greedy strategy based on the fitness function, and the community cleaning strategy is employed to avoid the nodes with negative fitness so as to improve the accuracy of community discovery. After that, the gravitational degree is used to expand the communities for the second time. Thus, all nodes in the network belong to at least one community. Finally, we calculate the distance between the communities and merge similar communities to obtain a less- undant community structure. Experimental results demonstrate that our algorithm outperforms other state-of-the-art algorithms.

Keywords: overlapping community discovery; gravitational degree; greedy strategy; two expansions

Citation: Li, Y.; He, J.; Wu, Y.; Lv, R. Overlapping Community Discovery Method Based on Two Expansions of Seeds. *Symmetry* **2021**, *13*, 18. https://dx.doi.org/10.3390/sym13010018

Received: 25 November 2020
Accepted: 18 December 2020
Published: 24 December 2020

1. Introduction

Many complex systems exist in the form of networks in the real world, such as social networks [1,2], traffic networks [3,4], network sparsification [5] and protein interaction networks [6,7]. These complex systems can be characterized as complex networks sto in symmetric matrix for analysis and research. Entities in the complex network are represented by nodes, and the relationships between the entities are represented by edges [8,9]. Many researches based on complex networks have been investigated, such as social computing [10], network computation [11], and community discovery [12]. The community structure (module or cluster) is an important feature of a complex network, which means that the network is composed of several communities. The connections between the nodes in the community are very close, while the connections between the communities are relatively sparse [13]. The purpose of community discovery (or community detection [14]) is to mine community structures in a complex network. Community discovery can reveal the universal features of a complex network and help in understanding its topology accurately, which provides guidance for the use and transformation of the network and promotes the practical application of the network. Hence, community discovery has become one of the hotspots of complex network research [15] and various researches have been investigated, such as disjoint community detection [16,17], overlapping community detection [18], and multiobjective community detection [19].

Early researches on community discovery mainly focused on nonoverlapping communities, which assumed that each node belongs to only one community and there is no overlap of any two communities. Many representative algorithms have been proposed,

such as the graph-partitioning-based method [20], label-propagation-based method [21], clustering method [22,23], and optimization method [24,25]. However, in the real world, things often have the characteristics of diversity. One thing often belongs to multiple categories and there may be overlap between communities. Therefore, overlapping community discovery has become a new research hotspot in recent years. Researches on overlapping community discovery can be divided into two categories: global-network-information-based and local-network-information-based methods.

The methods based on global network information aim to find the community structure in the whole network by optimizing a certain global objective function using whole connection information, which mainly include the link-based method [26,27] and the clique percolation method [28]. These methods can get better results in community discovery, but they have high time complexity and are not suitable for large-scale complex networks with numerous nodes. The methods based on local information aim to find the community structure starting from a node in the network by optimizing a certain local objective function using local connection information, which mainly include the label propagation method [29,30] and the local community expansion method [31,32]. Since the process of community discovery is only related to the local information in the network, the time complexity is low. Thus, these methods are suitable for large-scale complex networks. However, their disadvantage is that when the parameters of the algorithms change slightly, the results of community discovery change remarkably.

To tackle this problem, this paper proposes an overlapping community discovery method based on Two Expansions of Seeds (TES). The main features of this method are that the topological feature of the network (node degree centrality) is used to define the gravitational degree and the local maximum node is taken as the seed. The reason is that the greater the gravitational degree of the node, the greater its influence and the stronger its information transmission ability in the network is, which is beneficial for robust community discovery. Then, the seed is expanded by the greedy strategy based on the fitness function. When new nodes are added to the community, the community structure may be changed, thereby, there may be nodes with negative fitness. To avoid such nodes, this paper adopts the community cleaning strategy. After the expansion based on the fitness function, a community can cover most of the nodes in the network, but there are still a small number of nodes that cannot be assigned to any community because of the uction of community fitness. To solve this problem, this paper uses a gravitational function to expand the nodes that are not included in any community for the second time. Thus, all nodes belong to at least one community. Finally, by calculating the distance between communities and merging similar communities, we effectively uce the undant communities. The main contributions of this paper are as follows:

- We propose an overlapping community discovery algorithm named TES.
- TES employs the gravitational degree to find the local maximum nodes as the seeds and expands these seeds by the greedy strategy.
- Experimental results verify that our algorithm has better performance than other competitive algorithms.

The rest of this paper is organized as follows: Section 2 briefly summarizes the related work. Section 3 proposes our algorithm, named TES, which is composed of three parts: seed selecting, twice node expanding, and overlapping community merging. Section 4 reports the performance of TES. We draw the conclusion in Section 5.

2. Related Work

In this section, we will briefly review the categories of the overlapping community discovery methods first. Then, we will introduce the methods of local community optimization and expansion in detail and analyze the shortcomings of the-state-of-the-art algorithms. This paper aims to deal with the problem of unreasonable seed selection for local community optimization and expansion.

The overlapping community discovery methods can be divided into four categories: link-based method, clique percolation method, label-propagation-based method, and local-community-optimization-and-expansion-based method.

- The link-based method converts the cluste objects into network edges (or links) and deals with these edges by nonoverlapping partitions. Since a node is usually a vertex of multiple edges, if these edges belong to different linked communities, the node is an overlapping node. The LINK algorithm [27] is representative of this method. In addition, *k*-means was employed to expand seeds twice in dynamic community detection [33].
- The clique percolation method considers that a community is composed of a number of fully connected subgraphs. defined as a clique, and an adjacent clique forms a community. Since a node may belong to more than one clique, it is an overlapping node. However, the algorithm has higher constraints on interconnected conditions and depends on the selection of parameter *k*. The CPM algorithm [28] is representative of this method.
- The label-propagation-based method assigns a unique label to each node during initialization; updates the label and its membership by iteration; and finally, assigns the nodes with the same label to the same community. Apparently, if a node has multiple labels, the node is an overlapping node. The COPRA algorithm [29] is a representative of this method.
- The local-community-optimization-and-expansion-based method starts from the local communities, expands the communities gradually based on the optimization function, and forms cross-regions between multiple extensions, thus finding overlapping community nodes. The representative algorithms are LFM [31] and GCE [32]. In addition to the above algorithms, there are some classical methods, such as the semisupervised learning method [34]; deep learning method [35]; and the CONGA algorithm [36], which splits the clone node by itself and adds a virtual edge between the split nodes to find the overlapping nodes.

Among the abovementioned methods, the fourth one—local community optimization and expansion—becomes more and more popular. For example, the research in [21] found that taking the local maximum node defined by the degree centrality as the seed can discover higher quality communities and avoid instability at the same time. The research in [37] was about two methods to define the node influence: the community structure of social networks and the influence-based measure of node intimacy center, and took the nodes with great influence as the seeds. The EAGLE algorithm took the largest clique in the network as the seed and igno the second largest one, which has high time complexity [38]. Another paper [39] selected a group of nodes as seeds that were closely connected in the network, namely, an Egonet (hawk-eye network), but this method is more suitable for networks with a large global clustering coefficient. A seed set expansion method based on graph partitioning was proposed in [40] to find a group of nodes with low conductivity, and the node closest to the cluster was taken as a seed. The online social network (OSN) algorithm, as a multilever community discovery algorithm, combined user interests and cohesiveness to coarsen the initial network and found an initial community assignment using stochastic inference in the coarsest network [41]. All these methods use the local topology information of the network to optimize the local optimization function to find the community structure in the network. It does not need to know the global topology of the network, and shows certain advantages in large-scale networks. Therefore, seed selection is the foundation of this kind of method, which will affect the quality of community structure mining. The LFM algorithm [31] and the DEMON algorithm [32] expand the community by random seed selection, which inevitably causes the instability of community discovery. The GCE algorithm improved the LFM algorithm by mining *k*-cliques as the seed through the classic Bron–Kerbosch algorithm in the network [42]. In this method, cliques are fixed, but the seed selection depends on the selection of parameter *k*, which can easily cause the problem of low network coverage.

To solve the problem of unreasonable seed selection for local community optimization and expansion, this paper proposes an overlapping community discovery algorithm based on two expansions of seeds. A node with the local maximum gravitational degree defined by degree centrality is taken as a seed. This method has the advantages of a high-quality community and robust results, but the disadvantage is that these communities cannot cover the whole network. To overcome this problem, the communities are expanded for the second time to ensure that each node belongs to at least one community.

3. Proposed Method

In this section, we propose the TES algorithm, which is composed of three parts. The first part employs the gravitational degree defined by the network topological feature (degree centrality) to find the local maximum nodes as the seeds. The second part expands these seeds by the greedy strategy based on the fitness function. Then, the communities are expanded for the second time based on the gravitational function. The third part calculates the distance between the communities and merges the similar communities to get the final communities.

3.1. Seed Selection

In actual networks, some nodes are usually closely connected with other nodes, called central nodes, which contribute greatly to information transmission. They are usually scatte across the whole network and located in regions where the nodes are more closely connected. This is consistent with the fact that the nodes in a community are closely connected, while the connections between communities are sparse. Hence, the central nodes can be taken as the seeds. The centrality of a node reflects its centrality and importance in the network [43]. Inspi by the gravitational relationships in the dynamic social network [44], this paper proposes a gravitational degree based on degree centrality to measure the influence of the central nodes on other nodes.

Newton's law of universal gravitation holds that any two particles are attracted by a force in the direction of the line between them. The gravitation is proportional to the product of their masses and inversely proportional to the square of their distance, as shown in Equation (1).

$$F = g \times \frac{m_1 \times m_2}{r^2}, \tag{1}$$

where g is the gravitational constant, m_1 and m_2 are the masses of two particles, and r is the distance between two particles.

In this paper, a network is represented by an undirected graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of n vertices and $E = \{e_1, e_2, \ldots, e_m\}$ is a set of m edges.

Definition 1. *Node centrality is the degree of a node, denoted by* $d(v_i)$.

Definition 2. *If there is an edge between nodes* v_i *and* v_j*, then node* v_j *is a neighbor of node* v_i*. All neighbors of node* v_i *are denoted by* $n(v_i)$.

Definition 3. *To measure the similarity between nodes* v_i *and* v_j*, this paper employs the Jaccard similarity coefficient [45], denoted by* $s(v_i, v_j)$.

$$s(v_i, v_j) = \frac{|n(v_i) \cap n(v_j)|}{|n(v_i) \cup n(v_j)|}. \tag{2}$$

Definition 4. *The distance between node* v_i *and its neighbor* v_j *is* $d(v_i, v_j)$.

$$d(v_i, v_j) = 1 - s(v_i, v_j). \tag{3}$$

Definition 5. *The gravitation of node v_i to its neighbor v_j is $Gr(v_i, v_j)$.*

$$Gr(v_i, v_j) = g \times \frac{d(v_i) \times d(v_j)}{(1 - s(v_i, v_j))^2}. \quad (4)$$

Using the node degree to measure the quality of a node can reflect the ability of information transmission to its neighbor. The gravitational degree of v_i to its neighbor v_j is directly proportional to the node degree and inversely proportional to the distance between them.

Definition 6. *The gravitational degree of node v_i is the sum of its gravitation to all nodes in the network.*

$$GD(v_i) = \sum_{v_j \in N(v_i)} Gr(v_i, v_j) = g \times \sum_{v_j \in N(v_i)} \frac{d(v_i) \times d(v_j)}{(1 - s(v_i, v_j))^2}. \quad (5)$$

The greater the gravitational degree of node v_i, the greater its influence on the network. The stronger the information transmission ability of a node, the more likely it is to become a seed node.

An illustrative example is shown as follows:

Example 1. *In Figure 1, node v_1 has 7 neighbors, i.e., $n(v_1)$= {2, 3, 4, 5, 6, 7, 8}. Node v_4 has 3 neighbors, i.e., $n(v_4)$= {1, 2, 3}. Thus, node centrality of nodes v_1 and v_4 are $d(v_1)$ = 7 and $d(v_4)$ = 3, respectively. $n(v_1) \cap n(v_4)$ and $n(v_1) \cup n(v_4)$ are {2, 3} and {1, 2, 3, 4, 5, 6, 7, 8}, respectively. Thus, $s(v_1, v_4) = 2/8 = 0.25$ and $d(v_1, v_4) = 1 - 0.25 = 0.75$. Hence, $Gr(v_1, v_4) = 9.8 * 7 * 3/0.75/0.75 = 365.9$; $GD(v_4) = Gr(v_1, v_4) + Gr(v_2, v_4) + Gr(v_3, v_4) = 365.9 + 326.7 + 326.7 = 1019.2$.*

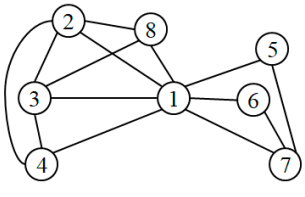


Figure 1. An illustrative network with 8 nodes and 14 edges.

Definition 7. *If the gravitational degree of a node is no less than that of all its neighbors, the node will be called the local maximum degree node of the network.*

The local maximum node has a large gravitational degree and strong information transmission ability. Most of them are scatte in the network. Therefore, this paper selects the local maximum nodes as the seeds. The seed selection algorithm is shown in Algorithm 1. First, all nodes are marked as 0 and the gravitational degree of each node is calculated. The node with the largest gravitational degree is put into the seed set. Then, the node with the local maximum degree is marked as 1, and the node and its neighbors are moved out of the vertex set. Search for the next seed iteratively until all nodes have been marked and moved out of the vertex set.

Algorithm 1 GetSeed.

Require: network $G = (V, E)$;

Ensure: seed set S;

$S \leftarrow \emptyset$;
for each $i \in n$ **do**
  $v_i.label \leftarrow 0$;
  $GD(v_i) = \sum_{v_j \in N(v_i)} Gr(v_i, v_j)$;
end for
while $V \neq \emptyset$ **do**
  $s \leftarrow argmax_{v \in V}(\{GD(v)\})$;
  if $s.label = 0$ **then**
    $S \leftarrow S \cup \{s\}$;
    $s.label \leftarrow 1$;
    $V \leftarrow V - \{s \cup N(s)\}$;
  end if
end while
return S

3.2. Community Discovery

For each seed in seed set S, this paper iteratively adds its neighbors to the community to discover natural communities. There are many ways to expand the community, including the minimum one norm [20], the label propagation method [29], and the fitness function method [31,42]. This paper employs the fitness function method since it can provide good results on real datasets.

Definition 8. *Community C is a subset of V. For community C in network $G = (V, E)$, its neighbor $N(C)$ is defined as*

$$N(C) = \{v_j | \forall e_{ij} \in E, v_i \in C, v_j \notin C\}. \tag{6}$$

Definition 9. *For community C in network $G = (V, E)$, its fitness $f(C)$ is defined as*

$$f(C) = \frac{d_{in}^C}{(d_{in}^C + d_{out}^C)^\alpha}, \tag{7}$$

where d_{in}^C and d_{out}^C are the sum of the degrees of the nodes that are inside and outside community C, respectively. $d_{in}^C = 2 * e(C)$ and $d_{out}^C = |E| - e(C)$, where $e(C)$ is the number of edges inside community C. $\alpha > 0$ is an adjustment parameter.

α in the fitness function is the resolution parameter, which can adjust the scale of the community discove. The smaller α is, the greater the influence of d_{in}^C. This will lead to a rapid increase of $f(C)$ after adding node v_i to community C. Therefore, community C can accept more nodes. When α tends to be 0, the community may expand to cover the entire network. On the contrary, the larger α is, the smaller the impact of d_{in}^C. This will lead to the tiny increase of $f(C)$ after adding node v_i. Therefore, a small community is formed. When $\alpha = 1$, $f(C) = d_{in}^C / (d_{in}^C + d_{out}^C)$. The more sparse the connection between community C and outside is, the smaller d_{out}^C is and the larger $f(C)$ is, which can reflect the local connection density of community C.

Example 2. *In Figure 1, suppose community C is composed of nodes 5, 6, and 7. $N(C) = \{1\}$ since node 1 connects with community C. Node 8 does not belong to $N(C)$ since it does not connect with*

community C. Suppose there is an edge between nodes 5 and 8, node 8 belongs to $N(C)$. d_{in}^{C} *is 4 since the degrees of nodes 5, 6, and 7 in community C are 1, 1, and 2, respectively, and 1 + 1 + 2 = 4. Another method is to count the number of edges in community C. There are 2 edges in community C, thus,* $d_{in}^{C} = 2 * 2 = 4$. *Similarly,* $d_{out}^{C} = 14 - 2 = 12$.

Definition 10. *The fitness* $f(v_i)$ *of node* v_i *can be obtained as follows:*

$$f(v_i) = \begin{cases} f(C \cup \{v_i\}) - f(C) & \forall v_i \in N(C) \\ f(C) - f(C - \{v_i\}) & \forall v_i \in C \end{cases}. \tag{8}$$

The disadvantage of this method is that although most of nodes can be assigned to the corresponding communities, some nodes fail to be assigned, thus resulting in low network coverage. Therefore, this paper expands the nodes that have not been assigned to the community for the second time. This is in accordance with the actual situation. For example, in a social network, everyone has friends and belongs to a circle of friends [37]. This paper assumes that each node belongs to at least one community. A gravitational function is defined by the ratio of the gravitation between nodes and the gravitational degree of nodes. The gravitation of node v_i is the sum of the gravitational degrees between node v_i and its neighbors. The more neighbors of node v_i the community C contains, the greater the gravitation between the community and node v_i is. The gravitational function is given as follows:

Definition 11. *The gravitation of community C to node* v_i *is measu by the gravitational degree, and the gravitational function* $GF(C, v_i)$ *is*

$$GF(C, v_i) = \frac{\sum_{(v_j \in C) \cap (v_j \in N(v_i))} Gr(v_i, v_j)}{GD(v_i)}. \tag{9}$$

When the seed set is found in the first stage, the seed is expanded by the greedy strategy, that is, the local objective function of the community is maximized by adding node to the temporal community or deleting it from the community. We will show the principle of the algorithm as follows: We put a seed into temporal community C first. Then, we calculate the fitness of all its neighbors and add the maximum fitness neighbor v_{max} into C, as shown in lines 3–7 of Algorithm 2. After adding the maximum fitness neighbor, the structure of the community will be changed. At this time, the fitness of each node for the new temporary community should be updated. If a node has a negative fitness, it will be removed from the community, as shown in lines 9–14 of Algorithm 2. Iterate the above expansion until the fitness decreases when any node is added. We store temporal community C into community set CS and remove these nodes from the network.

Obviously, when a community is expanded, the fitness of the nodes in the community and the neighbors need to be recalculated. To solve this problem, we adopt the following steps. If there is an edge between v_i and v_j, then d_{ij} is 1; otherwise, it is 0. The initials of d_{in}^{C} and d_{out}^{C} are 0. If node v_i is added into the community, we adopt Equations (10) and (11). If node v_i is removed from the community, we adopt Equations (12) and (13).

$$d_{in}^{C \cup \{v_i\}} = d_{in}^{C} + 2 \times \sum_{v_j \in C \cap N(v_i)} d_{ij}. \tag{10}$$

$$d_{out}^{C \cup \{v_i\}} = d_{out}^{C} - \sum_{v_j \in C \cap N(v_i)} d_{ij}. \tag{11}$$

$$d_{in}^{C - \{v_i\}} = d_{in}^{C} - 2 \times \sum_{v_j \in C \cap N(v_i)} d_{ij}. \tag{12}$$

$$d_{out}^{C-\{v_i\}} = d_{out}^{C} + \sum_{v_j \in C \cap N(v_i)} d_{ij}. \tag{13}$$

Example 3. *In Figure 1, according to Example 2, community C ={5, 6, 7}. Let node 1 be added into community C. According to Equation (10),* $d_{in}^{C\cup\{v_1\}} = 4 + 2 * 3 = 10$*, since when node 1 is added, there are three edges that are added into community C. According to Equation (11),* $d_{out}^{C\cup\{v_1\}} = 12 - 3 = 9$*. Let node 5 be removed from community C. According to Equation (12),* $d_{in}^{C-\{v_5\}} = 4 - 2 * 1 = 2$*, since there is one edge in community C that connects with node 5. According to Equation (13),* $d_{out}^{C-\{v_5\}} = 12 + 1 = 13$*.*

Equation (8) is used to update the community fitness. In this way, we only need to know the degree of node v_i and calculate d_{ij} of the nodes which are both in community C and neighbors of v_i. To further speed up the calculation, we store d_{in}^{C} and d_{out}^{C}, which will be updated when temporal community C adds a new node or removes a node, as shown in lines 7–8 and 11–12 of Algorithm 2.

Algorithm 2 GetNaturalcoms.

Require: network $G = (V, E)$, seed set S, and parameter α;
Ensure: community set CS;
1: $CS = \emptyset$,$d_{in}^{C} = d_{out}^{C} = 0$;
2: **for** each $s \in S$ **do**
3: $C \leftarrow \{s\}$;
4: **while** $C \neq \emptyset$ **do**
5: $v_{max} \leftarrow argmax_{v \in N(C)}(\{f(v)\})$;
6: **if** $f(v_{max}) > 0$ **then**
7: $C \leftarrow C \cup v_{max}$;
8: Update d_{in}^{C} and d_{out}^{C};
9: **for** each $v_j \in C$ **do**
10: **if** $f(v_j) < 0$ **then**
11: $C \leftarrow C - \{v_j\}$;
12: Update d_{in}^{C} and d_{out}^{C};
13: **end if**
14: **end for**
15: **else**
16: break;
17: **end if**
18: **end while**
19: $CS \leftarrow CS \cup C$;
20: $V \leftarrow V - C$;
21: **end for**
22: **return** CS

Finally, we expand nodes for the second time. If a node does not belong to any community, the node is merged into the community with the greatest gravitation, as shown in Algorithm 3.

Algorithm 3 ExpandingSecond.

Require: node set V, and community set CS;

Ensure: community set CS;

1: **if** $V \neq \emptyset$ **then**
2: **for** each $v_j \in V$ **do**
3: **for** $C_i \in CS$ **do**
4: $i_{max} \leftarrow argmax(\{GF(C_i, v_j)\})$;
5: **end for**
6: $C_{i_{max}} \leftarrow C_{i_{max}} \cup v_j$;
7: **end for**
8: **end if**
9: **return** CS

3.3. Merging Overlapping Communities

In a nonoverlapping community, a node belongs to only one community [46], while a node may belong to multiple communities in an overlapping community. Therefore, there may be similarities between two communities. When a certain similarity is reached, the excessive overlapping phenomenon will occur, resulting in a undant community [47]. Hence, after discovering the communities, this paper defines a measure of community distance which is used to discover and merge the overlapping communities to simplify the community structure.

Definition 12. *The distance between communities C_1 and C_2 is*

$$\delta_E(C_1, C_2) = 1 - \frac{|C_1 \cap C_2|}{min(|C_1|, |C_2|)}. \tag{14}$$

In this paper, ϵ is the threshold of the distance parameter. If $\delta(C_1, C_2) < \epsilon$, communities C_1 and C_2 are merged into one community since they overlap excessively. The Merge_Overlap algorithm is shown in Algorithm 4.

To avoid invalid calculations, we adopt the principle of inverted index to prune invalid detection of overlapping communities. Therefore, set $Cp(v_i)$ is used to store the communities in which node v_i belongs. An illustrative example is shown as follows:

Suppose we have 3 communities: $C_1 = \{a, b\}$, $C_2 = \{b\}$, and $C_3 = \{c\}$. We know that $Cp(a) = \{1\}$, $Cp(b) = \{1, 2\}$, and $Cp(c) = \{3\}$. To obtain the overlapping community of C_1, we calculate $Cp(a) \cup Cp(b) = \{1, 2\}$ since $C_1 = \{a, b\}$. Therefore, communities C_1 and C_2 are two overlapping communities. It is not necessary to calculate the distance between communities C_1 and C_3. Therefore, the inverted index is an effective pruning strategy.

According to the above example, we should create set $Cp(v_i)$ at first, as shown in lines 1–7 of Algorithm 4. Apparently, if the number of elements in $Cp(v_i)$ is greater than 1, it indicates that node v_i belongs to multiple communities and is an overlapping node. We determine whether the communities in $Cp(v_i)$ overlap or not, as shown in lines 8–19 of Algorithm 4.

To sum up, Algorithm 5 presents the overlapping community discovery algorithm based on two expansions of seeds.

Algorithm 4 MergeOverlap.

Require: network $G = (V, E)$, community set CS, and parameter ϵ;

Ensure: the new community set CS;

1: **for** $v_i \in V$ **do**
2: **for** each $C_j \in CS$ **do**
3: **if** $v_i \in C_j$ **then**
4: $Cp(v_i) \leftarrow Cp(v_i) \cup j$;
5: **end if**
6: **end for**
7: **end for**
8: **for** each $C_j \in CS$ **do**
9: **for** each $v_i \in C_j$ **do**
10: **if** $length(Cp(v_i)) > 1$ **then**
11: $Cv \leftarrow Cv \cup Cp(v_i)$;
12: **end if**
13: **end for**
14: **for** $cv_i \in Cv$ **do**
15: **if** $dis(C_j, C_{cv_i}) < \epsilon$ **then**
16: $C_{cv_i} \leftarrow C_j \cup C_{cv_i}$;
17: **end if**
18: **end for**
19: **end for**
20: **return** CS

Algorithm 5 TES.

Require: network $G = (V, E)$, parameter α, and parameter ϵ;

Ensure: community set CS, community set $c(v)$ to which the node belongs;

1: $S \leftarrow GetSeed(V, E)$; //Searching for the seed in the network
2: $CS \leftarrow GetNaturalcoms(V, E, S, \alpha)$; //Expand each seed according to the fitness function
3: $CS \leftarrow ExpandingSecond(V, CS)$; //Expand the nodes for the second time
4: $CS \leftarrow MergeOverlop(V, E, CS, \epsilon)$; //Merge the overlapping communities in the network
5: **return** CS

3.4. Theoretical Analysis

The space complexity and time complexity of TES are $O(k * n + m)$ and $O(k * n^2 + m))$, respectively, where k, n, and m are the number of seeds, nodes, and edges in G, respectively. The reason is shown as follows:

The space complexity of network G is $O(n + m)$. The space complexity of all neighbors of each node is $O(m)$ since each edge should be calculated. Thus, the time complexity of $n(v_i)$ of each node is also $O(m)$. Further, the space complexity and time complexity of $s(v_i, v_j)$, $d(v_i, v_j)$, and $Gr(v_i, v_j)$ are also $O(m)$. Obviously, the time complexity of $GD(v_i)$ of each node is $O(m)$ and the space complexity of $GD(v_i)$ is $O(n)$. Hence, the time complexity of lines 2–5 in Algorithm 1 is $O(m)$. Since each node will be checked once, the time complexity of lines 6–13 is $O(n)$. Therefore, both the space complexity and time complexity of Algorithm 1 are $O(n + m)$.

Suppose we find k seeds, where $k << n$. When a node is added into or removed from a community, no more than n edges are checked. Thus, the time complexity of Equations (10)–(13) are $O(n)$. Hence, the time complexity of Equations (7)–(9) are also $O(n)$. A node can be assigned into no more than k communities. Therefore, the time complexity of Algorithm 2 is $O(k * n^2)$.

Suppose there are t nodes which are expanded twice, where $t << n$. Each node will be added into each community once. Thus, the time complexity of lines 3–5 is $O(k * n)$. Therefore, the time complexity of Algorithm 3 is $O(t * k * n)$.

Obviously, the time complexity of lines 1–7 of Algorithm 4 is $O(k * n)$ since there are k communities and n nodes. Similarly, the time complexity of lines 8–19 of Algorithm 4 is also $O(k * n)$.

Apparently, each community has no more than n nodes. Thus, the space complexities of these communities are $O(k * n)$. Hence, the space complexities of Algorithms 2, 3, and 4 are $O(k * n)$.

Since $t << n$, the time complexity of TES is $O(n + m + k * n^2 + t * k * n + k * n) = O(k * n^2 + m)$ and the space complexity of TES is $O(n + m + k * n) = O(k * n + m)$.

4. Experimental Results and Analysis

4.1. Baseline Methods

To verify the performance of TES, three state-of-the-art algorithms are selected: CONGA [36], COPRA [29], and LFM [31]. In addition, the TES algorithm has three key steps: searching for seeds, discovering communities based on two expansions, and merging overlapping communities. The two expansions of communities include the first expansion of the community based on the fitness function and the second expansion of the community based on the gravitational function. The community expansion based on the fitness function includes community cleaning. To verify the reasonability of these parts, four comparative algorithms—TES_Seed, TES_Unclean, TES_Fitness, and TES_Unmerge—are constructed, and their specific descriptions are shown in Table 1.

Table 1. Comparative algorithms.

Algorithms	Description
TES_Seed	Nodes are randomly selected as seeds.
TES_Unclean	Community cleaning is not performed after the first expansion.
TES_Fitness	The community is expanded only once based on the fitness function.
TES_Unmerge	Overlapping communities are not detected.

4.2. Benchmark Datasets

In this paper, we compare the performance of the TES algorithm on five real network datasets. The real network datasets are shown in Table 2.

Table 2. Real network datasets.

Datasets	Number of Nodes	Number of Edges	Description
Karate	34	78	Karate club network [48]
Dolphins	62	159	Dolphins social network [49]
Les Miserables	77	508	Les Miserables network [50]
Football	115	616	American college football network [51]
Power	4941	6594	The US power grid network [52]

4.3. Evaluation Criteria

To evaluate the performance of the proposed algorithm, this paper employs extended modularity [38] and overlapping modularity [53] as the evaluation criteria.

The main idea of modularity (Q) is that if a subgraph is a community, the number of edges of its internal nodes is greater than that of a randomly generated subgraph [54].

Unfortunately, the Q function can only be used to evaluate nonoverlapping communities. To evaluate the overlapping community structure, extend modularity (EQ) was proposed based on the Q function [38]. The EQ function is shown as Equation (15).

$$EQ = \frac{1}{2m}\sum_{k=1}^{K}\sum_{v_j,v_j\in C_k}[A_{ij} - \frac{d_i d_j}{2m}]\frac{1}{O_i O_j}, \tag{15}$$

where m is the total number of edges of the network, K is the number of communities discove, d_i is the degree of node v_i, O_i is the number of communities to which node v_i belongs, and A is the adjacency matrix of the network. If there is an edge between v_i and v_j, then $A_{ij} = 1$; otherwise, $A_{ij} = 0$.

Overlapping modularity (Q_{ov}) is another method to evaluate the structure of overlapping communities [53], as shown in Equation (16):

$$Q_{ov} = \frac{1}{m}\sum_{c\in C}^{n}\sum_{i,j\in V}^{n}[A_{ij}\beta_{l(i,j),c} - \frac{\beta^{out}_{l(i,j),c}k^{out}_i\beta^{in}_{l(i,j),c}k^{in}_j}{m}], \tag{16}$$

where m is the total number of edges of the network, A is the adjacency matrix of the network, β is the strength of an edge $l = (i, j)$ which belongs to community C, k^{in}_j is the in-degree of node j, and k^{out}_i is the out-degree of node i.

EQ and Q_{ov} are both in the interval [0,1]. The greater they are, the better the community discovery results will be.

4.4. Parameter Selection

The selection of parameters will affect the results of community discovery. The TES algorithm has two parameters, α and ϵ. According to Equation (7), $\alpha = 1$ is a special value. According to Equation (14), ϵ is in the range of (0,1). Thus, we select α in the range of [0.8, 1.5] and ϵ in the range of [0.1, 0.9], and the step is 0.1. EQ is employed to evaluate the performance. The experimental results are shown in Figures 2 and 3.

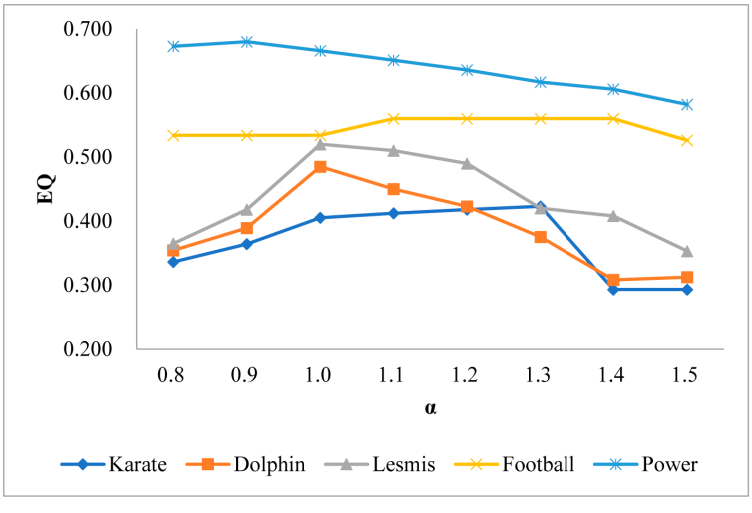


Figure 2. Comparison of extend modularity (EQ) with different α on real networks.

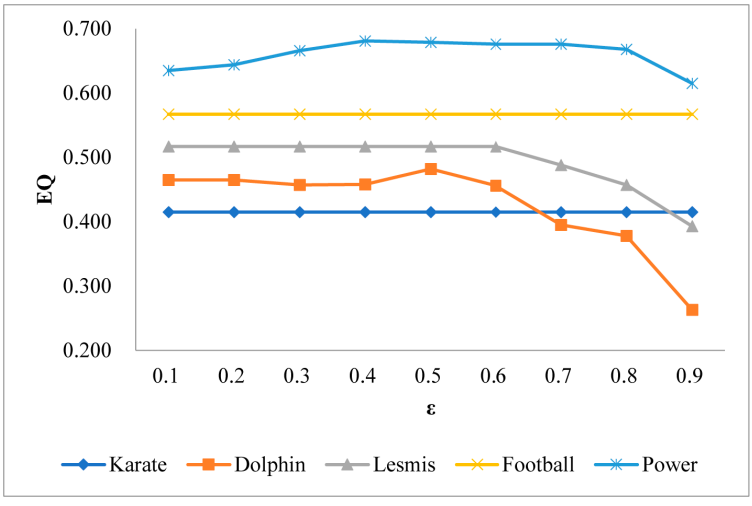


Figure 3. Comparison of EQ with different ϵ on real networks.

Figures 2 and 3 show the trend of EQ along with the increase of parameters α and ϵ, respectively. In general, the influence of α on community discovery is greater than that of ϵ. In Figure 2, it can be seen that with the increase of α, EQ increases first and then decreases. For different networks, the maximum EQ is obtained with different α. The maximum EQ values are achieved at $\alpha = 1.3$ on the Karate and Football networks, $\alpha = 1$ on the Dolphin and Lesmis networks, and $\alpha = 0.9$ on the Power network. From Figure 3, it can be seen that ϵ has little influence on the Karate network and the Football network, but has the greatest impact on the Dolphin network. EQ values of the Dolphin and Lesmis networks are not significantly changed when $\epsilon \in [0.1, 0.5]$. However, when $\epsilon \in [0.5, 0.9]$, EQ decreases rapidly as ϵ increases. All five networks obtain the maximum EQ when $\epsilon = 0.5$.

In conclusion, for five real networks, when $\alpha = 1.3$ and $\epsilon = 0.5$ for the Karate and Football networks, $\alpha = 1$ and $\epsilon = 0.5$ for the Dolphin and Lesmis networks, and $\alpha = 0.9$ and $\epsilon = 0.5$ for the Power network, the optimal community discovery results can be achieved. Therefore, in the rest of this paper, TES selects the above parameters for different networks.

4.5. Performance Evaluation

4.5.1. Module Performance Evaluation

In this subsubsection, we verify that each module has an effect on the improvement of the proposed algorithm. The experiments are carried out on five real networks, and evaluation criteria EQ is selected to evaluate the influence of each module on the TES algorithm. The parameters of TES_Seed, TES_Unclean, TES_Fitness, and TES_Unmerge are the same as those of TES. The experimental results are shown in Table 3. The coverage rates of the nodes with only one expansion and two expansions are calculated, respectively. Therefore, TES_Fitness with one expansion and TES with two expansions are selected. The coverage rates of the two algorithms are reported in Table 4.

From Table 3, it can be seen that all four parts of the TES algorithm have impacts on the TES algorithm and have different influence on different networks. Therefore, TES outperforms the other four algorithms. For example, TES gets 0.675 on Power dataset, which is larger than that obtained by the other four algorithms. According to Equation (15), we know that the greater EQ is, the better the community discovery results will be. The reasons are as follows: It should be noticed that the results of TES_Seed in Table 3 are not robust. The reason is that TES_Seed randomly selects the seed to expand, resulting in different community discovery results. Thus, the results are different even under the same

parameters. Hence, the results of TES_Seed in Table 3 are the average value of 20 times. After the first community expansion based on the fitness function, TES_Unclean does not clean the community. When the community structure changes, there may be negative fitness nodes in the community, which will effect the quality of the community discovery results. TES_Unmerge has the most significant impact on the algorithm, which proves that excessive overlapping between communities has a great impact on community structure.

Table 3. Comparison of *EQ*.

Algorithms	Karate	Dolphin	Lesmis	Football	Power
TES_Seed	0.402	0.413	0.467	0.512	0.490
TES_Unclean	0.411	0.425	0.474	0.511	0.601
TES_Fitness	0.383	0.417	0.482	0.507	0.649
TES_Unmerge	0.380	0.251	0.428	0.449	0.434
TES	**0.417**	**0.482**	**0.517**	**0.560**	**0.675**

Table 4. Comparison of the coverage rate.

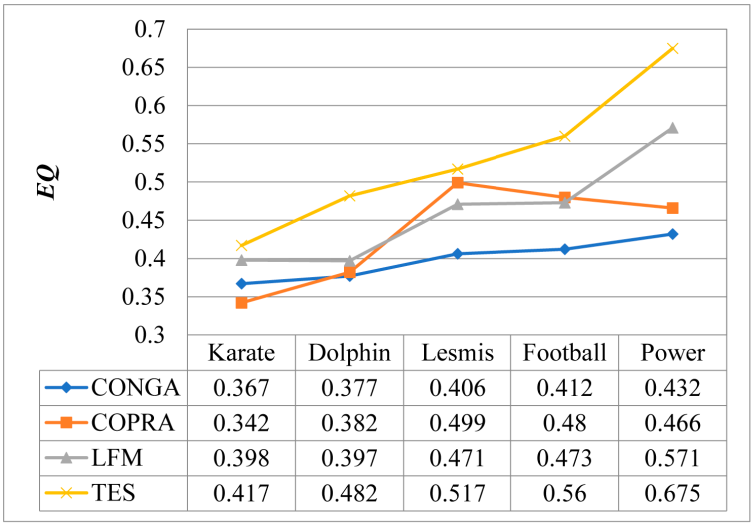
Algorithms	Karate	Dolphin	Lesmis	Football	Power
TES_Fitness	0.94	0.95	0.87	0.76	0.89
TES	1.00	1.00	1.00	1.00	1.00

TES_Fitness expands the community based on the seeds only once, which leads to the decrease of *EQ* and affects the coverage rate of network nodes. From Table 4, we know that the coverage rate of TES_Fitness are all less than 1. The reason is as follows: For complex networks with fewer nodes, the coverage rate of the nodes can be high with only one expansion. However, with the increase of the nodes, the network scale becomes larger and larger, and the coverage rate with only one expansion becomes lower and lower. After two expansions of the community, the TES algorithm can cover all nodes in the network completely, and a high coverage rate of 1.00 can be achieved for a large network such as Power.

Hence, we can safely say that the four parts of the TES algorithm are all very important. The community discovery result is robust since the local maximum node is selected as the seed based on the gravitational degree. Community cleaning can avoid negative fitness nodes when the community structure changes. The natural community can significantly increase the coverage rate of the network nodes through two expansions, and the merging of the overlapping communities can deal with undant communities effectively. The four parts can effectively improve the quality of community discovery.

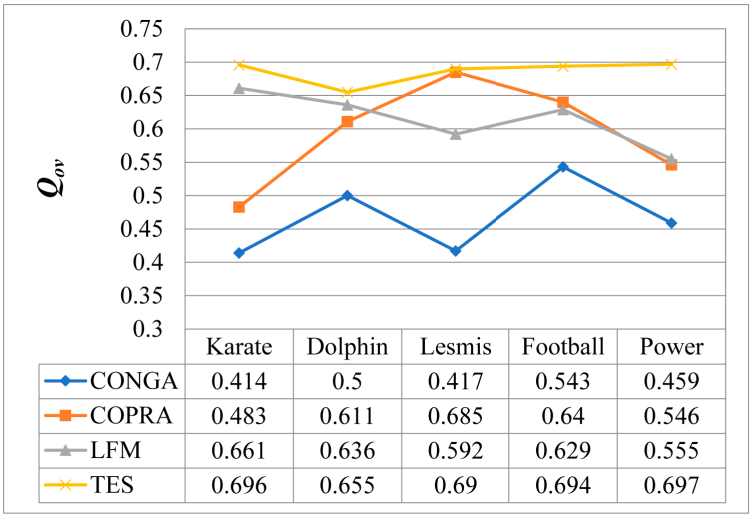
4.5.2. Algorithm Performance Evaluation

To report the performance of the TES algorithm, this paper selects three state-of-the-art algorithms: the CONGA algorithm, based on the splitting method for overlapping community discovery; the COPRA algorithm, based on the label propagation method; and the LFM algorithm, based on local community optimization and expansion. The parameter of CONGA is community number c, which needs to be determined according to the modularity degree function. The parameter of COPRA is the label length v, which is from 2 to 8 with steps of 1. The parameter of LFM is the resolution parameter α, which is from 0.8 to 1.5 with steps of 0.1. For each algorithm, we select the best results as the final results shown in Figures 4 and 5.



	Karate	Dolphin	Lesmis	Football	Power
CONGA	0.367	0.377	0.406	0.412	0.432
COPRA	0.342	0.382	0.499	0.48	0.466
LFM	0.398	0.397	0.471	0.473	0.571
TES	0.417	0.482	0.517	0.56	0.675

Figure 4. Comparison results of EQ on different networks.



	Karate	Dolphin	Lesmis	Football	Power
CONGA	0.414	0.5	0.417	0.543	0.459
COPRA	0.483	0.611	0.685	0.64	0.546
LFM	0.661	0.636	0.592	0.629	0.555
TES	0.696	0.655	0.69	0.694	0.697

Figure 5. Comparison results of overlapping modularity (Q_{ov}) on different networks.

From Figures 4 and 5, TES outperforms all three competitive algorithms since both EQ and Q_{ov} obtained by TES are better than those of the other three algorithms on the five datasets. For example, from Figure 4, we know that EQ of TES is 0.417 on the Karate network, while the other three algorithms are all less than 0.4. Similarly, Q_{ov} of TES is 0.697 on the Power network, while the other three algorithms are all less than 0.56. As we know, the greater EQ and Q_{ov} are, the better the community discovery results will be. Hence, the community discovery results of TES are significantly improved compa with the other three algorithms. The reason is that the natural community discovery is based on local community optimization, and the expansion is only related to the local topology structure of the network, not the global topology of the whole network. Although the LFM algorithm is based on local community optimization, EQ and Q_{ov} values achieved by the LFM algorithm are lower than that of the COPRA algorithm on the Lesmis and Football

networks, but higher than that of the COPRA algorithm on the other three networks. The reason is that the LFM algorithm randomly selects seeds, meaning that the community structure discove is not robust.

In summary, TES has better performance than all competing algorithms.

4.6. Case Study

To further clarify the performance of TES, the Karate network is employed to show the community discovery results. Figure 6 shows the community discovery results obtained by the TES algorithm. The seeds are nodes 1, 17, 26, and 34, and four communities are obtained. Node 10 is the overlapping node of the grey and yellow communities.

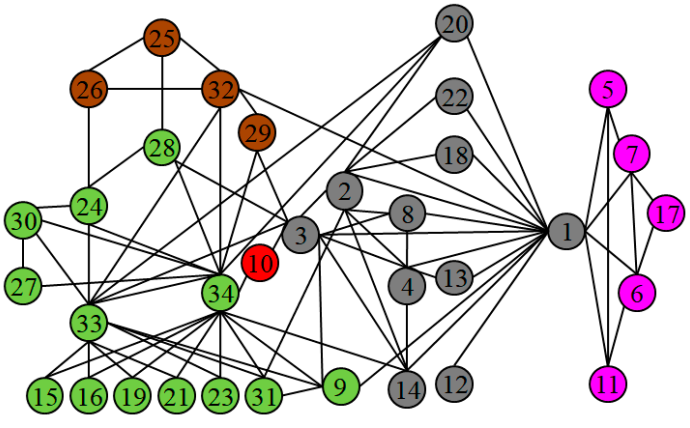


Figure 6. Community discovery result by TES algorithm on the Karate network. The seeds are nodes 1, 17, 26, and 34, and four communities are obtained. Node 10 is the overlapping node of the gray and yellow communities.

It can be seen from Figure 6 that the TES algorithm finds four communities, while the CONGA and COPRA algorithms both discover two communities and the LFM algorithm discovers five communities. Compa with the CONGA and COPRA algorithms, the partition of the network by TES algorithm is more detailed. For example, community {5,6,7,11,17} is closely related to node 1, but the nodes inside community {5,6,7,11,17} have stronger connection relationship with each other. The TES algorithm can mine small communities in large-scale communities, mainly because in the first part of the algorithm, the center node with strong information transmission ability is taken as the seed. Although the LFM algorithm discovers five communities, the seed does not have centrality since the LFM algorithm randomly selects seeds. The expanded community structure locality is poor, and a community is included in another community. The reason for this kind of situation is that the LFM algorithm does not detect the merged undant community, which illustrates the importance of detecting overlapping community in the TES algorithm.

5. Conclusions

In this paper, we propose an overlapping community discovery algorithm, named TES, which has three parts. In the first part, the local maximum node is taken as the seed based on the gravitational degree. The second part discovers the natural community by two expansions. The community is expanded based on the fitness function. After adding a new node, the community is cleaned. The second expansion is based on the gravitational function. The third part examines and merges the overlapping communities. To verify the reasonability of these parts, four comparative algorithms, TES_Seed, TES_Unclean, TES_Fitness, and TES_Unmerge, are proposed. Besides these four algorithms, three state-

of-the-art algorithms: CONGA, COPRA, and LFM, are employed. Experimental results on five real networks report that TES outperforms all these competitive algorithms.

Author Contributions: Conceptualization, Y.L. and Y.W.; methodology, Y.L. and J.H.; validation, J.H., Y.L. and Y.W.; investigation, R.L.; writing—original draft preparation, J.H.; writing—review and editing, Y.L. and Y.W.; supervision, Y.W. and R.L.; funding acquisition, R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Social Science Fund of China under grant number 18BGL191.

Informed Consent Statement: Informed written consent was obtained from the authors for publication of this paper.

Data Availability Statement: Data was obtained from http://www-personal.umich.edu/mejn/netdata/.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this article.

References

1. Gu, K.; Wang, L.; Yin, B. Social community detection and message propagation scheme based on personal willingness in social network. *Soft Comput.* **2019**, *23*, 6267–6285. [CrossRef]
2. He, J.; Liu, H.; Zheng, Y.; Tang, S.; He, W.; Du, X. Bi-labeled LDA: Inferring interest tags for non-famous users in social network. *Data Sci. Eng.* **2020**, 5, 27–47. [CrossRef]
3. Li, Y.; Zhang, H.; Zhu, H.; Li, J.; Yan, W.; Wu, Y. IBAS: Index based A-star. *IEEE Access* **2018**, *6*, 11707–11715. [CrossRef]
4. Dolgorsuren, B.; Xu, W.; Khan, K.U.; Jeong, B.S.; Lee, Y.K. SP2: Spanner construction for shortest path computation on streaming graph. In Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory, Jeju Island, Korea, 17–19 October 2016; pp. 43–50.
5. Batjargal, D.; Khan, K.U.; Lee, Y.K. EM-FGS: Graph sparsification via faster semi-metric edges pruning. *Appl. Intell.* **2019**, *49*, 3731–3748. [CrossRef]
6. Wu, Y.; Tong, Y.; Zhu, X.; Wu, X. NOSEP: Nonoverlapping sequence pattern mining with gap constraints. *IEEE Trans. Cybern.* **2018**, 48, 2809–2822. [CrossRef]
7. Hai, M.; Li, H.; Ma, Z.; Gao, X. Algorithm for detecting communities in complex networks based on Hadoop. *Symmetry* **2019**, *11*, 1382 [CrossRef]
8. Shi, Q.; Shan, J.; Yan, W.; Wu, Y.; Wu, X. NetNPG: Nonoverlapping pattern matching with general gap constraints. *Appl. Intell.* **2020**, *50*, 1832–1845. [CrossRef]
9. Wu, Y.; Shen, C.; Jiang, H.; Wu, X. Strict pattern matching under non-overlapping condition. *Sci. China Inf. Sci.* **2017**, *60*, 012101. [CrossRef]
10. Bu, Z.; Li, H.J.; Zhang, C.; Cao, J.; Li, A.; Shi, Y. Graph k-means based on leader identification, dynamic game and opinion dynamics. *IEEE Trans. Knowl. Data Eng.* **2019**. [CrossRef]
11. Li, H.J.; Bu, Z.; Wang, Z.; Cao, J.; Shi, Y. Enhance the performance of network computation by a tunable weighting strategy. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *2*, 214–223. [CrossRef]
12. Chen, D.; Fu, Y.; Shang, M. An efficient algorithm for overlapping community detection in complex networks. In Proceedings of the WRI Global Congress on Intelligent Systems, Xiamen, China, 19–21 May 2009; pp. 244–247.
13. Atzmueller, M.; Doerfel, S.; Mitzlaff, F. Description-oriented community detection using exhaustive subgroup discovery. *Inf. Sci.* **2016**, *329*, 965–984. [CrossRef]
14. Geng, X.; Lu, H.; Sun, J. Network structural transformation-based community detection with autoencoder. *Symmetry* **2020**, *12*, 944. [CrossRef]
15. Chen, J.; Liu, M.; Liu, X. Research on of overlapping community detection algorithm based on tag influence. *Clust. Comput.* **2019**, *22*, 6669–6679. [CrossRef]
16. Fortunato, S.; Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **2016**, *659*, 1–44. [CrossRef]
17. Javed, M.A.; Younis, M.S.; Latif, S.; Qadir, J.; Baig, A. Community detection in networks: A multidisciplinary review. *J. Netw. Comput. Appl.* **2018**, *108*, 87–111. [CrossRef]
18. Xie, J.; Kelley, S.; Szymanski, B.K. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Comput. Surv.* **2013**, *45*, 1–35. [CrossRef]
19. Guerrero, M.; Gil, C.; Montoya, F.G.; Alcayde, A.; Baños, R. Multi-objective evolutionary algorithms to find community structures in large networks. *Mathematics* **2020**, *8*, 2048. [CrossRef]
20. Li, Y.; He, K.; Kloster, K.; Bindel, D.; Hopcroft, J. Local spectral clustering for overlapping community detection. *ACM Trans. Knowl. Discov. Data* **2018**, *12*, 17. [CrossRef]
21. Chen, Y.; Shi, S.; Chen, G.; Yu, Z. Overlapping community discovery based on node hierarchy and label propagation gain. *Pattern Recognit. Artif. Intell.* **2015**, *28*, 289–298.

22. Liu, H.; Ling, H.; Jian, J.; Chen, L. Overlapping community discovery algorithm based on hierarchical agglomerative clustering. *Int. J. Pattern Recognit. Artif. Intell.* **2015**, *32*, 1850008. [CrossRef]
23. Xu, M.; Li, Y.; Li, R.; Zou, F.; Gu, X. EADP: An extended adaptive density peaks clustering for overlapping community detection in social networks. *Neurocomputing* **2019**, *337*, 287–302. [CrossRef]
24. Guerrero, M.; Baños, R.; Gil, C.; Montoya, F.G.; Alcayde, A. Evolutionary algorithms for community detection in continental-scale high-voltage transmission grids. *Symmetry* **2019**, *11*, 1472. [CrossRef]
25. Li, Y.; Wang, J.; Wang, X.; Zhao, Y.; Lu, X.; Liu, D. Community detection based on differential evolution using social spider optimization. *Symmetry* **2017**, *9*, 183. [CrossRef]
26. Sun, H.; Liu, J.; Huang, J.; Wang, G.; Jia, X.; Song, Q. LinkLPA: A link-based label propagation algorithm for overlapping community detection in networks. *Comput. Intell.* **2017**, *33*, 308–331. [CrossRef]
27. Ahn, Y.; Bagrow, J.; Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **2010**, *466*, 761. [CrossRef]
28. Palla, G.; Derenyi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, 435, 814–818. [CrossRef] [PubMed]
29. Gregory, S. Finding overlapping communities in networks by label propagation. *New J. Phys.* **2010**, *12*, 2011–2024. [CrossRef]
30. Kianian, S.; Khayyambashi, M.; Movahhedinia, N. Semantic community detection using label propagation algorithm. *J. Inf. Sci.* **2015**, *42*, 166–178. [CrossRef]
31. Lancichinetti, A.; Fortunato, S.; Kertész, J. Detecting the overlapping and hierarchical community structure of complex networks. *New J. Phys.* **2008**, *11*, 19–44. [CrossRef]
32. Coscia, M.; Rossetti, G.; Giannotti, F.; Pedreschi, D. DEMON: A local-first discovery method for overlapping communities. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 615–623.
33. Cheraghchi, H.S.; Zakerolhosseini, A. Toward a novel art inspi incremental community mining algorithm in dynamic social network. *Appl. Intell.* **2017**, *46*, 409–426. [CrossRef]
34. Yang, L.; Cao, X.; He, D.; Wang, C.; Wang, X.; Zhang, W. Modularity based community detection with deep learning. In Proceedings of the International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 2252–2258.
35. Yang, L.; Cao, X.; Jin, D.; Wang, X.; Meng, D. A unified semi-supervised community detection framework using latent space graph regularization. *IEEE Trans. Cybern.* **2017**, *45*, 2585–2598. [CrossRef] [PubMed]
36. Gregory, S. An algorithm to find overlapping community structure in networks. In In Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, 17–21 September 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 91–102.
37. Maryam, H.; Kamran, Z.; Ahmad, R. A community-based approach to identify the most influential nodes in social networks. *J. Inf. Sci.* **2017**, *43*, 204–220.
38. Shen, H.; Cheng, X.; Cai, K.; Hu, M. Detect overlapping and hierarchical community structure in networks. *Phys. A Stat. Mech. Appl.* **2009**, *388*, 1706–1712. [CrossRef]
39. Gleich, D.; Seshadhri, C. Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 597–605.
40. Whang, J.; Gleich, D.; Dhillon, I. Overlapping community detection using seed set expansion. In Proceedings of the 22nd ACM International Conference on Information Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 2099–2108.
41. Su, C.; Guan, X.; Du, Y.; Wang, Q.; Wang, F. A fast multi-level algorithm for community detection in directed online social networks. *J. Inf. Sci.* **2018**, *44*, 392–407. [CrossRef]
42. Lee, C.; Reid, F.; Mcdaid, A.; Hurley, N. Detecting highly overlapping community structure by greedy clique expansion. In Proceedings of the fourth SNA-KDD Workshop on Social Network Mining and Analysis, Washington, DC, USA, 25 July 2010.
43. Cai, G.; Wang, R.; Liu, G. Hierarchical overlapping community discovery algorithm based on node purity. In Proceedings of the International Conference on Intelligent Information Processing, Haikou, Hainan, China, 14–15 July 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 248–257.
44. Liu, J.; Du, Y.J.; Li, Q.; Fu, C. Social community evolution by combining gravitational relationship with community structure. *Intell. Data Anal.* **2018**, *22*, 1143–1161. [CrossRef]
45. Li, Y. A new vertex similarity metric for community discovery: A distance neighbor model. In *Asian Conference on Intelligent Information and Database Systems*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 228–237.
46. Wu, Y.; Zhu, C.; Li, Y.; Guo, L.; Wu, X. NetNCSP: Nonoverlapping closed sequential pattern mining. *Knowl. Based Syst.* **2020**, *196*, 105812. [CrossRef]
47. Chen, J.; Zhou, G.; Nan, Y.; Zeng, Q. Semi-supervised local expansion method for overlapping community detection. *Comput. Res. Dev.* **2016**, *53*, 1376–1388.
48. Zachary, W.W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **1997**, *33*, 452–473. [CrossRef]
49. Lusseau, D.; Schneider, K.; Boisseau, O.J.; Haase, P.; Slooten, E.; Dawson, S.M. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **2003**, *54*, 396–405 . [CrossRef]
50. Knuth, D.E. *The Stanford GraphBase: A Platform for Combinatorial Computing*; ACM Press: New York, NY, USA, 1993.

51. Girvan, M.; Newman, M.E.J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821–7826. [CrossRef]
52. Watts, D.J.; Strogatz, S.H. Collective dynamics of small-world networks. *Nature* **1998**, *93*, 440–442. [CrossRef] [PubMed]
53. Nicosia, V.; Mangioni, G.; Carchiolo, V.; Malgeri, M. Extending the definition of modularity to directed graphs with overlapping communities. *J. Stat. Mech. Theory Exp.* **2009**, *3*, 3166–3168. [CrossRef]
54. Newman, M.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113. [CrossRef] [PubMed]

symmetry

MDPI

Article

Symmetrical Model of Smart Healthcare Data Management: A Cybernetics Perspective

Wajdi Alhakami [1], Abdullah Baz [2], Hosam Alhakami [3], Abhishek Kumar Pandey [4] and Raees Ahmad Khan [4,*]

[1] Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; whakami@tu.edu.sa
[2] Department of Computer Engineering, College of Computer and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia; aobaz01@uqu.edu.sa
[3] Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia; hhhakam@uqu.edu.sa
[4] Department of Information Technology, Babasaheb Bhimrao Ambedkar University, Lucknow 226025, India; abhishekkumarpanday5@gmail.com
* Correspondence: khanraees@bbau.ac.in or khanraees@yahoo.com

Received: 19 November 2020; Accepted: 11 December 2020; Published: 16 December 2020

Abstract: Issues such as maintaining the security and integrity of data in digital healthcare are growing day-by-day in terms of size and cost. The healthcare industry needs to work on effective mechanisms to manage these concerns and prevent any debilitating crisis that might affect patients as well as the overall health management. To tackle such critical issues in a simple, feasible, and symmetrical manner, the authors considered the ideology of cybernetics. Working towards this intent, this paper proposes a symmetrical model that illustrates a compact version of the adopted ideology as a pathway for future researchers. Furthermore, the proposed ideology of cybernetics specifically focuses on how to plan the entire design concept more effectively. It is important for the designer to prepare for the future and manage the design structure from a product perspective. Therefore, the proposed ideology provides a symmetric mechanism that includes a variety of estimation and evaluation techniques as well as their management. The proposed model generates a symmetric, variety-issue, reduced infrastructure that can produce highly effective results due to an efficient usability, operatability, and symmetric operation execution which are the benefits of the proposed model. Furthermore, the study also performed a performance simulation assessment by adopting a multi-criteria decision-making approach that helped the authors compare the various existing and proposed models based on their levels of effectiveness.

Keywords: healthcare data; data management; digital services; cybernetics; symmetrical designing

1. Introduction

Management of smart services is a challenging and crucial task, especially in the current scenario of an unprecedented surge in the number of cyber-attacks. As cited by several studies, digital healthcare services are major targets of these cyber invasions and have been severely affected [1]. Hence, the management and security of smart healthcare facilities that directly depend on digital infrastructure have become essential. In an environment where cybercrime is on the rise, security and healthcare experts must work on more productive and efficacious mechanisms to strengthen smart services [1–4].

There is a need for the healthcare sector to define its digital transactions and infrastructure from a new perspective. In this line, several experts and researchers iterate on the techniques and methodologies that are in use at present, which subscribe to universally accepted standards such as the Health Insurance Portability and Accountability Act (HIPPA), and these techniques are perfectly

updated and secure for digital and smart healthcare infrastructure. However, the picture drawn by the cyber-attack census report is alarming, especially in the context of healthcare exploitation incidents. The figures clearly show that the policies described in these standards are not adequate in terms of security and failure management. Thus, the authors strongly believe that such a scenario calls for reconstructing or developing symmetric steps from different perspectives to ensure security and digital maintenance in healthcare infrastructure.

In addition, the present state of the pandemic that has arisen all over the world has exposed the fault lines that exist and are cropping up in smart healthcare services. In the event of emergency and massive cases, digital healthcare services have been found to be inadequate [5–10]. Managing such a major health catastrophe with foolproof security is a difficult and challenging task for healthcare organizations. Thus, the prevailing uncertainty in the present context has motivated the authors to work on redefining the healthcare infrastructure by identifying the possible drawbacks in the existing healthcare infrastructure. The research study conducted in this context cites the various ways cyber invaders are able to exploit the vulnerabilities present in the healthcare infrastructure [11–15]. These exploit possibilities are present in both the organizational assets and its structure.

Furthermore, conceptualizing these vulnerability points indicates that "variety" is the main difficulty in the management of healthcare infrastructure [16–20]. Variety, as a concept, means different states of the same thing. This can be explained with the analogy of a smart room heater. If a heater were automatically set to manage the maximum temperature of 5 °C but suddenly the temperature fell below 5 °C, then the heater would not function properly because it had not been configured to function in temperatures below 5 °C. This example clearly displays that variety is always an issue in any type of system because every system has its limitations, but the application of the systems has diversity in nature.

Additionally, to bridge the lack of understanding about this issue in the healthcare sector, the authors present a brief descriptive analysis of the issue of variety in healthcare. The study also proposes a conceptual framework that can be adopted for managing the healthcare infrastructure symmetrically. For the envisioned model, the authors adopted the cybernetics concept that has significant potential in providing effective solutions in healthcare service management.

2. Materials and Methods

2.1. Previous Similar Research Initiatives

Defining any new concept and methodology is a challenging task. It requires an extensive review of the literature and critical analysis of domain centric sources. To understand the scenario of healthcare variety issues and devise a conceptual model for a symmetric smart healthcare system by adopting cybernetics concepts in healthcare, the authors perused the relevant research initiatives completed in this field. The literature search process and information about the literature examination is illustrated in Appendix A. Despite a thorough search, the authors found only a few partially similar papers that discuss healthcare modeling and its management. There are no specific studies on the management of healthcare data through cybernetics ideology, which is the key objective of the proposed study.

The core concept and idea of the proposed work is to portray a symmetric model that would enable experts to make healthcare infrastructure more symmetric and secure it from variety-related issues. To achieve this goal, the authors adopted the cybernetics approach. However, after conducting the initial search, the authors concluded that there was no similar research work available in data repositories. Another important aspect that needs to be underlined in this context is that cybernetics is not an approach or technique with predefined symmetric steps. An ideology of thinking processes gives benefit in designing development steps. We used cybernetics in this form in this study, whereas several studies we refer to only align cybernetics with engineering and some type of system development. In addition, various other methods and techniques are available which specifically focus on the healthcare domain. Nevertheless, there is a need to redefine the complete healthcare data management

due to its complexity and vastness. Before proposing any specific technique that would solve any specific healthcare loopholes, it is necessary to define the complete healthcare data management from a novel design perspective. This type of approach gives an ideal pathway for creating a systematic and a simple system. The proposed study aimed to achieve this goal by portraying a systematic ideology for future researchers.

Some of the relevant studies that can be cited in the context of adapting cybernetics ideology are mentioned below:

Korotkova's article discusses the role of IT and cybernetics in Russian healthcare. The article describes the current situation and the scope of cybernetics in healthcare with the help of fundamental analysis and examples [4]. The main focus of the article is clearly on the evolution of information technology in the country's healthcare infrastructure. Korotkova stated that health is a domain which is never going to end because if humans are present on the earth then they will get ill or suffer from any disease because it is the basic nature of the human body. The researcher further adds that information technology association in healthcare is also going to be a big revolution because every aspect of daily life and business sectors are very frequently adopting computers. The author adds that this balance can be achieved by applying cybernetics ideology in between digitalization and healthcare.

Khayal et al. worked on personalized health service modeling. The proposed model in the paper uses engineering dynamic theories to portray organizational and personal healthcare needs of patients from personalized level [5]. The approaches and methods used in the paper have the ability to provide beneficial results in the healthcare sector. This study also motivated us to develop a symmetrical model for the healthcare sector.

Further, Faggini et al. proposed a model that was developed to maintain sustainability in healthcare by various digital infrastructures. The paper proposed a theoretical model named DocBox24, which is based on the online sustainable healthcare service delivery [18]. To validate the work of their study, the authors also portrayed real time examples and provided a comparison analysis based on various facts. The study illustrated the power and significance of digital infrastructure in healthcare very effectively.

To manage the digital infrastructure and data management of healthcare, another study discusses the blockchain based secure data management and travel in an Internet of Things (IoT) environment [19]. In the current era of digitalization, the demand and significance of this type of methodology and work is very high. The work portrays a blockchain-based model that discusses secure communication and works on all types of data layers in healthcare.

Moreover, another researcher, Yang et al., worked on data security and its validation in healthcare. The study assesses a data validation scenario and then proposes an effective and efficient security assured model, which deals with data validation issues in healthcare [20]. The findings of the study are a significant contribution to the research being conducted in this domain.

A technical report discusses the use of sociocybernetics in health management [6]. Though the application of cybernetics discussed in this paper is related to the social perspective, we were able to relate it to the scope of cybernetics in healthcare.

2.2. Symmetric Variety: A Significant Topic in Healthcare

Issues are a common and a frequent occurrence in healthcare infrastructures. However, inspite of being a critical aspect, the issue of variety has not been addressed or raised emphatically by the research community yet. Admittedly, exceptions are always available in society, so the authors believe that there might be some research teams that have worked previously, or are working, on the variety-related issues of healthcare. Variety, as a concept, is also used for addressing the failure of the infrastructure or breaches in healthcare [7].

From a general perspective though, variety is considered as the diversity of any system, process or product for a certain state of domain. This means that every resource of any system has its own variety or diversity. For example, an automated room heater has various resources, such as atmosphere changes,

room temperature, electronics of machines, etc. Every resource has its own diversity, for example, the atmosphere has a nature of changing, so it could be anything after a period of 1 h. Temperature can also be changed, depending on the atmospheric condition.

In this framework, the key question is whether the issues pertaining to variety can have any adverse effects. Most certainly, the issue of variety can cause exploitation and failure situations in any type of system and process. To understand this more clearly, let us allude to an example of a healthcare organization or a hospital that has a capacity of 100 beds. The digitalized infrastructure that is implanted in the organization also has the data capacity of carrying information of 100 patients. Now in a panic situation such as COVID-19 spread, if the patients count in healthcare organizations becomes more than 100, the infrastructure would not be able to maintain the data structure properly. However, consider that every system has some extra space to carry resources, so the infrastructure would also have the capacity of carrying approximately 200 patients' data at a given time. More specifically, in the context of COVID-19, when the count of COVID-19 positive cases in many countries of the world is not less than 5000 per day, unless the infrastructure is engineered to tackle variety related issues, it would be unable to carry the extra burden. Evidently, the storage capacity is modified in variety and its adverse effect is digital failure.

Moreover, to build a strong research base for the issue of variety, authors found a study that discussed about the attributes that affect the quality of healthcare. This study also illustrated the issue of variety in their work. The work discusses about the quality issues that happen in any healthcare scenario and the factors that lead to this situation [15]. According to the study, management and perfection from every end in the healthcare system is necessary for a faultless process. There are various factors such as resource disturbances, delay in availability, interruption in digital environment, etc., that portray a situation of failure or exploitation in the system. These factors simply illustrate a specific situation of fault or vulnerability, but the analysis of the causes behind these factors clearly shows that these issues are also associated with the issue of variety in healthcare.

Hence, the issue of variety is a critical aspect in healthcare. It is imperative for the research community to focus more intently on mechanisms that can provide effective solutions in this domain. Furthermore, to gauge the broader picture of this situation in real time scenarios, the following studies that mention infrastructure failures that have happened in the healthcare infrastructure in the recent past have been cited below:

Another technical report discusses that the systems failure in healthcare is a disaster for structural activities and services [7]. The article also focuses on the significance of diversity or variety in healthcare as well as its adverse effects. The non-stop increase in population is one of the biggest variety changes and a critical issue in healthcare. The article cites that capacity has a significant role in variety related issues. Such issues demand extra attention in healthcare infrastructures.

A news report talks about the weak points of healthcare infrastructure that were exposed to the public during the pandemic situation of COVID-19 [8]. The report shows that there are many infrastructural bug points that can cause a failure of any healthcare structure. Most of these issues, directly or indirectly, stem from the issue of variety.

A report from India points out that an increasing number of cases have already caused healthcare infrastructure failures in many cities in India [9]. The report highlights that Maharashtra, a state of India, had more than 100,000 Corona cases until the time the report was drafted. Moreover, the official figure for the number of beds in the government hospitals of Maharashtra is approximately 52,000. However, these data are for regular, non- pandemic situations. In the wake of a health emergency such as the COVID-19 pandemic, the situation would worsen because the health infrastructure of the hospitals would not be able to cater to the increased number of COVID-19 cases.

In addition to discussing the situation and its possible adverse effects on healthcare, it is also important to understand the scenario from a digital perspective. The abovementioned incidents and examples do not discuss the digital perspective of healthcare. However, it is important to understand that the incidents that are happening and affecting healthcare infrastructure physically, or from any

other perspective, are also affecting the digital infrastructure. The reason for this is that all the processes that are applicable in healthcare organizations, and the infrastructures that are employed, are implemented by digital platforms. This can be done only when the services are smart or digitalized. Physical or any other type of variety issue, directly or indirectly, affects the digital infrastructure.

In the context of analyzing other issues that impede the efficacy of smart healthcare services, the authors have recently presented the Census of Cyber Attacks. The data depict the pattern in which the attackers are penetrating the healthcare infrastructure to exploit various inherited vulnerability points for implementing attacks. A report on the financial information about healthcare breaches tells that the cost of a breach in healthcare is around USD 1.8 million [10]. This is a huge figure. Another survey reports that 53% of the healthcare organizations are currently suffering from breach incidents all around the world. To make this more specific and simple, we have enlisted only July's statistics [11] of healthcare data breaches in Figure 1.

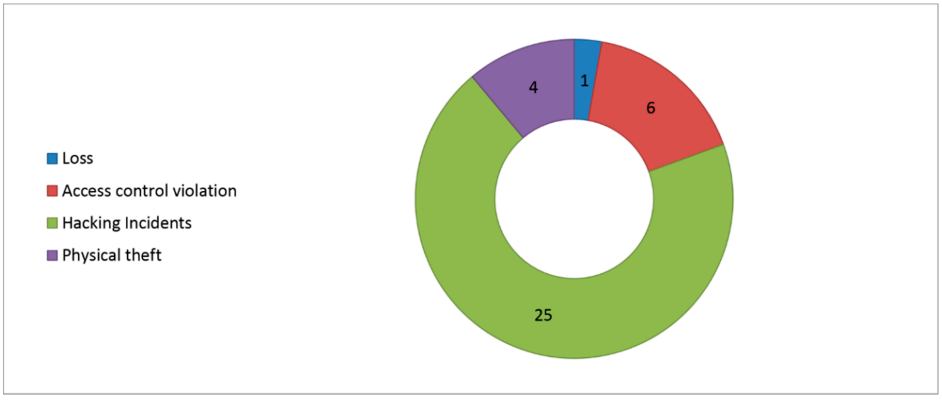


Figure 1. Health Record Breach Census (July 2020).

Figure 1 portrays the number of highly sensitive healthcare breach incidents that have been reported by the host agencies in the past. The analysis of the Census clearly shows that most of the breach incidents were performed and deployed by hacking. This shows the possibilities of exploitation and the loopholes in digital infrastructure of healthcare organizations. These types of incidents necessitate redefining the management of data and digital infrastructure in healthcare.

2.3. Tools and Techniques

Healthcare infrastructure is an association of various components and it has vast complexity in its various attributes. Thus, in order to acquire a method and the materials for redefining the digital infrastructure of healthcare, authors needed to focus on a large domain. The authors found that it is very complex to address all these attributes in one paper because the paper, as well as the concept, would be hard to understand. Therefore, the proposed paper addresses the infrastructure of healthcare from a data perspective.

2.4. Data Perspective Based Symmetrical Modeling of Healthcare Infrastructure

Defining any system and architecture is challenging, more so if it comprises a large amount of attributes and number of domains. The healthcare infrastructure is the most complex and difficult to understand system, particularly in the current situation. This is because several techniques, approaches and ideas are employed by an uncountable number of experts and researchers across the world at present. In such a scenario, it was difficult for the authors to portray a unified infrastructure and

incorporate every aspect of healthcare in one paper. To tackle this premise, the authors have defined the healthcare infrastructure from a data perspective in this paper.

Modeling healthcare from a data perspective enabled the authors to simplify every aspect and the complexity of healthcare. To understand the significance of data perspective modeling, it is imperative to pay attention to the changes that have been brought about by the current digitalized era. Every business and sector that is digitally connected is producing data and operating on the basis of data. In such a context, layering healthcare infrastructure from a data perspective would be beneficial for the research community and the authors of the proposed paper.

Thus, we have analyzed the data usage in healthcare for defining the healthcare model from a data perspective. Identifying data usage gives us an idea about how the healthcare data works and is produced or managed. As per the analysis, healthcare infrastructure can be categorized into a four layer model from the data perspective. The layers are: a data production layer, data transaction layer, data storage layer and data application layer. These layers handle the entire data management in healthcare infrastructure by assuring their layered work. A detailed description for better understanding is given in the ensuing paragraphs and a graphical representation of the same is given in Figure 2.

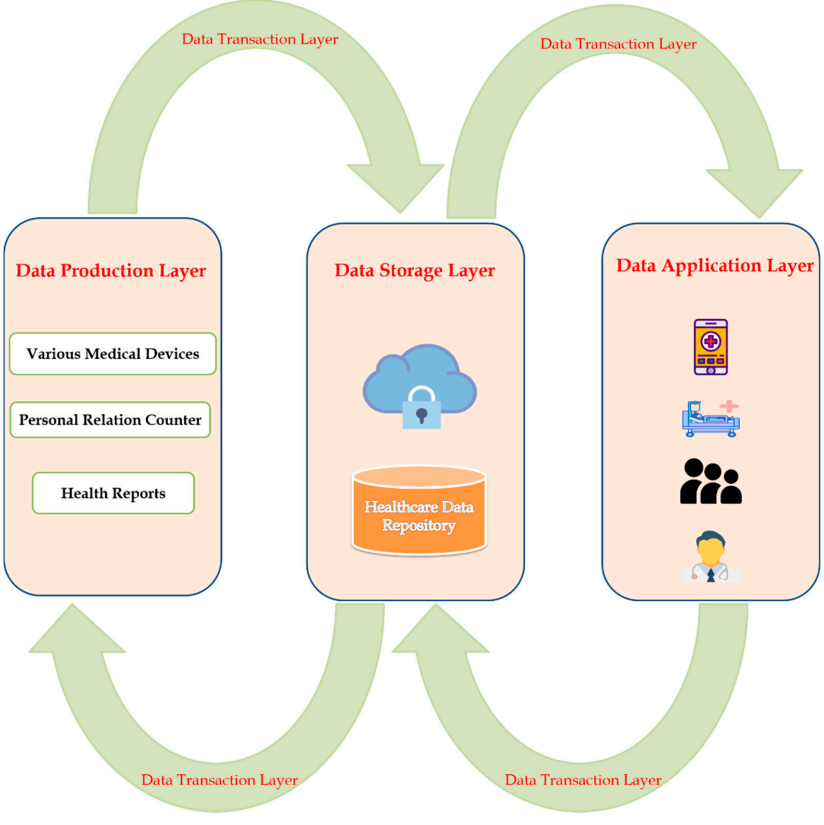


Figure 2. Layered Modeling of Healthcare Infrastructure.

The data production layer: this layer deals with the origin of healthcare data [1]. In simpler words, a data production layer generates or produces data by various types of its attributes such as personal relation counters where new patients come and discuss their previous and current medical history for obtaining the relevant medical assistance. Various medical devices are the next attribute

that include IoMT (Internet of Medical Things) and many other medical digital equipment that give health and medical information about the patient. Health reports generated from various types of tests also come under the production layer, which includes medical condition, and numerical data that are very sensitive for the patients' health. The data that come under this layer are generated or acquired by the healthcare organizations for the treatment of the patients.

The data storage layer: this layer deals with data storing techniques and their management. All types of data that are used and acquired from healthcare services are stored and managed by this layer through various databases and cloud platforms [1]. This layer is responsible for managing data storage and its security. It is imperative for the healthcare security experts to employ assured security on this layer because it is the second most penetrated layer by the attackers.

The data application layer: this layer deals with the post data usage domains. This implies that the application layer is responsible for the use of healthcare data from its different platforms, processes and vectors [1]. Different telemedicine projects such as health apps and SMS services regarding lab tests are considered in this layer of healthcare. Confidential data related to patients and healthcare services that can be accessed by various doctors and administrative bodies also come under this layer because they are directly using the acquired or stored information in healthcare organizations. This is the most vulnerable layer as it is frequently penetrated upon and exploited by the attackers.

The data transaction layer: transaction is associated with traffic or interchange in-between one place or node with another. Hence, as evident from its name, the data transaction layer is associated with various techniques and methods as well as ways of data travel in between the organizational infrastructure [1]. This type of layer often becomes exposed and unsecured in a non-managed healthcare organizational structure. Moreover, considering the possibilities of exploitation, the security of this layer requires approaches such as the blockchain and psudomation. Overall, this layer solely deals with the data that are transmitted from one node to another in the healthcare infrastructure.

The given backdrop as regards the layered view of the healthcare organization and its infrastructure from a data perspective, makes it easier to solve the variety issue in healthcare by adopting the cybernetics ideology.

2.5. Adopted Symmetrical Ideology

Selecting an approach and the idea for the intended research work is a very significant task. It becomes even more critical when the research subject pertains to an extra sensitive domain such as healthcare and working on its redefinition process. That is why the authors have adopted an *ideology of thinking* instead of adopting a specific approach or technique. The authors selected cybernetics ideology for redefining healthcare data's management infrastructure.

Cybernetics is an ideology or *path of thinking* for design structures and defining them [12]. The word cybernetics originated from a Greek ancient word called Kybernetike which means governance. However, many researchers and scientists strongly believe that the actual meaning of the word is "*steering*". More specifically, the term is defined as "*the art of steering*". This is the key point around which the whole concept of second order cybernetics works.

The authors adopted the cybernetics ideology called the second order cybernetics. Second order cybernetics is the next generation or version of cybernetics ideology that gives solutions to the problems that are wicked [13]. Wicked problems are defined as problems that are hard or impossible to solve completely in the system. Thus, for tackling these situations, a next second order cybernetics version was introduced by the researchers to minimize the negative effects of the problems. This type of ideology can be applied in various real time situations and problems.

The second order cybernetics clearly reflects the ideology that is based on the art of steering. We applied this concept on redefining healthcare. Further, to understand the idea of steering, it is important to assume that every design process or idea that the designers envision is done with a desired outcome or result or purpose of design in their mind. The art of steering or cybernetics provides them

the ability to produce the desired outcomes by associating conversation and feedback loops to tackle obstacles in the design process [13].

As mentioned in Figure 3, there is a process of designing which can be achieved simply by some desired steps represented by the green straight line. However, the cybernetics ideology believes that there are some obstacles that misguide the designers from the desired path and move the process towards a different line. The cybernetics ideology empowers the designers to reduce or correct this obstacle and achieve the desired objective through conversation and feedback loop applications. Conversation and feedback loops are processes that work on action–reaction rules.

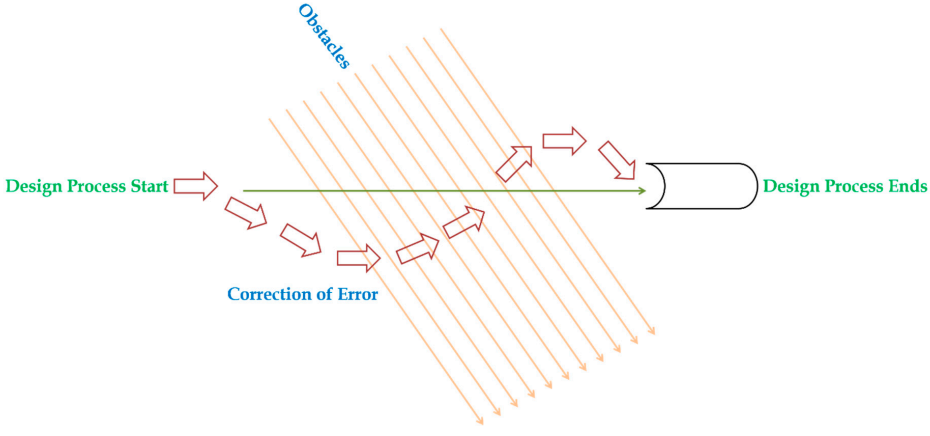


Figure 3. Cybernetics Ideology in Conceptual Way.

To explicate further, there is a simple and a proper path for achieving a goal that is necessary for the designers to follow so as to achieve the intended objective in a design process. However, maximum design processes are influenced by various obstacles that work as distractions and the precise goal is lost. Hence, to avoid any deviations, conversation based feedback loops provide continuous comparison options in a design process. This gives additional awareness to the designers about their path of designs and model.

The biggest and the most significant challenge for the designer during the framing of the design process is to understand the context along with the available attributes and their relation or use to achieve the desired goal in the end. A famous cybernetician, Rittel, states that conversation is always the best option and solution for improving, discussing and examining the design goal [13]. A great and iconic cybernetics genius in history, Ross Ashby, quoted in his book "An Introduction to Cybernetics" that cybernetics is not a concept that deals with "*what is this thing*" in design process, rather than that, it deals with "*what does it do*" [14]. In simple words, cybernetics teaches the way of proper designing instead of teaching a specific technique or approach. Therefore, it is always beneficial for a designer to understand the soul of designing instead of understanding a specific approach.

Furthermore in the context of healthcare, the authors have applied the conversational feedback loop based structuring of models that can be employed at any layer of the healthcare infrastructure. Cybernetics based models that can be achieved by this adopted ideology will help the designers to develop a more systematic healthcare infrastructure by minimizing the issue of variety and various other issues that are associated with the cybernetics concept. The adopted ideology is a way of achieving things by applying true and appropriate thinking.

2.6. Symmetrical Model

Redefining the data management of healthcare infrastructure demands a model instead of any process and approach. There is a specification related difference between a model and process. The proposed model defines a pathway for designing healthcare infrastructure from every layer.

The current healthcare infrastructure does not have any standardized set of models that can be adopted by every healthcare organization for developing its digital structure. To solve this issue and give a new perspective to the healthcare sector, the authors have devised a model premised on the cybernetics ideology. The model summarizes the infrastructure from a data perspective and works on the idea of conversation directed by cybernetics.

Figure 4 illustrates the model premised on cybernetics. The model is named ARAR based on the authors' names who have developed this model: Abhishek, Rajeev, Alka and Raees. The model has simple conversation based steps that intend to produce effective healthcare digital infrastructure from a data perspective. The first step, as formulated by the ARAR model, is data initiation. This is a primary step in any healthcare data infrastructure. After the successful initiation of data from various data sources based on the nature of the layer, the next step is to classify the variety of every source with respect to the data carrying and management in between the healthcare design processes. This type of classification gives us an idea about the position of the source in various possible conditions. After classifying the variety of sources, the next step is to evaluate the severity of various classified varieties. This gives the designer an overview of the possible conditions of sources. The evaluated results from these steps are compared with the desired outcome or goals of the design as a next step. This type of information is more of a reality check for the designers, helping them to ensure that the design process is on the right track. If the results are compatible and the designers as well as the experts find a perfect alignment between the development and the planned target, only then does the model allows the designer for the next step. The next step is to decide on the evaluated variety that is compatible with the goal of the design. Similarly, if the variety classification and its severity levels do not have the potential to achieve the goal of the design process, that model sends the process again to the first initial stage. There are two possibilities for designers at the decision stage of a model. The possibilities are whether to choose ignorance or the minimization/extract option. The decision stage is solely based on the severity of the variety and goal of the design process. The decision taken, or the choices made, at this stage directly affect the design of structure.

Moreover, to make it more understandable and clear, the authors have classified the whole framed model into five steps. These steps are: initiation; classification; assessment; result; application. The proposed model is a combination of these five steps, respectively, as displayed in Figure 5.

Further, as clearly portrayed in the framed model, it reflects and works on the cybernetics ideology in order to remediate variety related issues from digital infrastructure of healthcare. As observed in the earlier section of this study, variety is a wicked problem. A wicked problem is a problem that cannot be extracted or removed completely from the system because it is a part of the system. Such problems are related to the inevitable characteristics that are built into the system. To overcome issues that are variety-centric, the proposed model can be adopted by experts for redefining digital healthcare infrastructure.

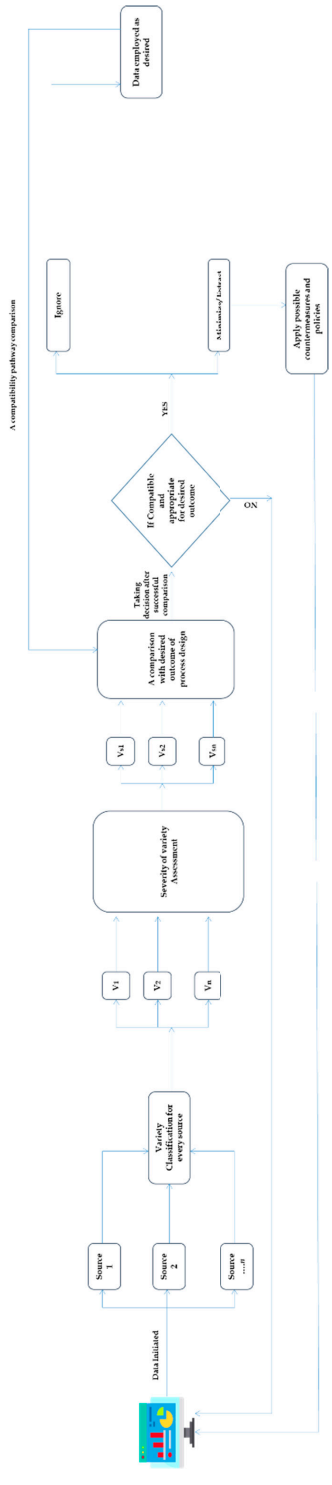


Figure 4. Symmetrical ARAR Model.

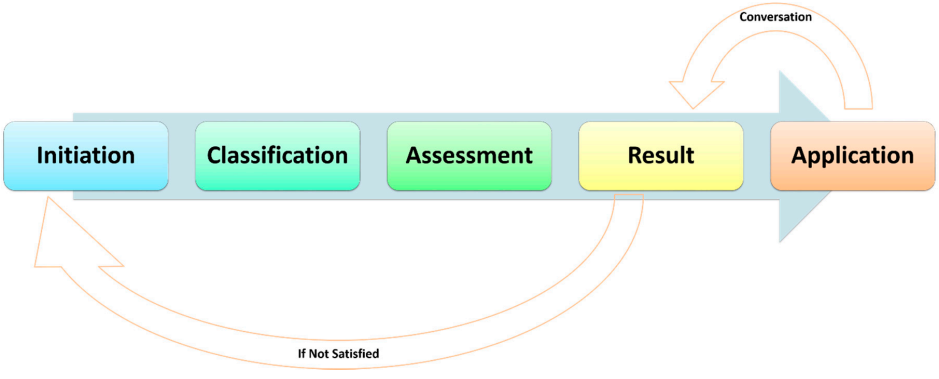


Figure 5. Conceptual Idea of Symmetrical Model.

2.7. *Performance Simulation*

To establish the efficacy of the proposed ARAR mode, the authors conducted a comparative analysis by simulating the effectiveness of the results of the proposed and other similar frameworks. Such an analysis and numerical quantification portrays an evident understanding about the effectiveness of the proposed model. The authors selected six different models for evaluation, including the proposed one. The names of the selected models are P5 Cybernetics Health (FW1); Personalized Health Dynamic System Model (FW2); Data Validation Model (FW3); DocBox 24 (FW4); Proposed ARAR Model (FW5) and DITrustblockchainIoHT Model (FW6). The initial serial numbers for all the selected models were assigned by the experts.

Further, to perform a numerical simulation of performance and prioritization of models, the authors adopted the popular hybrid multi criteria decision-making methodology, named the analytical hierarchy process combined with fuzzy set theory (Fuzzy-AHP) [16,17]. This methodology was used as a mechanism to simulate the effectiveness of the proposed model vis-à-vis the various selected models. This gives an idea about the performance of the models. The methodology has been detailed in the supplementary material of this paper. Fuzzy-AHP is a process that gives accurate results that are widely acceptable and validated. It is a well-established methodology. Simulation of the performance will help the future researchers in identifying the most effective model. The adopted methodology of prioritization works on the membership functions, and the numerical evaluation is adopted by [21–23]. To conduct the numerical evaluation, Table 1 illustrates the triangular fuzzy number for every specific model in a pair-wise comparison matrix. After identifying the pair-wise fuzzy numbers, the examiners defuzzified the values by adopting an α cut approach [24]. Table 2 portrays the value acquired by the α cut approach and the defuzzified value of the triangular fuzzy numbers. The final weights and ranking are described in Table 3 and Figure 6.

Table 1. Triangular Fuzzy Numbers for Every Specific Model.

	FW1	FW2	FW3	FW4	FW5	FW6
P5 Cybernetics Health (FW1)	1.00000, 1.00000, 1.00000	0.97000, 1.25000, 1.61000	1.05900, 1.58500, 2.22100	0.77300, 1.01200, 1.28800	0.76100, 0.91200, 1.09700	1.12800, 1.55400, 1.98800
Personalized Health Dynamic System Model (FW2)	-	1.00000, 1.00000, 1.00000	0.63500, 0.91400, 1.34300	0.42700, 0.63400, 0.96600	0.34800, 0.49000, 0.87300	0.56900, 0.72000, 0.97000
Data Validation Model (FW3)	-	-	1.00000, 1.00000, 1.00000	0.51500, 0.65800, 0.78500	0.52100, 0.66000, 0.91900	0.62700, 0.81200, 1.07200

Table 1. *Cont.*

	FW1	FW2	FW3	FW4	FW5	FW6
DocBox 24 (FW4)	-	-	-	1.00000, 1.00000, 1.00000	0.55600, 0.64500, 0.81200	1.48300, 1.95800, 2.52900
Proposed ARAR Model (FW5)	-	-	-	-	1.00000, 1.00000, 1.00000	0.56900, 0.78600, 1.15600
DITrustblockchainIoHT Model (FW6)	-	-	-	-	-	1.00000, 1.00000, 1.00000

Table 2. Defuzzified TFN value.

	FW1	FW2	FW3	FW4	FW5	FW6
P5 Cybernetics Health (FW1)	1.00000	1.26900	1.61200	1.02100	0.92100	1.55600
Personalized Health Dynamic System Model (FW2)	0.78800	1.00000	1.26900	0.66500	0.55000	0.74500
Data Validation Model (FW3)	0.62000	0.78800	1.00000	0.65400	0.69000	0.83100
DocBox 24 (FW4)	0.97900	1.50400	1.53000	1.00000	0.66500	1.98200
Proposed ARAR Model (FW5)	1.08700	1.81700	1.44900	1.50500	1.00000	0.82400
DITrustblockchainIoHT Model (FW6)	0.64300	1.34300	1.20400	0.50500	1.21300	1.00000
C.R. = 0.006129						

Table 3. Final Ranking.

S. No.	Models/Frameworks	Weights	Ranks
1	P5 Cybernetics Health (FW1)	0.193206	3
2	Personalized Health Dynamic System Model (FW2)	0.129727	5
3	Data Validation Model (FW3)	0.121105	6
4	DocBox 24 (FW4)	0.197429	2
5	Proposed ARAR Model (FW5)	0.204062	1
6	DITrustblockchainIoHT Model (FW6)	0.154471	4

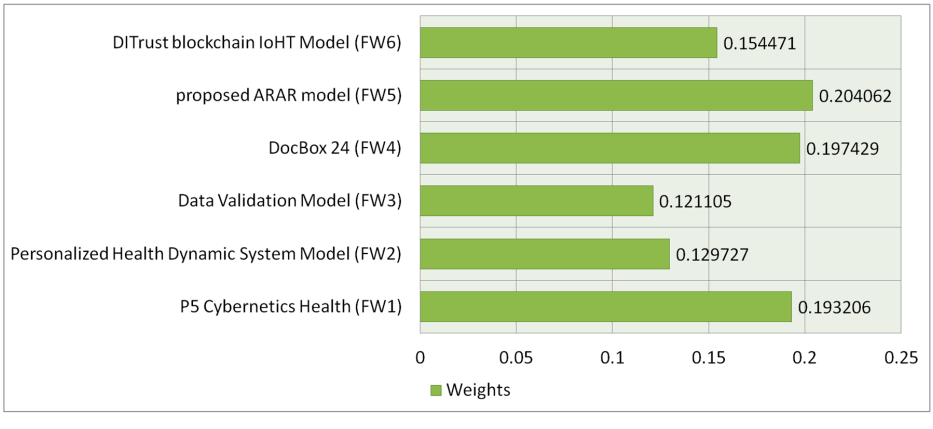


Figure 6. Graphical Illustration of Priority.

However, defining healthcare and its functionality for achieving a secure and systematic workflow is a challenging task, yet as discussed in this paper, various effective models are present in relevant fields to accomplish this ideal goal. Hence, it is a confusing job for researchers and experts to decide which model is effective and which is not. To simplify this challenge and portray a systematic and numerically assessed effectiveness order, this section illustrates an ideal pathway. The result obtained

from numerical assessment shows that the proposed model ARAR has the highest effect ratio in all of the selected models with the weight of 0.204062 and Consistency Ratio (CR) value of 0.006129.

The descending order of the prioritization for the selected model is FW5 > FW4 > FW1 > FWW6 > FW2 > FW3. The results discussed in this section of performance simulation portray that FW5 has highest effectiveness and priority and FW3 has the lowest. Overall, the proposed model is one of the most suitable frameworks and ideologies that can be adopted by the researchers and future practitioners in their domain.

2.8. Significance of Symmetrical Model towards Healthcare

A conceptualized ideology can only be validated if it is proven efficacious. A proper discussion and analysis can only convince the industry about the potential of the model. The authors have provided an ideology of cybernetics from a new perspective for the research community. In this context, it is equally important to cite the significance and potential of the suggested ideology for the betterment of the healthcare sector. Thus for making this task little easier and more specific, the authors have discussed the significance of the proposed ideology for mapping the design process for healthcare.

The proposed cybernetics ideology states that there is a need to apply every step during the design process as a conversation approach [13]. Conversation approach is the best way to understand the gaps between the current and the next phase that need to be bridged [13].

The proposed model is the best example of the ideology discussed in the article about how this thinking can change the whole process of designing. Cybernetics in digital healthcare infrastructure designing will be a revolutionary step, if applied as discussed in this article, i.e., based on the conversation approach. The concept has been illustrated more emphatically through Figure 7.

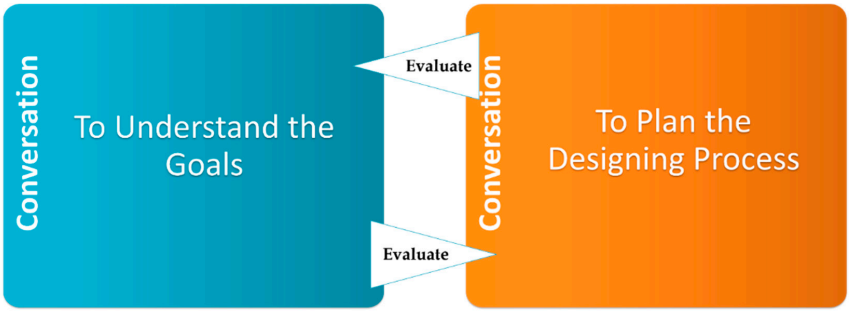


Figure 7. Summary of Proposed Symmetrical Ideology.

The above Figure 7 clearly describes a mutual conversation based technique that needs inter evaluation between the design processes. Designing the healthcare infrastructure from cybernetics perspective demands an initial but thorough understanding of the designer's idea. As shown in Figure 7, whenever the designers think about designing a process, they need to follow this basic understanding of clarifying the goals before designing and then preparing the whole design process according to the conversation-based results. Based on the assessment of the cybernetics methodology, it is very significant to apply the above ideology of conversation for effective results in symmetric designing.

Furthermore, if we talk about the significance of the proposed model in healthcare, then we need to analyze the usability of this model in layered categorization of healthcare. The most significant advantage of using the proposed model in healthcare layered categorization is that of the scope and space for possible variety of situations. As cited in the present study, variety affects almost every aspect in healthcare infrastructure and the experts must maintain requisite variety in healthcare for effective operatability of the system. Symmetric operation continuity is always important and required for any type of a system.

Moreover, numerical analysis always portrays a significant role in understanding the effect and usefulness of any proposed system. Thus, the authors have conducted a numerical analysis of the performance of the proposed model by conducting a comparative analysis. The results, as discussed in the previous section, clearly illustrate that the proposed ARAR model has the highest priority in all of the selected models. The analysis was done by using the well-established Multi Criteria Decision Making (MCDM) approach fuzzy AHP which provided accurate and conclusive results. This type of comparison and performance estimation is a novel concept and would prove to be a corroborative mechanism for researchers and practitioners working in this domain.

The cybernetics ideology that the authors have suggested in this study demands a conversation based process that firstly identifies the goal, and then prepares a symmetric designing process based on the requisites for achieving the planned goal. Hence, the proposed model would not only give optimum productivity but also save on the time, cost and other resources invested in developing the healthcare infrastructure. Moreover, this procedure demands a unique but conceptual thinking that is called the *modern day cybernetics.* The suggested model would provide symmetric operational system that is inherited from variety. The management of these issues would give usability without any interruption because the proposed model analyzes and effectively reduces the possibilities of failure by its conversation-based prioritization and remediation approach. The symmetric operatability thus obtained, produces efficient operation execution in the system because if usability is maintained in any system, without any interruption, then high operatability can be achieved automatically by any system [17].

3. Conclusions

It is a difficult task to manage the healthcare infrastructure smoothly in the current era. Every new update and patch in the system creates a loophole or possibilities for failure, thus posing a huge risk to the infrastructure. A thorough perusal of the issues in healthcare, as discussed in this study, cited that variety is one of the most critical issues. A symmetric understanding and perspective analysis is needed to tackle this issue in the present day healthcare infrastructure which is becoming even more complex because of digital deployment. Researching this premise, the authors proposed an ideology of cybernetics that has the potential to be an efficacious mechanism. Moreover, the variety issue of healthcare can only be minimized by managing its wicked nature. The suggested model would be successful in achieving the target of reducing the issues of variety by managing and comparing its different sources and its possible estimated variety. The propositioned model is only a conceptual overview of the vast work that is under development by the authors.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-8994/12/12/2089/s1.

Author Contributions: W.A. and A.B. contributed to the motivation, the interpretation of the method effects, and the results; H.A. and A.K.P. provided the concept, prepared the draft versions, performed the evaluation, and provided the conclusions; A.K.P. proposed minor suggestions and R.A.K. supervised the study. All authors have read and agreed to the published version of the manuscript.

Funding: The Deanship of Scientific Research at Taif University, the Kingdom of Saudi Arabia.

Acknowledgments: This project was supported by Taif University Researchers Supporting Project number (TURSP-2020/107), Taif University, Taif, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The authors also applied keywords such as "Symmetrical Healthcare Model", "Cybernetics in Healthcare", and "Digital Model Healthcare", associating the arithmetic search operators AND and OR on various research data repositories such as Google Scholar, IEEE Xplore, ResearchGate, etc. The total number of search results found on these data repositories was around 10,000, but, unfortunately, many of them were duplicate and irrelevant to the topic. To exclude the irrelevant and duplicate

papers from the search results, the authors performed various stage-wise analyses. In the first stage of the search, we excluded or included the studies based on their title. In the second step, we analyzed the abstracts of the papers and matched their requirements such as cybernetics and symmetrical modeling. Overall, the authors adopted a thorough review process of the previous research studies [16]. Among these as well, the authors found very few articles that were partially similar to the topic or its core intent.

References

1. Kumar, R.; Pandey, A.K.; Baz, A.; Alhakami, H.; Alhakami, W.; Agrawal, A.; Khan, R.A. Fuzzy-Based Symmetrical Multi-Criteria Decision-Making Procedure for Evaluating the Impact of Harmful Factors of Healthcare Information Security. *Symmetry* **2020**, *12*, 664. [CrossRef]
2. Health Infra Struggles to Keep Pace as Covid-19 Cases Surge. Available online: https://www.livemint.com/news/india/how-coronavirus-left-india-s-health-infrastructure-creaking-11591525975138.html (accessed on 2 October 2020).
3. Disaster of System Failure: Challenges for Health Care in 21st Century. Available online: https://insightplus.mja.com.au/2020/1/disaster-of-system-failure-challenges-for-health-care-in-2020/ (accessed on 2 October 2020).
4. Korotkova, O.M.; Korotkova, O.M.; Belokoneva, I.V.; Belokoneva, I.V. Development of It and Cybernetics in Russian Healthcare: Past, Present, Future. *Молодежныйинновационныйвестник* **2019**, *8*, 107–108.
5. Khayal, I.S.; Farid, A.M. A Dynamic System Model for Personalized Healthcare Delivery and Managed Individual Health Outcomes. *arXiv Prepr.* **2019**, arXiv:1910.09104.
6. Efthymiou, I.P.; Vozikis, A.; Sidiropoulos, S. Application of Sociocybernetic model in the field of Health Management. *Int. J. Sci. Eng. Res.* **2019**, *10*, 451–460.
7. Infrastructure Failure. Available online: https://medicine.llu.edu/sites/medicine.llu.edu/files/docs/infrastructure-failure.pdf (accessed on 6 October 2020).
8. The Coronavirus Exposes Our Health Care System's Weaknesses. We Can Be Stronger. Available online: https://www.statnews.com/2020/03/02/the-coronavirus-exposes-our-health-care-systems-weaknesses-we-can-be-stronger/ (accessed on 6 October 2020).
9. Agrawal, A.; Zaroor, M.; Alenezi, M.; Kumar, R.; Khan, R.A. Security durability assessment through Fuzzy Analytic Hierarchy process. In *PeerJ Computer Science*; PeerJ Inc.: Corte Madera, CA, USA, 2019; pp. 1–43. [CrossRef]
10. 53% of Healthcare Organizations Have Experienced a PHI Breach in the Past 12 Months. Available online: https://www.hipaajournal.com/53-of-healthcare-organizations-have-experienced-a-phi-breach-in-the-past-12-months/ (accessed on 11 October 2020).
11. July 2020 Healthcare Data Breach Report. Available online: https://www.hipaajournal.com/july-2020-healthcare-data-breach-report/ (accessed on 11 October 2020).
12. Glanville, R. A (cybernetic) musing: Design and cybernetics. *Cybern. Hum. Knowing* **2009**, *16*, 175–186.
13. Dubberly, H.; Pangaro, P. Cybernetics and design: Conversations for action. In *Design Cybernetics*; Springer: Cham, Switherland, 2019; pp. 85–99.
14. Ashby, W.R. *An Introduction to Cybernetics*; Chapman & Hall Ltd.: London, UK, 1961.
15. Mosadeghrad, A.M. Factors Affecting Medical Service Quality. *Iran. J. Public Health* **2014**, *43*, 210–220. [PubMed]
16. Pandey, A.K.; Khan, A.I.; Abushark, Y.B.; Alam, M.M.; Agrawal, A.; Kumar, R.; Khan, R.A. Key Issues in Healthcare Data Integrity: Analysis and Recommendations. *IEEE Access* **2020**, *8*, 40612–40628. [CrossRef]
17. Kumar, R.; Zarour, M.; Alenezi, M.; Agrawal, A.; Khan, R.A. Measuring security durability of software through fuzzy-based decision-making process. *Int. J. Comput. Intell. Syst.* **2019**, *12*, 627–642. [CrossRef]
18. Faggini, M.; Cosimato, S.; Nota, F.D.; Nota, G. Pursuing Sustainability for Healthcare through Digital Platforms. *Sustainability* **2019**, *11*, 165. [CrossRef]
19. Abou-Nassar, E.M.; Iliyasu, A.M.; El-Kafrawy, P.M.; Song, O.Y.; Bashir, A.K.; Abd El-Latif, A.A. DITrust chain: Towards blockchain-based trust models for sustainable healthcare IoT systems. *IEEE Access* **2020**, *8*, 111223–111238. [CrossRef]

20. Yang, P.; Stankevicius, D.; Marozas, V.; Deng, Z.; Liu, E.; Lukosevicius, A.; Min, G. Lifelogging data validation model for internet of things enabled personalized healthcare. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *48*, 50–64. [CrossRef]
21. Kumar, R.; Khan, S.A.; Khan, R.A. Analytical network process for software security: A design perspective. *CSI Trans. ICT* **2016**, *4*, 255–258. [CrossRef]
22. Sahu, K.; Shree, R. Stability: Abstract roadmap of software security. *Am. Int. J. Res. Sci. Eng. Math.* **2015**, *15*, 183–186.
23. Kumar, R.; Khan, A.I.; Abushark, Y.B.; Alam, M.M.; Agrawal, A.; Khan, R.A. An Integrated Approach of Fuzzy Logic, AHP and TOPSIS for Estimating Usable-Security of Web Applications. *IEEE Access* **2020**, *8*, 50944–50957. [CrossRef]
24. Agrawal, A.; Pandey, A.K.; Baz, A.; Alhakami, H.; Alhakami, W.; Kumar, R.; Khan, R.A. Evaluating the security impact of healthcare Web applications through fuzzy based hybrid approach of multi-criteria decision-making analysis. *IEEE Access* **2020**, *8*, 135770–135783. [CrossRef]

symmetry

MDPI

Article

Smart Root Search (SRS): A Novel Nature-Inspired Search Algorithm

Narjes Khatoon Naseri [1,*], Elankovan A. Sundararajan [1], Masri Ayob [2] and Amin Jula [2]

[1] Centre of Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; elan@ukm.edu.my

[2] Data Mining and Optimization Research Group (DMO), Centre for Artificial Intelligent (CAIT), Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia; masri@ukm.edu.my (M.A.); amin.jula@petronas.com (A.J.)

* Correspondence: narjes@thecads.com

Received: 16 November 2020; Accepted: 4 December 2020; Published: 7 December 2020

Abstract: In this paper, a novel heuristic search algorithm called Smart Root Search (SRS) is proposed. SRS employs intelligent foraging behavior of immature, mature and hair roots of plants to explore and exploit the problem search space simultaneously. SRS divides the search space into several subspaces. It thereupon utilizes the branching and drought operations to focus on richer areas of promising subspaces while extraneous ones are not thoroughly ignored. To achieve this, the smart reactions of the SRS model are designed to act based on analyzing the heterogeneous conditions of various sections of different search spaces. In order to evaluate the performance of the SRS, it was tested on a set of known unimodal and multimodal test functions. The results were then compared with those obtained using genetic algorithms, particle swarm optimization, differential evolution and imperialist competitive algorithms and then analyzed statistically. The results demonstrated that the SRS outperformed comparative algorithms for 92% and 82% of the investigated unimodal and multimodal test functions, respectively. Therefore, the SRS is a promising nature-inspired optimization algorithm.

Keywords: combinatorial optimization problem; heuristics method; nature-inspired algorithm; NP-hard problem; plant root

1. Introduction

Nature is replete with intelligent and disciplined phenomena with impressive capabilities that are being continuously discovered. The behavior of animals and insects has been observed for centuries; however, the observed behaviors, which are represented by physics and particle dynamics, represent only a small portion of the intelligent processes that exist in nature. The intelligent behavior found in nature has inspired many intelligent search algorithms, such as genetic algorithms (GAs) [1], particle swarm optimization (PSO) [2,3], ant colony optimization [4,5], the artificial fish swarm algorithm [6], the artificial immune system [7], bacterial foraging optimization algorithm [8,9], bat-inspired algorithm [10,11], imperialist competitive algorithm [12–14], and the gravitational attraction search [15]. Although each of the intelligent search algorithms exhibits their own set of efficiencies and can solve many types of optimization problems, there are some issues they have in common for solving large-scale problems especially those possessing search spaces with staggering high dimensions. Amongst these issues, local optima problem [16–18] and the absence of inborn exploitation operations [19–21] are seemingly impossible to overcome. Thus, many researchers are searching for unique methods to tackle these issues. Hybrid heuristic search and memetic algorithms [19,22,23] have been proposing since years ago to tackle these problems. Innovative combinations of the Bee

colony algorithm with other heuristics are also showing promising results in properly exploring search spaces [24,25].

In the process of searching the soil for nutrients, such as minerals and water, plant roots demonstrate highly intelligent behavior [26]. This intelligence must be creative to ensure the survival of the plant in environments with low levels of water and insufficient nutrients. In these limited-resource communities, roots, as explained in Section 1.1.1, often adapt and search a broad area before plant dies due to absence of food or water. Therefore, the intelligent behavior of roots could be the basis for a brand new, swift, effective, and appealing intelligent search algorithm able to efficiently search large spaces.

A few attempts have been made to imitate plant root behavior and develop new search algorithms. These studies resulted in four algorithms: the root growth model (RGM) [27], root mass optimization (RMO) [28], the artificial root foraging model (ARFO) [29,30] and artificial root mass (ARM) [31]. Despite researchers' efforts, the proposed algorithms have been unable to utilize plant intelligence in a way that overcomes the weaknesses that affect other heuristic search algorithms (refer to Section 1.1.2 for more details).

Hence, to achieve high efficiency and overcome problems, a new, independent, growth-inspired Smart Root Search (SRS) algorithm was proposed by the authors of this paper in [32]. The SRS is equipped with unique features and well-defined operators that are extracted precisely from intelligence of plant, and, unlike previous plant root-based search algorithms, is in absolute alignment with optimization principles. The novelty of SRS can be summarized in three items including (1) dividing the search space into a number of subspaces helping the algorithm to quickly find the more potential areas of the search space, and control local convergence of the algorithm in those areas; (2) defining three different types of roots, immature, mature and hair roots, that use different exploration approaches together with a mechanism to convert immature to mature, in order to search more in promising areas and provide embedded local search mechanism; and (3) proposing root drouth operator to control local and global convergence of the algorithm simultaneously.

As the proposed SRS was a brief preliminary model, it required supplementary parts, revision, implementation, test and comparison. To this end, in this paper the final version of the model is proposed and explained in detail and represented in the form of flowchart and pseudocode. Supported by graphical examples, a new structure is employed for roots for getting better performance, and root drought equation is improved to protect promising roots more accurately. In addition to that, a clear parameter initialization is added to the algorithm, and a precise explanation is provided for Immature-to-Mature mechanism of the roots. Most importantly, a complete experimental test has designed and applied to evaluate the performance of the algorithm and compare with introduced comparative algorithms followed by an in-detail statistical test.

The remainder of the paper is structured according to the following. The literature on the intelligent behavior of roots is reviewed in Section 1.1. A detailed explanation of the SRS algorithms is then described in Section 2, followed by the experimental test and results in Section 3. Finally, conclusion and suggestions for future work are provided in Section 4, followed by references.

1.1. Literature Review

From a general perspective, living things can be divided into two main groups: animals and plants. Animals sense their environment using their senses, such as sight and touch. Similarly, plants perceive their surrounding environment using a series of senses and reactions. To become more familiar with plant physiology and their senses and reactions, this section provides information on how plant roots sense their environment and the consequences of these reactions.

1.1.1. Plant Senses and Reactions

Plants use the same five senses as humans: hearing [33], touching, tasting, seeing and smelling. Furthermore, plants have evolved to use more senses than humans and, in fact, have approximately twenty distinct senses. Indeed, plants are able to detect moisture, gravity, minerals, humidity, light, wind, soil composition and structure, snow melt, pressure, temperature, and infection. Plants can decide to react against environmental stimuli based on the information obtained by these senses. Thus, plants are considered prototypical intelligent creatures [26,34] and exhibit their intelligence via shoot and root growth.

A thorough review on the architecture of root system and the pathways and networks forming root behavior was provided in [35]. Those authors showed that root growth is a reaction to the nutrients of the soil depending on several changing factors. Hydrotropism, nutrient tropism, cell memory and electrical impulse are some of the most interesting behaviors that demonstrate root intelligence. These behaviors are summarized below.

Hydrotropism: Plant survival depends on the ability of roots to find water in soil. Hence, plant root growth curves correspond to the moisture gradient (higher water potential) called hydrotropism [36,37].

In addition, when they encounter moisture in soil, roots absorb and store water to support all plant activities. The plant loses stored water during plant growth or evapotranspiration. Roots can also transfer water to dry parts of the soil and release it to promote root survival [38–40]. This release occurs when the absorbed water is not sufficient for root survival because roots that cannot survive will dry out.

Nutrient tropism: Nitrates and phosphates are considered the most important elements for plant growth [41]. Important developmental processes, such as lateral root (LR) and hair root (HR) formation as well as primary root (PR) elongation (length), are to a great extent sensitive to the nutrient concentration changes.

Strong evidence shows that the nitrate concentration affects LR formation: development of LR is hindered by high nitrate concentrations and stimulated by low concentrations of nitrate, respectively [35,42]. PR elongation under the inhibitory impact of high nitrate concentrations is also discussed in [43]. Accessibility and distribution of phosphate and Nitrate have been shown to have contrasting effects on PR elongation and LR density but comparable impacts on LR elongation [44]. PR elongation is known to decrease with increasing nitrate availability but increase as the phosphate supply increases. The LR density remains constant across varying concentrations of nitrates but decreases as the phosphate supply increases. In contrast, LR elongation is suppressed by high concentrations of nitrate as well as phosphate.

In this regard, Ref. [45] demonstrated that phosphate starvation enhances HR elongation and density. Furthermore, research conducted at the Pennsylvania State University shows that *Arabidopsis thaliana* roots grow more condensed and longer reacting to lower availability of phosphate [46].

Memory: Although plants do not have a neural network, many studies show that they can recall some conditions, which suggests that plants exhibit memory. Ref. [47] addressed traumatic plant memories, related facts, and potential mechanisms. Stress factors make the plant impervious to subsequent exposures. This stress-related feature indicates that every plant has a memory capacity. In addition, plants also possess "stress memory" and "drought memory". Surprisingly, the proportion of live biomass after a late drought is higher in plants that were exposed to drought earlier in their growing season contrasted with single-stressed plants [48].

Electrical Impulse: Plants also use a message-passing system [49]. Research on plants has shown that electrical communication plays a significant role in root-to-shoot contact in the plants under water stress. Furthermore, Ref. [50] showed that when one organ of a seedling is stimulated (i.e., the root region), a characteristic response (electrical stimulus) is produced and would be recorded upward in another organ from the stimulating area.

1.1.2. Plant-Imitating Methods

A comprehensive analysis at plant root domains shows that only a few research studies have focused on using the inherent intelligence of the plants as a search algorithm. The studies have been leaded and conducted by Zhu Yunlong and his several research teams, respectively. In this section, the small number of proposed plant root algorithms and the main ideas are assessed and discussed.

RGM is a proposed algorithm for numerical function optimization that simulates the interactions between HR growth and the soil [27,51]. In each iteration of growth in RGM, high-functioning roots, which have higher Morphactin (fitness function) values, are selected to branch areas distant from the selected roots. New branches are called HRs. HRs follow random growth directions, and their growth length depends on their growth direction. HRs are added to a set of roots, and then a set of non-selected branching roots is removed. Accordingly, RGM could be known as a local search algorithm that does not take advantage of the well-extracted root intelligence and is not suitable to search in large search spaces. Furthermore, when local optima become trapped, the RGM method fails.

In 2013, an RGM for numerical optimization—the RMO algorithm—was proposed [28]. RMO is the primary inspiration for two other algorithms: ARFO and ARM.

ARFO [29,30] was proposed for image segmentation problems and then generalized to address other optimization problems. This algorithm uses the Auxin hormone levels of roots as the objective function and employs branching, re-growing, hydrotropism and gravity-tropism operations. The ARFO root system consists of three groups of roots, including main roots and lateral roots (large and small elongated length units, respectively) and dead roots. Two main shortcomings of ARFO are evident: first, absence of precise extraction and accurate modeling of plant intelligence in terms of root growth. Second, optimization-averse behavior is inherent in the algorithm. A list of shortcomings of the ARFO is presented in Table 1.

Table 1. Intelligent and optimization-averse behaviors of artificial root foraging model (ARFO) algorithm.

Optimization-Adversative Behaviors	**Common Intelligent Behaviors**
Increasing root length in promising main root areas	Using short-step movements/changes in promising areas to identify additional search locations
Decreasing root length in non-promising LR areas	Using large-step movements/changes in non-promising areas to escape non-promising areas
LR are exploited in non-promising areas	Exploitation occurs in promising areas
Applying short-length branches causes very fast convergence	Avoiding fast convergence
No chance for enhancing bad solutions	Bad solutions have more chances for enhancement

ARM optimization was proposed to solve the data clustering problem. ARM is based on a harmony-like search algorithm [52] that simulates plant root growth strategies, such as proliferation and decision making, that depend on the growth direction [31]. ARM generates a set of roots randomly. Some of the roots with better fitness values can continue their growth, while the rest stop growing. For every root, one neighbor is selected randomly. If the fitness of the neighbor is better than that of the root, then a new root will be generated between them in the search space; otherwise, a new root will be generated randomly. The new root will be added to the set of roots with better fitness values than its parent. Therefore, no intelligent root behaviors are applied in ARM.

2. Smart Root Search Algorithm

2.1. Intelligence of Plant Root Growth

Root growth intelligence can be outlined as follows:

(a) Roots grow in the direction of the nutrient sources in the soil.
(b) Root growth accelerates and generates new branches and HRs depending on the nutrient concentration of the soil.
(c) Each part of a root uses electrical impulses to send information about its current situation to the other parts.
(d) Plants can memorize and respond to information.
(e) Water stress states cause roots to dry up.

These intelligence mechanisms motivated us to design the SRS optimization algorithm. The SRS is described in detail in Section 2.2. Table 2 presents a mapping of the optimization with real plant root growth (botany) terms.

Table 2. Mapping optimization terms and plant root mechanism.

Botany Terms	Optimization Terms
Soil	Search Space
Plant Root Set	Solutions' Vector
Root	Solution
Nitrate Concentration	Objective Function
Location of the Highest Nitrate Concentration	Optimal Solution
Growth Step	Iteration
Hair Roots Germination	Local Search Operator
Root Growth	Solution Movement
Root Drouth	Solution Elimination
Root Growth Speed	Velocity of Movement
Branching	Solution Reproduction
Immature Root	Limited-move Solution
Growth Direction	Movement Coefficient Set

2.2. SRS Algorithm

The SRS has some characteristics that distinguish it from similar algorithms.

I. SRS divides the search space into several subspaces and distributes the first generation of roots equally among them. This helps SRS to apply different search policies to different parts of the search space. Similar algorithms do not provide such functions in their standard versions.
II. SRS-generated roots are immature upon germination but become mature after a few iterations. Thus, the algorithm can apply different search policies by using the same roots based on their age. In contrast, other algorithms use fixed exploratory policies during their execution.
III. SRS utilizes an embedded local search mechanism applied by a group of roots called HRs.
IV. SRS utilizes a dynamic population size. This gives the SRS the capability to decrease the number of solutions in non-promising subspaces and to increase the number of solutions in promising areas.

Knowing that, the main procedures of the SRS algorithm are Parameter Initialization, Dividing the Search Space, Initialization of the First Generation, Evaluation, Sorting and Ranking of the Roots, Root Growth, Root Drouth and Root Branching followed by HRG and Termination Criteria that are described in the following sections and visualized in Figure 1. In addition, the time complexity of the algorithm is discussed in Appendix B.

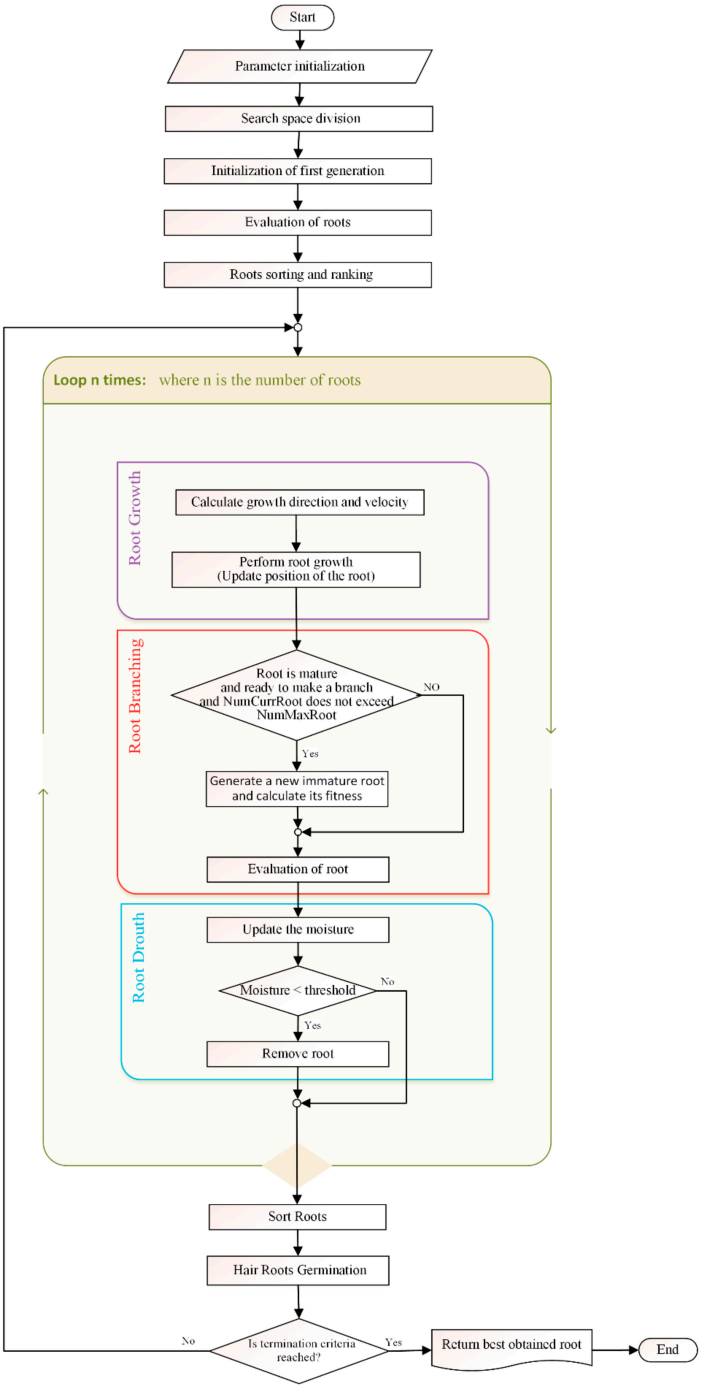


Figure 1. The smart root search (SRS) flowchart.

2.2.1. Parameter Initialization

SRS search elements and operations can be adjusted by setting a few guiding parameters. The parameter values depend on the problem specifications. Studying and understanding these parameters and how they affect the algorithm are crucial for applying the algorithm successfully. The parameters are defined in Sections 2.2.2–2.2.9 and initialized in Section 3.1.1 for the investigated test functions.

2.2.2. Dividing the Search Space

In extensive search spaces, search algorithms must probe into a huge number of space points. Total number of space points is considerably more than the number of initial solutions in the first iteration of running the algorithm. As the initial solutions of the algorithm generate randomly, the distribution of the solution in the search space is not often uniform throughout the search space, and therefore, the search space would not be inspected thoroughly.

Thus, based on a divide-and-conquer strategy [53], SRS algorithm divides the problem search space into *Ns* subspaces. The number of subspaces directly affects the SRS convergence speed. Therefore, an effective *Ns* initialization value depends on the problem specifications including the structure of the solutions, number of dimensions and sizes of the various search space dimensions. The user can use any approach to divide the search space such that the boundary of each subspace can be determined.

2.2.3. Initialization of First Generation

The SRS randomly generates *NumMinRoot* number of solutions for initial generation so that the number of solutions in every subspace is equal. If there are some remaining unassigned solutions, they will be randomly assigned to the subspaces. Every solution in the SRS is mapped to a root, and the location of root i in a D-dimensional problem search space is presented below (Equation (1)), where x_i^d represents the location of root i in dimension d. A basic structure of a root is also illustrated in Figure 2. Once this step ends, all subspaces possess the same number of roots generated.

$$x_i = \left(x_i^1, x_i^2, \ldots, x_i^d, \ldots, x_i^D\right) \tag{1}$$

x^1	x^2	$x^3 \ldots x^{D-1}$	x^D	Nitrate Concentration	Velocity	Type	Age
$Value_1$	$Value_2$	$Value_3 \ldots Value_{D-1}$	$Value_D$	Objective Function Value	Velocity Value	Type of Root	Age Value

Figure 2. The basic structure of an SRS root. The first row represents the structure while the second row explains what should be assigned to every element of the structure.

2.2.4. Evaluation of Roots

Nitrate is the most important factor of growth in plants, followed by phosphate [41]. Botany research has demonstrated that concentration of nitrate and phosphate play critical roles in the root growth speed, and the density of branches and HRs [35,41,44–46,54–56]. These roles are summarized in Table 3. In terms of the similar effects of high nitrate and low phosphate concentration on root growth speed, an aggregated effect of nitrate and phosphate can be extracted from Table 3, as shown in Table 4. To simplify the proposed model, those combinations in which the nitrate and phosphate concentrations exert the same effects on root growth speed are considered. Due to the importance of the simplicity of the SRS model, we suppose that as the concentration of nitrate increases, the concentration of phosphate decreases (Equation (2)). This assumption facilitates defining the concepts that are more compatible with the terminology of the botany. Accordingly, as shown in Equation (3), the only nutrient

that affects root life is nitrate, and the objective function (f(x)) value of each root is considered as the *Nitrate Concentration* of that root.

$$\text{Nitrate Concentration} = \frac{1}{\text{Phosphate Concentration}}. \tag{2}$$

$$f(x) = \text{Nitrate Concentration} \tag{3}$$

Table 3. Nitrate and Phosphate effects on root behaviors.

Nutrients Concentration	Effects		
	Root Growth Speed	**Hair Roots Density**	**Branching Density**
High Nitrate	↓	Nothing	Nothing
Low Phosphate	↓	↑	↑
Low Nitrate	↑	Nothing	Nothing
High Phosphate	↑	↓	↓

↑ Represents Increasing and ↓ represents Decreasing.

Table 4. Aggregated Nitrate and Phosphate effects on root behaviors.

Nutrients Concentration	Effects		
	Root Growth Speed	**Hair Roots Density**	**Branching Density**
High Nitrate & Low Phosphate	↓	↑	↑
Low Nitrate & High Phosphate	↑	↓	↓

↑ Represents Increasing and ↓ represents Decreasing.

2.2.5. Root Sorting and Ranking

The SRS sorts its current roots at the end of initialization step as well as all execution iterations. It lets SRS identifying the best roots of every subspace, ranking each root in the root list, and locate the global best root at the top of the list of roots.

2.2.6. Root Growth

The roots of plants expand at varying rates and in various directions in order to locate richer areas of nutritional elements. In the same way, SRS roots grow (move) within the problem search space to find superior locations. This growth occurs at a predesignated velocity in a given direction. Therefore, the velocity and direction of every root must be determined beforehand. The root growth mechanism leads to a controlled local convergence in every subspace. Further details of the root growth are provided in Sections 2.2.6.2 and 2.2.6.3

From lifetime standpoint, SRS divides roots into two groups: permanent and temporary roots. Permanent roots are immature and mature roots that are defined based on their age. These roots fall into a temporary root category known as JRs. Figure 3 depicts the types of roots in a plant.

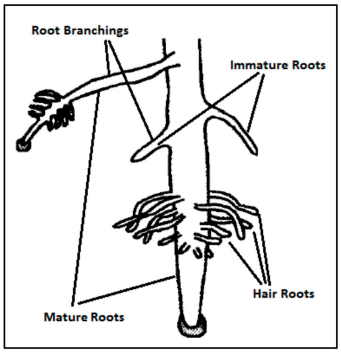


Figure 3. Types of roots [57].

2.2.6.1. Types of Roots

I. Mature Roots

Mature roots can change their growth direction and velocity to improve exploration. Once a mature root reaches a good location in the search space, it reduces its growth speed to explore the area more efficiently in increments. Consistently, such roots attempt to escape poor locations by increasing their growth speed. In addition, mature roots also make new roots (i.e., by branching) to facilitate searching alternate locations and directions. Once a mature root creates a new branch, it becomes the parent root of that new branch. As will be explained in Section 2.2.7, when growing in relatively appropriate areas of the search space, a root will form additional new branches. Therefore, mature roots are very flexible and exhibit different intelligent behavior during the search process.

II. Immature Roots

The second group of permanent roots consists of immature roots. These types of roots are not old enough to create new branches or make changes in their velocity and direction. Instead, they retain their original characteristics until transforming into mature ones. Therefore, every immature root can just receive velocity from the parent root. Immature roots also continue growing into the location of the parent based on random directions determined during germination. SRS utilizes immature roots to explore more of the search space that has not yet been reached. Accordingly, they play an important role in SRS by avoiding the trapping of local optima.

III. HR

Exploitation is a searching mechanism that can be utilized dependently or through hybridization with exploration methods to efficiently search the neighborhoods of the best generated solutions [23,58,59]. Many searching methods do not have an exploiting operation built-in and an additional local search method is needed to construct a hybrid method [19]. In contrast, in its structure, SRS employs hair roots to incorporate a fast but efficient exploiting mechanism.

HRs are short in size and age in the nature [35,46]. These roots support mature roots to gather more water and nutrients while they are in a rich part of soil. The natural behavior is simulated by SRS by an operator called HRG. HRs play their exploitation role without having to make new branches or grow in the search space; thus, the HR velocity and direction are not well defined.

2.2.6.2. Growth of Mature Roots

Definition 1. *Best Root Set (BRS).*

In every subspace, the first k best roots of the subspace create the BRS. For doing that, k will be specified by utilizing "Roulette Wheel Selection via Stochastic Acceptance" [60]. Every root of the subspace grows to the nearest root of the BRS. For finding the best nearest root to root i, $density_{j,i}$ is defined for every root j in BRS. Equation (4) gives the $density_{j,i}$, where the dominator is the Euclidean distance between roots i and j, NC_j is the nitrate concentration of root j, and x_i^d and x_j^d are the locations of roots i and j in dimension d, respectively.

$$density_{j,i} = \frac{NC_j}{\sqrt{\sum_{d=1}^{D}\left(x_j^d - x_i^d\right)^2}} \tag{4}$$

Therefore, among the BRS roots, root j is the best nearest root to root i, and $density_{j,i}$ is maximized (Equation (5)).

$$best\_closest_i = \left\{j \middle| \, density_{j,i} is\, MAX\right\} \tag{5}$$

I. Velocity of Mature Roots

A user-defined maximum velocity, v_{max}, is used to be a baseline of calculating the velocity of mature roots. The velocity of roots cannot exceed v_{max}. Then, in accordance with the rank of every root among all roots, the velocity of the root is determined as a fraction of v_{max}, such that the velocity is lower with increasing rank. This policy helps roots located in promising areas to grow slower and achieve more precise exploration while forcing the rest of the roots to move away from non-promising areas. The velocity of root i can be obtained by Equation (6), where glb_rank_i and $NumCurrRoot$ are the global rank of root i and the number of roots that currently exist, respectively.

$$v_i(t) = v_{max} - \left[v_{max}\left(1 - \frac{glb\_rank_i}{NumCurrRoot}\right)\right] \tag{6}$$

II. Direction of Mature Roots

For every dimension, a coefficient is required for the current velocity of the root to grow in the direction of its best closest root by applying a proper dimensional velocity. In addition, the coefficients should take values so that the growing root does not grow beyond the best closest one. To obtain these important movement coefficients, geometric relationships are helpful. Let the growth angle of root i toward its best, closest root $best\_closest$, be θ. The growth angle cosine of root i in dimension d, $cos\theta_i^d$, will be calculated simply by using Equation (7).

$$cos\theta_i^d = \frac{x_{best\_closest}^d - x_i^d}{\sqrt{\sum_{d=1}^{D}\left(x_{best\_closest}^d - x_i^d\right)^2}} \tag{7}$$

Therefore, assuming the current location and growth velocity of root i in dimension d are $x_i^d(t)$ and $v_i^d(t)$, respectively, the next location of the root will be determined by Equation (8), for which $v_i^d(t)$ should be calculated by (9). To make root growth easier to realize and apply, a complete example of how a mature root grows from its current location ($x_i(t)$) to its next location ($x_i(t+1)$) is presented in Appendix A.

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t) \tag{8}$$

$$v_i^d(t) = \left[v_i(t) \times cos\theta_i^d\right] \tag{9}$$

2.2.6.3. Root Growth of Immature Roots

As mentioned in Section 2.2.6.1, immature roots are newborn roots that are unable to change their growth velocity and direction or to make new roots. Regarding their growth velocity, every immature root gets its parent root velocity and will follow the growth behavior of the parent. Additionally, growth occurs in a randomly selected direction. To this end, every immature root that is generated randomly selects a point in its related subspace and considers that point its best, closest root. Then, Equation (7) is utilized to set the coefficients and determine the growth direction. Similarly, Equations (8) and (9) are applied to obtain the dimensional velocities and new locations.

2.2.6.4. Immature to Mature Transformation Mechanism

By providing a glimpse of the independent and complementary roles played by different types of roots, the mechanisms of SRS reveal how important it is to define a mechanism that transforms an immature root into a mature one. A simple maturation mechanism is proposed in SRS. As shown in Figure 1, every root has an attribute called Age. Age is initialized as 0 once a new root is generated. In each iteration of the algorithm, the root's Age increases by one. If Age value of any of the immature roots reaches a threshold, *Mature_Age*, status of the root changes to "mature". *Mature_Age* should be defined by the user based on the adopted exploration and exploitation policy such that higher *Mature_Age* values correspond to more exploration and less exploitation.

2.2.7. Root Branching

Branching is a mechanism in root growth that produces new roots to increase the search rate in those parts of the soil that have not yet been investigated. Similarly, SRS utilizes a *Branching* operation that generates new immature roots in the search space. Every new generated root in its early stages is adjacent to a mature root which is considered its parent.

In plant roots, two to five branches exist in each centimeter, depending on the phosphate concentration [44]. For SRS, a mechanism was designed to allow better roots to generate more branches. In this mechanism, branching is intended to encourage mature roots in every iteration by granting two to five nitrate concentration-based scores. Therefore, each mature root will branch after several iterations if the sum of collected scores (SCS) meets *Minimum Required Ability* (*MRA*) that is a predefined threshold value. The *MRA* can be set dynamically throughout the execution of the algorithm or determined by a user; higher *MRA* values correspond to fewer newly generated roots. Once a root reaches *MRA* and generates a new root, the SCS resets to 0, and the scores are re-collected.

This mechanism divides all the available roots in every subspace into four groups, each of which corresponds to one of the values of the range [2,5]. To avoid generating many ungainly new roots, higher scores should be assigned to small portions of roots having higher nitrate concentrations. Equation (10) is solved and the obtained value is used to dedicate the minimum possible percentage of roots needed to achieve this aim. The percentages of roots that should be assigned to other groups will be obtained using the next coefficients provided in the equation. Table 5 presents the root-dedication percentages of groups and scores. Additionally, Figure 4 shows how two different roots behave when collecting scores to reach *MRA* and generate new branches when located in two different parts of the search space.

$$x + 2x + 4x + 8x = 100 \;\Rightarrow\; x \cong 6.66 \tag{10}$$

Table 5. Root group scoring for branching.

	Coefficient	Ratio of Group Roots to All	Score
Group 1	1	$1 \times 6.66 = 6.66\%$	5
Group 2	2	$2 \times 6.\ 66 = 13.33\%$	4
Group 3	4	$4 \times 6.\ 66 = 26.66\%$	3
Group 4	8	$8 \times 6.\ 66 = 53.33\%$	2

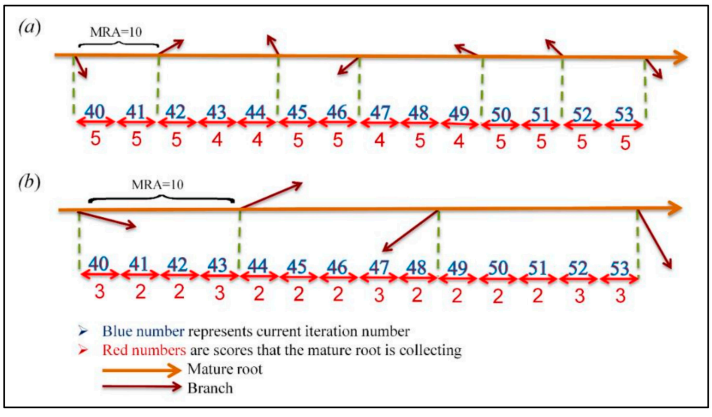


Figure 4. An example of NC-based grouping affects parent root branching when locating in a (**a**) promising part of search space, and (**b**) unpromising part of search space.

2.2.8. Root Drouth

In order to remove improper roots while the number of roots in the search space increases due to the branching operator, another operator is required. This operator must be designed such that the proportion of newly generated roots and removed roots gets controlled and global convergence of the algorithm is guaranteed. As this operator simulates the root drouth mechanism of roots, the same name is employed. To simplify describing the Root Drouth operator, a new term will be defined.

Definition 2. *To implement the Root Drouth operator, every root gains a certain volume of moisture, Moisture Percentage (MP), at initialization. As a moderate value, MP is initially 50. Whereas the MP changes as a mature root grows, immature roots have constant MP values to help them sustain throughout pre-pubertal development.*

As mentioned previously, botany research has demonstrated that plant roots absorb and store water from the areas in soil that contain higher level of moisture and transfer this stored water for use in drier areas [38–40]. Root drouth occurs provided that a root encounters dry soil that lacks a water supply. In the growth process of a root in SRS, *MP* increases as much as the *Encourage_Value* or decreases as much as the *Penalty_Value* if the root arrives at a location that is better than its current location or does not, respectively. If *MP* decreases to a predetermined drouth threshold, the root dries out. To clarify the drouth process, a comparative sample is presented in Figure 5. This figure shows how two different roots are penalized or encouraged after maturing in iteration 40.

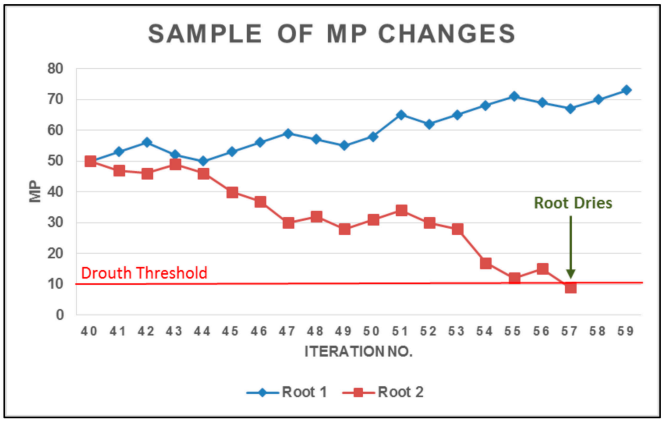


Figure 5. Changes in moisture percentage (MP) value of roots over puberty.

Encourage_Value is a constant, however *Penalty_Value* varies in different subspaces depending on the subspace rank. Thus, weak subspaces possess higher *Penalty_Values*, and accordingly, SRS dries the roots of weak subspaces earlier. To this end, a maximum penalty rate is considered, called *Max_Penalty*. thereupon, the *Penalty_Value* of subspace *j*, $Penalty\_Value_j$, will be calculated using Equation (12) where *j*, is the index and also rank of the subspace among all available subspaces that are sorted according to their best root. *Penalty_change* is calculated using Equation (11), where *Ns* is the number of available subspaces, and the decimal parameter called *Penalty_Rate* has a value in the range [0, 1]. *Penalty_Rate* strongly affects the convergence speed of the algorithm and should be defined by the user with respect to the problem characteristics. Smaller values of the *Penalty_Rate*, smaller *Penalty_values* for strong subspaces, and higher values of *Penalty_Rate*, smaller differences between the *Penalty_values* of weak and strong subspaces. We note that the rank of the best root of a subspace in the list of all existing roots determines the rank of that subspace.

$$Penalty\ \_Change = \frac{Max\_Penalty\ (1 - Penalty\_Rate)}{Ns} \tag{11}$$

$$Penalty\_Value_j = \begin{cases} Max\_Penalty * \ Penalty\_Rate, & j = 1 \\ Penalty\_Value_{j-1} + Penalty\ \_Change, & otherwise \end{cases} \tag{12}$$

Although Root Branching and Drouth cooperate to control the number of available roots, the SRS will likely encounter an overloaded root set, leading to a slow down. Thus, an auxiliary control mechanism is needed to initiate branching. Here, *NumMaxRoot* is defined as a threshold for the maximum number of roots currently active in the search space. If the number of active roots reaches *NumMaxRoot*, which is defined by the user, none of the roots can generate a branch, even they are otherwise eligible to do so, until vacancies are produced by the drying out of deficient roots drying.

2.2.9. HRG

HRG consists of the five following steps that begin with determining the number of roots can generate HRs and end with growth of the parent toward the best neighbor location found by the HRs.

1. A set of the best mature roots (let us call them *m*) of every subspace will be picked by "*Roulette Wheel Selection* via *Stochastic Acceptance*" (RWSSA) [60] for generating plenty of HRs in their neighborhood regions.
2. A random number *l* in the range (1, *a*) will be generated, where *a* is the neighborhood radius defined by the user based on the problem characteristics.

3. For every selected mature root i in step 1, RWSSA will be used to generate a random number k in the range (1, D).

    a. k of D dimensions of root i will be selected randomly.

    b. For every selected dimension d in 3.a, two new roots, $HR1_i^d = \left(x_i^1, x_i^2, \ldots, x_i^d + l, \ldots, x_i^D\right)$ and $HR2_i^d = \left(x_i^1, x_i^2, \ldots, x_i^d - l, \ldots, x_i^D\right)$, will be generated, and their nitrate concentration needs will be calculated.

4. If HRj_i^d is one of the generated HRs with the greatest nitrate concentration value among the other HRs and their parents, then the parent will grow to reach the location of HRj_i^d.
5. The generated HRs are no longer required and will dry immediately.

There are two HRG points that must be clarified. First, based on the 3rd step in the HRG, the $2k$ HRs will be generated in the neighboring area of a selected mature root in a D-dimensional space. Accordingly, in every iteration, a total of $2mk$ HRs will be generated. Second, the best rate of executing the HRG (i.e., the so-called HRG Rate) can be predefined by the user or calculated dynamically. Finding the best HRG Rate requires new dependent research, which we suggest as a possibility for the future.

2.2.10. Termination Criteria

Consistent with all other heuristic search algorithms, SRS execution will be stopped if at least one of the following criteria is met: (1) reaching the expected solution or (2) exceeding the maximum number of iterations. Once the algorithm stops, the best-found root will be shown as the final result of the algorithm. The SRS pseudo code is represented in Figure 6.

begin
Initialize parameters
Divide search space into *Ns* subspaces
Initialize population of subspaces
Evaluate of roots (Nitrate calculation) using Equation (3)
Sort roots
while (the termination criteria are not reached)
    for **(**each root)
        - Calculate velocity of the root using Equation (6)
        - Calculate θ angles of the root according to the best closest set of roots in each subspace using Equation (7)
        - **Grow the root** (Update position of the root using velocity and θ angles) using Equations (8) and (9)
        - **Evaluate of the root**
        - if ((root is mature) && (SCS >= MRA) && (NumCurrRoot < NumMaxRoot)) then
        - **Generate an immature root** (Based on parent location, velocity & random angles)
        - **Evaluate of the immature root**
        - end if
        - Update encouragement or punishment values of the root moisture using Equations (11) and (12)
        - if (MP < threshold) then
        - **Drouth of the root**
    end for
    Sort Roots
    Germinate Hair roots for the best closest set of roots for all subspaces based on HairRootRate
end while
return best obtained root
end

SCS: sum of collected scores for branching
MRA: a user-predefined threshold value for branching
MP: a volume of moisture supplied

Figure 6. Pseudo code for SRS algorithm.

2.2.11. On the Convergence of the SRS

All elements of a search algorithm should cooperate to provide an average convergence rate that is not too fast or slow to reach the global best solution [61,62]. The SRS mechanisms are designed so that convergence is supported by two different but complementary approaches. First, each subspace has a local convergence rate when roots converge toward a few of their best local solutions. Here, the aim is finding the local optimal solution of every subspace. Second, global convergence occurs when, in the execution process, the roots of worse existing subspaces dry out based on the dynamic punishment process explained in Section 2.2.7 until there is only one subspace remaining at the last iterations of the algorithm. Finally, the local convergence of the last subspace together with the root drouth process will be leading to the identification of the global optimal solution of the problem. Figure 7 simply illustrates how available roots in different subspaces converge to their corresponding local optima and consequently facilitate reaching global optima in the best subspace by drying all the roots in the other subspaces.

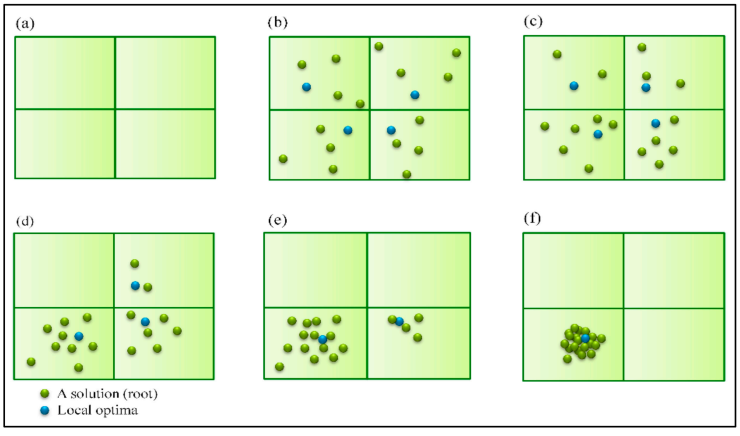


Figure 7. Converges of SRS toward global optima; (**a**) shows how search space gets divided into four subspaces; (**b**) shows process of generating random solutions in subspaces and choosing optimal solution of each of them; (**c**) demonstrates how number of solutions gets less in subspaces that could not find proper solutions while it gets more in promising subspaces; (**d**,**e**) represent the worst subspaces having no solution contrasted with the best subspace determined gradually; (**f**) SRS found the global optima in the best subspace.

3. Experimental Test, Results and Discussion

3.1. Experimental Tests: SRS vs. GA, PSO, Independent Component Analysis (ICA) and Differential Evolution (DE)

This section presents the results obtained by the SRS while searching for the optimal solution of the test functions described in Section 3.1.2. Here, performance of SRS is analyzed and compared with that of the employed comparative algorithms. A standard SRS is presented, and standard versions of GA, PSO, ICA and DE are employed as comparative algorithms. The best solution reached by each algorithm, the standard deviations of different solutions achieved in different runs by every algorithm, and the rank of each among all comparative algorithms are presented in the tables prepared for every test function.

This section consists of three subsections that explain the experimental methodology and settings, test functions and their specifications, and obtained results.

3.1.1. Settings

All algorithms were implemented in Microsoft Visual C# .NET 2015. To provide fair conditions for all comparative algorithms, common parameters, such as population size and the total number of evaluations of the objective function, were chosen to be the same for each algorithm. Referring to [63], the population size was selected to be 125. The maximum number of evaluations of the objective function was 1,000,000 to allow all algorithms to achieve the best possible solution. The other parameters are given below:

For GA, second point crossover operation, which influences the variation in generations, with a rate of 0.85 was employed as recommended by [64]. Mutation operation, which controls genetic diversity, was also set to 0.01 based on [64].

In PSO, *C1* and *C2* are constant coefficients that change the weighting of personal and global experiences, respectively, and both were set to 2 in our experiments. The inertia weight that demonstrates how particles' previous velocity influences the subsequent velocity, was chosen to be 0.9, as recommended by [65].

In ICA, the imperialist rate states how many countries will be selected as imperialists. In addition, the competition rate is the second parameter of ICA, and it determines number of times that imperialists participate in a competition to take the weakest colony of the weakest emperor. To follow the research methodology described by [22], the mentioned parameters were set to 10 and 15, respectively.

F is a constant that can be used to manipulate the differential variation between two solutions in DE. It was selected to be 0.5 in our setting. The crossover rate value, which controls the changes in the diversity of the population, was set to 0.9, as recommended by [66].

SRS has effective parameters in its different parts. Assigning proper values for or defining equations to calculate these parameters is outside of the scope of this paper and requires independent studies. Hence, in this study, constant values are given to these parameters, as shown in Table 6, where the MDV is the Max Domain Value of the problem. This table shows that the assigned values of the Mature Age and Penalty Rate parameters are different for unimodal and multimodal functions. Given that the risk of falling into the trap of local optima increases for multimodal functions as the number of dimensions increases, assigning higher and lower values to the Mature Age and Penalty Rate parameters relative to those used for unimodal functions helps the SRS to explore areas that are slightly more distant from the current regions of active roots more thoroughly. Reversing these assignments for unimodal functions improves the ability of SRS to perform more exploitation around current roots and thus find better solutions.

Table 6. SRS parameters setting.

	Ns	**NumMinRoot (Population Size)**	**NumMaxRoot**	**Vmax**	**MRA**	**Max_Penalty**	**Encourage_Value**	**Penalty_Rate**	**Mature_Age**
Unimodal Functions	8	125	2000	0.33 * MDV	20	10	2	0.75	4
Multimodal Functions	8	125	2000	0.33 * MDV	20	10	2	0.15, 0.25	15

3.1.2. Benchmark Functions

Many benchmark test functions are presented in the literature that are designed to provide identical comparison environments to evaluate the performance of optimization algorithms [67–71]. As a common methodology, these test functions are being used to test and validate the performance and efficiency of new optimization algorithms. Most known test functions can be categorized into unimodal and multimodal groups. Compared to unimodal functions, which have one and only one optimal solution, multimodal test functions have a number of optimal solutions, including global and local ones, making them suitable to evaluate the ability of an algorithm to avoid becoming trapped in local optima. For multimodal test functions, if an algorithm has a poor exploration process that cannot search the whole search space efficiently, it will inevitably fall into the trap of local optima.

Among the many different test functions available, a set of 24 most common ones were chosen in this research for comparing the performance of the GA, PSO, ICA, DE and SRS. This set includes 12 unimodal (f_1–f_{12}) and 11 multimodal (f_{13}–f_{23}) functions, which are listed in Tables 7 and 8, respectively. The dimensions (D), domain ranges (Range), minimum value (f_{min}) and formulations of the employed test functions are listed in the mentioned tables.

Table 7. Unimodal Test Functions.

No.	Function	D	Range	f_{min}	Formulation
f_1	Cigar	30	$[-100, 100]$	0	$x_1^2 + 10^6 \sum_{i=2}^{D} x_i^2$
f_2	Dixon-Price	30	$[-10, 10]$	0	$(x_1 - 1)^2 + \sum_{i=2}^{D} i\left(2x_i^2 - x_{i-1}\right)^2$
f_3	Quartic	30	$[-1.28, 1.28]$	0	$\sum_{i=1}^{D} ix_i^4 + random[0, 1)$
f_4	Rosenbrock	30	$[-5, 5]$	0	$\sum_{i=1}^{D-1} 100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2$
f_5	Schwefel 1.2	30	$[-100, 100]$	0	$\sum_{i=1}^{D}\left(\sum_{j=1}^{i} x_j\right)^2$
f_6	Schwefel 2.22	30	$[-100, 100]$	0	$\sum_{i=1}^{D} \lvert x_i \rvert + \prod_{i=1}^{D} \lvert x_i \rvert$
f_7	Schwefel 2.23	30	$[-10, 10]$	0	$\sum_{i=1}^{D} x_i^{10}$
f_8	Sphere	30	$[-100, 100]$	0	$\sum_{i=1}^{D} x_i^2$
f_9	Step	30	$[-100, 100]$	0	$\sum_{i=1}^{D} (\lfloor x_i + 0.5 \rfloor)^2$
f_{10}	SumSquares	30	$[-10, 10]$	0	$\sum_{i=1}^{D} ix_i^2$
f_{11}	Trid 10	10	$\left[-D^2, D^2\right]$	0	$210 + \sum_{i=1}^{D} (x_i - 1)^2 - \sum_{i=2}^{D} x_i x_{i-1}$
f_{12}	Zakharov	10	$[-5, 10]$	0	$\sum_{i=1}^{D} x_i^2 + \left(0.5 \sum_{i=1}^{D} ix_i\right)^2 + \left(0.5 \sum_{i=1}^{D} ix_i\right)^4$

Column D represents the number of dimensions of the problem search space.

Table 8. Multimodal Test Functions.

No.	Function	D	Range	f_{min}	Formulation
f_{13}	Ackley	30	[−32, 32]	0	$20 + e - 20e^{\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right)} - e^{\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)}$
f_{14}	CosineMixture	30	[−500, 500]	0	$D \times 1.643788341 - 0.1\sum_{i=1}^{D}\cos(5\pi x_i) - \sum_{i=1}^{D} x_i^2$
f_{15}	Griewank	30	[−600, 600]	0	$\frac{1}{4000}\left(\sum_{i=1}^{D} x_i^2\right) - \left(\prod_{i=1}^{D}\cos\left(\frac{x_i}{\sqrt{i}}\right)\right) + 1$
f_{16}	Perm	4	[−D, D]	0	$\sum_{i=1}^{D}\left(\sum_{j=1}^{D}\left(j^i + \beta\right)\left(\left(\frac{x_j}{j}\right)^i - 1\right)\right)^2,\ (\beta > 0)$
f_{17}	Qing	30	[−10, 10]	0	$\sum_{i=1}^{D}\left(x_i^2 - i\right)^2$
f_{18}	Quintic	30	[−1, 1]	0	$\sum_{i=1}^{D}\left\lvert x_i^5 - 3x_i^4 + 4x_i^3 + 2x_i^2 - 10x_i - 4\right\rvert$
f_{19}	Rastrigin	30	[−5.12, 5.12]	0	$\sum_{i=1}^{D}\left(x_i^2 - 10\cos(2\pi x_i) + 10D\right)$
f_{20}	Schwefel	30	[−500, 500]	0	$D \times 418.9829 + \sum_{i=1}^{D} -x_i \sin\left(\sqrt{\lvert x_i\rvert}\right)$
f_{21}	Schwefel 2.25	30	[0, 10]	0	$\sum_{i=2}^{D}\left[(x_i - 1)^2 + \left(x_1 - x_i^2\right)^2\right]$
f_{22}	Styblinski_Tang	30	[−5, 5]	0	$D \times 39.16599 + \frac{1}{2}\sum_{i=1}^{D}\left(x_i^4 - 16x_i^2 + 5x_i\right)$
f_{23}	Xin-She Yang 02	30	[−2π, 2π]	0	$\frac{\sum_{i=1}^{D}\lvert x_i\rvert}{exp\left(\sum_{i=1}^{D}\sin\left(x_i^2\right)\right)}$

Column D represents the number of dimensions of the problem search space.

3.2. Obtained Results and Discussion

The SRS and all comparative algorithms were repeated 40 times to solve every test function by employing different seed values. As all investigated test functions are minimization problems, the minimum objective function values obtained at the end of every execution were used to calculate the average value of obtained results (mean) and standard deviation (Std) of the results of the algorithms. The obtained mean values were also used to determine the ranks of the algorithms for different test functions. The mean and standard deviation of unimodal and multimodal function values obtained by GA, PSO, ICA, DE and SRS are shown in Tables 9 and 10, respectively, together with the achieved ranks.

Table 9 shows that SRS outperforms all the comparative algorithms and reaches rank 1 in 11 of 12 test functions for unimodal functions. The DE algorithm reaches rank 1 in solving the Quartic function, while SRS reaches rank 4. Thus, SRS successfully achieved rank 1 in 91.67% of the times it attempted to solve unimodal test functions and reached rank 1.25 on average (Table 11). Hence, for unimodal test functions, the overall search performance is SRS > DE > PSO > ICA > GA. The comparison is illustrated in Figure 8 too. To compare consistency of the algorithms in solving a problem in different executions, standard deviation values are also provided in Table 9. The table indicates that the SRS has obtained the lowest standard deviation value in solving all the problems but the one it could not gain rank 1.

Definition 3. *Distance Score (DS) shows how many times of the best solution of a base algorithm the best obtained solution of an algorithm is far from the global optimal solution of the problem. For instance, if the base algorithm is SRS, and the global optimal solution, the best solution of the SRS and the best solution of DE are 0, 1 and 10, respectively, then DS of DE is* $\frac{10}{1} = 10$ *. To avoid possible very large numbers, we can use the* Log_{10} *of Distance Score (LDS). Accordingly, LDS of DE in the given example is* $Log_{10}^{\left(\frac{10}{1}\right)} = 1$.

Table 9. Comparison of SRS with GA, PSO, ICA and DE on Unimodal test functions. All results have been averaged over 40 runs.

No.	Function		GA	PSO	ICA	DE	SRS
f_1	Cigar	Mean	6.62×10^8	1.12×10^9	8.44×10^8	9.33×10^8	1.11×10^7
		Std	2.51×10^8	1.5×10^8	4.93×10^8	1.37×10^9	2.62×10^6
		Rank	2	5	3	4	**1**
f_2	Dixon-Price	Mean	677.43	276.84	2330.86	900.74	1.73
		Std	420.64	86.26	2929.02	1330.92	0.63
		Rank	3	2	5	4	**1**
f_3	Quartic	Mean	18.88	11.94	11.03	9.53	14.003
		Std	2.91	0.78	0.64	0.29	1.4
		Rank	5	3	2	**1**	4
f_4	Rosenbrock	Mean	528.93	437.82	545.02	258.45	25.55
		Std	315.32	106.66	418.31	139.21	2.85
		Rank	4	3	5	2	**1**
f_5	Schwefel 1.2	Mean	37,415.48	7301.03	1652.85	1109.23	883.58
		Std	9582.99	1350.01	520.67	730.18	204.17
		Rank	5	4	3	2	**1**
f_6	Schwefel 2.22	Mean	94.55	236.15	60.625	85	21
		Std	14.01	15.89	22.01	56.19	3.96
		Rank	4	5	2	3	**1**
f_7	Schwefel 2.23	Mean	25.05	7.48	5127.09	104,269.3	0
		Std	46.41	4.46	11,643.33	542,120.6	0
		Rank	3	2	4	5	**1**
f_8	Sphere	Mean	885.05	3195.13	800.43	727.2	11.75
		Std	287.48	525.35	395.63	1287.79	2.8
		Rank	4	5	3	2	**1**
f_9	Step	Mean	811.78	3227.32	921.43	760.93	11.78
		Std	315.9	399.79	569.69	1298.97	2.96
		Rank	3	5	4	2	**1**
f_{10}	SumSquares	Mean	104.36	74.59	125.65	107.42	0.52
		Std	40.67	16.14	43.7	136	0.31
		Rank	3	2	5	4	**1**
f_{11}	Trid 10	Mean	454.98	22.78	133.88	95.05	2.88
		Std	444.18	13.52	68.36	215.89	8.8
		Rank	5	2	4	3	**1**
f_{12}	Zakharov	Mean	58.9	0.032	2.31	1.28	0.003
		Std	27.98	0.01	2.04	3.31	0.002
		Rank	5	2	4	3	**1**

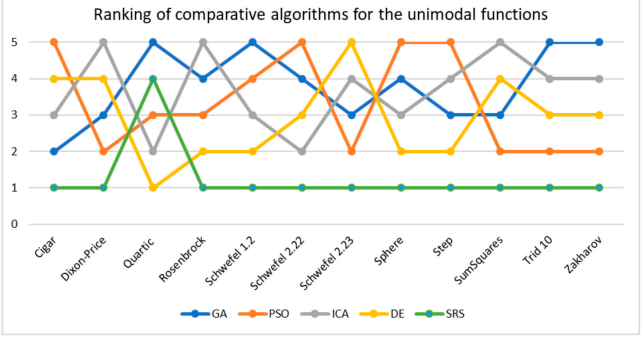


Figure 8. How SRS outperforms the other algorithms in benchmarking unimodal test function.

Table 10. Comparison of SRS with GA, PSO, ICA and DE on Multimodal test functions. All results have been averaged over 40 runs.

No.	Function		GA	PSO	ICA	DE	SRS
f_{13}	Ackley	Mean	7.88	5.64	8.67	10.43	0.64
		Std	0.79	0.36	1.21	1.47	0.16
		Rank	3	2	4	5	**1**
f_{14}	CosineMixture	Mean	2.993	2.999	3.296	2.978	2.975
		Std	0.019	0.016	0.063	0.015	0
		Rank	3	4	5	2	**1**
f_{15}	Griewank	Mean	6.43	10.61	13.38	6.7	0.99
		Std	1.65	1.83	6.27	4.16	0.05
		Rank	2	4	5	3	**1**
f_{16}	Perm	Mean	4.941	0.057	0.286	0.292	0.012
		Std	10.35	0.075	0.19	1.16	0.017
		Rank	5	2	3	4	**1**
f_{17}	Qing	Mean	3.29×10^7	8.71×10^7	3.02×10^8	2.38×10^8	2.67×10^3
		Std	2.91×10^7	3.50×10^7	2.83×10^8	4.47×10^8	4.90×10^2
		Rank	2	3	5	4	**1**
f_{18}	Quintic	Mean	50.4	86.05	149.16	63.38	15.91
		Std	9.33	8.69	84.74	63.89	5.02
		Rank	2	4	5	3	**1**
f_{19}	Rastrigin	Mean	46.04	132.34	56.02	60.66	39.92
		Std	9.03	9.21	13.7	16.68	6.69
		Rank	2	5	3	4	**1**
f_{20}	Schwefel	Mean	1499.5	4553.8	7530.7	5204.3	4314.7
		Std	336.51	339.53	595.61	587.37	848.04
		Rank	**1**	3	5	4	2
f_{21}	Schwefel 2.25	Mean	1745.65	98,272.93	25,246.87	7037.74	754.08
		Std	637.03	17,093.3	7069.6	3546.7	523.5
		Rank	2	5	4	3	**1**
f_{22}	Styblinski-Tang	Mean	24.49	63.27	408.48	220.37	137.1
		Std	5.437	14.68	37.13	38.46	26.71
		Rank	**1**	2	5	4	3
f_{23}	Xin-She Yang 02	Mean	9.15×10^{-12}	8.88×10^{-10}	7.41×10^{-10}	2.67×10^{-11}	4.33×10^{-12}
		Std	1.64×10^{-12}	5.91×10^{-10}	9.15×10^{-10}	2.60×10^{-11}	5.72×10^{-13}
		Rank	2	5	4	3	**1**

Table 11. The average of achieved rank, and percentage of reaching rank 1 by the comparative algorithms.

	GA	PSO	ICA	DE	SRS
Average Rank for Unimodal test functions	3.83	3.33	3.66	2.91	1.25
Average Rank for Multimodal test functions	2.27	3.54	4.36	3.54	1.27
Average Rank for all test functions	3.09	3.43	4	3.22	1.26
Percentage of reaching rank 1 for Unimodal test functions	0	0	0	8.33%	91.67%
Percentage of reaching rank 1 for Multimodal test functions	18.18%	0	0	0	81.81%
Percentage of reaching rank 1 for all test functions	8.7%	0	0	4.35%	86.96%

To further investigate the obtained results, LDS of all comparative algorithms are calculated by taking SRS as the base algorithm and visualized in Figure 9. As the graph indicates GA possesses the highest LDS at 2.08 on average for all unimodal functions followed by ICA and DE with 2.03 and 2.02, respectively. The closest result to SRS belongs to PSO at 1.68 on average. To make sure that the results of Schwefel 2.23 does not impact the analysis, we have included the median of the results as well. The median of LDS of all algorithms shows that GA and ICA have the highest values at 1.86 on average followed by DE while PSO shows relatively lower number at 1.62. Thus, we can conclude that the SRS has reached to the closest solutions to the global optimal solution in benchmarking the unimodal functions.

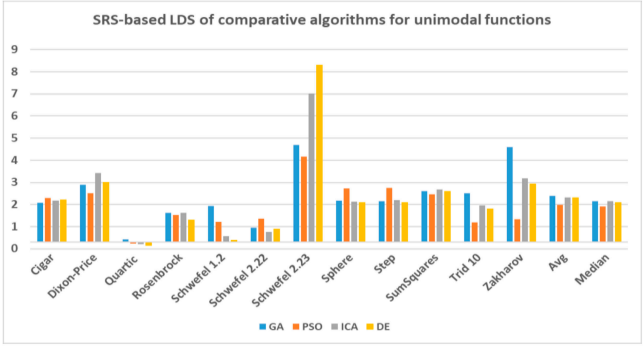


Figure 9. LDS of the comparative algorithms for unimodal functions based on SRS.

On the other hand, Table 10 demonstrates that SRS performs better than all the comparative algorithms and reaches rank 1 in 9 of the 11 multimodal test functions. As Figure 10 also represents, GA and PSO reached rank 1 in solving the Styblinski-Tang and Levy 8 functions, respectively. Hence, as shown in Table 11, 81.81% of the time, SRS achieves rank 1 in solving multimodal test functions and rank 1.27 on average. Accordingly, the overall search performance can be concluded as SRS > GA > PSO > DE > ICA. Furthermore, consistency of the algorithms' behavior in different runs can be assessed by comparing achieved standard deviation values. Table 10 shows that the consistency of the SRS was premier whenever it has reached rank 1.

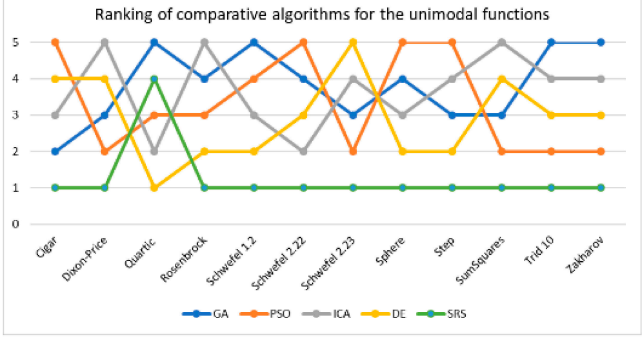


Figure 10. How SRS outperforms the other algorithms in benchmarking multimodal test function.

Furthermore, LDS of the all comparative algorithms are also computed by taking SRS as the base algorithm and illustrated in Figure 11. As the visual shows ICA has the highest LDS at 1.3 on average for all multimodal functions followed by PSO and DE with 1.14 and 1.1, respectively. The closest result to SRS belongs to GA at 0.78 on average. We can use median of results rather than average to remove impacts of Qing solutions. The median of LDS of all algorithms shows that ICA and DE have the highest values at 1.13 and 0.79 on average, respectively, followed by PSO while GA reaches significantly lower number at 0.36. Accordingly, it can be concluded that the SRS has reached to the closest solutions to the global optimal solution in benchmarking the multimodal functions, however it requires better parameter tuning for these functions.

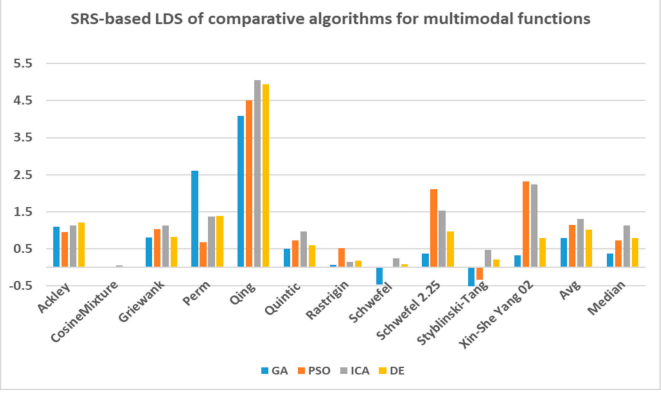


Figure 11. LDS of the comparative algorithms for unimodal functions based on SRS.

The aforementioned results indicate that the SRS can be considered a strong search algorithm for solving both unimodal and multimodal test functions. The exploration ability of SRS is stronger than those of the comparative algorithms for solving multimodal functions, while its effective exploitation features make it the best algorithm for finding the optimal values of unimodal functions. Tables 9 and 10 indicate that the SRS obtained the nearest to optimal solutions for most of the test functions, and the differences between the solutions achieved by the SRS and those of the comparative algorithms were significant for some test functions, such as Dixon-Price, Rosenbrock, Schwefel 2.23, Sphere, Step, SumSquares, Trid 10, Ackley and Qing.

Eventually, to compare the performance of SRS with the comparative algorithms for unimodal and multimodal functions altogether, we used all obtained mean values of all algorithms, what are shown in Tables 9 and 10, for clustering the algorithms to show whether the SRS results fall in separate cluster from other algorithms or not. To this aim, Agglomerative Hierarchical clustering algorithm [72,73] was used in Python programming language to conduct clustering and resulting clusters plotted as a Dendrogram diagram [74] and shown in Figure 12. According to the graph, the four comparative algorithms have fallen in one cluster together, cluster A whish is colored by orange, while SRS is lonely in the cluster B, in blue. Furthermore, the distance of cluster B from cluster A is as much as we can conclude that these two clusters are significantly far from each other. Accordingly, SRS performance is dramatically different form the other algorithms used in this experiment and as SRS solutions are closer to the global optimal solution of the test functions, we can conclude that SRS performance is significantly better than the others.

In addition to the experimental tests, statistical tests can also be useful to show significant differences between the results obtained using SRS and the comparative algorithms. To this end, Friedman [75] and one-way analysis of variance (ANOVA) [76] tests were conducted. One-way ANOVA is also used to determine whether the means of two or more independent sets of data are statistically significantly different. The p-value computed by the Friedman test for the algorithm was 39.269. This p-value is greater than the critical value of 9.4877, which represents $\alpha = 0.05$ and 4 degrees of freedom (DFs) in the Chi-Square distribution table [77]. Therefore, significant differences existed among the results obtained by the algorithms. The results of the one-way ANOVA are presented in Tables 12 and 13. These results also revealed that there was a statistically significant difference in the mean value of the SRS relative to the five investigated algorithms for every utilized test function.

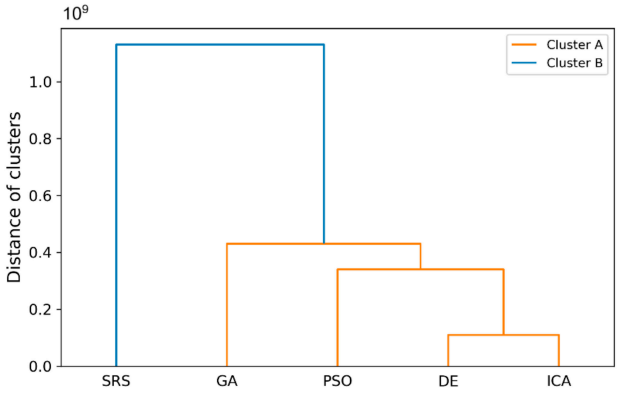


Figure 12. Dendrogram diagram plotted using Agglomerative hierarchical clustering results of comparative algorithms.

Table 12. The one-way ANOVA test results for Unimodal test functions.

No.	Function	Source of Variation	Sum of Squares (SS)	df	Mean Square (MS)	F	*p*-Value	F Criteria
f_1	Cigar	Between Groups Within Groups	3.29×10^{23} 1.33×10^{24}	4 26,940	8.23×10^{22} 4.95×10^{19}	1661.663	0.0000	2.372262
f_2	Dixon-Price	Between Groups Within Groups	8.68×10^{13} 5.21×10^{14}	4 26,973	2.17×10^{13} 1.93×10^{10}	1123.734	0.0000	2.372262
f_3	Quartic	Between Groups Within Groups	1,686,230 2,148,251	4 26,980	421,557.5 79.62382	5294.365	0.0000	2.372261
f_4	Rosenbrock	Between Groups Within Groups	1.63×10^{12} 9.17×10^{12}	4 26,981	4.07×10^{11} 3.4×10^{8}	1196.571	0.0000	2.372261
f_5	Schwefel 1.2	Between Groups Within Groups	1.18×10^{13} 1.1×10^{12}	4 26,953	2.95×10^{12} 40,871,333	72,289.63	0.0000	2.372262
f_6	Schwefel 2.22	Between Groups Within Groups	2.45×10^{8} 5.9×10^{8}	4 26,859	61,363,422 21,971.96	2792.806	0.0000	2.372263
f_7	Schwefel 2.23	Between Groups Within Groups	4.55082×10^{20} 4.69513×10^{21}	4 26,947	1.14×10^{20} 1.74×10^{17}	652.9695	0.0000	2.372262
f_8	Sphere	Between Groups Within Groups	4.45×10^{11} 1.68×10^{12}	4 26,938	1.11×10^{11} 62,294,092	1787.44	0.0000	2.372262
f_9	Step	Between Groups Within Groups	4.03×10^{11} 1.57×10^{12}	4 26,987	1.01×10^{11} 58,023,788	1738.198	0.0000	2.372261
f_{10}	SumSquares	Between Groups Within Groups	8.83×10^{9} 2.76×10^{10}	4 26,945	2.21×10^{9} 1,024,844	2154.738	0.0000	2.372262
f_{11}	Trid 10	Between Groups Within Groups	8.92×10^{9} 1.69×10^{10}	4 26,954	2.23×10^{9} 626,415.4	3561.234	0.0000	2.372262
f_{12}	Zakharov	Between Groups Within Groups	48487053 1.02×10^{9}	4 27,699	12,121,763 36,707.12	330.2292	0.0000	2.372253

Table 13. The one-way ANOVA test results for Multimodal test functions.

No.	Function	Source of Variation	Sum of Squares (SS)	df	Mean Square (MS)	F	*p*-Value	F Criteria
f_{13}	Ackley	Between Groups	207,007	4	51,751.75	8462.465	0.0000	2.372261
		Within Groups	165,251.6	27,022	6.115446			
f_{14}	CosineMixture	Between Groups	181.8325	4	45.45813	12,153.57	0.0000	2.372261
		Within Groups	100.9847	26,999	0.00374			
f_{15}	Griewank	Between Groups	30,220,789	4	7,555,197	1710.236	0.0000	2.372262
		Within Groups	1.19×10^{8}	26,970	4417.633			
f_{16}	Perm	Between Groups	285,060.8	4	71,265.21	11,057.39	0.0000	2.372257
		Within Groups	176,497.1	27,385	6.445029			
f_{17}	Qing	Between Groups	9.84×10^{23}	4	2.46×10^{23}	868.8985	0.0000	2.372261
		Within Groups	7.65×10^{24}	27,008	2.83×10^{20}			
f_{18}	Quintic	Between Groups	1.3×10^{12}	4	3.26×10^{11}	797.0792	0.0000	2.372261
		Within Groups	1.1×10^{13}	26,984	4.08×10^{8}			
f_{19}	Rastrigin	Between Groups	34,201,633	4	8,550,408	2962.64	0.0000	2.372261
		Within Groups	77,851,945	26,975	2886.078			
f_{20}	Schwefel	Between Groups	2.68×10^{10}	4	6.71×10^{9}	4296.281	0.0000	2.372261
		Within Groups	4.21×10^{10}	26,991	1,561,568			
f_{21}	Schwefel 2.25	Between Groups	2.1×10^{14}	4	5.26×10^{13}	2194.452	0.0000	2.372261
		Within Groups	6.46×10^{14}	26,978	2.4×10^{10}			
f_{22}	Styblinski-Tang	Between Groups	1.86×10^{8}	4	46,522,011	5330.031	0.0000	2.372261
		Within Groups	2.36×10^{8}	27,001	8728.281			
f_{23}	Xin-She Yang 02	Between Groups	1.99×10^{-7}	4	4.97×10^{-8}	77.35719	0.0000	2.372261
		Within Groups	1.73×10^{-5}	26,996	6.42×10^{-10}			

4. Conclusions and Future Work

In this paper, a high-performance combinatorial search algorithm called SRS is introduced. SRS was inspired by the plant root growth in soil that occurs to find higher densities of nutrition and water. By mapping solutions to the root and then utilizing three different types of roots with different search characteristics, the algorithm shows high efficiency in both exploration and exploitation activities. Mature roots are responsible for exploration, while immature roots help the SRS escape from local optima traps and non-promising points. Meanwhile, hair roots as very short life searching elements try to search around the best-found solutions for finding higher satisfying points. To evaluate the performance of the SRS and compare its efficiency with those of other algorithms, a complete experimental test was conducted. Twenty-four unimodal and multimodal test functions were employed, and GA, PSO, ICA and DE were applied as comparative algorithms for SRS to find the optimal values of the test functions. To ensure that the collected results were reliable and accurate, every algorithm was executed 40 times per test function, and the average of the results was used in the comparisons.

In investigating unimodal test functions, the collected results demonstrated that the SRS performed significantly better than the comparative algorithms, except for the quartic function. Therefore, the SRS won the competition for 91.67% of the functions. For the multimodal test functions, the achieved results indicated that the SRS prevailed 81.81% of the time. Overall, the aggregated results for the unimodal and multimodal test functions indicated that SRS is superior to the comparative algorithms for 86.96% of the test functions used. The higher efficiency in addressing unimodal and multimodal functions demonstrates that the SRS has strong ability to coordinate exploitation and exploration in a way that local optima points are not able to catch it into the traps. Briefly, based on the experimental results and discussion provided here, it can be concluded that the SRS is a formidable competitor for currently well-known combinatorial search algorithms in solving different types of optimization search problems. Well-defined structures and carefully tuned operators are the strengths of the SRS to solving np-hard optimization problems.

Despite the discussed capabilities of the SRS, there are some aspects of the algorithm that can still be upgraded. In terms of functional settings, the SRS enjoys many parameters to customize the

functionality of the algorithm. Different values for these parameters may cause different effects in solving various types of optimization problems. Therefore, in the future, efforts should be focused on finding effective values for these parameters. Furthermore, these parameters exert mutual effects on the performance and final achieved solutions. Additional research is needed to regulate the relations among these parameters to be able to run the algorithm effectively by setting a smaller number of parameters. In addition, SRS must be applied in solving several benchmarks as well as real-world optimization problems to determine its strengths and weaknesses as a further matter, the HRG operation was not utilized in the conducted experimental test to provide a more suitable test environment for examining the core exploration and exploitation behaviors of the mature and immature roots. In the future, the performance of the SRS will be achieved by applying a well-organized HRG when solving np-hard problems.

Author Contributions: Methodology, N.K.N.; software, N.K.N. and A.J.; validation, E.A.S. and M.A.; formal analysis, N.K.N.; investigation, N.K.N.; resources, N.K.N. and E.A.S.; writing—original draft preparation, N.K.N.; writing—review and editing, N.K.N., E.A.S. and M.A.; visualization, N.K.N. and A.J.; supervision, E.A.S. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Kebangsaan Malaysia, Research University grant numbered DIP-2018-041 and grant numbered FRGS/1/2014/ICT07/UKM/02/1.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. An Example of Root Growth

A complete example on how a mature root grows from its current location ($x_i(t)$) to next location ($x_i(t+1)$) is presented in this section. The example shows how SRS calculates velocity and dimensional velocities of a particular root, determines direction of growth for different dimensions and eventually identifies next location of the root. The example starts by presenting current parameters of the SRS as well as current properties of the investigated root in Table A1. Then, in Figure A1, it shows the calculations step by step until next location of the root is identified in the last row of the table.

Step 1: Next velocity of the root using Equation (6)	
$v_i(t) = 10 - \left[10\left(1 - \frac{36}{50}\right)\right] = 8$	

Step 2: Growth direction of all dimensions using Equation (7)	
$cos\theta^1 = \frac{35-83}{126.6} = -0.3791$	$cos\theta^2 = \frac{93-15}{126.6} = 0.6161$
$cos\theta^3 = \frac{66-7}{126.6} = 0.466$	$cos\theta^4 = \frac{17-43}{126.6} = -0.2054$
$cos\theta^5 = \frac{21-79}{126.6} = -0.4581$	$cos\theta^6 = \frac{4-15}{126.6} = -0.0869$
Step 3: Dimensional velocities of the root using Equation (9)	
$v_i^1(t) = 8 \times (-0.3791) = -3.03 \cong -3$	$v_i^2(t) = 8 \times (0.6161) = 4.93 \cong 5$
$v_i^3(t) = 8 \times (0.466) = 3.73 \cong 4$	$v_i^4(t) = 8 \times (-0.2054) = -1.6 \cong -2$
$v_i^5(t) = 8 \times (-0.4581) = -3.6 \cong -4$	$v_i^6(t) = 8 \times (-0.0869) = -0.7 \cong -1$

Step 4: New coordinates of the root to grow toward using Equation (8)	
$x_i^1(t+1) = 83 - 3 = 80$	$x_i^2(t+1) = 15 + 5 = 20$
$x_i^3(t+1) = 7 + 4 = 11$	$x_i^4(t+1) = 43 - 2 = 41$
$x_i^5(t+1) = 79 - 4 = 75$	$x_i^6(t+1) = 15 - 1 = 14$
$x_i(t+1) = (80, 20, 11, 41, 75, 14)$	

Figure A1. Step by step calculations of the example.

Table A1. SRS Parameters and current properties of the investigated root in the root growth example.

SRS Parameters			**The Investigated Root Current Properties**		
Max velocity	*NumCurrRoot*	Dimension	glb_rank_i	Current location $x_i(t)$	Best closest $x_{best\_closest\ i}(t)$
10	50	6	36	(83, 15, 7, 43, 79, 15)	(35, 93, 66, 17, 21, 4)

Appendix B. On the Time Complexity of the SRS

In computer science, time complexity is used to show the amount of time an algorithm requires to run. Time complexity is calculated by counting the basic operations of the algorithm and usually is a function of the input size of the algorithm. For sufficiently large values if input size, let us say n, time complexity can be expressed by $O(f(n))$ notation in which $f(n)$ is a function of n represents the total number of basic operations of the algorithm having n input values [53]. In this section, we provide the calculated time complexity of the SRS to ease the time estimation of solving different problems.

In one hand, there are three main factors impacting the runtime of the SRS. Number of roots being used to solve the problem, Number of Dimensions of the problem and the number of times we want the algorithm iterates to get to the expected results. On the other hand, the algorithm consists of two main parts including Initialization that initialize the subspaces roots, and the Body that performs the optimization process of the algorithm.

Knowing that, assume that n is the number of roots, m is the number of dimensions and k is the total number of iterations of the algorithm. According to the conducted analysis and calculations, the Initialization is of order $O(m \times n) + O(n \log n)$ and the Body is of order $O(m \times n \times k) + O(k \times n \log n)$.

References

1. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; University of Michigan Press: Ann Arbor, MI, USA, 1975.
2. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95—International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 1944, pp. 1942–1948.
3. Ma, Z.Y.; Yuan, X.; Han, S.; Sun, D.; Ma, Y. Improved Chaotic Particle Swarm Optimization Algorithm with More Symmetric Distribution for Numerical Function Optimization. *Symmetry* **2019**, *11*, 876. [CrossRef]
4. Dorigo, M. Optimization, Learning and Natural Algorithms. Ph.D. Thesis, Politecnico di Milano, Milan, Italy, 1992.
5. Zhao, H.G.; Gao, W.; Deng, W.; Sun, M. Study on an Adaptive Co-Evolutionary ACO Algorithm for Complex Optimization Problems. *Symmetry* **2018**, *10*, 104. [CrossRef]
6. Li, X.L.; Shao, Z.J.; Qian, J.X. An optimizing method based on autonomous animals: Fish-swarm algorithm. *Syst. Eng. Theory Pract.* **2002**, *22*, 32–38.
7. Farmer, J.D.; Packard, N.H.; Perelson, A.S. The immune system, adaptation, and machine learning. *Phys. D* **1986**, *2*, 187–204. [CrossRef]
8. Passino, K.M. Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Syst.* **2002**, *22*, 52–67.
9. Chen, Y.-P.; Li, Y.; Wang, G.; Zheng, Y.-F.; Xu, Q.; Fan, J.-H.; Cui, X.-T. A novel bacterial foraging optimization algorithm for feature selection. *Expert Syst. Appl.* **2017**, *83*, 1–17. [CrossRef]
10. Yang, X.-S. A New Metaheuristic Bat-Inspired Algorithm. In *Nature Inspired Cooperative Strategies for Optimization, Studies in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 284, pp. 65–74.
11. Meng, X.-B.; Gao, X.Z.; Liu, Y.; Zhang, H. A novel bat algorithm with habitat selection and Doppler effect in echoes for optimization. *Expert Syst. Appl.* **2015**, *42*, 6350–6364. [CrossRef]

12. Atashpaz-Gargari, E.; Lucas, C. Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. In Proceedings of the 2007 IEEE Congress on Evolutionary Computation, Singapore, 25–28 September 2007; pp. 4661–4667.
13. Jula, A.; Othman, Z.; Sundararajan, E. Imperialist competitive algorithm with PROCLUS classifier for service time optimization in cloud computing service composition. *Expert Syst. Appl.* **2015**, *42*, 135–145. [CrossRef]
14. Jula, A.; Naseri, N.K.; Rahmani, A.M. Gravitational Attraction Search with Virtual Mass (GASVM) to solve Static Grid Job scheduling Problem. *J. Math. Comput. Sci.* **2010**, *1*, 305–312. [CrossRef]
15. Webster, B.; Bernhard, P.J. A Local Search Optimization Algorithm Based on Natural Principles of Gravitation. In Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas, NV, USA, 23–26 June 2003; pp. 255–261.
16. Lee, Z.-J.; Su, S.-F.; Chuang, C.-C.; Liu, K.-H. Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. *Appl. Soft Comput.* **2008**, *8*, 55–78. [CrossRef]
17. Ying, S.; Zengqiang, C.; Zhuzhi, Y. New Chaotic PSO-Based Neural Network Predictive Control for Nonlinear Process. *Neural Netw. IEEE Trans.* **2007**, *18*, 595–601. [CrossRef]
18. Zhang, Z.; Zhang, N.; Feng, Z. Multi-satellite control resource scheduling based on ant colony optimization. *Expert Syst. Appl.* **2014**, *41*, 2816–2823. [CrossRef]
19. Moscato, P. Memetic algorithms: A short introduction. In *New Ideas in Optimization*; David, C., Marco, D., Fred, G., Dipankar, D., Pablo, M., Riccardo, P., Kenneth, V.P., Eds.; McGraw-Hill Ltd.: Maidenhead, UK, 1999; pp. 219–234.
20. Yong, W.; Lin, L. Heterogeneous Redundancy Allocation for Series-Parallel Multi-State Systems Using Hybrid Particle Swarm Optimization and Local Search. *Syst. Man Cybern. Part A Syst. Hum. Ieee Trans.* **2012**, *42*, 464–474. [CrossRef]
21. Yang, P.; Yang, H.; Qiu, W.; Wang, S.; Li, C. Optimal approach on net routing for VLSI physical design based on Tabu-ant colonies modeling. *Appl. Soft Comput.* **2014**, *21*, 376–381. [CrossRef]
22. Jula, A.; Othman, Z.; Sundararajan, E. A Hybrid Imperialist Competitive-Gravitational Attraction Search Algorithm to Optimize Cloud Service Composition. In Proceedings of the 2013 IEEE Workshop on Memetic Computing (MC), Singapore, 16–19 April 2013; pp. 37–43.
23. Jula, A.; Naseri, N.K. A Hybrid Genetic Algorithm-Gravitational Attraction Search algorithm (HYGAGA) to Solve Grid Task Scheduling Problem. In Proceedings of the International Conference on Soft Computing and its Applications(ICSCA'2012), San Francisco, CA, USA, 24–26 October 2012; pp. 158–162.
24. Amato, F.; Castiglione, A.; Moscato, V.; Picariello, A.; Sperlì, G. Multimedia summarization using social media content. *Multimed. Tools Appl.* **2018**, *77*, 17803–17827. [CrossRef]
25. Zhu, X.; Wang, N. Cuckoo search algorithm with onlooker bee search for modeling PEMFCs using T2FNN. *Eng. Appl. Artif. Intell.* **2019**, *85*, 740–753. [CrossRef]
26. Garz, P.C.; Keijzer, F. Plants: Adaptive behavior, root-brains, and minimal cognition. *Adapt. Behav. Anim. Animat. Softw. Agents Robot. Adapt. Syst.* **2011**, *19*, 155–171. [CrossRef]
27. Zhang, H.; Zhu, Y.; Chen, H. Root Growth Model for Simulation of Plant Root System and Numerical Function Optimization. In *Intelligent Computing Technology*; Huang, D.-S., Jiang, C., Bevilacqua, V., Figueroa, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7389, pp. 641–648.
28. Qi, X.; Zhu, Y.; Chen, H.; Zhang, D.; Niu, B. An Idea Based on Plant Root Growth for Numerical Optimization. In *Intelligent Computing Theories and Technology*; Huang, D.-S., Jo, K.-H., Zhou, Y.-Q., Han, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7996, pp. 571–578.
29. Ma, L.; Hu, K.; Zhu, Y.; Chen, H.; He, M. A Novel Plant Root Foraging Algorithm for Image Segmentation Problems. *Math. Probl. Eng.* **2014**, *2014*, 16. [CrossRef]
30. Ma, L.; Zhu, Y.; Liu, Y.; Tian, L.; Chen, H. A novel bionic algorithm inspired by plant root foraging behaviors. *Appl. Soft. Comput.* **2015**, *37*, 95–113. [CrossRef]
31. Qi, X.; Zhu, Y.; Zhang, H.; Zhang, D.; Wu, J. A novel bio-inspired algorithm based on plant root growth model for data clustering. In Proceedings of the 35th Control Conference (CCC), Chengdu, China, 27–29 July 2016; pp. 9183–9188.
32. Naseri, N.K.; Sundararajan, E.; Ayob, M.; Jula, A. Smart Root Search (SRS): A New Search Algorithm to Investigate Combinatorial Problems. In Proceedings of the 2015 Seventh International Conference on Computational Intelligence, Modelling and Simulation (CIMSim), Kuantan, Malaysia, 27–29 July 2015; pp. 1–16.

33. Hill, P.S. *Vibrational Communication in Animals*; Harvard University Press: Cambridge, MA, USA, 2008.
34. Buhner, S.H. *Plant Intelligence and the Imaginal Realm beyond the Doors of Perception into the Dreaming of Earth*; Inner Traditions Bear and Company: Rochester, VT, USA, 2014; p. 576.
35. Jung, J.K.H.; McCouch, S.R. Getting to the roots of it: Genetic and hormonal control of root architecture. *Front. Plant Sci.* **2013**, 4. [CrossRef]
36. Takahashi, N.; Yamazaki, Y.; Kobayashi, A.; Higashitani, A.; Takahashi, H. Hydrotropism Interacts with Gravitropism by Degrading Amyloplasts in Seedling Roots of Arabidopsis and Radish. *Plant Physiol.* **2003**, *132*, 805–810. [CrossRef]
37. Moriwaki, T.; Miyazawa, Y.; Kobayashi, A.; Takahashi, H. Molecular mechanisms of hydrotropism in seedling roots of Arabidopsis thaliana (Brassicaceae). *Am. J. Bot.* **2013**, *100*, 25–34. [CrossRef]
38. Prieto, I.; Armas, C.; Pugnaire, F.I. Hydraulic lift promotes selective root foraging in nutrient-rich soil patches. *Funct. Plant Biol.* **2012**, *39*, 804–812. [CrossRef]
39. Hultine, K.R.; Cable, W.L.; Burgess, S.S.; Williams, D.G. Hydraulic redistribution by deep roots of a Chihuahuan Desert phreatophyte. *Tree Physiol.* **2003**, *23*, 353–360. [CrossRef]
40. Blum, A. Plant Water Relations, Plant Stress and Plant Production. In *Plant Breeding for Water-Limited Environments*; Springer: New York, NY, USA, 2011; pp. 11–52. [CrossRef]
41. López-Bucio, J.; Cruz-Ramírez, A.; Herrera-Estrella, L. The role of nutrient availability in regulating root architecture. *Curr. Opin. Plant Biol.* **2003**, *6*, 280–287. [CrossRef]
42. Signora, L.; De Smet, I.; Foyer, C.H.; Zhang, H. ABA plays a central role in mediating the regulatory effects of nitrate on root branching in Arabidopsis. *Plant J.* **2001**, *28*, 655–662. [CrossRef]
43. Tian, Q.; Chen, F.; Zhang, F.; Mi, G. Possible Involvement of Cytokinin in Nitrate-mediated Root Growth in Maize. *Plant Soil* **2005**, *277*, 185–196. [CrossRef]
44. Linkohr, B.I.; Williamson, L.C.; Fitter, A.H.; Leyser, H.M.O. Nitrate and phosphate availability and distribution have different effects on root system architecture of Arabidopsis. *Plant J.* **2002**, *29*, 751–760. [CrossRef]
45. Jiang, C.; Gao, X.; Liao, L.; Harberd, N.P.; Fu, X. Phosphate Starvation Root Architecture and Anthocyanin Accumulation Responses Are Modulated by the Gibberellin-DELLA Signaling Pathway in Arabidopsis. *Plant Physiol.* **2007**, *145*, 1460–1470. [CrossRef]
46. Bates, T.R.; Lynch, J.P. Plant growth and phosphorus accumulation of wild type and two root hair mutants of Arabidopsis thaliana (Brassicaceae). *Am. J. Bot.* **2000**, *87*, 958–963. [CrossRef]
47. Bruce, T.J.A.; Matthes, M.C.; Napier, J.A.; Pickett, J.A. Stressful "memories" of plants: Evidence and possible mechanisms. *Plant Sci.* **2007**, *173*, 603–608. [CrossRef]
48. Walter, J.; Nagy, L.; Hein, R.; Rascher, U.; Beierkuhnlein, C.; Willner, E.; Jentsch, A. Do plants remember drought? Hints towards a drought-memory in grasses. *Environ. Exp. Bot.* **2011**, *71*, 34–40. [CrossRef]
49. Fromm, J.; Fei, H. Electrical signaling and gas exchange in maize plants of drying soil. *Plant Sci.* **1998**, *132*, 203–213. [CrossRef]
50. Mishra, N.S.; Mallick, B.N.; Sopory, S.K. Electrical signal from root to shoot in Sorghum bicolor: Induction of leaf opening and evidence for fast extracellular propagation. *Plant Sci.* **2001**, *160*, 237–245. [CrossRef]
51. Zhang, H.; Zhu, Y.; Chen, H. Root growth model: A novel approach to numerical function optimization and simulation of plant root system. *Soft Comput.* **2014**, *18*, 521–537. [CrossRef]
52. Geem, Z.W.; Kim, J.H.; Loganathan, G. A new heuristic optimization algorithm: Harmony search. *Simulation* **2001**, *76*, 60–68. [CrossRef]
53. Neapolitan, R.; Naimipour, K. *Foundations of Algorithms*; Jones & Bartlett Learning, LLC: Sudbury, MA, USA, 2010.
54. Rubio, V.; Bustos, R.; Irigoyen, M.L.; Cardona-López, X.; Rojas-Triana, M.; Paz-Ares, J. Plant hormones and nutrient signaling. *Plant Mol. Biol.* **2009**, *69*, 361–373. [CrossRef]
55. López-Bucio, J.; Hernández-Abreu, E.; Sánchez-Calderón, L.; Nieto-Jacobo, M.F.; Simpson, J.; Herrera-Estrella, L. Phosphate Availability Alters Architecture and Causes Changes in Hormone Sensitivity in the Arabidopsis Root System. *Plant Physiol.* **2002**, *129*, 244–256. [CrossRef]
56. Nacry, P.; Canivenc, G.; Muller, B.; Azmi, A.; Van Onckelen, H.; Rossignol, M.; Doumas, P. A Role for Auxin Redistribution in the Responses of the Root System Architecture to Phosphate Starvation in Arabidopsis. *Plant Physiol.* **2005**, *138*, 2061–2074. [CrossRef]
57. Beckett, M.; Garnett, S.; Hay, M.; Makings, E.; Marriott, H.; Nieuwenhuizen, N.; Sabela, G.; Scheffler, C.; Steenkamp, L.; Viljoen, K.; et al. *Siyavula: Life Sciences Grade 10*; Siyavula, Ed.; Connexions Rice University: Houston, TX, USA, 2011.

58. Vafaee, F.; Turan, G.; Nelson, P.C.; Berger-Wolf, T.Y. Balancing the Exploration and Exploitation in an Adaptive Diversity Guided Genetic Algorithm. In Proceedings of the 2014 IEEE Congress on Evolutionary Computation (CEC), Beijing, China, 6–11 July 2014; pp. 2570–2577.
59. Chen, J.; Xin, B.; Peng, Z.H.; Dou, L.H.; Zhang, J.A. Optimal Contraction Theorem for Exploration-Exploitation Tradeoff in Search and Optimization. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2009**, *39*, 680–691. [CrossRef]
60. Lipowski, A.; Lipowska, D. Roulette-wheel selection via stochastic acceptance. *Phys. A Stat. Mech. Its Appl.* **2012**, *391*, 2193–2196. [CrossRef]
61. Rudolph, G. *Convergence Properties of Evolutionary Algorithms*; Verlag Dr. Kovac: Hamburg, Germany, 1997.
62. He, J.; Lin, G.M. Average Convergence Rate of Evolutionary Algorithms. *IEEE Trans. Evol. Comput.* **2016**, *20*, 316–321. [CrossRef]
63. Karaboga, D.; Basturk, B. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. *J. Glob. Optim.* **2007**, *39*, 459–471. [CrossRef]
64. Homaifar, A.; Qi, C.X.; Lai, S.H. Constrained optimization via genetic algorithms. *Simulation* **1994**, *62*, 242–253. [CrossRef]
65. Kennedy, J.; Kennedy, J.F.; Eberhart, R.C.; Shi, Y. *Swarm Intelligence*; Morgan Kaufmann: San Francisco, CA, USA, 2001.
66. Karaboga, D.; Akay, B. A comparative study of Artificial Bee Colony algorithm. *Appl. Math. Comput.* **2009**, *214*, 108–132. [CrossRef]
67. Adorio, E.P.; Diliman, U. *Mvf-Multivariate Test Functions Library in C for Unconstrained Global Optimization*; University of the Philippines Diliman: Quezon City, Philippines, 2005.
68. Gavana, A. Test Functions Index. Available online: http://infinity77.net/global_optimization/test_functions.html (accessed on 27 April 2016).
69. Jamil, M.; Yang, X.-S. A literature survey of benchmark functions for global optimisation problems. *Int. J. Math. Model. Numer. Optim.* **2013**, *4*, 150–194. [CrossRef]
70. Yang, X.S. Appendix A: Test problems in optimization. *Eng. Optim.* **2010**, *1*, 261–266.
71. Jula, A.; Nilsaz, H.; Sundararajan, E.; Othman, Z. A new dataset and benchmark for cloud computing service composition. In Proceedings of the 2014 5th International Conference on Intelligent Systems, Modelling and Simulation, Langkawi, Malaysia, 27–29 January 2014; pp. 83–86.
72. Everitt, B.S. Cluster analysis of subjects, hierarchical methods. *Encycl. Biostat.* **2005**, 2. [CrossRef]
73. Day, W.H.E.; Edelsbrunner, H. Efficient algorithms for agglomerative hierarchical clustering methods. *J. Classif.* **1984**, *1*, 7–24. [CrossRef]
74. CaliŃski, T. Dendrogram. *Wiley Statsref Stat. Ref. Online* **2014**. [CrossRef]
75. Wells, C.S. Encyclopedia of Research Design. In *Encyclopedia of Research Design*; Salkind, N.J., Ed.; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2010. [CrossRef]
76. Kim, H.-Y. Analysis of variance (ANOVA) comparing means of more than two groups. *Restor. Dent. Endod.* **2014**, *39*, 74–77. [CrossRef] [PubMed]
77. McHugh, M.L. The Chi-square test of independence. *Biochem. Med.* **2013**, *23*, 143–149. [CrossRef] [PubMed]

symmetry

MDPI

Article

Evolving Hierarchical and Tag Information via the Deeply Enhanced Weighted Non-Negative Matrix Factorization of Rating Predictions

Alpamis Kutlimuratov [1], Akmalbek Abdusalomov [1] and Taeg Keun Whangbo [2,*]

[1] Department of IT Convergence Engineering, Gachon University, Sujeong-Gu, Seongnam-Si, Gyeonggi-Do 461-701, Korea; alpamis92@gc.gachon.ac.kr (A.K.); akmaljon@gachon.ac.kr (A.A.)
[2] Department of Computer Science, Gachon University, Sujeong-Gu, Seongnam-Si, Gyeonggi-Do 461-701, Korea
* Correspondence: tkwhangbo@gachon.ac.kr

Received: 5 October 2020; Accepted: 18 November 2020; Published: 23 November 2020

Abstract: Identifying the hidden features of items and users of a modern recommendation system, wherein features are represented as hierarchical structures, allows us to understand the association between the two entities. Moreover, when tag information that is added to items by users themselves is coupled with hierarchically structured features, the rating prediction efficiency and system personalization are improved. To this effect, we developed a novel model that acquires hidden-level hierarchical features of users and items and combines them with the tag information of items that regularizes the matrix factorization process of a basic weighted non-negative matrix factorization (WNMF) model to complete our prediction model. The idea behind the proposed approach was to deeply factorize a basic WNMF model to obtain hidden hierarchical features of user's preferences and item characteristics that reveal a deep relationship between them by regularizing the process with tag information as an auxiliary parameter. Experiments were conducted on the MovieLens 100K dataset, and the empirical results confirmed the potential of the proposed approach and its superiority over models that use the primary features of users and items or tag information separately in the prediction process.

Keywords: recommendation system; weighted non-negative matrix factorization; hierarchical information; tag information; deep factorization

1. Introduction

Recently, with the increase in the availability of data from online content providers, delivering valuable information that gratifies and holds a consumer's interest has attracted significant attention; thus, modeling an effective recommendation system is essential. The primary objective of a recommendation system is to offer suggestions based on user preferences, which are solicited from historical data, such as ratings, reviews, and tags. Recommendations help in accelerating searches and enable users to access more pertinent content. Therefore, web service providers have extensively cogitated about developing recommender systems that analyze and harness user–item interactions to increase customer satisfaction, profits, and personalized suggestions for their services. Several modern-day internet applications have integrated recommendation systems, including Google, Netflix, eBay, and Amazon.

Recommendation systems are designed based on the type of information obtained such that the diversity of information influences their implementation and structure. To this effect, two traditional approaches exist for building recommendation systems: content-based filtering (CBF) and collaborative filtering (CF) [1,2]. The former approach generates recommendations by analyzing the availability of

the user–item interaction data, which largely requires collecting explicit information [3–5]. For instance, content-based movie recommendations accommodate the features of a movie that match those of a user's past preferences. Thus, identifying a connection between the items and users is highly important. However, in recent years, owing to the limitations of this approach, such as privacy concerns and the dearth of supplementary information for items, web services have adopted the CF architecture in recommendation systems. The algorithms of this method utilize the items rated by a user to predict unrated items when offering recommendations and subsequently automate these predictions by acquiring user perceptions among a niche audience [2,3,6–10].

Memory- and model-based techniques are commonly used to elucidate CF recommendations [3,11–15]. Past studies have demonstrated the benefits of memory-based CF, wherein rating predictions are computed from the preferences of similar users via a rating matrix [12,16–19]. Conversely, the model-based CF technique leverages a user–item rating matrix to initially build a predictive model using deep learning methods and then source the rating predictions from it [3,20]. CF-based recommendation systems are susceptible to data sparsity and the cold-start problem, which are open issues in the recommendation system research area and put the responsibility on any kind of recommendation system algorithms and methods to avoid and solve them [21]. First, fewer user interactions with items in a user–item rating matrix invokes data sparsity; specifically, the input rating matrix is not sufficient to train a model to make predictions. Thus, only 10–25% of the matrix is populated with ratings. Second, the cold-start problem arises when information about new users or items and their interactions is insufficient to garner suitable recommendations.

One of the most effective implementations of model-based CF is the matrix factorization (MF) method. This method deconstructs the user–item rating matrix into two less latent factor sub-matrices of user preferences and item characteristics, respectively, and then a vector constituting an item and a user feature is generated to predict the user's rating for an item [3,14,22,23]. Moreover, MF spontaneously integrates a mix of implicit and explicit information related to users or items. Factorization methods have since demonstrated substantial efficiency when resolving the issues of data sparsity and the cold start in recommendation systems.

In this study, we aimed to address the two aforementioned issues using hierarchical and tag information through enhanced matrix factorization to eventually improving the performance of recommender systems. Hierarchical information helps with meaningfully concealing information regarding items, such as categories of movie genres on streaming websites (e.g., Netflix and Disney+) or product catalogs on shopping websites (e.g., Amazon, Alibaba, and eBay).

Users and items of the real practical recommendation systems could exhibit certain hierarchical structures. For example, a user (girl) may usually select movies from the main category "romance," or more exactly, the user watches movies under the sub-category of romance drama. Similarly, the item (the Apple Watch Series 5) can be placed in the main category "electronics," or more specifically, the item is tantamount to the sub-category "smart watches." The classification of an item into appropriate lower-level categories or nodes is conducted sequentially. Items in the same hierarchical level are likely to share similar attributes, thus they are likely to get similar rating scores. Equivalently, users in the same hierarchical level are likely to share similar preferences, thus they are likely to rate certain items similarly [24]. For this reason, recently, evolving hierarchical structures of items or users have been developing to improve recommendation system performances. The priority of hierarchical structures and their unavailability also motivated us to research hierarchical structures of users and items for recommendation systems. During the research, evolving the hierarchical structures of items and users simultaneously and mathematically modeling them for recommendation systems were studied. Along with the above, integrating tag information with mathematically modeled hierarchical structures of items and users into a systematic model that puts a basis for a recommendation system was also investigated.

In contrast, tag information comprises words or short phrases assigned to items by a user that reflects their associations or behavior, and in turn, facilitates predictions by passing it as a value to the

prediction algorithm. Researchers have previously reported on the benefits of making recommendations using tags and generating hierarchical information to not only improve results but also tackle issues of data sparsity and cold starts [8,13,24–28]. Furthermore, to the best of our knowledge, despite the significant amount of research that has been conducted to explicate the use of matrix factorization via hierarchical and tag information individually in recommendation systems, the two have rarely been applied in a combination.

In this study, we developed a novel MF-based methodology to predict ratings by incorporating both hierarchical and tag information simultaneously. The rationale behind the proposed approach was to deeply enrich a basic MF model to obtain hierarchical relationships for predicting the ratings and then regularize it using tag information. Our main contributions using this approach included the following:

- Deeply extending the basic MF model to identify hierarchical relationships that facilitate the rating predictions.
- Regularizing the resultant model with tag information, as well as hierarchical data.
- Conducting experiments on the MovieLens 100K dataset (https://grouplens.org/datasets/movielens/) to evaluate the proposed methodology.
- Reducing data sparsity and cold-start issues encountered by other CF methods.

The remainder of this paper is structured as follows. In Section 2, works pertaining to tag-based recommendation systems, generating hierarchical features, and existing MF methods are reviewed. In Sections 3 and 4, we discuss the proposed methodology in detail and validate its accuracy via experiments and comparisons with other MF methods. Section 5 presents the conclusions and scope of future work; finally, the reviewed materials are referenced, where many of which are recent publications.

2. Related Work

Several studies in the recent past have harnessed hierarchical and tag information as auxiliary features to address issues related to data sparsity and cold starts in recommendation systems [13,25,28,29]. CF-based recommender systems are commonly employed to predict ratings based on user histories; however, they ignore costly features, which introduce data sparsity and cold starts, which in turn hampers performance. Therefore, various studies have integrated auxiliary information in the recommendation process [30,31].

Auxiliary features often maintain a rich knowledge structure i.e., a hierarchy with dependencies. Yang et al. [13] proposed an MF-based framework with recursive regularization that analyzes the impacts of hierarchically organized features in user–item interactions to improve the recommendation accuracy and eliminate the cold-start problem. Lu et al. [32] developed a framework that exploited these hierarchical relationships to identify more reliable neighbors; moreover, the framework modeled the hierarchical structure based on potential users' preferences. The hierarchical itemspace rank (HIR) algorithm utilizes the intrinsic hierarchical structure of an itemspace to mitigate data sparsity that may affect the quality of recommendations [33].

Most modern recommender systems trawl both explicit and implicit data for useful information, including ratings, images, text (tags), social information, items, and user characteristics, to offer recommendations. We can thus infer that analyzing tag information is important in recommender systems, as they not only recap the characteristics of items but also help in identifying user preferences. For example, food recommendations are made by a model trained on a dataset comprising user preferences that are collated from ratings and tags specified in product forms to indicate their preferred food components and features [25]. Karen et al. [27] proposed a generic method that modifies CF algorithms to accommodate tags and deconstructs 3D correlations into three 2D correlations. Moreover, Wang et al. [34] formulated a novel approach that combined tags and ratings-based CF to discern similar users and items.

Our proposed methodology deviates from these methods in that the tags obtained from user–item interactions are used to regularize the MF process, whereas hierarchical information delivers the rating predictions. In summary, existing MF models that use hierarchical and tag information individually have delivered satisfactory results despite the complexity. However, to the best of our knowledge, there is no available advantageous work that seamlessly incorporates the hierarchical and tag information.

3. Methodology

This section is devoted to illustrating our proposed methodology that predicts rating scores by evolving hierarchical structures of items and users simultaneously with a mathematically modeled combination of tag information. Specifically, the notations that are used in this paper are first introduced, and then, a basic model that builds the basis of the proposed model is described. After that, we go into the details of the model components that mathematically model the hierarchical structures of items and users simultaneously and the integration of tag information, respectively, the combination of which leads to an optimization problem. Lastly, we come up with an efficient algorithm to solve it.

3.1. Notations

Table 1 enumerates the notations used in this paper.

Table 1. Notation definitions.

Notation	Description
$\mathbf{H}$	Matrices are denoted by boldface capital letters
$\mathbf{h}$	Vectors are denoted by boldface lowercase letters
$\|H\|_F$	Frobenius norm of matrix
$\odot$	Hadamard product
λ	Regularization parameter
$tr(\cdot)$	Trace of a matrix
β	Extra regularization parameter

3.2. Basic Matrix Factorization

We modeled our approach on a basic weighted non-negative matrix factorization (WNMF) method owing to its feasible and easy implementation in recommendation systems with large inputs and sparse data. This method factorizes an input rating matrix into two non-negative sub-matrices $\mathbf{P}$ and $\mathbf{Q}$ of sizes $n \times r$ and $r \times m$, respectively.

$$\mathbf{R'} \approx \mathbf{PQ} = \begin{bmatrix} p_1 \\ p_2 \\ \cdots \\ \cdots \\ p_n \end{bmatrix} \begin{bmatrix} q_1 q_2 \cdots, q_m \end{bmatrix} \tag{1}$$

The rating score given by p_i to q_j is then computed as $\mathbf{R}^{'}(i,j) = \mathbf{P}(i,:)\mathbf{Q}(:,j)$. $\mathbf{P}$ and $\mathbf{Q}$ are evaluated by solving the following optimization problem:

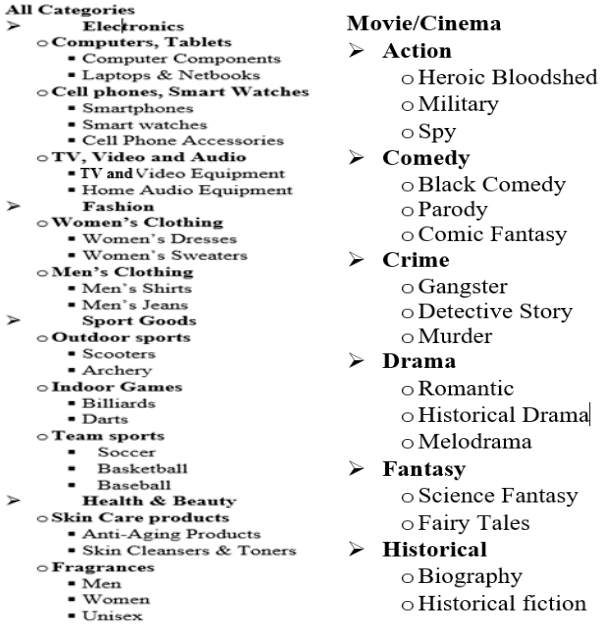
$$\underbrace{\min}_{P,Q} \|W \odot (\mathbf{R} - \mathbf{PQ})\|_F^2 + \lambda(\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2) \tag{2}$$

where W is the hyperparameter that regulates the contribution of $\mathbf{R}^{'}(i,j)$ in the learning process such that $W(i,j) = 1$ for $\mathbf{R}^{'}(i,j) > 0$; else, $W(i,j) = 0$. $\odot$ is the Hadamard element-wise multiplication

operator, λ is the regularization parameter used to moderate the complexity and overfitting during learning, and $\|\mathbf{P}\|_F^2$ and $\|\mathbf{Q}\|_F^2$ are the Frobenius norms of the corresponding matrices [27].

3.3. Acquiring the Hierarchical Structured Information

Some features of users and items are hierarchically structured. For instance, as shown in Figure 1b, the genres of movies can be organized into a hierarchical structure. It is very likely that movies that are associated with the detailed genres are more similar than those in subgenres. For this reason, it should be suitable to recommend a movie that is in the same detailed genre as one that has got a high rating score from the user. Hierarchical structures of users and items involve complementary information and capturing them simultaneously can further improve the recommendation performance. Therefore, in this subsection, acquiring the hierarchically structured information of users and items is introduced by enhancing the basic weighted non-negative matrix factorization model.



(**a**) eBay products categories (**b**) Movie genre categories

Figure 1. (**a**) Hierarchical structure of eBay products and (**b**) an illustration of movie genre categories.

One of the most significant challenges for recommendation systems is to elicit valuable information from the features of highly correlated users and items in a user–item interaction that forms the basis of the prediction process. Typically, this is modeled using the flat attributes of users (for example, gender and age) or items (in the case of a movie, this can include an actor, a producer, release date, language, and country). However, these features may often be represented in a multilevel structure, i.e., a hierarchy, in the form of a tree with nested nodes (for example, movie genres and user occupations). Simple representations of a hierarchical structure include movie genres and product categories on e-commerce websites, as shown in Figure 1.

For example, the movie Godfather (an item) can be classified by traversing the hierarchical tree nodes as follows: main genre→subgenre, per Figure 1b, which then resembles crime→gangster. Similarly, the Apple Watch Series 5 (an item) can be placed in a hierarchical structure, per Figure 1a, as main category→subcategory→explicit subcategory, which is tantamount to electronics→cell phones,

smart watches→smart watches. The classification of an item into appropriate lower-level categories or nodes is conducted sequentially.

User preferences are similarly structured. For instance, a user who chooses to rate movies in the crime genre may prefer the gangster subgenre over others, and those who shop for items belonging to a particular hierarchical level of the product catalog may express coincidental preferences by consistently rating items that exhibit similar characteristics.

From Section 3.2, WNMF was adopted as the core model to acquire implicit hierarchical information and thereby predict rating scores. The user–item rating matrix, $\mathbf{R}$, was deconstructed into two lower-dimensional non-negative submatrices, $\mathbf{P}$ and $\mathbf{Q}$, constituting user preferences and item characteristics, respectively, and expressed as the flat structures of features. Because $\mathbf{P}$ and $\mathbf{Q}$ are non-negative, we applied the non-negative matrix factorization to them to interpret the corresponding hierarchically structured information, which then served to predict the rating scores given by Equation (1).

$\mathbf{P}$ and $\mathbf{Q}$ were extracted such that $\mathbf{P} \in \mathbb{R}^{n\times r}$ and $\mathbf{Q} \in \mathbb{R}^{r\times m}$ to indicate the latent representations of n users and m items in an r-dimensional latent category (space). $\mathbf{P}$ and $\mathbf{Q}$ were further factorized to model the hierarchical structure owing to their non-negativity.

Therefore, in a particular embodiment, $\mathbf{P}$ was factorized into two matrices, $\mathbf{P}_1 \in \mathbb{R}^{n\times n_1}$ and $\tilde{\mathbf{P}}_2 \in \mathbb{R}^{n_1\times r}$, as follows:

$$\mathbf{P} \approx \mathbf{P}_1\tilde{\mathbf{P}}_2 \tag{3}$$

where n is the number of users, r is the number of latent categories (space) in the first hierarchical level, and n_1 and is the number of subcategories in the second hierarchical level. Thus, $\mathbf{P}_1 \in \mathbb{R}^{n\times n_1}$ is the relationship of n users to n_1 subcategories. $\tilde{\mathbf{P}}_2$ denotes the second level of the hierarchical structure of users obtained from the relationship between the number of latent categories (space) in the first hierarchical level and n_1, i.e., the number of latent subcategories in the second hierarchical level. To compute the third level of a hierarchical structure of users, as given in Equation (4), $\tilde{\mathbf{P}}_2$ is further factorized as $\mathbf{P}_2 \in \mathbb{R}^{n_1\times n_2}$ and $\tilde{\mathbf{P}}_3 \in \mathbb{R}^{n_2\times r}$:

$$\mathbf{P} \approx \mathbf{P}_1\mathbf{P}_2\tilde{\mathbf{P}}_3 \tag{4}$$

where n_2 is the number of subcategories in the third hierarchical level. Therefore, deep factorization on $\mathbf{P}$ serves to obtain the xth level of the hierarchical structure of users, $\mathbf{P}_x$, which is accomplished by factorizing $\tilde{\mathbf{P}}_{x-1}$, the latent category relationship matrix of the $(x-1)$th level of the hierarchical structure, into non-negative matrices, as follows:

$$\mathbf{P} \approx \mathbf{P}_1\mathbf{P}_2 \ldots\ldots\ldots\ldots \mathbf{P}_{x-1}\mathbf{P}_x \tag{5}$$

where $\mathbf{P}_i \geq 0$ for $i \in \{1, 2, .., x\}$, $\mathbf{P}_1$ is an $n \times n_1$ matrix such that $\mathbf{P}_i$ is an $n_{i-1} \times n_i$ matrix, and $\mathbf{P}_x$ is $n_{x-1} \times r$ matrix.

The above factorization process as illustrated in Figure 2 is repeated for $\mathbf{Q}$ to obtain the level of the hierarchical structure of items. For this, the relationship of m items with r-dimensional latent categories (space) is represented as $\mathbf{Q} \in \mathbb{R}^{r\times m}$, which is further factorized into $\mathbf{Q}_1 \in \mathbb{R}^{m_1\times m}$ and $\tilde{\mathbf{Q}}_2 \in \mathbb{R}^{r\times m_1}$ to describe the second level of items in the hierarchy given by:

$$\mathbf{Q} \approx \tilde{\mathbf{Q}}_2\mathbf{Q}_1 \tag{6}$$

where m_1 is the number of sub-categories in the second hierarchical level and $\mathbf{Q}_1 \in \mathbb{R}^{m_1\times m}$ is the relationship of m items to the m_1 latent subcategories. The latent category relationship of the non-negative matrix $\tilde{\mathbf{Q}}_2 \in \mathbb{R}^{r\times m_1}$ of the second hierarchical level is defined as the affiliation between r-dimensional latent categories (space) in the first hierarchical level and m_1 latent subcategories in the

second hierarchical level. Equation (7) gives the third level of the hierarchical structure of items, where $\tilde{\mathbf{Q}}_2$ is also factorized as $\mathbf{Q}_2 \in \mathbb{R}^{m_2 \times m_1}$ and $\tilde{\mathbf{Q}}_3 \in \mathbb{R}^{r \times m_2}$, where m_2 is the number of subcategories in the third hierarchical level:

$$\mathbf{Q} \approx \tilde{\mathbf{Q}}_3 \mathbf{Q}_2 \mathbf{Q}_1 \tag{7}$$

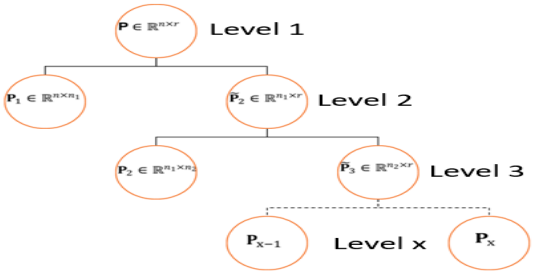


Figure 2. Obtaining the hierarchical structure of users.

Deep factorization on $\mathbf{Q}$, as illustrated in Figure 3, secures the yth level of a hierarchical structure of items, $\mathbf{Q}_y$, which is accomplished by factorizing $\tilde{\mathbf{Q}}_{y-1}$, in the $(y-1)$th level of the hierarchy, as follows:

$$\mathbf{Q} \approx \mathbf{Q}_y \mathbf{Q}_{y-1} \dots \mathbf{Q}_2 \mathbf{Q}_1 \tag{8}$$

where $\mathbf{Q}_j \geq 0$ for $j \in \{1, 2, .., y\}$, $\mathbf{Q}_1$ is an $m_1 \times m$ matrix such that $\mathbf{Q}_j$ is an $m_j \times m_{j-1}$ matrix, and $\mathbf{Q}_y$ is an $r \times m_{y-1}$ matrix.

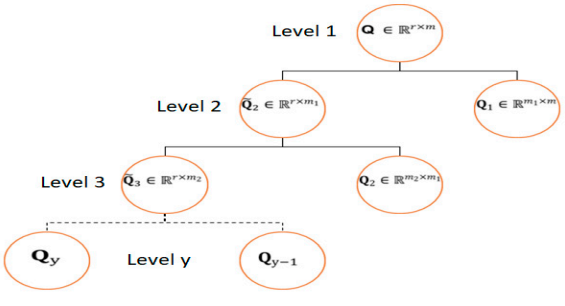


Figure 3. Obtaining a hierarchical structure of items.

Finally, the below optimization problem needs to be effectively solved for building a model that outlines the hierarchical structures of users and items:

$$\underbrace{\min}_{\mathbf{P}_1, \dots \mathbf{P}_x, \mathbf{Q}_1 \dots \mathbf{Q}_y} \|W \odot \left(\mathbf{R} - \mathbf{P}_1 \dots \mathbf{P}_x \mathbf{Q}_y \dots \mathbf{Q}_1\right)\|_F^2 + \lambda \left(\sum_{i=1}^{x} \|\mathbf{P}_i\|_F^2 + \sum_{j=1}^{y} \|\mathbf{Q}_j\|_F^2\right) \tag{9}$$

where $\mathbf{P}_i \geq 0$ for $i \in \{1, 2, .., x\}$ and $\mathbf{Q}_j \geq 0$ for $j \in \{1, 2, .., y\}$.

The rating prediction process that involves acquired user's and item's hierarchically structured information is represented in Figure 4.

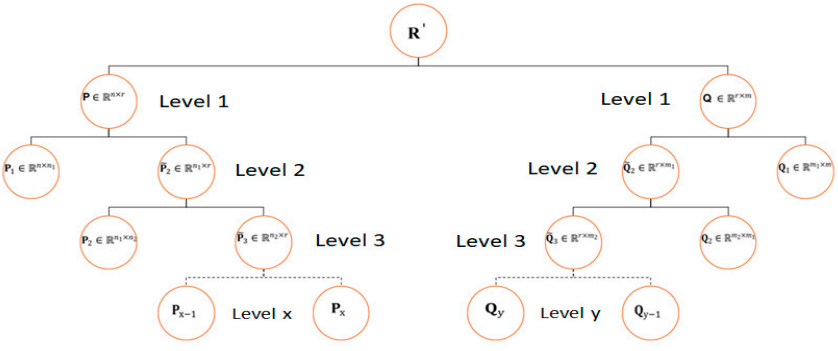


Figure 4. An illustration of predicting a rating score based on hierarchical structures of users and items.

3.4. Incorporating Tag Information

Tag information was incorporated uniquely into our proposed methodology for deriving an association between the supplementary information solicited from WNMF and tag repetitiveness in items [3]. For example, an "organized crime" tag assigned to the movie "The Godfather" (item) by a user may also apply to other items with similar characteristics, which is reflected in the degree of repetitiveness. Therefore, the matrix factorization process of a basic WNMF model is regularized using the tag information to complete our prediction model. In short, we aimed to form two item-specific latent feature vectors from the MF process of our WMNF model that are similar in nature and contain items with common tag information. For a tag information matrix $\mathbf{T}$, each of its components $\mathbf{T}_{it}$ for item i and tag t is a tf $*$ idf value [35]:

$$\mathbf{T}_{it} = \mathrm{tf}(i,t) * \log_2\left(\frac{m}{\mathrm{df}(t)}\right) \tag{10}$$

where $\mathrm{tf}(i,t)$ is the normalized frequency of t occurring in i, $\mathrm{df}(t)$ is the number of items that contain t, and m is the total number of items. Thus, the similarity between items i and j is computed using the cosine similarity metric given, as follows:

$$S_{i,j} = \frac{\sum_{t \in T^{ij}} T_{it} T_{jt}}{\sqrt{\sum_{t \in T^{ij}} T_{it}^2} \sqrt{\sum_{t \in T^{ij}} T_{jt}^2}} \tag{11}$$

where T^{ij} is the index of tags occurring in both i and j. The two item-specific latent feature vectors that are most similar are then obtained by affixing an item similarity regularization criterion function to the WNMF model, as follows:

$$\begin{aligned} \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{i,j} \|q_i - q_j\|_F^2 &= \frac{\beta}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[S_{i,j} \sum_{r'=1}^{r} \left(q_{r'i} - q_{r'j} \right)^2 \right] \\ &= \frac{\beta}{2} \sum_{r'=1}^{r} \mathbf{Q}_{r'*} \mathbf{L} \mathbf{Q}_{r'*}^T = \frac{\beta}{2} \mathrm{tr}\left(\mathbf{Q} \mathbf{L} \mathbf{Q}^T \right) \end{aligned} \tag{12}$$

where $S_{i,j}$ defines the similarity between i and j; $q_1 q_2 \ldots, q_m$ are latent characteristic vectors that populate $\mathbf{Q}$; r is the dimension of each item in the vector, i.e., $q_{r'i}$ and $q_{r'j}$ are the values of vector items i and j of the r'th dimension; $\mathbf{L}$ denotes the Laplacian matrix given by $\mathbf{L} = \mathbf{D} - \mathbf{S}$ for a diagonal matrix $\mathbf{D}$ such that $\mathbf{D}_{ij} = \sum_j S_{ij}$. $\mathrm{tr}(\cdot)$ is a trace of the matrix; β is an extra regularization parameter that controls the contribution of the tag information [36].

The rating predictions were made by combining Equations (9) and (12) and utilizing the following objective function for the minimization task:

$$\underbrace{\min}_{\mathbf{P}_1,\ldots\mathbf{P}_x,\mathbf{Q}_1\ldots\mathbf{Q}_y} \|W \odot \left(\mathbf{R} - \mathbf{P}_1 \ldots \mathbf{P}_x\mathbf{Q}_y \ldots \mathbf{Q}_1\right)\|_F^2 + \lambda\left(\sum_{i=1}^{x}\|\mathbf{P}_i\|_F^2 + \sum_{j=1}^{y}\|\mathbf{Q}_j\|_F^2\right) + \frac{\beta}{2}tr\left(\mathbf{Q}\mathbf{L}\mathbf{Q}^T\right) \quad (13)$$

where $\mathbf{P}_i \geq 0$ for $i \in \{1, 2, .., x\}$ and $\mathbf{Q}_j \geq 0$ for $j \in \{1, 2, .., y\}$.

3.5. Optimization Problem

The optimization problem is complicated owing to the non-convexity of the objective function, but solving for it also helps in validating the method that is administered in a recommendation system. Our optimization method modified the approach in [37] in that all variables of the objective function given in Equation (13) were updated interchangeably such that the function becomes convex, which does not occur otherwise.

3.5.1. The Basis of Updating P_i

When $\mathbf{P}_i$ is updated, terms unrelated to $\mathbf{P}_i$ are discarded by fixing the other variables, and the resulting objective function is expressed as:

$$\underbrace{\min}_{P_i \geq 0} \|W \odot (\mathbf{R} - \mathbf{A}_i\mathbf{P}_i\mathbf{H}_i)\|_F^2 + \lambda \|\mathbf{P}_i\|_F^2 \quad (14)$$

where $\mathbf{A}_i$ and $\mathbf{H}_i$ for $1 \leq i \leq x$, are defined as:

$$\mathbf{A}_i = \begin{cases} \mathbf{P}_1\mathbf{P}_2 \ldots\ldots\ldots\ldots\mathbf{P}_{x-1} & \text{if } i \neq 1 \\ \mathbf{I} & \text{if } i = 1 \end{cases} \quad (15)$$

$$\mathbf{H}_i = \begin{cases} \mathbf{P}_{i+1} \ldots \mathbf{P}_x\mathbf{Q}_y \ldots \mathbf{Q}_1 & \text{if } i \neq x \\ \mathbf{Q}_y \ldots \mathbf{Q}_1 & \text{if } i = x \end{cases} \quad (16)$$

The Lagrangian function in Equation (14) is:

$$L(\mathbf{P}_i) = \|W \odot (\mathbf{R} - \mathbf{A}_i\mathbf{P}_i\mathbf{H}_i)\|_F^2 + \lambda\|\mathbf{P}_i\|_F^2 - Tr\left(M^T\mathbf{P}_i\right) \quad (17)$$

where M is the Lagrangian multiplier. The derivative of $L(\mathbf{P}_i)$ with respect to $\mathbf{P}_i$ is then given by:

$$\frac{\partial L(\mathbf{P}_i)}{\partial \mathbf{P}_i} = 2\mathbf{A}_i^T\|W \odot (\mathbf{A}_i\mathbf{P}_i\mathbf{H}_i - \mathbf{R})\|\mathbf{H}_i^T + 2\lambda\mathbf{P}_i - M \quad (18)$$

By setting the derivative to zero and employing the Karush–Kuhn–Tucker complementary condition [37], i.e., $M(s,\ t)\mathbf{P}_i(s,t) = 0$, we obtain:

$$\left[\mathbf{A}_i^T[W \odot (\mathbf{A}_i\mathbf{P}_i\mathbf{Q} - \mathbf{R})]\mathbf{H}_i^T\ +\ \lambda\mathbf{P}_i\right](s,\ t)\mathbf{P}_i(s,t) =\ 0 \quad (19)$$

Finally, the updated rule of $\mathbf{P}_i$ is computed using:

$$\mathbf{P}_i(s,t) \leftarrow \mathbf{P}_i(s,t)\sqrt{\frac{\left[\mathbf{A}_i^T(W \odot \mathbf{R})\mathbf{H}_i^T\right](s,\ t)}{\left[\mathbf{A}_i^T(W \odot (\mathbf{A}_i\mathbf{P}_i\mathbf{H}_i))\mathbf{H}_i^T\ +\ \lambda\mathbf{P}_i\right](s,\ t)}} \quad (20)$$

3.5.2. The Basis of Updating Q_i

Similarly, for $\mathbf{Q}_i$, the unrelated terms are initially discarded by fixing the other variables, and the resulting objective function is expressed as:

$$\underbrace{\min}_{\mathbf{Q}_i \geq 0} \|W \odot (\mathbf{R} - \mathbf{B}_i\mathbf{Q}_i\mathbf{K}_i)\|_F^2 + \lambda\|\mathbf{Q}_i\|_F^2 + \frac{\beta}{2}\mathrm{tr}\left(\mathbf{Q}\mathbf{L}\mathbf{Q}^T\right) \tag{21}$$

where $\mathbf{B}_i$ and $\mathbf{K}_i$ for $1 \leq i \leq x$, are defined as:

$$\mathbf{B}_i = \begin{cases} \mathbf{P}_1 \ldots \mathbf{P}_x\mathbf{Q}_y \ldots \mathbf{Q}_{y+1} & \text{if } i \neq y \\ \mathbf{P}_1 \ldots \mathbf{P}_x & \text{if } i = y \end{cases} \tag{22}$$

$$\mathbf{K}_i = \begin{cases} \mathbf{Q}_{y-1} \ldots \mathbf{Q}_1 & \text{if } i \neq 1 \\ \mathbf{I} & \text{if } i = 1 \end{cases} \tag{23}$$

We can then compute the updated rule for $\mathbf{Q}_i$ in the same way as $\mathbf{P}_i$:

$$\mathbf{Q}_i(s,t) \leftarrow \mathbf{Q}_i(s,t)\sqrt{\frac{\left[\mathbf{B}_i^T(W \odot \mathbf{R})\mathbf{K}_i^T + \frac{\beta}{2}\mathrm{tr}\left(\mathbf{Q}\mathbf{L}\mathbf{Q}^T\right)\right](s,\ t)}{\left[\mathbf{B}_i^T(W \odot (\mathbf{B}_i\mathbf{Q}_i\mathbf{K}_i))\mathbf{K}_i^T\ +\ \lambda\mathbf{Q}_i + \frac{\beta}{2}\mathrm{tr}\left(\mathbf{Q}\mathbf{L}\mathbf{Q}^T\right)\right](s,\ t)}} \tag{24}$$

The optimization with the above updating rules for $\mathbf{P}_i$ and $\mathbf{Q}_j$ tries to unveil the approximation of the factors in the proposed model. Each hierarchical level is pre-trained to get an initial approximation of the matrices $\mathbf{P}_i$ and $\mathbf{Q}_j$. The input user–item rating matrix is factorized into $\tilde{\mathbf{P}}_1$ $\tilde{\mathbf{Q}}_1$ by solving Equation (2). Then, $\tilde{\mathbf{P}}_i$ and $\tilde{\mathbf{Q}}_i$ are further factorized into $\tilde{\mathbf{P}}_i \approx \mathbf{P}_1\tilde{\mathbf{P}}_2$ and $\tilde{\mathbf{Q}}_i \approx \tilde{\mathbf{Q}}_2\mathbf{Q}_1$, respectively. The factorization step is continued up until the pth user and qth item hierarchical levels are obtained. The fine-tuning process is performed by updating $\mathbf{P}_i$ and $\mathbf{Q}_i$ using Equations (20) and (24) separately. The step first involves updating $\mathbf{Q}_i$ in sequence and then $\mathbf{P}_i$ in sequence. Finally, the predicted rating matrix will be equal to $\mathbf{R}^{'} = \mathbf{P}_1 \ldots \mathbf{P}_x\mathbf{Q}_y \ldots \mathbf{Q}_1$.

3.6. Convergence Analysis

The examination of the convergence of the proposed model was conducted as follows. The assistant function in [38] was used to prove the convergence of the model.

Definition 1. *The assistant function* [38] *is defined as* $G(h, h')$ *for* $F(h)$ *if the conditions:*

$$G(h, h') \geq F(h),\ G(h, h) = F(h) \tag{25}$$

are satisfied.

Assumption 1. *If G [38] is an assistant function for F, then F is non-increasing under the update:*

$$h^{(t+1)} = \arg\min G\left(h, h^{(t)}\right) \tag{26}$$

Proof.

$$F\left(h^{t+1}\right) \leq G\left(h^{(t+1)}, h^{(t)}\right) \leq G\left(h^{(t)}, h^{(t)}\right) \leq G\left(h^{(t)}\right) \tag{27}$$

□

Assumption 2. [39] *For any matrices* $A \in \mathbb{R}_+^{n\times n}$,$B \in \mathbb{R}_+^{k\times k}$, $S \in \mathbb{R}_+^{k\times k}$, *and* $S' \in \mathbb{R}_+^{k\times k}$, *where* ***A*** *and* ***B*** *are symmetric, the following inequality holds:*

$$\sum_{s=1}^{n}\sum_{t=1}^{k}\frac{(AS'B)(s,t)S^2(s,t)}{S'(s,t)} \geq \mathrm{Tr}\left(S^T ASB\right) \tag{28}$$

The objective function in Equation (14) can be written in the following form by developing the quadratic terms and removing terms that are unrelated to $\boldsymbol{P}_i$:

$$J(\mathbf{P}_i) = \mathrm{Tr}\left(-2\mathbf{A}_i^T(W\odot\mathbf{R})\mathbf{H}_i^T\mathbf{P}_i^T\right) + \mathrm{Tr}\left(\mathbf{A}_i^T\left(W\odot\left(\mathbf{A}_i^T\mathbf{P}_i\mathbf{H}_i\right)\right)\mathbf{H}_i^T\mathbf{P}_i^T\right) + \mathrm{Tr}\left(\lambda\mathbf{P}_i\mathbf{P}_i^T\right) \tag{29}$$

Theorem 1.

$$\begin{aligned}\mathbf{G}\left(\mathbf{P},\mathbf{P}'\right) &= -2\sum_{s,t}\left(\mathbf{A}_i^T(W\odot\mathbf{R})\mathbf{H}_i^T\right)(s,t)\mathbf{P}_i(s,t)\left(1+\log\frac{\mathbf{P}_i(s,t)}{\mathbf{P}_i'(s,t)}\right)\\ &+\sum_{s,t}\frac{(\mathbf{A}_i^T(W\odot(\mathbf{A}_i^T\mathbf{P}_i\mathbf{H}_i))\mathbf{H}_i^T)(s,t)\mathbf{P}_i^2(s,t)}{\mathbf{P}_i'(s,t)} + \mathrm{Tr}\left(\lambda\mathbf{P}_i\mathbf{P}_i^T\right)\end{aligned} \tag{30}$$

The above function is an assistant function for $J(\boldsymbol{P}_i)$. Moreover, it is a convex function in $(\boldsymbol{P}_i)$ and its global minimum is:

$$\mathbf{P}_i(s,t) \leftarrow \mathbf{P}_i(s,t)\sqrt{\frac{\left[\mathbf{A}_i^T(W\odot\mathbf{R})\mathbf{H}_i^T\right](s,t)}{\left[\mathbf{A}_i^T(W\odot(\mathbf{A}_i\mathbf{P}_i\mathbf{H}_i))\mathbf{H}_i^T+\lambda\mathbf{P}_i\right](s,t)}} \tag{31}$$

Proof. The proof is similar to that in [40] and thus the details are omitted. □

Theorem 2. *Updating $\boldsymbol{P}_i$ with Equation (20) will monotonically decrease the value of the objective in Equation (13).*

Proof. With Assumption 1 and Theorem 1, we have:

$$J\left(\mathbf{P}_i^{(0)}\right) = G\left(\mathbf{P}_i^{(0)},\mathbf{P}_i^{(0)}\right) \geq G\left(\mathbf{P}_i^{(1)},\mathbf{P}_i^{(0)}\right) \geq J\left(\mathbf{P}_i^{(1)}\right) \tag{32}$$

That is, $J(\mathbf{P}_i)$ decreases monotonically. Equivalently, the update rule for $\mathbf{Q}_i$ will also monotonically decrease the value of the objective in Equation (13). Since the value of the objective in Equation (13) is at least edged by zero, we can have shown that the optimization technique of the proposed method converges. □

3.7. Time Complexity Analysis

The most expensive operations in the proposed model are the initialization and fine-tuning process that leads to increasing the efficiency of the model. Namely, the time complexity of the decomposition of $\widetilde{\mathbf{P}_i} \in \mathbb{R}^{n_{i-1}\times r}$ to $\mathbf{P}_i \in \mathbb{R}^{n_{i-1}\times n_i}$ and $\widetilde{\mathbf{P}_{i+1}} \in \mathbb{R}^{n_i\times r}$ is $O(kn_{i-1}n_ir)$ for $1 < i < x$ and $O(knn_1r)$ for $i = 1$, where k is the number of iterations in the decomposition process. Hence, the cost of initializing the $\mathbf{P}_i$'s is $O(kr(nn_1 + n_1n_2 + \ldots + n_{x-2}n_{x-1}))$. Likewise, the cost of initializing the $\mathbf{Q}_i$'s is $O\left(kr\left(mm_1 + m_1m_2 + \ldots + m_{y-2}m_{y-1}\right)\right)$. The computational costs of fine-tuning $\mathbf{P}_i$ and $\mathbf{Q}_i$ in each iteration are $O(nn_{i-1}n_i + nn_im + n_{i-1}n_im)$ and $O(mm_{i-1}m_i + mm_in + m_{i-1}m_in)$. Let $n_0 = n$, $m_0 = m$, $n_x = m_y = r$, then the time complexity of fine-tuning is $O\left(k_f\left[(n+m)\left(\sum_{i=1}^{x} n_{i-1}n_i + \sum_{j=1}^{y} m_{j-1}m_j\right) + nm\left(\sum_{i=1}^{x} n_i + \sum_{j=1}^{y} m_j\right)\right]\right)$, where k_f is the number of iterations in the fine-tuning process. The time complexity of computing the item similarities and $\mathbf{L}$ is $O\left(m^2t\right)$, where m is the total number of items and t is the total number of tags. Hence, the total time complexity is the sum of the cost of the initialization, fine-tuning, and computing the item similarities. It is interesting to note that in practice, two hierarchical levels of users and items, $x = 2$ and $y = 2$, give better performance advancement over MF and WNMF. When $x > 2$ and $y > 2$, the performance of

the proposed model is also better than that of x = 2 and y = 2, but the time complexity grows. Therefore, the optimal value of x and y is chosen to be 2 practically because the time complexity is not larger than for MF and WNMF.

4. Experiment

4.1. Dataset

To evaluate the performance of our model, an experiment was performed with the latest small MovieLens 100K dataset. The dataset comprises 100,000 movie ratings and 3683 tags that are essentially user-generated metadata (a single word or short phrase) about movies. The ratings are scored on a scale of 0.5 to 5.0 stars, and movies and users are selected from a total of 19 genres and 21 occupation categories, respectively. While the genres and occupations are leveraged for hierarchical information of the movies and users, the tags lend to tag information.

4.2. Measurement Metric

The dataset was randomly divided into 60% and 80% for training, and the remaining instances were split as 40% and 20% for testing. The prediction accuracy of the proposed model was measured using the popular mean absolute error (MAE) metric. MAE returns the average absolute deviation of the prediction from the ground truth:

$$\mathrm{MAE} = \frac{\sum_{(i,j)\in\tau}\left|R_{ij} - R'_{ij}\right|^2}{|\tau|} \tag{33}$$

where τ is a set of ratings, and R and R′ are the true and predicted ratings, respectively. The smaller the value of MAE, the more accurate the prediction; hence, MAE is preferred when the indicator values are small.

4.3. Results

We evaluated the model using two indicators: the rating prediction error (i.e., MAE) for the predetermined weights of the tag information and the extent of mitigating the item cold-start problem.

4.3.1. Prediction Accuracy with Tag Information Weights

It is worth noting that the proposed method completed the entire workflow for the rating prediction only in the case of items constituting tag information, while for the rest of the instances, it morphed into a basic WNMF model, i.e., without solving for Equations (10)–(13). To prove the superiority of our approach, two baseline methods were selected for comparison, where the results are summarized in Table 2.

1. Matrix factorization: Proposed by Koren et al. [3], this method factorizes a user–item rating matrix and learns the resultant user and item latent feature vectors to minimize the error between the true and predicted ratings.
2. Weighted non-negative matrix factorization: This was also chosen as the base model for the proposed approach, where WNMF attempts to factorize a weighted user–item rating matrix into two non-negative submatrices to minimize the error between the true and predicted ratings.

The results were taken when the parameter r was defined as 20 and the size of n_1 ranged according to {50, 100, 150, 200, 250}, while m_1 ranged according to {100, 200, 300, 400, 500}. The values of the hierarchical layers x and y were equal to 2. $W \odot \mathbf{R} \approx W \odot (\mathbf{P}_1\mathbf{P}_2\mathbf{Q}_1\mathbf{Q}_2)$ where the matrices are given as follows $\mathbf{P}_1 \in \mathbb{R}^{n\times n_1}$, $\mathbf{P}_2 \in \mathbb{R}^{n_1\times r}$ and $\mathbf{Q}_1 \in \mathbb{R}^{r\times m_1}$, $\mathbf{P}_2 \in \mathbb{R}^{m_1\times m}$. Overall, when the values of the dimensions rose, the model performance tended to grow at first and consequently fell.

Table 2. Comparison of the mean absolute error (MAE) results of the rating predictions between different methods and the proposed approach.

Training Set Size (%)	MAE				
	MF	WNMF	Proposed		
			MAE	Number of Optimal Hierarchical Levels	
				x (Users)	y (Items)
60	0.7635	0.7820	0.7386	2	2
80	0.7586	0.7657	0.7309	2	2

MF: matrix factorization, WNMF: weighted non-negative matrix factorization.

The extra regularization parameter, β, controls the contribution of tag information in learning the item latent feature vector. In other words, for $\beta = 0$, our methodology adopted the basic WNMF to compute Equation (13) and thereafter predict ratings, whereas, for non-zero β values, the weight of the tag information manifested its effects on the predictions, as illustrated in Figure 5. Although this reflects a certain degree of reliance on β for the proposed approach, it also proved the efficiency of using a combination of hierarchical and tag information. The correlation between MAE and β for β in the range of 0.05–3.0 was plotted for the 80% training dataset and the accuracy increased proportionally with β, peaking between 1.0 and 2.1 (lowest MAE recorded).

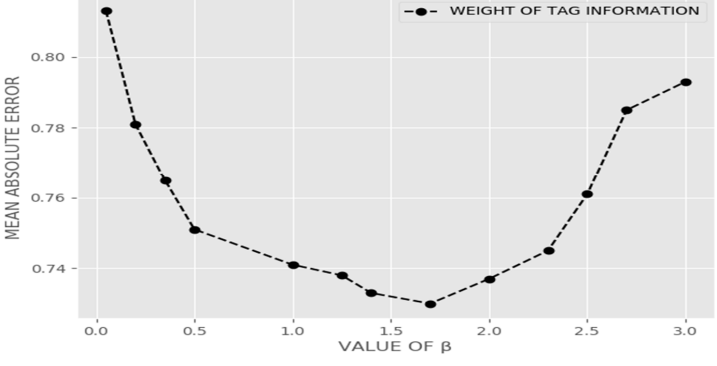


Figure 5. The weight of the tag information in the recommendation system.

4.3.2. Mitigation of the Item Cold Start

One of the main challenges encountered when building a recommendation system is the cold-start problem, which arises when a new user or item is introduced for which no past interactions are available. In particular, collaborative filtering algorithms are more prone to the cold-start problem. As basic models, matrix factorization algorithms (WNMF and MF) have poor performances in the case of the cold-start problem due to a lack of preference information [26,27,41]. Supposing the tag information is accessible for use, our proposed model can mitigate the cold-start problem by seamlessly incorporating the tag information to provide a recommendation. Tag information not only contains an explanation of the items but also provides the sentiment of users. In particular, the proposed method tries to make two item-specific latent feature vectors as similar as possible if the two items have a similar tagging history. It can give recommendations to new users who have no preference for any items. In such cases, the proposed approach helped in alleviating the cold-start problem by integrating tag information, where other comparable methods failed.

To test this, the ratings of 50 and 100 randomly selected items from the 80% training dataset were discarded such that they were viewed as new items (cold-start items) in the recommendation system. In the cold-start experiments, the results of the proposed model performance were taken when the parameters of the model were set to the optimal values of $\beta = 1.8$, $r = 20$, and the number of hierarchical layers x and y were equal to 2. The comparative results are presented in Table 3, which shows that the proposed method outperformed the MF and WNMF models, validating the conducted test, showing that tag information could be used to execute recommendations for cold-start items. It is evident that in both instances, the proposed methodology helped with mitigating the cold-start problem for new items significantly better than its competitors.

Table 3. MAE performance comparisons for the item cold-start problem.

Cold-Start Case	50 Cold-Start Items					100 Cold-Start Items				
	MF	WNMF	Proposed			MF	WNMF	Proposed		
			MAE	Number of Optimal Hierarchical Levels				MAE	Number of Optimal Hierarchical Levels	
				x (Users)	y (Items)				x (Users)	y (Items)
All items	0.8894	0.8461	0.8096	2	2	0.9135	0.8836	0.8740	2	2
Cold-start items	0.9247	0.8613	0.8287	2	2	0.9591	0.9165	0.9107	2	2

4.3.3. Top-N Recommendation Results

Along with providing superior MAE results for rating predictions, the proposed model also showed its superiority when performing the top-N recommendation task. Experiments on the proposed model for top-N recommendation identified the items that best fit the user's personal tastes obtained from their hierarchically structured features and tagging history. To evaluate the top-N performance of the proposed model, an 80% training dataset was used to generate a ranked list of size N items for each user. The proposed method and the other two baseline cutting edge methods were compared using the most widely used MovieLens 100K dataset, as indicated in Figure 6. The comparison task was performed for three sizes of N: the first was the top-5, the second was the top-10, and the final one was the top-15. When the size of N was equal to 5, the MAE of the MF method was 0.748, while the MAE of the WNMF method was higher by 0.01 than the MF method. However, the proposed model outperformed both the MF and WNMF methods and accomplished the lowest error rate of 0.736 for the top-5 and 0.752 for the top-10, whereas the other two methods (MF and WNMF) showed 0.757 and 0.772 for the top-10, respectively. Our suggested approach required expensive operations for the initialization and fine-tuning process. For this reason, the proposed method had a slightly higher error rate compared to the MF method, as indicated for the top-15. From these experiments, the proposed method still worked successfully and the superiority was clearly verified.

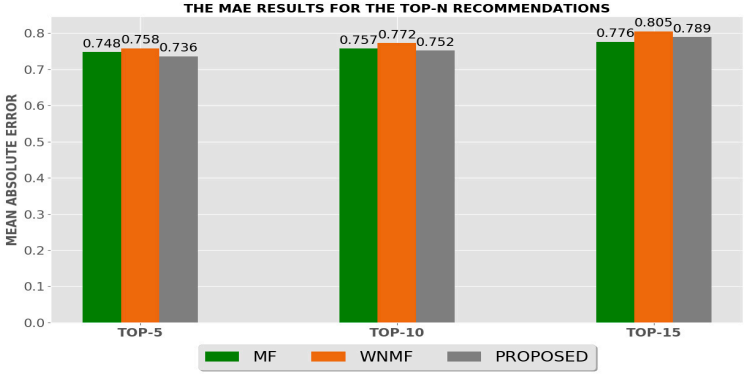


Figure 6. The MAE results for the top-N recommendations.

5. Conclusions

Presently, while the development of personalized recommendation systems has been continuing to grow to a high degree, data sparsity, cold starts, and improving recommendation system performances are still open challenges that need to be solved in the recommendation system area. In this study, we proposed a novel rating prediction model with enhanced matrix factorization using hierarchical and tag information that addressed the above issues. Experimental results revealed the significant influence of the hierarchical and tag information used in combination to alleviate the issues of data sparsity and item cold starts compared to established MF techniques. The entire workflow of our proposed model for rating predictions was completed only in the case of items constituting tag information with the hierarchical information of users and items. In particular, deep factorization on the user preference and item characteristic matrices was accomplished due to their non-negativity to get hidden-level hierarchical structured features, while tag information was used to regularize the matrix factorization process of a basic WNMF model to complete our prediction model. During the experimental testimony process, we concluded that if the values of the dimensions increased, the proposed model performance tended to increase at first and then decrease. Despite the superiority of the proposed approach, several problems were encountered, especially with the advances in the domain that focus on the high volume of data available for making recommendations. Therefore, future research could explore more sophisticated models for estimating the importance of the hidden features of users and items that the features represented as hierarchical structures, as well as tag information preference, by using recent deep learning methods and algorithms. Additionally, future research work might similarly also develop an explainable and interpretable recommendation system based on the above hidden features.

Author Contributions: This manuscript was designed and written and the experiments were performed by A.K. A.A. helped to revise and improve the manuscript. The theory and experiments were analyzed and commented on by T.K.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: R2020040243).

Acknowledgments: The authours A.K. and A.A. would like to express their sincere gratitude and appreciation to the supervisor, Taeg Keun Whangbo (Gachon University) for his support, comments, remarks, and engagement over the period in which this manuscript was written. Moreover, the authors would like to thank the editor and anonymous referees for the constructive comments in improving the contents and presentation of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutierrez, A. Recommender systems survey. *Knowl. Based Syst.* **2013**, *46*, 109–132. [CrossRef]
2. Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P.B. *Recommender Systems Handbook*; Springer: Berlin, Germany, 2011; ISBN 978-0-387-85819-7.
3. Koren, Y.; Bell, R.; Volinskiy, C. Matrix factorization techniques for recommender systems. *IEEE Comput.* **2009**, *42*, 30–37. [CrossRef]
4. Zhang, S.; Yao, L.; Sun, A.; Tay, Y. Deep Learning based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* **2018**, *52*, 1–38. [CrossRef]
5. Ortega, F.; Hurtado, R.; Bobadillla, J.; Bojorque, R. Recommendation to groups of users the singularities concept. *IEEE Access* **2018**, *6*, 39745–39761. [CrossRef]
6. Tuzhilin, A.; Adomavicius, G. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749.
7. Su, X.; Khoshgoftaar, T.M. A survey of collaborative filtering techniques. *Adv. Artif. Intell.* **2009**. [CrossRef]
8. Chatti, M.A.; Dakova, S.; Thus, H.; Schroeder, U. Tag-based collaborative filtering recommendation in personal learning environments. *IEEE Trans. Learn. Technol.* **2012**, *6*, 337–349. [CrossRef]
9. Goldberg, D.; Nichols, D.; Oki, B.M.; Terry, D. Using collaborative filtering to weave an information tapestry. *Commun. ACM* **1992**, *35*, 61–70. [CrossRef]

10. Liu, J.; Tang, M.; Zheng, Z.; Liu, X.; Lyu, S. Location-Aware and Personalized Collaborative Filtering for Web Service Recommendation. *IEEE Trans. Serv. Comput.* **2016**, *9*, 686–699. [CrossRef]
11. Herlocker, J.L.; Konstan, J.A.; Borchers, A.; Riedl, J. An algorithmic framework for performing collaborative filtering. In Proceedings of the SIGIR'99: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 230–237.
12. Guo, X.; Yin, S.-C.; Zhang, Y.-W.; Li, W.; He, Q. Cold start recommendation based on attribute-fused singular value decomposition. *IEEE Access* **2019**, *7*, 11349–11359. [CrossRef]
13. Yang, J.; Sun, Z.; Bozzon, A.; Zhang, J. Learning hierarchical feature influence for recommendation by recursive regularization. In Proceedings of the Recsys: 10th ACM Conference on Recommender System, Boston, MA, USA, 15–19 September 2016; pp. 51–58.
14. Koren, Y.; Bell, R. Advances in collaborative filtering. In *Recommender Systems Handbook*; Springer: Berlin, Germany, 2011; pp. 145–186.
15. *Unifying User-Based and Item-Based Collaborative Filtering Approaches by Similarity Fusion*; SIGIR '06; ACM: New York, NY, USA, 2006.
16. Zarei, M.R.; Moosavi, M.R. A Memory-Based Collaborative Filtering Recommender System Using Social Ties. In Proceedings of the 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), Tehran, Iran, 6–7 March 2019.
17. Stephen, S.C.; Xie, H.; Rai, S. Measures of similarity in memory-based collaborative filtering recommender system: A comparison. In Proceedings of the 4th Multidisciplinary International Social Networks Conference, 4th Multidisciplinary International Social Networks Conference (MISNC), Bangkok, Thailand, 17–19 July 2017.
18. Al-bashiri, H.; Abdulgabber, M.A.; Romli, A.; Kahtan, H. An improved memory-based collaborative filtering method based on the TOPSIS technique. *PLoS ONE* **2018**, *13*, e0204434. [CrossRef] [PubMed]
19. Li, X.; Li, D. An Improved Collaborative Filtering Recommendation Algorithm and Recommendation Strategy. *Mobile Inform. Syst.* **2019**. [CrossRef]
20. Fang, Y.; Si, L. Matrix co-factorization for recommendation with rich side information and implicit feedback. In *Hetrec 11*; ACM: New York, NY, USA, 2011.
21. Kumar, A.; Sodera, N. Open problems in recommender systems diversity. In Proceedings of the International Conference on Computing, Communication and Automation (ICCCA2017), Greater Noida, India, 5–6 May 2017.
22. Salakhutdinov, R.; Mnih, A. Probabilistic matrix factorization. In Proceedings of the NIPS'07: 20th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007.
23. Seo, S.; Huang, J.; Yang, H.; Liu, Y. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Recsys '17*; ACM: New York, NY, USA, 2017.
24. Maleszka, M.; Mianowska, B.; Nguyen, N.T. A method for collaborative recommendation using knowledge integration tools and hierarchical structure of user profiles. *Knowl. Based Syst.* **2013**, *47*, 2013. [CrossRef]
25. Ge, M.; Elahi, M.; Tobias, I.F.; Ricci, F.; Massimo, D. Using tags and latent factors in a food recommender system. In Proceedings of the DH '15: 5th International Conference on Digital Health, Florence, Italy, 18–20 May 2015.
26. Garg, N.; Weber, I. Personalized, interactive tag recommendation for flickr. In Proceedings of the 2nd ACM International Conference on Recommender Systems, RecSys'08, Lausanne, Switzerland, 23–25 October 2008; pp. 67–74. [CrossRef]
27. Tso-Sutter, K.H.L.; Marinho, L.B.; Schmidt-Thieme, L. Tag-aware recommender systems by fusion collaborative filtering algorithms. In Proceedings of the SAC '08: 2008 ACM Symposium on Applied Computing, Fortaleza, Brazil, 16–20 March 2008.
28. Schein, A.I.; Popescul, A.; Ungar, L.H.; Pennock, D.M. Methods and metrics for cold-start recommendations. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 11–15 August 2002; pp. 253–260.
29. Vall, A.; Skowron, M.; Schedl, M. *Improving Music Recommendations with a Weighted Factorization of the Tagging Activity*; ISMIR: Montreal, QC, Canada, 2015.
30. Shi, C.; Liu, J.; Zhuang, F.; Yu, P.S.; Wu, B. Integrating Heterogeneous Information via Flexible Regularization Framework for Recommendation. *Knowl. Inform. Syst.* **2016**, *49*, 835–859. [CrossRef]

31. Wu, J.; Chen, L.; Yu, Q.; Han, P.; Wu, Z. *Trust-Aware Media Recommendation in Heterogeneous Social Networks*; Springer: Berlin, Germany, 2015.
32. Lu, K.; Zhang, G.; Li, R.; Zhang, S.; Wang, B. Exploiting and exploring hierarchical structure in music recommendation. In *AIRS 2012: Information Retrieval Technology*; Springer: Berlin, Germany, 2012; pp. 211–225.
33. Nikolakopoulos, N.; Kouneli, M.A.; Garofalakis, J.D. Hierarchical Itemspace Rank: Exploiting hierarchy to alleviate sparsity in ranking-based recommendation. *J. Neurocomput.* **2015**, *163*, 126–136. [CrossRef]
34. Wang, Z.; Wang, Y.; Wu, H. Tag meet ratings: Improving collaborative filtering with tag-based neighborhood method. In Proceedings of the SRS'10 ACM, Hong Kong, China, 7 February 2010.
35. Shepitsen, A.; Gemmell, J.; Mobasher, M.; Burke, R. Personalized recommendation in social tagging systems using hierarchical clustering. In Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys, Lausanne, Switzerland, 23–25 October 2008.
36. Chung, F. *Spectral Graph Theory*; American Mathematical Society: Providence, RI, USA, 1997.
37. Trigeorgis, G.; Bousmalis, K.; Zaferiou, S.; Schuller, B. A deep semi-nmf model for learning hidden representations. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), Beijing, China, 21–26 June 2014; pp. 1692–1700.
38. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **2001**, *13*, 556–562.
39. Ding, C.; Li, T.; Peng, W.; Park, H. Orthogonal nonnegative matrix t-factorizations for clustering. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 126–135.
40. Gu, Q.; Zhou, J.; Ding, C.H.Q. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In Proceedings of the 2010 SIAM International Conference on Data Mining, Columbus, OH, USA, 29 April–1 May 2010; pp. 199–210.
41. Lam, X.N.; Vu, T.; Le, T.D.; Duong, A.D. Addressing cold-start problem in recommendation systems. In Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, Suwon, Korea, 31 January 2008; pp. 208–211.

symmetry

MDPI

Article

Toward Social Media Content Recommendation Integrated with Data Science and Machine Learning Approach for E-Learners

Zeinab Shahbazi and Yung Cheol Byun *

Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Provience, Korea; zeinab.sh@jejunu.ac.kr
* Correspondence: ycb@jejunu.ac.kr

Received: 31 August 2020; Accepted: 19 October 2020; Published: 30 October 2020

Abstract: Electronic Learning (e-learning) has made a great success and recently been estimated as a billion-dollar industry. The users of e-learning acquire knowledge of diversified content available in an application using innovative means. There is much e-learning software available—for example, LMS (Learning Management System) and Moodle. The functionalities of this software were reviewed and we recognized that learners have particular problems in getting relevant recommendations. For example, there might be essential discussions about a particular topic on social networks, such as Twitter, but that discussion is not linked up and recommended to the learners for getting the latest updates on technology-updated news related to their learning context. This has been set as the focus of the current project based on symmetry between user project specification. The developed project recommends relevant symmetric articles to e-learners from the social network of Twitter and the academic platform of DBLP. For recommendations, a Reinforcement learning model with optimization is employed, which utilizes the learners' local context, learners' profile available in the e-learning system, and the learners' historical views. The recommendations by the system are relevant tweets, popular relevant Twitter users, and research papers from DBLP. For matching the local context, profile, and history with the tweet text, we recognized that terms in the e-learning system need to be expanded to cover a wide range of concepts. However, this diversification should not include such terms which are irrelevant. To expand terms of the local context, profile and history, the software used the dataset of Grow-bag, which builds concept graphs of large-scale Computer Science topics based on the co-occurrence scores of Computer Science terms. This application demonstrated the need and success of e-learning software that is linked with social media and sends recommendations for the content being learned by the e-Learners in the e-learning environment. However, the current application only focuses on the Computer Science domain. There is a need for generalizing such applications to other domains in the future.

Keywords: data science; DBLP platform; Twitter; deep reinforcement learning; machine learning; recommendation system

1. Introduction

The recommendation system (RS) creates possible options for users based on user interest [1]. The proposed recommendation system is based on information which the user gave to the system in the past. The given information may have many ratings which show that the aim of the user is to get information from a particular domain—e.g., research area, documents, tweets, etc. Based on the recommendation system architecture, the given information can be used as training data, either supervised learning or unsupervised learning—e.g., clustering or document classification problems [2,3]. In recent years, the recommendation system becomes a popular engine to implement on many websites to

find the preference of users. Two main techniques which present this system are known as content-based recommendation (CB) and collaborative filtering recommendation (CF) [4–6]. Comparing these two systems, CF is the most often used technique. The CB system is processed as recommending similar options to the user based on user choices, and extract the target from user input information which is accessible to the user profile and also the output profile [7].

The DBLP is known as the Digital Bibliographic Library Project, which indexes more than 2 million research publications. The dump of the website is freely available for download. This dataset will help to identify top-rated articles based on learners profile, context, and history. Twitter is a microblogging website where the social community generates a large number of small text messages of 140 characters. It is noted that, daily, more than 300 million tweets are generated by the social community, expressing the opinions and sentiments related to different things. The proposed system will identify the top-rated and most popular Twitter user relevant to the user's profile. For this, the system will perform network analysis and identify the most popular Twitter users using different social network analysis techniques, such as by measuring in-degree centrality. The centrality of a node describes its popularity. These users are matched with the current learner activity. Not getting relevant recommendations from social networks for e-learners based on the user's context, historical data, and profiles are the problem statement.

The developed system provides recommendations from the social network for e-learners. The recommendations are based on the users' local context, profiles, and historical data. The system is a web application that searches the required data from the Web using different sources, such as the Twitter user network and DBLP. The developed system recommends top-ranked information from these sources depending upon the context, profile, and history of e-Learners. The application also provides an interface to e-learners, where they can share their content and can read the contents shared by other users, which is relevant to them. The reinforcement learning algorithm is presented as a recommendation platform in the proposed system. Reinforcement learning is a learning algorithm which works based on user feedback. The quality control of the system improves based on the recommendation rating. As we know, the social media contents are unstructured and have a lack of trust. Based on the main architecture of the reinforcement learning, when the user gets a good recommendation then the feedback will be good and the reward sent to system is good. In this case, the system is learning the user's request and the recommendations are also trained based on that. Similarly, if the user feedback is not good, then the system learns to avoid recommending bad articles. To look for the history of the recommendation system, first, it must find the solution to solving the problem of information overloading in Internet sources. Along with the number of uploaded files, Internet sources are relatively high, and users do not know how to control it and also spend more time and energy looking for and extracting the topic which they need. Based on the Eugene search result, the information overloading issue was discovered in the information retrieval system, and in 1950 was submitted by Moors [8]. In the proposed model, the recommendation system brings the following contribution for e-learners:

The main contribution of this paper summarized as below:

- A real-time system which provides top-ranked Twitter user networks to e-learners from Twitter, according to their context, history and profiles.
- The proposed system recommends top-ranked articles according to the e-learner's context, e-learner's history, and e-learner's profiles from DBLP.
- The system also makes recommendations to e-learners from a local database.
- The main objective of this study is the use of data mining and machine learning approaches for social media content recommendation.

In this work, we proposed a social media content recommendation which provides the learning material to e-Learners. The developed system is a real-time application that identifies the required data from the Web using different sources, such as Twitter and DBLP. The designed system will recommend

top-ranked information from Twitter and DBLP sources depending upon the context, profile, and history of e-Learners. The application will also provide research articles related to the users searched topic. Moreover, we have used data mining and machine learning approach to improve the accuracy of social media content recommendation. Reinforcement learning is used as a machine learning algorithm which combined with data mining techniques to extract the hidden knowledge from users tweets. Finally, we illustrate the constructiveness of reinforcement learning, which applied for prediction and recommendation of social media contents. The remainder of this paper is organized as follows: Section 2 gives the literature review of the recommendation system. Section 3 explains the data analysis for social media contents recommendation. Section 4 presents the predictive analysis of Twitter and DBLP dataset using a reinforcement learning algorithm. Section 5 presents the prediction result of Twitter and DBLP platform and, finally, we conclude the paper in Section 6.

2. Literature Review

In this section, we discuss the pros and cons of the existing recommendation system [9]. Moreover, we will also investigate the state-of-art approaches for Twitter recommendation, DBLP recommendation and recommendation based on reinforcement learning.

2.1. Recommendation System

The recommendation system is a service to help users for easy access to their request in different areas [10,11]. Tan and He [12] presented a procedure of physical resonance that is famous for resonance similarity (RES). This approach shows the comparison of superior prediction and traditional similarity based on user evaluation. Similarly, there are many IoT-based platforms, such as healthcare [13–15], indoor localization [16,17], and many other IoT systems [18–22], which have improving possibilities based on integration with the functionality of the recommendation system [23,24]. Hwang et al. [25] execute the hotel reviews for a hotel management system based on Trip Advisor review information and a Latent Dirichlet Allocation (LDA) semantic-based process to recognize and capture the performance of the Term Frequency-Inverse Document Frequency (TF-IDF) process. In the presented approach, all features related to hotels are extracted. The final results show that the LDA has less precision than word-based LDA. To make the System more accurate, Jannach et al. [26] presented regression-based and item-based recommendation. The developed recommendation systems with different researchers used collaborative filtering techniques and algorithms [27–33]. Collaborative filtering gets information based on user input knowledge and evaluates the relationship between different users to accomplish specific deductions of feature spaces.

2.2. Twitter Recommendation

Twitter is one of the social media platforms based on sharing, uploading user opinions and providing information about new studies, interests, etc. [34–36]. There are many research articles related to Twitter classification in various goals. Some tweet recommendation systems proposed Twitter as a reliable information spreader. Tweet recommendation and also Twitter users are the main research direction in this topic too [37]. Based on the proposed methodology, there are three main options to find the user influence on the Twitter platform that are named as followers, re-tweets and page rank [38]. Building the recommendation system of followers to find the differences between tweets and user profile, page rank or tweet rank estimate the efficiency of user influence to find the similarity between shared link and user profile structure. All the proposed methods in previous studies were based on the content-based recommendation to propose tweets without reflection of joint view. Tweet recommendation is to target user, using a latent factor, collaborate ranking and specific feature. User interest re-tweets are collected and estimated to establish user preference and make a recommendation. The latent factor is the improved version of collaborating ranking for ranking criterion. The latent factor is used as a parameter to increase the accuracy of system [39].

2.3. DBLP Recommendation

DBLP is one of the online and open source references for published articles in the computer science area. Based on the need for the user by visiting the DBLP website, it is comfortable and easy to access recently published or any specific articles. DBLP was developed from small experiments on web servers to famous open-data access servers in the computer science research area [40]. One of the critical parts in DBLP recommendation, based on users searching for information-related articles on their own interests, is also recommended. The user searching process is explained step by step in Figure 1. Articles contain complete information about authors, publication time, access pages, etc. [41].

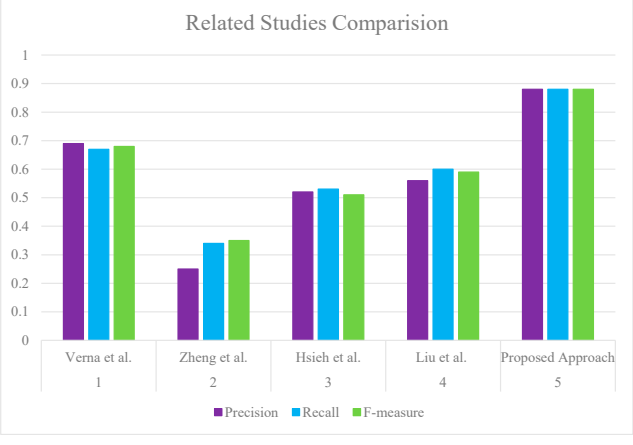


Figure 1. Comparison of related studies' accuracy.

2.4. Reinforcement Learning Recommendation

Recommendation, based on reinforcement learning (RL), is simplified to the Markov Decision Process (MDP). This model works as a long-run performance system. Most of the RL based systems have challenges from large-scale separated action space. There are some proposed systems for solving this issue, such as the strength of previous information about the actions around them, which generate proto-action using the k-nearest neighbor search system. This method rejects the dimensions from negative influences, which user does not care about, and replaces them through convenient action [42–49]. Moreover, MDP is used to model the recommendation process in RL. Compared with the Multi-Armed Bandit (MAB) based system, MDP cannot obtain the running frequency of reward. They try to define the state as an n-gram or model the item in MDP and define the action as the recommendation between items. This process cannot apply to the large datasets. If the candidate set item becomes more significant at the same time as the size of the state, the space also increases and transition data face sparsity problems and can just apply them on related parameters in a specified state [50–54]. Table 1 shows the comparison of various recommendation systems and their objectives and advantages. In the mentioned table, ten various recommendation systems are measured.

Table 1. The advantages and objective of the various recommendation systems.

Authors	Objective	Advantage
Long Zhao et al., Zemin Liu et al. (2020) [55]	Solving the large action space issue based on the Reinforcement Learning Algorithm.	Reinforcement learning solves the large action space issue.
Bushra Alhijawi et al. [56], Yousef Kilani et al. (2020)	Recommendation system classification on MRS, TPCRS, SRS and CRS	Classifying recommendation system to avoid the overloading issue.
Ruotsalo et al. [57] (2013)	Raising the digital cultural heritage accessibility	It is useful to apply to any data.
Braunhofer et al. [58] (2014)	Place of interest (POI) recommendation	Generate related recommendations with higher useability.
Elahi et al. (2013) [59]	POI-based user personality recommendation	Desist the cold start issue.
Ostuni et al. [60] (2013)	Movie theatre recommendation	Clear the content-based recommendation results.
Braunhofer et al. [61] (2014)	User personality recommendation based on contact preferences	Presenting more related recommendations based on the higher rating.
Noguera et al. (2012) [62]	Users' physical locations recommendation	Useful in e-tourism. The ability to have a 3D map.
Bouneffouf et al. [63] (2012)	Dynamic exploration recommendation	Optimal value selection while avoiding the traditional algorithms. Increasing business success probability.
Ge et al. (2010) [64]	Parking position recommendation	Providing various optimal driving routes based on online processing time.

3. Social Media Content Recommendation for E-Learners

The proposed recommendation system is comprised of two main modules—a recommendation system and the predictive analysis of social media content recommendation.

3.1. E-Learners Recommendation System

The proposed recommendation system is comprised of three-parts—presentation layer, business layer, physical layer—which are shown in Figure 2. The presentation layer is responsible for exposing the services to the front end through the user interface. The business layer represents the core functionality of the recommendation system, which is categorized into two modules—i.e., the e-learning system and the reinforcement learning-based social media content recommendation system. The e-learning system is responsible for providing relevant recommendations from social media to the e-learner. The e-learner can get top-ranked articles on the Twitter user network, which is according to the user's interest. Similarly, the reinforcement learning-based social media content recommendation system is to use data mining and machine learning approach to improve the accuracy of social media content recommendation. Reinforcement learning is used as a machine learning algorithm, which is combined with data mining techniques to extract the hidden knowledge from users tweets. Lastly, the physical layer represents the back-end database, which is responsible for storing the data.

The data collection phase is one of the primary tasks in the knowledge discovery process. The knowledge discovery process identifies hidden patterns from an enormous amount of data. We perform knowledge discovery by identifying user profiles from the DBLP and Twitter website. We targeted published articles and uploaded tweets for our experiments and the data crawling process was customized accordingly. Table 2 shows the detailed information of collected Twitter and DBLP datasets. Figure 3 shows the class diagram of the data collection process.

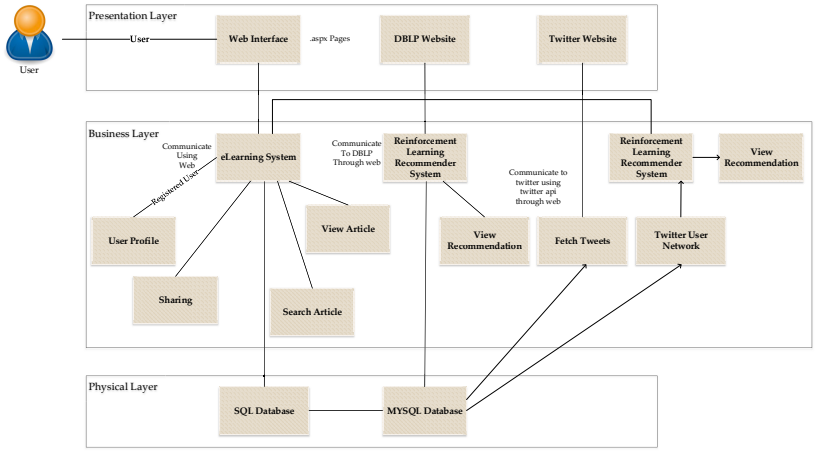


Figure 2. E-learners recommendation system.

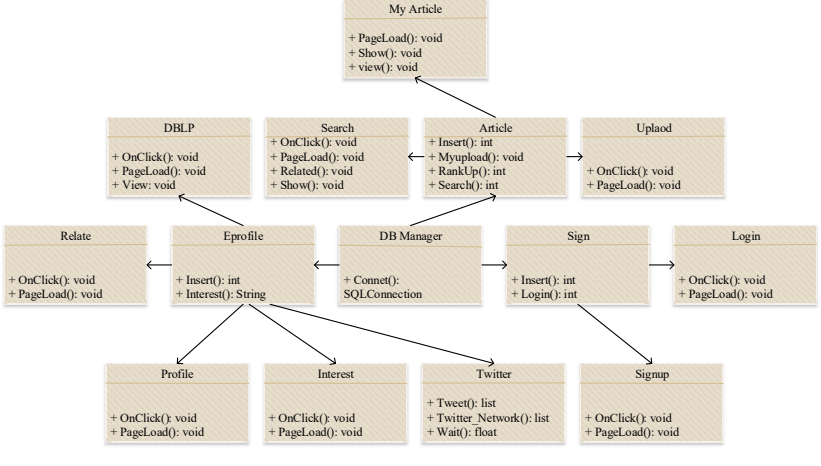


Figure 3. Class diagram of the collected dataset.

Table 2. Data information.

Statistics	Numerical Values
Number of directions	744.456
Number of articles	567.916
Number of Users	582.933
Avg/Max/Min of time	3.5/6.4/1.6
Training Data	70%
Test Data	30%

In total, two open-access social media websites were selected for this process, which contain comments, tweets, short texts and research articles. A total of 70% of the dataset was used in the training set and 30% for the test set. The primary block diagram of the proposed system is shown

in Figure 4. The designed block diagram of the proposed data and predictive analysis model based on Twitter and DBLP platform is composed of four main sections. The first section is designed as a data collection layer. The data collection section contains two social media platform datasets, named Twitter and DBLP library. The collected dataset from the Twitter platform includes tweets, projects, comments, photos, news and conferences. The collected dataset from the DBLP platform includes articles, publication access point, publication time and publication date. To process the collected dataset for further steps, two data analysis techniques were applied in this process which are in the second section or the pre-processing data section. Data analysis and predictive analysis technique was applied to the input dataset. The data analysis technique contains the time series analysis, statistical analysis, tweet analysis and article analysis. The predictive analysis contains reinforcement learning prediction techniques. The next section is the recommendation layer. It presents the output information of the previous steps and relevant recommendation results based on user preferences. The final section is the user feedback, which is the main point of this system to improve the quality of the recommendation. Based on using the reinforcement learning algorithm as a recommendation technique, the system learns from user positive and negative responses to the agent and by repeating this process, improving the system recommendation and trust quality.

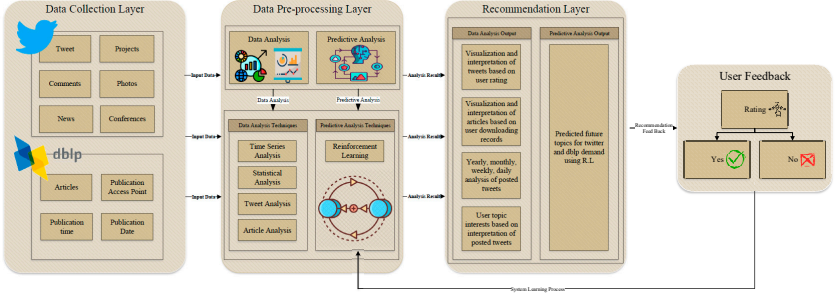


Figure 4. Block diagram of the proposed data and predictive analysis model, based on Twitter and DBLP platform.

3.2. Dataset

In this system, the collected dataset is from "Twitter social media platform records (Twitter API)" and "DBLP research library history" to analyze and explore the hidden information for improving the recommendation system. Data mining approaches and techniques were applied in the proposed dataset to clean and pre-process it to refine the performance and stability of dataset. Moreover, the following steps were performed to process a better service for e-learners on the social media platform:

- Collecting data;
- Cleaning data;
- Manipulate missing values;
- Missing value extraction;
- Discovering the available features.

After managing the social media dataset and enterprising the information, data pre-processing was applied for further process for normalizing dataset and for keeping the necessary information. Data normalization was needed for changing the data form and structure to make it convenient for further steps. The following Table 3 presents the extracted information and features from a dataset.

Table 3. Twitter and DBLP platform features and description.

#	Features	Description
1	Tweet	The information which users share together
2	Projects	The information of various projects (question and answer)
3	Comments	Comments for shared topic
4	Photos	Shared photos by various users
5	News	Shared daily news
6	Conferences	Upcoming conferences or opinions about previous conferences
7	Articles	Published articles
8	Publication Access Point	Reference pages or article access information
9	Publication Time	Article publication time
10	Publication Date	Article publication date

3.3. Data Mining and Visualization

In the proposed system, we applied data mining techniques to determine the necessary and useful information from the dataset for a suitable recommendation, based on user interest. The mentioned analysis below exploits the collected dataset:

- Twitter and DBLP article recommendation for e-learners based on tweet frequency, selected articles and user preference;
- Time series analysis based on Monthly and daily analysis;
- Twitter and DBLP platform analysis based on e-learner preferences;
- Twitter and DBLP platform analysis based on e-learner clicked links;

3.3.1. Time Series Analysis

Time series analysis applied in this process to produce the new information for article recommendation. The selected analysis is based on the date and time of sharing information which is available in the dataset. To inform the time series analysis, the duration of data is for (2019) crawled information from mentioned platforms. To start the analysis data segregated into two sections (monthly and daily) to produce the Twitter and DBLP frequency. Figures 5 and 6 present the daily and monthly basis of recommendation to e-learners. Daily basis records show the total record of the user activities in one day, and monthly basis shows the total record of the monthly user activities.

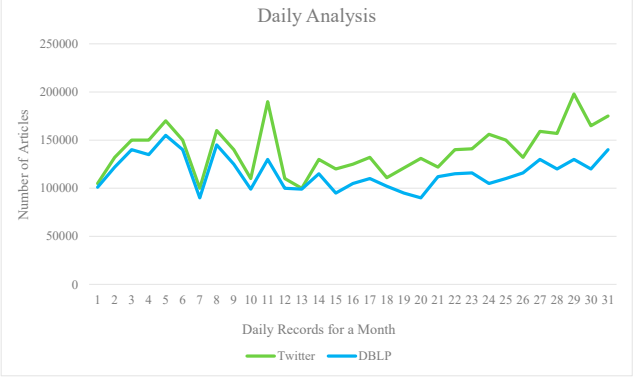


Figure 5. Time series analysis of Twitter and DBLP platform (daily basis).

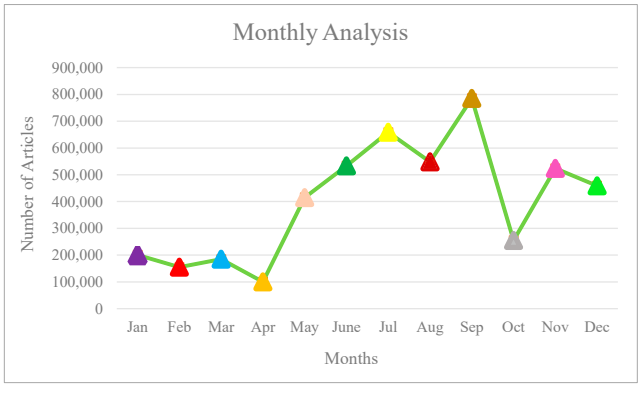


Figure 6. Time series analysis of Twitter and DBLP platform (Monthly basis).

3.3.2. Twitter API Analysis Based on Profile Address

In this part, profile address analysis accomplishes analyzing the Twitter API dataset. To visualize Twitter API, based on profile address, street names are extracted from profiles and apply them as location labels to visualize the Twitter frequency. The following parameters—e.g., profile address and tweet topics— are used as inputs of visualization based on Twitter API profile addresses. The following Figure 7 describes the minimum and maximum updates based on profile addresses. The following address is randomly selected from the dataset. Each location presenting one area based on the profile addresses.

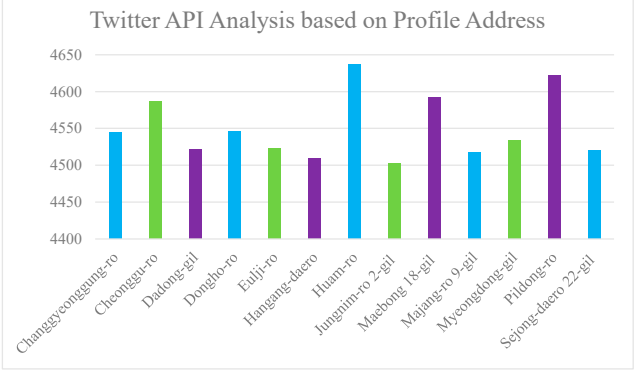


Figure 7. Twitter API analysis according to profile address.

3.4. Discover Patterns and Features

Using data mining techniques improve the process to extract the hidden information from the generated results. Table 4 presents the details of the extracted features from a dataset. The prediction process in the proposed system causes improved system performance and, similarly, recommends highly related information to the e-learner.

3.5. Interaction Model for the Proposed Recommendation Platform

The work-flow of the proposed RL recommendation model is illustrated in Figure 8. The developed system is comprised of the technical infrastructure of the system. Tweets will be fetched from Twitter using Twitter API (tweet environment). Whenever any user uploads an article in the System, the System will remove the noise data from the title of the article. Afterwards, the system will make a list of words which contain the title words and terms related to these words. The grow bag will provide the related terms. Against each word from the list, tweets will be fetched and will be saved in the database. When any e-learner wants the recommendations from Twitter against any article, the System will make two lists of words. The first list will contain the interest and terms related to the interest of an e-learner. The second list will contain the local context and terms related to the local context of the e-learner. The system will get all the tweets which were saved in the database against that article. The system will match the sub-strings of each tweet with the words of both lists. Whenever any word matches with a tweet, one score will be added to the score of that tweet and will recommend the top tweet to the user.

Table 4. List of discovered features.

#	Features	Description
1	Time series	Applying time series analysis in this system causes us to extract the information related to visited links per day or download and sharing information per day and, similarly, total average per month
2	E-learner profile details	Based on the e-learners profile, the major interest of the user on various topics and user clicks and shared tweets, news and articles are extracted.
3	statistical features	Extract the histogram, error rate, etc. from raw dataset for articles frequency.
4	Article types	Generate various article topics, titles, etc.
5	Tweet types	Generate different tweet information, comments, news and shared links.

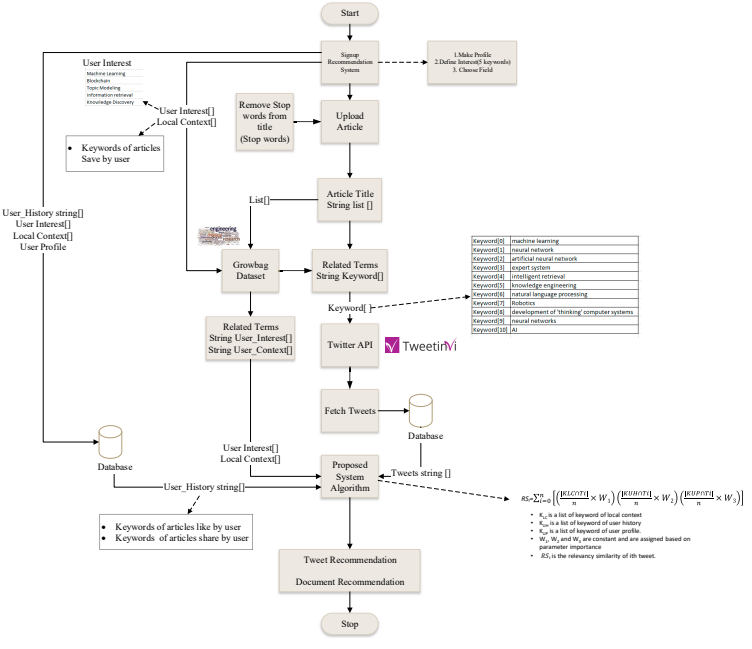


Figure 8. System work-flow of the proposed recommendation platform.

Tweets are ranked by using the following Equation (1).

$$RS_i = \sum_{i=0}^{n} [(\frac{K_{LC} \cap T_i}{n} * W_1)(\frac{K_{UH} \cap T_i}{n} * W_2)(\frac{K_{UP} \cap T_i}{n} * W_3)] \quad (1)$$

K_{LC} is a list of keywords of local context. K_{UH} is a list of keywords from the user history. K_{UP} is a list of keyword from the user profile. W_1, W_2 and W_3 are constants and are assigned based on parameter importance. RS_i is the relevancy similarity of the tweet.

4. Predictive Analysis of Twitter and DBLP Data Using Reinforcement Learning

The availability of a considerable amount of digital articles and tweets pose a challenge to discover highly relevant contents for e-learners. Current search approaches have inherited problems and use a limited set of parameters for searching the meta-data, mainly based upon the indexed keywords only. This is not enough for the users, and users are often frustrated, mainly due to the availability of a huge number of search results for a searched query. There is a need for the system, especially for the e-learners community, which can provide online real-time information from social media to the e-learner. E-learners can get top-ranked articles from the Twitter user network, which is according to the user's interest. A recommendation of social networks for the e-learner is a system that will provide the learning material to the e-Learners. The system will be a web application that will search for the required data from the Web using different sources, such as the Twitter user network and DBLP. The developed system will recommend top-ranked information from these sources depending upon the context, profile, and history of e-Learners. The application will also provide research articles related to the users searched topic. The system inputs would be the usage and viewing history of users and user profiles built by the users and the user local context. Based on this information, the system will find research articles from DBLP and Twitter users from the Twitter microblogging website.

This section presents the predictive analysis related to generated knowledge and details based on the previous sections. The presented tweet and article recommendations for e-learners are shown in Figure 9. The Applied Reinforcement Learning machine learning technique is the proposed system for recommending tweets and articles to e-learner users. The presented process predictive analysis is divided into three main sections. The first section contains the input data collected from social media platforms. The data provide the information related to e-learners IP, article title, access page, e-learner preference, e-learner click information, access date, article ID, access day, article category, article type and access time. They move to the second section, before training the dataset, pre-processing, feature engineering, data transformation and feature selection is applied to make the dataset ready for further process. After splitting the data in the train and test set, a reinforcement learning technique is applied for recommending the information, based on user preferences.

Reinforcement Learning Optimization

Reinforcement learning recommendation system contains various methods to optimize user interest. In this process, user interest directly optimizes by using FeedRec [65] through the simulation process. To do this, we need the find to "ground truth" of the system to get the maximum user engagement. Based on this, the processing algorithm shows if it is possible to get the optimal policy and maximized user engagement delay. The procedure of simulating is defining $S(z_t, i_t; \beta_z)$ with a mini-batch SGD by applying the predicted dataset. This dataset prepared based on prediction policy π_b and is immediately used as a manufacturing simulator. To get the efficiency of prediction policy π_b, the main loss in weight is minimized as in Equations (2) and (3).

$$\gamma(\beta_z) = \sum_{t=0}^{T-1} \frac{1}{m} \sum_{j=1}^{m} (v_{0:t}, D)\delta_t(\beta_z) \tag{2}$$

$$\delta_t(\beta_z) = \lambda_g * \Psi(g_t, \hat{g}_t) + \lambda_x * (x_t - \hat{x}_t)^2 + \lambda_y * \Psi(y_t, \hat{y}_t) + \lambda_w * (w^r - \hat{w}^r)^2 \tag{3}$$

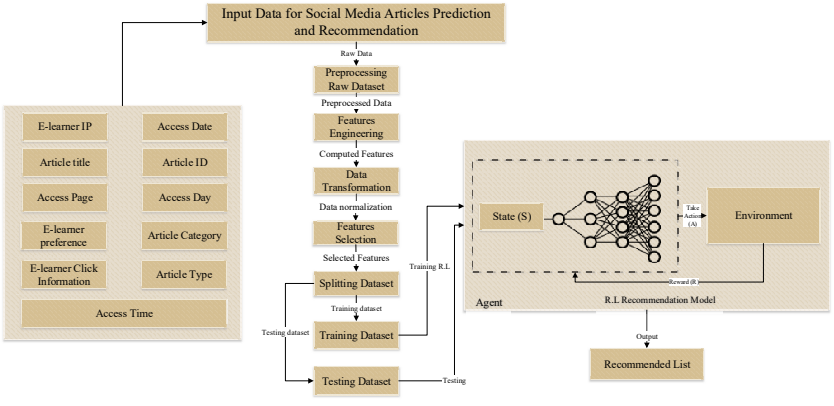


Figure 9. Architectural diagram for predictive analysis.

N is defined as the total number of directions in the predicted dataset. To reduce the disparity, $v_{0:t}$ is defined as the significant ratio between π. π is the policy which extracted from the Q-network—e.g., $\in$-greedy. Cross entropy is defined as Ψ, which shows the loss function. c is defined as the hyper-parameter to avoid from the large ratio. The multi-task loss function is defined as $\delta_t(\beta_z)$ for evaluating the comparison between regression loss and multi-task loss and λ is defined as a hyper-parameter controller to evaluate the various tasks. Based on the updates from β_z, the extracted π from the Q-network is continuously changing. To keep the standard policies adaptive, as well as

π, the S network also saves changes to ensure the optimal accuracy. To improve user satisfaction in previous research, diversity was mentioned as an effective process for the recommendation. Similarly, it is an unintentional system to optimize user engagement. Based on the above-mentioned FeedRec framework, it is possible to optimize user engagement through various means in diversity immediately. To generate the simulation data, two types of lists are defined as user engagement and a list of recommendations.

1. Linear Style: In this process, the most satisfying results belong to a linear relationship with higher entropy. Based on this, the user can get more information and also use the system often. The probability of the user in using the system and searching for articles is defined as Equation (4).

$$p(Use|\phi_1, ..., \phi_n) = x\alpha(\phi_1, ..., \phi_n) + y, x > 0 \tag{4}$$

The article recommendation system is defined as $(\phi_1, ..., \phi_n)$, and the meaning of entropy is defined as $x\alpha(\phi_1, ..., \phi_n)$. x and y are used in the range of $0, 1$.

2. Quadratic Style: The highly user satisfaction made by moderate entropy. The probability of user in using a system and searching for articles is defined as Equation (5).

$$p(Use|\phi_1, ..., \phi_n) = exp(-\frac{(\alpha(\phi_1, ..., \phi_n - \mu)^2)}{\theta}) \tag{5}$$

The above evaluations show the relationship between the user and system agent. The output of this process shows that FeedRec contains the ability to fit various types of dispensation among the entropy of recommendation list and user engagement.

5. Prediction Result of Twitter and DBLP Platform

In this section, the development environment, prediction results and implementation process of the proposed recommendation system for e-learners are presented in detail.

5.1. Experimental Environment and Setup

The implementation of the proposed model structure and environment is presented in this section. Table 5 summarizes the experimental set up of the proposed model. All experiments and results of the system are carried out using Intel(R) Core(TM) i7-8700 CPU @3.20 GHz 3.19 GHz processor with 32 GB memory. The reinforcement learning technique used for the recommendation system. Similarly, the library and framework used in the proposed system is Jupyter notebook. The programming language used in the designing of this System is WinPython–3.6.2.

Table 5. System's components and specification.

Component	Description
Programming language	WinPython–3.6.2, IDE Jupyter Notebook
Operating system	Windows 10 64bit
Browser	Google Chrome, opera
GPU	Nvidia GForce 1080
Library and framework	Web Service
CPU	Intel(R) Core(TM) i7-8700 CPU @3.20 GHz
Memory	32 GB
Recommendation Modules	Reinforcement Learning
Optimization Algorithm	Model Free optimization

5.2. Performance Evaluation

The selected users are bachelors, masters and PhD students. The ranked tweets by the system are given to users for evaluation. For each query, nine tweets are given to the users—three out of nine belong to categories: context-based recommendations, profile-based recommendations,

and history-based recommendations. In this evaluation a form is provided to each user. The evaluation form consists of user's personal information, and a scenario, keyword and ranked tweet relevant to that keyword. For the evaluation form, the user reads the scenario and checks the keyword relevance with the tweets. The user reads the first tweet if this tweet is relevant to the given keyword then the user marks this tweet as relevant or otherwise irrelevant. The user does the same steps for all tweets. According to the result, the context has high weight over profile and history. The mentioned weight is calculated by using the following Equations (6)–(8).

$$Context = (Context/(Context + History + Profile)) \tag{6}$$

$$Profile = (Profile/(Context + History + Profile)) \tag{7}$$

$$History = (History/(Context + History + Profile)) \tag{8}$$

Performance evaluation describes the formal procedure to estimate the model performance results. To specify the movement of our model, we applied three statistical evaluation method listed as Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

- Mean Square Error This statistical evaluation measure the relationship between predicted value and actual value based on the mentioned Equation (9).

$$MAE = \frac{\sum_{n=1}^{m} |A_n - \hat{A}_n|}{m} \tag{9}$$

- Mean Absolute Error

  This statistical evaluation measures the square of differences between predicted value and actual value based on the mentioned Equation (10).

$$MSE = \frac{\sum_{n=1}^{m} (A_n - \hat{A}_n)^2}{m} \tag{10}$$

- Root Mean Square Error

  This statistical evaluation measure the error rate, error size based on the target value which mentioned in Equation (11).

$$RMSE = \sqrt{\frac{\sum_{n=1}^{m} (A_n - \hat{A}_n)^2}{m}} \tag{11}$$

5.3. Prediction Results

Prediction results contain the output of the experiments based on the above-mentioned machine learning regression algorithms. This process contains 10 top related values based on e-learner requests and activities. All the experiments and machine learning algorithms were implemented in winpython programming environment. Figure 10 presents the efficiency of the operated models.

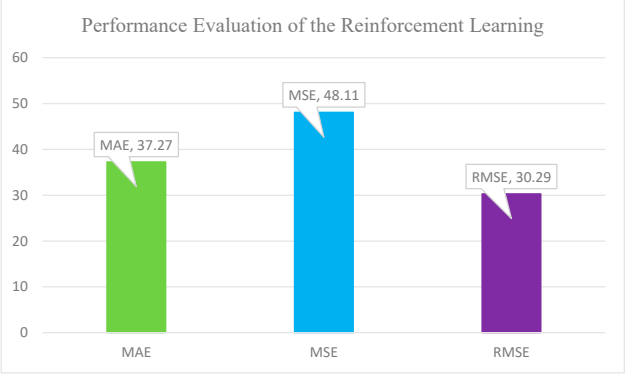


Figure 10. Performance evaluation of predictive R.L algorithm.

5.4. Recommendation Results

Based on the proposed process of the recommendation system on social media contents for e-learners, the reinforcement learning recommendation shows the system output on state, reward, loss and model frequency. Table 6 presents system response time detail information. Response time shows the output for three-timing information, containing loading time, searching time and execution time. Loading time is the time that it takes the user to load the web page. Searching time is the time it takes that user to search for the unique contents, and execution time is the time it takes to show the final search result.

Table 6. Response time of system.

Number	Loading T. (sec)	Searching T. (sec)	Execution T. (sec)
1	2.0883	0.0350	2.4505
2	0.5101	0.0348	0.5601
3	0.0012	0.0377	0.0401
4	1.7510	0.0627	1.8237
5	2.0883	0.0344	2.1331
6	2.0883	0.0344	2.4433
7	1.7510	0.0616	1.8226
8	1.7510	0.0358	1.8068
9	2.0883	0.0013	2.1017
10	1.7510	0.0400	1.8058

The main interface is shown in Figure 11. It gives two options “Sign in” and “Sign up”. The user selects the appropriate option as per their requirement. When a user selects the “Sign up” option, the user is redirected to the figure shown as (2), sign up. The registration process of users is shown in (2). Clicking on the “Sign up” option requires the user to fill in the personal information. Once the user selects the “Save” button, the form is sent to the server, and the user account has been made. The profile information of users is shown in (3). It gives multiple options “Edit”, “Choose File”, “View Profile”, “My Articles”, etc. The user selects the appropriate option as per their requirement. Selecting the “Edit” option allows them to update their profile information. Selecting the “My Articles” option shows the uploaded articles by the user shown in (4). When the user clicks on the “My Articles” button, it gives multiple options, such as, “View”, “DBLP”, “View Profile”, “My Articles”, “Upload Articles”, etc. The user selects the appropriate option based on their requirement. Selecting the “View” option redirects them to the same page where the user can read the selected article. When the user selects the “DBLP” option, it gives multiple options, such as, “Search (Button)”, “DBLP”, “View Profile”, “My

Articles", "Upload Articles", etc. The user selects the appropriate option as per their requirement. The user needs to fill the text box with the appropriate article name. When the user selects the "Search (Button)", they are redirected to the (9), where the user can see the recommendation from "DBLP". When the user selects the "Upload Articles" option, they are redirected to the (6), in which they can click on the "Upload Articles" button. It gives multiple options, such as, "Save (Button)", "Choose File", "DBLP", "View Profile", "My Articles", "Upload Articles", etc. In (6), the user needs to fill in the article information to upload articles in the system. Once the user selects the "Save" button, the form is sent to the server, and the articles are uploaded successfully. When the user selects the "Search Articles" option, then the system redirects to (5), where they can search for articles from the system. The articles against user query are shown in (8). It gives multiple options "View", "DBLP", "View Profile", "My Articles", "Upload Articles", etc. When the user selects "View", they are redirected to the same page where the user can read the selected article. The recommendation from DBLP is shown in (9).

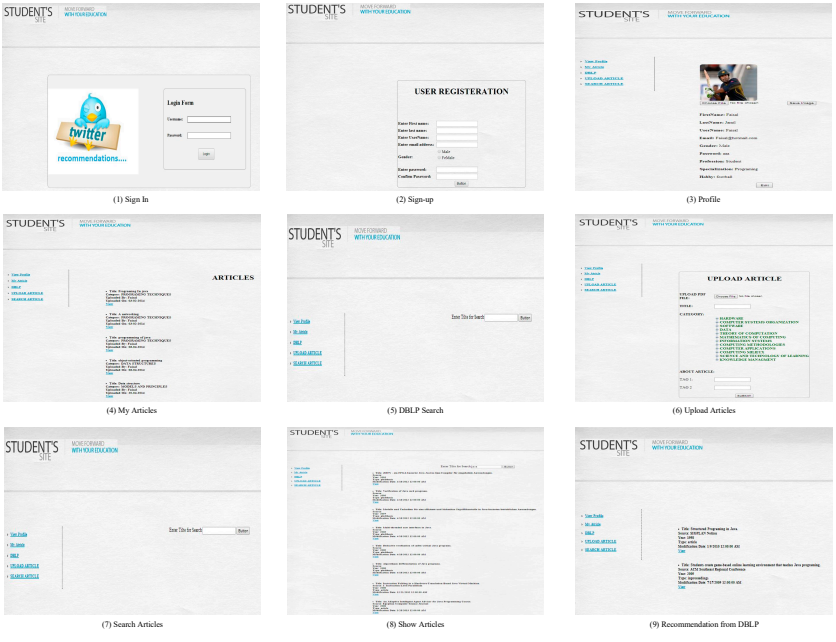


Figure 11. Social media content recommendation on web server.

Comparison and Baseline

Based on the proposed recommendation system, various algorithms compare together to show the system performance. Using the future reward (DDQN) increases the RL recommendation result above the (DN). Similarly, (DBGD) applied as an exploration system using $\in$-greedy to pass the system loss. Figure 12 shows the detail of system performance. In total, ten techniques are compared to get the system performance result. The applied techniques are defined as LR, FM, W&D, LinUCB, HLinUCB, DN, DDQN, DDQN+U, DDQN+U+EG and DDQN+U+DBGD. Based on the comparison, DDQN+U+DBGD has the highest score. Applying EG to DDQN+U, did not have much effect on improving the accuracy of the system.

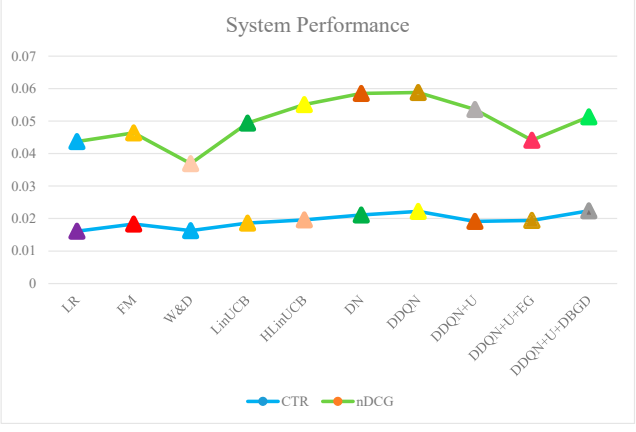


Figure 12. System performance.

Table 7 presents the diversity of user clicks that were measured by using cosine similarity. The smallest output represents better diversity. Similarly, some baseline methods—e.g., HLineUCB—achieve relatively equivalent recommendation diversity, which demonstrates UCB can get sensible result too.

Table 7. User click diversity results.

Technique	Recommendation Diversity
LR	0.2944
FM	0.3125
W&D	0.1758
LinUCB	0.3747
HLinUCB	0.2434
DN	0.2657
DDQN	0.2146
DDQN + U	0.2824
DDQN + U + EG	0.2118
DDQN + U + DBGD	0.2327

Figure 13 shows the preferences based on the system rewards. The presented reward is based on the recommended articles and total available articles.

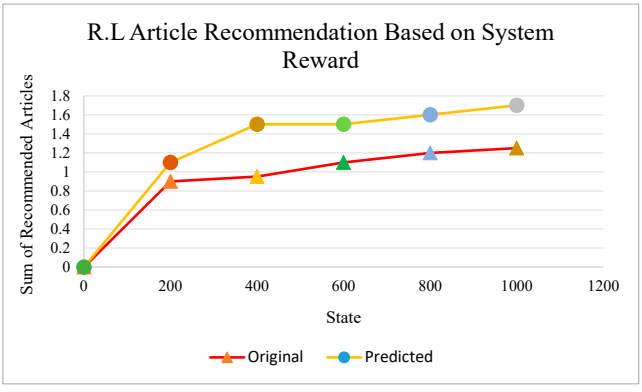


Figure 13. User preference based on the system reward.

Figure 1 shows the comparison of different studies related to our topic. In this figure, we compare the presented result with four recent research articles on the recommendation system, and it shows the proposed result that an F-measure output of 88% has a better consequence. The mentioned studies are proposed by Verma et al. [66], Zhang et al. [67], Hsieh et al. [68] and Liu et al. [69].

6. Conclusions and Future Work

In this paper, we present a reinforcement learning framework to customize online Twitter and DBLP article recommendation. The main differences between the proposed method and other methods are the efficient modeling of the articles, comments, user's feature, and also the design of explicitly reach a great reward. Based on the user clicks on URLs and user searching process, the system obtains more information from user feedback. Similarly, using the effective exploration strategy in this framework increases the recommendation diversity and also gets more reward recommendations. Experimental results suggest that the proposed system has higher accuracy for recommendation diversity and can distribute in other recommendation systems too. The system quality control and trust rely on user rating and feedback, which is the main concept of reinforcement learning. In the future, we are planning to develop the offline recommendation evaluation and generate other types of methods using the proposed framework.

Author Contributions: Data curation, Z.S.; Funding acquisition, Y.C.B.; Investigation, Z.S.; Methodology, Z.S.; Project administration, Y.C.B.; Supervision, Y.C.B. All authors have read and agreed to the published version of the manuscript.

Funding: Following are results of a study on the "Leaders in INdustry-university Cooperation+" Project, supported by the Ministry of Education and National Research Foundation of Korea

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rabiu, I.; Salim, N.; Da'u, A.; Osman, A. Recommender System Based on Temporal Models: A Systematic Review. *Appl. Sci.* **2020**, *10*, 2204. [CrossRef]
2. Pornwattanavichai, A.; Jirachanchaisiri, P.; Kitsupapaisan, J.; Maneeroj, S. Enhanced Tweet Hybrid Recommender System Using Unsupervised Topic Modeling and Matrix Factorization-Based Neural Network. In *Supervised and Unsupervised Learning for Data Science*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–143.
3. Yan, L.; Liu, Y. An Ensemble Prediction Model for Potential Student Recommendation Using Machine Learning. *Symmetry* **2020**, *12*, 728. [CrossRef]

4. Jun, H.J.; Kim, J.H.; Rhee, D.Y.; Chang, S.W. "SeoulHouse2Vec": An Embedding-Based Collaborative Filtering Housing Recommender System for Analyzing Housing Preference. *Sustainability* **2020**, *12*, 6964. [CrossRef]
5. Sánchez-Moreno, D.; López Batista, V.; Muñoz Vicente, M.D.; Sánchez Lázaro, Á.L.; Moreno-García, M.N. Exploiting the User Social Context to Address Neighborhood Bias in Collaborative Filtering Music Recommender Systems. *Information* **2020**, *11*, 439. [CrossRef]
6. Bai, Y.; Jia, S.; Wang, S.; Tan, B. Customer Loyalty Improves the Effectiveness of Recommender Systems Based on Complex Network. *Information* **2020**, *11*, 171. [CrossRef]
7. Jebur, A.A.; Atherton, W.; Al Khaddar, R.M.; Loffill, E. Settlement prediction of model piles embedded in sandy soil using the Levenberg–Marquardt (LM) training algorithm. *Geotech. Geol. Eng.* **2018**, *36*, 2893–2906. [CrossRef]
8. Luh, D.; Yang, T. Museum recommendation system based on lifestyles. In Proceedings of the 2008 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, Kunming, China, 22–25 November 2008; pp. 884–889.
9. Molnár, G. Challenges and opportunities in virtual and electronic learning environments. In Proceedings of the 2013 IEEE 11th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 26–28 September 2013; pp. 397–401.
10. Kim, J.; Wi, J.; Jang, S.; Kim, Y. Sequential Recommendations on Board-Game Platforms. *Symmetry* **2020**, *12*, 210. [CrossRef]
11. Cintia Ganesha Putri, D.; Leu, J.S.; Seda, P. Design of an Unsupervised Machine Learning-Based Movie Recommender System. *Symmetry* **2020**, *12*, 185. [CrossRef]
12. Tan, Z.; He, L. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle. *IEEE Access* **2017**, *5*, 27211–27228. [CrossRef]
13. Jamil, F.; Hang, L.; Kim, K.; Kim, D. A novel medical blockchain model for drug supply chain integrity management in a smart hospital. *Electronics* **2019**, *8*, 505. [CrossRef]
14. Jamil, F.; Iqbal, M.A.; Amin, R.; Kim, D. Adaptive thermal-aware routing protocol for wireless body area network. *Electronics* **2019**, *8*, 47. [CrossRef]
15. Jamil, F.; Ahmad, S.; Iqbal, N.; Kim, D.H. Towards a Remote Monitoring of Patient Vital Signs Based on IoT-Based Blockchain Integrity Management Platforms in Smart Hospitals. *Sensors* **2020**, *20*, 2195. [CrossRef] [PubMed]
16. Jamil, F.; Kim, D.H. Improving Accuracy of the Alpha–Beta Filter Algorithm Using an ANN-Based Learning Mechanism in Indoor Navigation System. *Sensors* **2019**, *19*, 3946. [CrossRef]
17. Jamil, F.; Iqbal, N.; Ahmad, S.; Kim, D.H. Toward Accurate Position Estimation Using Learning to Prediction Algorithm in Indoor Navigation. *Sensors* **2020**, *20*, 4410. [CrossRef]
18. Ahmad, S.; Jamil, F.; Khudoyberdiev, A.; Kim, D. Accident risk prediction and avoidance in intelligent semi-autonomous vehicles based on road safety data and driver biological behaviours. *J. Intell. Fuzzy Syst.* **2020**, *38*, 4591–4601. [CrossRef]
19. Jamil, F.; Kim, D. Payment Mechanism for Electronic Charging using Blockchain in Smart Vehicle. *Korea* **2019**, *30*, 31.
20. Shahbazi, Z.; Byun, Y.C. Towards a Secure Thermal-Energy Aware Routing Protocol in Wireless Body Area Network Based on Blockchain Technology. *Sensors* **2020**, *20*, 3604. [CrossRef]
21. Khan, P.W.; Byun, Y. A Blockchain-Based Secure Image Encryption Scheme for the Industrial Internet of Things. *Entropy* **2020**, *22*, 175.
22. Khan, P.W.; Byun, Y.C.; Park, N. IoT-Blockchain Enabled Optimized Provenance System for Food Industry 4.0 Using Advanced Deep Learning. *Sensors* **2020**, *20*, 2990. [CrossRef]
23. Shahbazi, Z.; Hazra, D.; Park, S.; Byun, Y.C. Toward Improving the Prediction Accuracy of Product Recommendation System Using Extreme Gradient Boosting and Encoding Approaches. *Symmetry* **2020**, *12*, 1566. [CrossRef]
24. Shahbazi, Z.; Byun, Y.C. Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. *Int. J. Adv. Sci. Technol.* **2019**, *29*, 6979–6988.
25. Hwang, S.Y.; Lai, C.Y.; Jiang, J.J.; Chang, S. The identification of noteworthy hotel reviews for hotel management. *Pac. Asia J. Assoc. Inf. Syst.* **2014**, *6*, 1. [CrossRef]

26. Jannach, D.; Gedikli, F.; Karakaya, Z.; Juwig, O. Recommending Hotels Based on Multi-Dimensional Customer Ratings. *ENTER. aau.at*. 2012; pp. 320–331. Available online: https://link.springer.com/chapter/10.1007/978-3-7091-1142-0_28 (accessed on 30 August 2020)
27. Ishtiaq, S.; Majeed, N.; Maqsood, M.; Javed, A. Improved scalable recommender system. *Nucleus* **2016**, *53*, 200–207.
28. Jazayeriy, H.; Mohammadi, S.; Shamshirband, S. A fast recommender system for cold user using categorized items. *Math. Comput. Appl.* **2018**, *23*, 1. [CrossRef]
29. Kanimozhi, K.S.M.L. Item Based Collaborative Filtering Approach for Big Data Application. *Semantic Scholar.* 2014. Available online: https://www.semanticscholar.org/paper/Item-based-Collaborative-filtering-approach-for-Big-Sudha-Lavanya/ffacdc02904cb34614a59c26645e031af32c4a28?p2df (accessed on 30 August 2020)
30. Linden, G.; Smith, B.; York, J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* **2003**, *7*, 76–80. [CrossRef]
31. Manu, M.; Ramesh, B. Single-criteria collaborative filter implementation using Apache Mahout in big data. *Int. J. Comput. Sci. Eng. Open Access* **2017**, *5*, 7–13.
32. Morozov, S.; Zhong, X. The evaluation of similarity metrics in collaborative filtering recommenders. In Proceedings of the Hawaii University International Conferences, Honolulu, HI, USA, 10–12 June 2013.
33. Shambour, Q.; Hourani, M.; Fraihat, S. An item-based multi-criteria collaborative filtering algorithm for personalized recommender systems. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 274–279. [CrossRef]
34. Gupta, V.; Hewett, R. Real-Time Tweet Analytics Using Hybrid Hashtags on Twitter Big Data Streams. *Information* **2020**, *11*, 341. [CrossRef]
35. Doulamis, A.; Voulodimos, A.; Protopapadakis, E.; Doulamis, N.; Makantasis, K. Automatic 3D Modeling and Reconstruction of Cultural Heritage Sites from Twitter Images. *Sustainability* **2020**, *12*, 4223. [CrossRef]
36. Resende de Mendonça, R.R.d.; Felix de Brito, D.F.d.; de Franco Rosa, F.d.F.; dos Reis, J.C.; Bonacin, R. A Framework for Detecting Intentions of Criminal Acts in Social Media: A Case Study on Twitter. *Information* **2020**, *11*, 154. [CrossRef]
37. Magdy, W.; Sajjad, H.; El-Ganainy, T.; Sebastiani, F. Distant supervision for tweet classification using youtube labels. In Proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015.
38. Jaeel, A.J.; Al-wared, A.I.; Ismail, Z.Z. Prediction of sustainable electricity generation in microbial fuel cell by neural network: Effect of anode angle with respect to flow direction. *J. Electroanal. Chem.* **2016**, *767*, 56–62. [CrossRef]
39. Nguyen-Truong, H.T.; Le, H.M. An implementation of the Levenberg–Marquardt algorithm for simultaneous-energy-gradient fitting using two-layer feed-forward neural networks. *Chem. Phys. Lett.* **2015**, *629*, 40–45. [CrossRef]
40. Ley, M. DBLP: Some lessons learned. *Proc. VLDB Endow.* **2009**, *2*, 1493–1500. [CrossRef]
41. Laender, A.H.; de Lucena, C.J.; Maldonado, J.C.; de Souza e Silva, E.; Ziviani, N. Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM SIGCSE Bull.* **2008**, *40*, 135–145. [CrossRef]
42. Tan, H.; Lu, Z.; Li, W. Neural network based reinforcement learning for real-time pushing on text stream. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan, 7–11 August 2017; pp. 913–916.
43. Zheng, G.; Zhang, F.; Zheng, Z.; Xiang, Y.; Yuan, N.J.; Xie, X.; Li, Z. DRN: A deep reinforcement learning framework for news recommendation. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 167–176.
44. Hu, Y.; Da, Q.; Zeng, A.; Yu, Y.; Xu, Y. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 368–377.
45. Zhao, X.; Xia, L.; Zhang, L.; Ding, Z.; Yin, D.; Tang, J. Deep reinforcement learning for page-wise recommendations. In Proceedings of the 12th ACM Conference on Recommender Systems, Vancouver, BC, Canada, 2–7 October 2018; pp. 95–103.

46. Zhao, X.; Zhang, L.; Ding, Z.; Xia, L.; Tang, J.; Yin, D. Recommendations with negative feedback via pairwise deep reinforcement learning. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1040–1048.
47. Chen, X.; Li, S.; Li, H.; Jiang, S.; Qi, Y.; Song, L. Generative adversarial user model for reinforcement learning based recommendation system. *arXiv* **2018**, arXiv:1812.10613.
48. Zhao, X.; Xia, L.; Tang, J.; Yin, D. " Deep reinforcement learning for search, recommendation, and online advertising: A survey" by Xiangyu Zhao, Long Xia, Jiliang Tang, and Dawei Yin with Martin Vesely as coordinator. *ACM SIGWEB Newsl.* **2019**, *4*, 1–15. [CrossRef]
49. Dulac-Arnold, G.; Evans, R.; van Hasselt, H.; Sunehag, P.; Lillicrap, T.; Hunt, J.; Mann, T.; Weber, T.; Degris, T.; Coppin, B. Deep reinforcement learning in large discrete action spaces. *arXiv* **2015**, arXiv:1512.07679.
50. Lu, Z.; Yang, Q. Partially observable Markov decision process for recommender systems. *arXiv* **2016**, arXiv:1608.07793.
51. Mahmood, T.; Ricci, F. Learning and adaptivity in interactive recommender systems. In Proceedings of the Ninth International Conference on Electronic Commerce, Minneapolis, MN, USA, 19–22 August 2007; pp. 75–84.
52. Rojanavasu, P.; Srinil, P.; Pinngern, O. New recommendation system using reinforcement learning. *Spec. Issue Intl. J. Comput. Internet Manag.* **2005**, *13*, pp. 23–28.
53. Shani, G.; Heckerman, D.; Brafman, R.I. An MDP-based recommender system. *J. Mach. Learn. Res.* **2005**, *6*, 1265–1295.
54. Taghipour, N.; Kardan, A.; Ghidary, S.S. Usage-based web recommendations: A reinforcement learning approach. In Proceedings of the 2007 ACM Conference on Recommender Systems, Minneapolis, MN, USA, 19–20 October 2007; pp. 113–120.
55. Zhao, L.; Liu, Z. A genetic algorithm for reinforcement learning. In Proceedings of the International Conference on Neural Networks (ICNN'96), Washington, DC, USA, 3–6 June 1996; Volume 2, pp. 1056–1060.
56. Alhijawi, B.; Kilani, Y. The recommender system: A survey. *Int. J. Adv. Intell. Paradig.* **2020**, *15*, 229–251. [CrossRef]
57. Ruotsalo, T.; Haav, K.; Stoyanov, A.; Roche, S.; Fani, E.; Deliai, R.; Mäkelä, E.; Kauppinen, T.; Hyvönen, E. Smartmuseum: A mobile recommender system for the Web of Data. *J. Web Semant.* **2013**, *20*, 50–67. [CrossRef]
58. Braunhofer, M.; Elahi, M.; Ricci, F. Usability assessment of a context-aware and personality-based mobile recommender system. In *International Conference on Electronic Commerce and Web Technologies, Proceedings of the EC-Web 2014: E-Commerce and Web Technologies, Munich, Germany, 1–4 September 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 77–88.
59. Elahi, M.; Braunhofer, M.; Ricci, F.; Tkalcic, M. Personality-based active learning for collaborative filtering recommender systems. In *Congress of the Italian Association for Artificial Intelligence, Proceedings of the AI*IA 2013: AI*IA 2013: Advances in Artificial Intelligence, Turin, Italy, 4–6 December 2013*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 360–371.
60. Ostuni, V.C.; Di Noia, T.; Di Sciascio, E.; Mirizzi, R. Top-n recommendations from implicit feedback leveraging linked open data. In Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, China, 12–16 October 2013; pp. 85–92.
61. Braunhofer, M.; Elahi, M.; Ge, M.; Ricci, F. Context dependent preference acquisition with personality-based active learning in mobile recommender systems. In *International Conference on Learning and Collaboration Technologies, Proceedings of the LCT 2014: Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Collaboration, Heraklion, Crete, Greece, 22–27 June 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 105–116.
62. Noguera, J.M.; Barranco, M.J.; Segura, R.J.; MartíNez, L. A mobile 3D-GIS hybrid recommender system for tourism. *Inf. Sci.* **2012**, *215*, 37–52. [CrossRef]
63. Bouneffouf, D.; Bouzeghoub, A.; Gançarski, A.L. A contextual-bandit algorithm for mobile context-aware recommender system. In *International Conference on Neural Information Processing, Proceedings of the ICONIP 2012: Neural Information Processing, Doha, Qatar, 12–15 November 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 324–331.

64. Ge, Y.; Xiong, H.; Tuzhilin, A.; Xiao, K.; Gruteser, M.; Pazzani, M. An energy-efficient mobile recommender system. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–28 July 2010; pp. 899–908.
65. Zou, L.; Xia, L.; Ding, Z.; Song, J.; Liu, W.; Yin, D. Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2810–2818.
66. Verma, A.; Virk, H. A hybrid genre-based recommender system for movies using genetic algorithm and knn approach. *Int. J. Innov. Eng. Technol.* **2015**, *5*, 48–55.
67. Zhang, J.; Peng, Q.; Sun, S.; Liu, C. Collaborative filtering recommendation algorithm based on user preference derived from item domain features. *Phys. A Stat. Mech. Appl.* **2014**, *396*, 66–76. [CrossRef]
68. Hsieh, M.Y.; Chou, W.K.; Li, K.C. Building a mobile movie recommendation service by user rating and APP usage with linked data on Hadoop. *Multimed. Tools Appl.* **2017**, *76*, 3383–3401. [CrossRef]
69. Liu, H.; He, J.; Wang, T.; Song, W.; Du, X. Combining user preferences and user opinions for accurate recommendation. *Electron. Commer. Res. Appl.* **2013**, *12*, 14–23. [CrossRef]

symmetry

MDPI

Article

User Behavior on Online Social Networks: Relationships among Social Activities and Satisfaction

Daniel Mican, Dan-Andrei Sitar-Tăut * and Ioana-Sorina Mihuţ

Faculty of Economics and Business Administration, Babeş-Bolyai University, 400591 Cluj-Napoca, Romania; daniel.mican@econ.ubbcluj.ro (D.M.); ioana.mihut@econ.ubbcluj.ro (I.-S.M.)
* Correspondence: dan.sitar@econ.ubbcluj.ro; Tel.: +40-(0)-264-418-652

Received: 8 September 2020; Accepted: 2 October 2020; Published: 10 October 2020

Abstract: Social networking sites (SNSs) are now ubiquitous communities for constant online interpersonal interactions that trigger symmetric or asymmetric effects on our everyday life. Recent studies advocate in favor of the significant role that SNSs have in promoting well-being and, more importantly, in disseminating reliable information during a global crisis, such as the current COVID-19 pandemic. Based on the growing importance of SNSs to the global framework, the main purpose of this study is to empirically assess the link between the use of symmetric social networks such as Facebook, or asymmetric social networks, like Instagram, and the level of satisfaction, employing the methodology of structural equation modeling. The results of the research validate the hypothesis that SNS activities increase the level of satisfaction, and therefore, that there is a direct link between the number of posts and comments and the level of satisfaction. Furthermore, based on the reversible and significant link between the level of satisfaction and the importance attributed to SNSs, the main conclusion of the study is that the higher the importance of the SNS, the greater the level of dissatisfaction experienced by users. Also, public activities on social networks positively affect social network satisfaction, while private activities have a direct negative relationship with the importance of social networks.

Keywords: social networks; behavior analysis; social behavior; social networking satisfaction

1. Introduction

Currently, economies around the world are being challenged by massive asymmetric shocks generated by the COVID-19 outbreak, even though its symmetric character is well recognized in the literature in the field [1]. Due to the high degree of existing interdependencies between countries, mediated through the flow of goods, services, technologies, people, and ideas, the negative effects of this critical public health emergency have spread at an unimaginable speed. The responses of various countries have been aligned with a multitude of information disseminated by institutions and governments, using both traditional methods of communication, as well as the new, revolutionary channel of social networks. Previous articles in this area have recognized the powerful role social media plays in sharing and obtaining information, especially during the COVID-19 pandemic [2]. Moreover, social networking sites (SNSs) have become extremely popular, not only during critical global situations but also more broadly, due to their general communication features that allow both online interpersonal interactions to occur, as well as providing the opportunity to engage in conversations with people who share the same interests. Many scientists from the fields of sociology, communication, informatics, and so on have been focused on online social network research [3]. In recent years, SNSs have expanded from computer sites to smart-phone applications, beginning to exert an even more profound influence on the social lives and economic activities of many

people [4]. The use of social networking sites has seen a significant increase, both among adolescents and adults. As indicated by Lenhart et al. [5], 73% of American teenagers use social networks, with an upward trend from year to year. According to Ilakkuvan et al. [6], young people spend an average of three hours per day on SNSs. There are several advantages to using social networks; most notably, the opportunity to share information with a large audience at a low cost [7]. Also, especially when it comes to the young generation, there is a strong need for acceptance and a powerful desire to be considered valuable and worthy by other members of the community; these urges can be satisfied through the use of SNSs [8].

With the rapid development of information networking technology, social networks such as Facebook, Twitter, and YouTube have emerged as an alternative to the traditional, face-to-face communication, especially in the case of young individuals, enriching and broadening people's social activities. Since the popularity of social networks is reaching a peak [9], the theories developed around this topic are becoming more and more complex [10]. Using matrix theory in social network analysis, Paul and Friginal [11] established a dichotomy concerning the two most widely used online platforms, namely Facebook and Twitter, i.e., considering Facebook as a symmetric social network and Twitter as an asymmetric one. This view was shared by the study of Conejero et al. [12], who consider Facebook and LinkedIn as symmetric social networks and Instagram and Twitter as asymmetric ones.

SNSs permit users to connect to each other by creating profiles through which they can provide a variety of personal information. Personal profiles can include a series of information and information types, including photos or videos. According to Kaplan and Haenlein [13], "the higher the social presence, the larger the social influence that the communication partners have on each other's behavior". Moreover, many organizations use SNSs to build communities for professional collaborations to share knowledge and learning materials among their employees [14]. SNSs represent nowadays a new innovative trend [13] whose framework is characterized by complexity, continuous transformation, and adjustment to world dynamics.

Understanding the components of the architecture of this system requires panoramic introspection. The "social" aspect invokes its existence in a social space that can be used for individual, professional, and/or entertainment purposes, and can be linked to the level of satisfaction an individual may experience. Furthermore, social networks have allowed users to be permanently connected, engage in content creation, or be updated with the most relevant information, without having to take a closer look at them. According to Kaplan and Haenlein [13], there are two key elements that social media embodies, namely, self-presentation and self-disclosure. The concept of self-presentation states that regardless of the type of social interaction, people manifest the tendency to seek to control the impressions others develop about them. In a study elaborated by Schlosser [15], it was acknowledged that to manage others' perception about themselves, individuals often present an edited version of their life on social media, that might, contrary to their expectations, reduce the level of enjoyment they actually experience in their daily life. This action is triggered by the desire to create an image that is consistent with personal identity. Usually, such a presentation is by self-disclosure, namely, the conscious or unconscious disclosure of personal information (for example, thoughts, feelings, pleasures, and dislikes). Self-disclosure is consistent with the image of themselves that people want to present and is an essential step in the development of close relationships.

In light of the increasing impact that social networking sites are having on our lives, especially among young people, it is important to acknowledge the fact that SNSs make a valuable contribution in shaping individuals' beliefs and to the construction of self-identity [16]. Taking these points into consideration, the main objective of the current study was to investigate the relationships among user online activities (looking at what friends have posted, posting different things, commenting on other users' photos or posts, receiving updates, chatting with others), usage intensity (number of friends or hours spent on SNSs) on specific social networks (Facebook, Instagram, Twitter, LinkedIn, and YouTube), importance, and level of satisfaction. First, the study extensively investigates all possible lower-level associations among items (correlation analysis). Second, several conceptual

links among compound factors (constructs) are examined via SEM analysis. Even though the influence of social media on individuals' behavior has captured the attention of researchers for many years [17], the novelty of the current study consists of its developing a relational model for a better assessment of users' social activities and their influence on the level of satisfaction users experience with SNSs.

The conceptual model includes constructs that reflect private activities on social networks, public activities on social networks, and the perceived importance of various social network sites. To our knowledge, this model has not been previously examined by other researchers in this respect. All correlations were studied both from professional (e.g., job/internship searching, or looking for instructive videos) and personal points of view (chatting, connecting with other people that share the same interests). Moreover, a relational model was developed to assess user social activities and their influence on the level of satisfaction experienced by SNS users.

The remainder of the paper is structured as follows: Section 2 offers a comprehensive presentation of the related literature in the field; Section 3 details the research methodology, providing a short description of the survey framework and research model, and continuing with the elaborated hypotheses, as well as describing the data collection and measurement process. Section 4 presents the data analysis and results, while Section 5 comprises a discussion and notes the main limitations of the present study. The final part of the paper presents the main conclusions.

2. Literature Review

According to the definition elaborated by Boyd and Ellison [18], social networking sites are web-based services where individuals may create different public or semipublic profiles in order to find and interact with other users. Since SNSs emerged, numerous studies have been published to provide a more in-depth understanding of user behavior. The research conducted by Shane-Simpson et al. [19] examined the social network framework by exploring questions such as who is attracted to social media sites, why they prefer a given site, and what are the social consequences of each site preference. After analyzing data collected from 663 students, Instagram was found to be the most frequently used platform among students, especially women. It was also found that most participants who preferred Twitter had both a public profile and a private one, and reported higher levels of self-disclosure. Participants who favor Facebook reported lower levels of self-disclosure, but a higher level of social ties.

Uses and gratifications theory (UGT) was put forward by Phua et al. [20], and the main SNS platforms like Facebook, Twitter, Instagram, and Snapchat were examined. The influence of several variables on the relationship between frequent use of each SNS and the bonding of social capital was analyzed. It was found that Twitter users had the largest bridging social capital, followed by Instagram, Facebook, and Snapchat users, while Snapchat users had the largest social-equity bonding, followed by Facebook, Instagram, and Twitter users. The effects of engaging in different SNS activities (broadcasting activities, directed communication including private messaging, commenting) by older users, compared to younger adults, were studied in an article by Kim and Shen [21]. The authors found that compared to younger adults, older adults benefit less from having a large network and more from engaging in directed communication activities.

Although the literature in the field describes a multitude of advantages associated with the use of social networking sites, some studies also emphasize the negative aspects these platforms may embody. SNSs are seen—especially by adolescents—as a universal solution for the vast majority of dilemmas. They are a continuous source of enjoyment, entertainment, social support, and well-being, even while humanity is facing a severe global health emergency threat, like the COVID-19 outbreak [22]. Moreover, it is important to note that SNSs give rise to negative behaviors. Paradoxically, SNSs are considered the point of origin for a wide variety of disorders like anxiety, loneliness, depression [23], and anger [24]. Therefore, there seems to be a growing number of social networking users who feel overwhelmed by the use of social networks [4], and it was found that social interaction overload and intrusion into work and private life were significant contributors to technology stress. Consequently, using a group of 180 college students, the study of Tafesse [25] confirmed the hypothesis according to which the use

of social networking sites negatively affects students' performance. Also, the perceived satisfaction with SNSs and technostress have been found to have a significant positive impact on rational use. The relationship between consumer socialization within SNSs, the consumer's need for uniqueness, and consumer satisfaction was examined by Abosag et al. [26]. Their findings showed that satisfaction with SNSs is enhanced by friend liking and is undermined by the need among users for uniqueness.

The dynamics of employee behavior and motives for using a variety of social media platforms, for personal and professional reasons, were explored by Lee [27]. The results demonstrated that employee communicative behaviors on social networking are unique, i.e., individual, interpersonal, and organizational-level factors collectively and jointly affect employees' positive and negative intentions. Also, social interaction overload negatively impacts performance perception, and the intrusion of private life affects performance and happiness. More and more studies similar to that of Moqbel and Kock [28] reveal that the degree of SNS dependence is growing and that this may have consequences on personal and work environments. SNS dependence reduces positive emotions that enhance performance and improve health. Also, it stimulates distraction, which is a performance inhibitor.

The relationship between social media and employee innovation via the mediation of organizational learning and knowledge sharing was studied by Khan and Khan [29]. Their results provided insights by demonstrating that knowledge sharing and social media can help facilitate employee innovation in the public sector.

3. Research Methodology

3.1. Research Model and Hypotheses Development

The research methodology applied in this paper is based on exploratory studies, with primary and secondary research being anchored on data collected through questionnaires, a literature review, and on specific statistical analyses and tests. The study investigates different aspects, e.g., how individuals are using SNSs for professional purposes vs. personal reasons. For each category, some examples were provided in the introduction section. Moreover, the correlations between the number of friends, hours spent, SNS usage and the level of satisfaction, importance, and SNS activities are assessed. Also, a relational model that evaluates social behavior, the impact of SNS activities on SNS importance and the level of satisfaction, is developed. A graphical representation of the model is illustrated in Figure 1. The logical assumption that drives the research hypotheses is that user activities on social networks influence the importance of, and satisfaction with, social networks. The popularity of SNSs and their use have increased in recent years, especially among young people [5]. Engagement with SNSs provides satisfaction, social support, and increased perceived importance [30]; however, exaggerated behavior in this direction or a selfish attitude [31] regarding peers (i.e., using chat instead of broadcasting opinions or just looking for "like" updates, having poor reactions or no reactions at all) may have the opposite effect on satisfaction [28] and can lead to negative mental outcomes [4,23,24]. Based on these considerations and an intensive literature review, the following hypotheses are put forward:

Hypothesis 1 (H1). *Private activities on social networks (PRASN) positively impact the importance of the social network (SNI).*

Hypothesis 2 (H2). *Public activities on social networks (PUASN) have a positive influence on the importance of social networks (SNI).*

Hypothesis 3 (H3). *Private activities on social networks (PRASN) are correlated inversely with satisfaction with social networks (SNSa).*

Hypothesis 4 (H4). *Public activities on social networks (PUASN) have a negative influence on satisfaction with social networks (SNSa).*

Hypothesis 5 (H5). *The importance of social networks (SNI) alters the degree of perceived satisfaction with social networks (SNSa).*

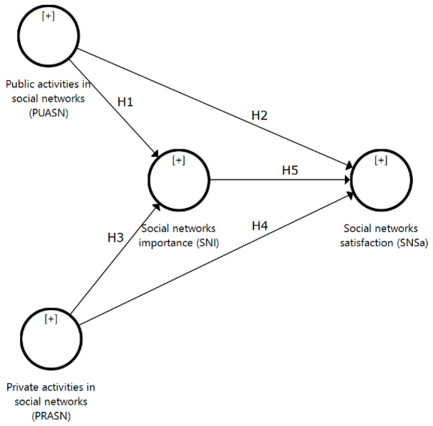


Figure 1. Research model depicting the influence of SNS activities on SN importance and satisfaction.

Private activities on social networks (PRASN) largely comprise two items (e.g., "I spend a lot of time receiving updates from liked pages"). Public activities on social networks (PUASN) were measured using three items (e.g., "I spend a lot of time commenting on other users' posts"). Perceived social networks importance (SNI) was measured using two items, taking into consideration professional and personal life (e.g., "Social networks are important for my professional life"). The overall perceived satisfaction with social networks (SNSa) was measured using two items, taking professional and personal life (e.g., "Social networks bring professional satisfaction") into consideration. The constructs used in the model, as well as the used items, are detailed in Table 1.

Table 1. Questionnaire (model) items, constructs, and descriptive statistics.

Latent Reflective Variable	Reflective Indicators	Description	Mean	Standard Deviation
Private activities on social networks (PRASN)	PRASN1	I spend a lot of time receiving updates from liked pages	3.117	1.168
	PRASN2	I spend a lot of time chatting with others	2.910	1.119
Public activities on social networks (PUASN)	PUASN1	I spend a lot of time looking at other users' posts	3.234	1.031
	PUASN2	I spend a lot of time posting things	3.622	0.987
	PUASN3	I spend a lot of time commenting on other users' posts	3.766	0.958
Social networks importance (SNI)	SNI1	Social networks are important for my professional life	3.505	1.106
	SNI2	Social networks are important for my personal life	3.559	1.096
Social networks satisfaction (SNSa)	SNSa1	Social networks bring professional satisfaction	1.973	0.716
	SNSa2	Social networks bring personal satisfaction	1.892	0.752

3.2. Data Collection and Measurements

In order to fulfill the main objectives of the current research, an anonymous online questionnaire was developed and disseminated via Google Forms in November 2018 to nontechnical students enrolled in the e-learning platform at one of the most prestigious faculties in Central and Eastern Europe. The questionnaire integrated a total of 26 items of various types, i.e., short answer questions, multiple-choice questions with one or n answers, or single or two-dimensional questions; all items were measured using a five-point Likert scale. To capture behaviors that could be extrapolated to a wider user community, we avoided respondents with technical profiles.

Using students as a sample group is a widely applied practice in academic research. Nonetheless, Sitar and Mican used managers and site owners [32]. Student sampling, also used in our research, is a reliable technique, as observed by Pollet and Saxton [33]. Since our potential subjects have different backgrounds in terms of family, environment, and previous exposure to SNSs, they were deemed to

be sufficiently diverse to provide generalized results. Acknowledging that a 100% response rate is unrealistic, a sample was used. The representativeness aspect was not neglected; thus, the minimum sample size was computed with Cochran's formula [34], based on a confidence interval of 5% and a confidence level of 95%. The simple random sampling method was used to ensure the same probability of being chosen during the entire sampling process for each individual [35]. Each student was chosen randomly, but that student could then choose to respond or not. Our survey pool consisted of 153 subjects enrolled in our courses, and the size of the computed representative sample was 111. The survey was distributed randomly and gradually to a subset from which 111 students ultimately agreed to fill in the questionnaire.

This sample size is above the minimal threshold described by Bonett and Wright [36] for correlation analyses and is satisfactory for PLS-SEM [37]. A total of 111 individuals responded to this questionnaire, all of whom were undergraduate students, aged 18–22. The gender distribution was as follows: 67.6% women and 32.4% men. They connect to the Internet through the following types of mobile devices: smartphone 83.8%, tablet 22.5%, notebook/ultrabook 11.7%, laptop 52.3%, and desktop 55.86%. The average reported weekly social networking time is 12.847 h, i.e., 110 min per day.

The descriptive statistics (mean and standard deviation which shows the spread of the values) of the items related to the constructs from the questionnaire included in the model are presented in Table 1. The rest of the variables included in our survey are illustrated in Tables 2–6.

4. Data Analysis and Results

The collected data were analyzed using the statistical software SPSS to measure the correlations that describe the relationship between the variables. Before performing the correlations, the normality Kolmogorov-Smirnov test was used with the quantitative data to decide whether parametric or nonparametric tests needed to be further applied. Because data series do not follow a normal distribution, Spearman's rho coefficients were computed. Also, ANOVA tests were considered to identify or reject behavior similarities among groups.

For the second analysis type, SmartPLS [38] was considered. This approach employs partial least squares structural equation modeling (PLS-SEM) and has become a quasi-standard method used in marketing and management research for analyses of the cause-effect relationships between latent constructs [39]. PLS-SEM is appropriate when the research goal is prediction oriented and/or when the research is exploratory in nature, i.e., researchers developing theories by doing multivariate analyses and focusing on explaining variance in the dependent variables [37]. PLS-SEM was suitable for the present research because the authors evaluated the relationships among several latent variables; this method combines the advantages of path analysis, factor analysis, and multiple regression analysis, and at the same time, helps researchers to examine relationships among variables in terms of explained variation. Due to the exploratory nature of the present study and the scarcity of previous research on the topic, and considering the predictive nature of the current research goals, PLS-SEM was considered a suitable and adequate data analysis method.

4.1. Correlation Analysis

The analysis of the correlations among the activities carried out on SNSs is presented in Table 2. There is a strong direct correlation between posting things and commenting on other users' content (e.g., photos, posts) (rho = 0.519, $p < 0.001$), and between looking at what friends have posted and commenting on other users' content (photos, posts), (rho = 0.509, $p < 0.001$). Both are highly significant. In the case of receiving updates from pages you like, there are three correlations, namely, a weak one with looking at what friends have posted (rho = 0.237, $p < 0.05$), another with posting things (rho = 0.379, $p < 0.001$), and a third with comments on other users' content (photos, posts), (rho = 0.292, $p < 0.001$). Between chatting with others and receiving updates from pages you like, there is a weak but significant link (rho = 0.324, $p < 0.001$). The ANOVA test did not confirm any significant gender differences regarding chatting or receiving updates. Additionally, the same analysis type indicated that

even if the younger users spend more time on social networks, no significant age or gender variances in this behavior were observed.

Table 2. Spearman's rho correlation between social networking activities.

	Private Activities		Public Activities		
Dimension	**Look**	**Chat**	**Updates**	**Post**	**Comment**
Look	1.000	0.146	0.237 *	0.465 ***	0.509 ***
Chat	0.146	1.000	0.324 ***	0.170	0.109
Updates	0.237 *	0.324 ***	1.000	0.379 ***	0.292 **
Post	0.465 ***	0.170	0.379 ***	1.000	0.519 ***
Comment	0.509 ***	0.109	0.292 **	0.519 ***	1.000

Note: Look = looking at what friends have posted; Chat = chatting with others; Updates = receiving updates from pages you like; Post = posting things; Comment = comments on other users'(photos, posts); Correlation is significant at the * = 0.05/** = 0.01/*** = 0.001 level (2-tailed).

The data regarding correlations between time spent on social media sites for professional and personal purposes are presented in Table 3. Even if, in general terms, there are no significant differences between the time allocated to private and professional activities, there is a slight distinction between each used SNS. Thus, there is a good correlation between the frequency of Facebook usage in professional activities and use in personal activities (rho = 0.501, $p < 0.001$). As for Instagram, there is a very high and highly significant correlation between the two usage types (rho = 0.786, $p < 0.001$). In the case of Twitter, we observed a strong link between the two usage types (rho = 0.647, $p < 0.001$), and a weak link between using Twitter for personal purposes and Instagram (rho = 0.200, $p < 0.05$) or LinkedIn for professional activities (rho = 0.295, $p < 0.05$). In the case of LinkedIn, there was a correlation between the two sides of LinkedIn use (rho = 0.755, $p < 0.001$) and a weak link between using LinkedIn for personal purposes and using Twitter for professional purposes (rho = 0.238, $p < 0.05$). For YouTube, there was a strong correlation between the frequency of use for the two domains (rho = 0.400, $p < 0.001$) and a weak one between using YouTube for personal use and LinkedIn at work (rho = 0.233, $p < 0.05$).

Table 3. The use of SNSs for professional vs. personal reasons.

Dimension	**FW**	**IW**	**TW**	**LW**	**YW**
FP	0.501 ***	−0.011	0.083	−0.011	0.082
IP	−0.066	0.786 ***	−0.177	−0.092	0.059
TP	0.122	−0.200*	0.647 ***	0.295 **	0.115
LP	−0.067	−0.020	0.238 *	0.755 ***	0.120
YP	0.004	−0.160	0.063	0.233 *	0.400 ***

Note: F = Facebook; I = Instagram; T = Twitter; L = LinkedIn; Y = YouTube; W = work/professional reasons; P = personal reasons; Correlation is significant at the * = 0.05/** = 0.01/*** = 0.001 level (2-tailed).

Table 4 illustrates the correlations between the number of current friends, the number of hours spent on SNSs, and the frequency of SNS use for professional and personal purposes. There are several weak and highly significant inverse correlations between the number of current friends and the frequency of Facebook use for professional purposes (rho = −0.325, $p < 0.001$), Facebook for personal purposes (rho = −0.249, $p < 0.01$), or Twitter for personal purposes (rho = −0.265, $p < 0.01$), and a weak and inverse relationship with the frequency of Twitter use for professional purposes (rho = −0.237, $p < 0.05$).

In terms of the number of hours spent on SNSs, three weak but significant inverse relationships were found with frequent use of LinkedIn for professional activities (rho = −0.249, $p < 0.01$), frequency of use LinkedIn for personal activities (rho = −0.195, $p < 0.05$), and the frequency of YouTube usage for personal purposes (rho = −0.203, $p < 0.05$).

Table 4. Correlation among the current number of friends, hours spent and SNS usage.

Dimension	FW	IW	TW	LW	YW	FP	IP	TP	LP	YP
FR	−0.325 ***	0.082	−0.237 *	−0.128	−0.059	−0.249 **	0.166	−0.265 **	−0.110	0.093
HR	−0.085	−0.004	−0.074	−0.249 **	−0.183	−0.138	0.007	−0.034	−0.195 *	−0.203 *

Note: FR = how many "friends" you currently have on SNSs; HR = hours spent on SNSs; F = Facebook; I = Instagram; T = Twitter; L = LinkedIn; Y = YouTube; W = work /professional reasons; P = personal reasons; Correlation is significant at the * = 0.05/** = 0.01/*** = 0.001 level (2-tailed).

The correlations between the current number of friends, the number of hours spent on SNSs, and satisfaction, importance, and SNS activities are presented in Table 5. Two weak and inverse correlations were observed between the current number of friends and comments on other users' pages (rho = −0.245, $p < 0.01$) and receiving updates from pages you like (rho = −0.198, $p < 0.05$). Regarding the number of hours spent on SNSs, there was a weak but significant link with chatting with others (rho = −0.246, $p < 0.01$).

Table 5. Correlation among the current number of friends, hours spent and satisfaction, importance, and SNS activities.

Dimension	SW	SP	IW	IP	Look	Updates	Post	Comment	Chat
FR	−0.115	−0.025	0.092	0.075	−0.162	−0.198 *	−0.067	−0.245 **	−0.088
HR	−0.018	−0.085	0.029	0.122	0.098	−0.066	0.108	−0.020	−0.246 **

Note: FR = how many "friends" you currently have on SNSs; HR = hours spent on SNSs; SW = SNSs satisfaction for professional reasons; SP = satisfaction for personal reasons; IW = SNSs importance for professional reasons; IP = SNSs importance for personal reasons; Correlation is significant at the * = 0.05/** = 0.01 level (2-tailed).

Observed correlations between professional and personal satisfaction and importance are presented in Table 6. Concerning satisfaction with SNSs for professional reasons, there were three correlations. The first one is a strong and highly significant correlation with satisfaction with SNSs for personal use (rho = 0.611, $p < 0.001$). The other two correlations were weak and negatively correlated with SNS importance for personal reasons (rho = −0.278, $p < 0.01$) and SNS importance for professional reasons (rho = −0.194, $p < 0.05$). In terms of satisfaction with SNSs for personal use, this was correlated weakly, but with high significance, with SNS importance for personal use (rho = −0.326, $p < 0.001$). We also noticed that SNS importance for professional use is strongly correlated with SNS importance for personal use (rho = 0.713, $p < 0.001$).

Table 6. Spearman's rho correlation between satisfaction, importance, and SNS activities.

					Private Activities			Public Activities	
Dimension	**SW**	**SP**	**IW**	**IP**	**Look**	**Chat**	**Updates**	**Post**	**Comment**
SW	1.000	0.611 ***	−0.194 *	−0.278 **	0.072	0.114	0.072	0.252 **	0.233 *
SP	0.611 ***	1.000	−0.149	−0.326 ***	0.109	0.066	0.109	0.259 **	0.258 **
IW	−0.194 *	−0.149	1.000	0.713 ***	−0.164	−0.242 *	−0.164	−0.127	0.031
IP	−0.278 **	−0.326 ***	0.713 ***	1.000	−0.186	−0.186	−0.186	−0.192 *	−0.225 *

Note: SW = SNSs satisfaction for professional reasons; SP = satisfaction for personal reasons; IW = SNSs importance for professional reasons; IP = SNSs importance for personal reasons; Correlation is significant at the * = 0.05/** = 0.01/*** = 0.001 level (2-tailed).

Regarding the mechanisms through which satisfaction with SNSs is influenced by SNS activities, the following correlations were established. Satisfaction with SNSs for professional reasons correlated poorly, but significantly, with posting things (rho = 0.252, $p < 0.01$) and comments to other users' (photos, posts) (rho = 0.233, $p < 0.05$). In the case of satisfaction with SNSs for personal use, the same weak and significant correlations with posting things (rho = 0.259, $p < 0.01$) and commenting on other users' material (photos, posts) (rho = 0.258, $p < 0.01$) was identified.

SNSs importance for professional reasons correlates poorly, inversely, and significantly with chatting with others (rho = −0.242, $p < 0.01$). Also, SNSs importance for personal reasons correlates

poorly, inversely, and significantly with posting things (rho = −0.192, $p < 0.05$) and comment on other users' (photos, posts) (rho = −0.225, $p < 0.05$).

4.2. Structural Equation Modeling Analysis

In order to reveal more regarding the immense complexity of the social network phenomenon and user behavior, a new approach is proposed. The degrees of importance and satisfaction with social networks were analyzed concerning activity types, i.e., public or private. To build the model and test the research hypotheses, the statistical software SmartPLS version 3.3.2 (SmartPLS GmbH, Boenningstedt, Germany) was used [38], which applies modeling using the partial structural equation with the smallest squares PLS-SEM, also called PLS path modeling. It requires two sets of validation procedures, i.e., for the outer and also for the inner/structural model.

4.2.1. Reflective Measurements Assessment

An assessment of the reflective measurements is presented in Table 7, which contains the model constructs with latent reflective variables and reflective indicators. The reliability and convergent validity results of the model can be observed by looking at the outer loadings, Cronbach's alpha, composite reliability, and average variance extracted (AVE). The outer loadings indicate the relationships between constructs and indicator variables in reflective measurement models and are computed for all measurement model constructs. For the outer loadings indicator, the values were higher than the minimum score of 0.70 [40], which infers adequate levels of indicator reliability. The indicators SNI2 (outer loading: 0.949) and SNSa2 (outer loading: 0.929) had the highest reliability, while the indicators ISN1 (outer loading: 0.712) and PUASN1 (outer loading: 0.703) had the lowest reliability.

Table 7. Convergent validity and internal consistency assessment of the reflective variables.

Latent Reflective Variable	Reflective Indicators	Outer Loadings	Cronbach's Alpha	Composite Reliability	Average Variance Extracted (AVE)
Private activities on social networks (PRASN)	PRASN1 PRASN2	0.747 0.875	0.498	0.796	0.662
Public activities on social networks (PUASN)	PUASN1 PUASN2 PUASN3	0.703 0.874 0.830	0.746	0.846	0.649
Social networks importance (SNI)	SNI1 SNI2	0.887 0.949	0.821	0.915	0.844
Social networks satisfaction (SNSa)	SNSa1 SNSa2	0.914 0.929	0.822	0.918	0.849

Regarding the internal consistency reliability, Cronbach's alpha was considered as the lower bound and composite reliability as the upper bound [41]. In this respect, almost all the values for the Cronbach's alpha coefficient were considerably above the recommended value of 0.70 [41]. The highest values were for SNSa (0.822) and SNI (0.821). On the other hand, for composite reliability, all the values could be considered adequate. In this case, the highest values were for the same constructs as in the case of Cronbach's alpha coefficient.

The average variance extracted (AVE) is a traditional measure to prove the convergent efficacy of the construct level. It is calculated as the grand mean value of the squared loadings of the indicators linked with the construct [42]. Regarding the convergent validity, the AVE was far above the minimum threshold of 0.5 [42], which indicates that convergent validity was confirmed for all the constructs. In this model, the highest values were for constructs SNSa (0.849) and SNI (0.844).

The heterotrait-monotrait ratio (HTMT) has been proven to be a reliable basis for the statistical discriminant validity test [43]. For the current case, as presented in Table 8, all HTMT values were lower than the conservative threshold value of 0.85 [43], and discriminant validity was confirmed among all pairs of constructs.

Table 8. Discriminant validity assessment for the reflective variables (HTMT criterion).

	PRASN	PUASN	SNI	SNSa
Private activities on social networks (PRASN)				
Public activities on social networks (PUASN)	0.5704			
Social networks importance (SNI)	0.4367	0.1983		
Social networks satisfaction (SNSa)	0.2092	0.3383	0.3283	

4.2.2. Collinearity Issues Assessment

To further evaluate the structural model, a collinearity assessment was needed among the values of all sets of predictor constructs (Inner VIF values) [40]. Table 9 shows the inner VIF values of all combinations of endogenous constructs and their corresponding exogenous constructs. As can be observed, all values were below the ideal conditions threshold value of 3 [40], which means that collinearity between the predictor constructs did not represent an issue in the present structural model.

Table 9. Collinearity assessment among the predictor constructs (Inner VIF values).

	PRASN	PUASN	SNI	SNSa
Private activities on social networks (PRASN)			11,325	12,055
Public activities on social networks (PUASN)			11,325	11,366
Social networks importance (SNI)				10,888
Social networks satisfaction (SNSa)				

4.2.3. Structural Model Relationships

To evaluate the structural model, the significance level of the relationships established between the constructs was analyzed. Figure 2 shows a graphic representation of the structural model, while Table 10 presents a summary of the results. We can see that 3 out of the 5 hypotheses that we developed were confirmed by this custom model. According to the *p*-values, H2 and H5 were supported with $p < 0.01$, and H3 with $p < 0.05$.

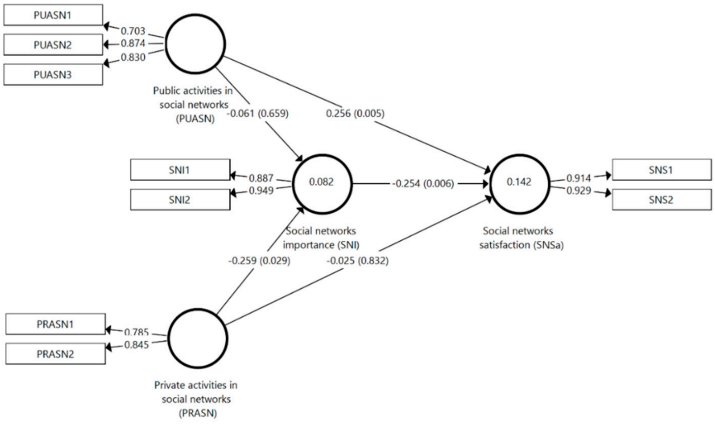


Figure 2. Graphic representation of structural model relationships.

Table 10. Direct and total effects.

	Effect	Deviation	T Statistics	*P* Values	Hypothesis
H1: PUASN → SNI	−0.0757	0.1382	0.4416	0.6590	Infirmed
H2: PUASN → SNSa	0.2612	0.0918	2.7897	0.0055	Confirmed
H3: PRASN → SNI	−0.2612	0.1186	2.1841	0.0294	Confirmed
H4: PRASN → SNSa	−0.0226	0.1161	0.2120	0.8322	Infirmed
H5: SNI → SNSa	−0.2577	0.0925	2.7465	0.0062	Confirmed

4.2.4. Predictive Relevance

To evaluate the predictive relevance [40] of the current model, R^2 and Q^2 values for the target variables were calculated and the PLSpredict algorithm was employed. The R^2 value for the final dependent variable in the model (SNSa) had a value of 0.142, which denotes a low predictive accuracy. Similarly, regarding the Q^2 value, the final dependent variable in the model (SNSa) had a value of 0.0824, which equally denotes a weak predictive power.

The PLSpredict algorithm [44] was employed to achieve a better assessment of the model's predictive power. The method uses training and holdout samples to generate and evaluate predictions from PLS path model estimations. As the data included in Table 11 reveal, all the Q^2 values were positive and the prediction error of the PLS-SEM results (RMSE and MAE) was smaller than that obtained by simply using the mean values; this indicates that this model offers better predictive performance. Taking into account the RMSE and MAE, in comparison with the LM results, the PLS-SEM results had lower prediction errors in terms of RMSE and MAE than the LM for all the items, which indicates that the model has high predictive power.

Table 11. Predictive power assessment using the PLSpredict procedure.

Indicator	PLS			LM		$RMSE_{PLS} < RMSE_{LM}$	$MAE_{PLS} < MAE_{LM}$	Predictive Power
	RMSE	MAE	Q^2_predict	RMSE	MAE			
SNSa1	0.7186	0.4985	0.0134	0.7310	0.5122	Yes	Yes	High
SNSa2	0.7494	0.5398	0.0262	0.7619	0.5601	Yes	Yes	

Note: PLS = prediction using PLS-SEM; LM = prediction using a linear model; RMSE = root mean squared error; MAE = mean absolute error.

5. Discussion and Limitations

5.1. Discussion

Existing studies by economists, psychologists, or sociologists agree on the significant role that the quality and quantity of social relationships have on the well-being of individuals [45]. Previous work confirmed the positive correlation between the density of social networks and the level of life satisfaction [46]. This link is explained by the fact that a larger number of social networks is positively correlated with a high degree of stability and security, and therefore, with a high level of life satisfaction [47]. As we experience a continuous evolution in the area of social networks, new forms of interconnections with other people are emerging day by day. The appearance of social media has substantially modified the range of criteria people use to evaluate their own well-being and life satisfaction [48]. Increasing the level of life satisfaction is an overall goal people want to achieve; such an undertaking reflects a subjective and global evaluation of the personal quality of life [49].

In this research, we investigated the link between the use of SNSs by students and the level of satisfaction they experience, both professionally and personally. The elaborated research model took into consideration the fact that SNS use by young people is increasing exponentially, day by day, and therefore, that it requires a further assessment of the channels of dissemination regarding the attributed importance or the satisfaction/dissatisfaction it may provide. The purpose of social networks is to extend the general framework of traditional communication channels. Many scholars

have expressed concern regarding the fact that online interaction techniques, such as the use of SNSs, are slowly becoming a substitute for face-to-face connections, and as a result, we may be confronted with a decline in individual social capital [18,50].

Moreover, it was argued that online interactions are less social than face-to-face ones; many studies in the field have confirmed this assumption. In particular, Kross et al. [51] observed that an increase in the number of hours spent on SNSs such as Facebook correlates with a decline in the level of life satisfaction over a certain period. These results were also supported by a study of Sabatini and Sarracino [52], which argued that the effect of SNSs, especially Facebook and Twitter, on individual well-being is negative. Interesting results are to be found in the work of Burke and Kraut [53], which suggests a positive link between online communication and well-being, as long as such communication includes people who are significantly important in the individual's life, or it incorporates characteristics which the individual considers to be suitable to their personality.

In the current research, the correlation coefficients denote positive significant relationships among all of the SNS activities depicted in Table 2, with the most powerful being between posting and commenting on other users' posts or photos. The behavior of students may also be transposed in the consumer community, where peers are strongly connected to a specific cause—like sharing food safety information—and trust each other [54]. Concerning these activities, we found there were no significant gender differences in terms of chatting or receiving updates, contrary to other studies that confirmed a higher involvement in this respect among males than females [55]. The time spent on social media sites for professional use is highly correlated with the related time spent for personal use on the same platform for the case of Instagram, followed by LinkedIn, and Facebook. The correlation among time spent on the same SNS but for different purposes is always positive; however, if both SNSs and purpose are different, inverse associations may be observed (e.g., Twitter for work and Instagram for private use). Overall social media intensity—i.e., number of friends and time spent on SNSs—is not quite a straight indicator for the time spent on a given SNS for professional/personal activities. In fact, our study indicates that this intensity correlates inversely with each of these times. The most prominent gap is between the number of friends on SNSs and the time spent on Twitter. Surprisingly, similar or inconclusive behavior was observed between social media intensity and satisfaction with SNSs, SNS importance, and activity type. There is an inverse correlation between the hours spent and chat activities, followed closely by number of friends and receiving updates from the liked pages. Going further, we noted a highly positive interdependence between the importance of SNSs (for professional and personal uses) and satisfaction. Surprisingly, importance was shown to correlate inversely with satisfaction, and possibly—although most were not significant—with all SNS activity types. Satisfaction was is positively impacted by public SNS activities (posting, commenting, and looking at posts) and indefinitely so by private ones (chatting and receiving updates from favorite pages). Since these results were not quite conclusive, they were studied further at a higher conceptual level by using a different technique.

After applying structural equation modeling, our results showed that public activities on social networks positively affect user satisfaction with social networks. Also, private activities on social networks have a direct negative relationship with the perceived importance of social networks. The perceived importance of social networks has a direct negative relationship with user satisfaction.

Socialization mediated through technology means that bidirectional social capital is exchanged between individuals and the network. Active public behavior on SNs contributes positively to increased perceived satisfaction (H2) since it is considered irrelevant for perceived importance (H1 infirmed). Selfish behavior on a SN, i.e., using it purely as a private communication channel (chatting) or to obtain only personal benefits (receiving updates from favorite pages), alters the perceived importance of the social network, since the obtained satisfaction is unquantifiable (H4 infirmed). Even if the social networks appear to be a panacea for both personal and professional concerns, their misuse may generate the opposite effect [23]. A higher level of perceived importance regarding a social network tends to yield lower perceived satisfaction (H5 confirmed). This relationship may look paradoxical,

but it was also confirmed by Yao and Cao [4]. The proposed model confirmed three relationships with similar strengths in terms of effect, but only one, i.e., between public activities and satisfaction, was positive.

Social networks offer social support, satisfaction, and a sense of well-being in day to day life, but also during word emergencies, like the current one created by the COVID-19 virus [56]. On the other hand, they may play a negative role by creating or reinforcing mental health problems or radical ideas [57]. These contradictory aspects reveal how vast, complex, and unpredictable this phenomenon is, and how far the literature has to go to completely understand it. Moderation, as in anything with addictive potential, may be the key to optimizing the positive-negative balance.

5.2. Limitations and Future Directions of Research

Despite the overall implications of this study, certain research limitations should be mentioned. First of all, even though the number of students that participated in the survey fulfills the minimum size requirements of both the analysis methods and the representativeness of the studied population, the sample could have been extended to students enrolled in other programs of study for more complex analysis. However, we must acknowledge that this sample size was severely reduced due to the GDPR constraints regarding the number of students enrolled in our e-learning system. Even if the student sampling method is widely used in research, our model and analyses may lack generalization. Region, gender, age, timing, profession are the main factors to be considered in larger samples relying on a multi-strata sampling method. Nevertheless, sophisticated models/analyses may fail during emergency crises, such as the COVID-19 lockdown, social distancing, or post-social distancing.

Considering the current crisis, one future direction would be to study and model the effects of the media during periods in which social distancing measures are implemented. Another would be to investigate more deeply how the complexity of the magnitude of social network use and social capital types facilitate transitional social support/satisfaction during lockdown for all involved parties [17].

6. Conclusions

Social networks are part of the everyday lives of people around the world; they are used for a multitude of purposes, considering the typology of the industry, from communication to informatics, and other fields [3]. As mentioned, the primary objective of this study was to assess, by using structural equation modeling, the link between SNS use and the level of satisfaction, in particular among young people. The obtained results indicate that the average amount of time students spend weekly on SNSs is 12.85 h, with no significant discrepancies between men and women. The dominant positions in terms of usage are attributed to Facebook and YouTube, with LinkedIn and Twitter being situated at the opposite end of the pole. Generally, there are no significant differences in terms of time spent for professional activities versus personal ones. The most popular and time-consuming activities that users carry out within SNSs are talking privately to others and tracking the news posted on their favorite pages, without revealing any significant differences between males and females.

The present study provides insights into social networking site use and the effects that this may have on an individual's life, in particular, in the case of students. As users spend more time looking at their friends' posts, they also spend more time posting things themselves and commenting on other users' content (photos, posts). Therefore, the current research also contributes to the existing literature by identifying correlations between these three activities. Similarly, as users spend more time receiving updates from pages they like, they spend more time chatting with others. SNS activities bring satisfaction, both in the professional and personal lives of most of the respondents. Of all the activities, posting things and commenting on other users' content (photos, posts) are statistically significant and correlated, i.e., the higher the number of posts and comments, the higher is the level of satisfaction, both professionally and personally. Similar results can be found in the work of Valenzuela et al. [58]. Paradoxically, following the analysis of the correlation between importance and satisfaction, it turned out that there is a reversible and significant link between these two. Therefore, if the importance

of SNSs increases, satisfaction decreases. In other words, the higher the importance, the more the discontent grows.

The model derived from structural equation modeling analysis confirms more intuitively a part of the correlation analysis findings. Public activities (posting things, commenting on other users' posts, and looking at posts) positively impact satisfaction, but the effect on the importance of SNSs is inconclusive. On the other hand, the output offered by the private activities (chatting with friends and receiving updates from favorite pages) is asymmetric; these variables have an indefinite effect on satisfaction and a negative one on importance. SEM analysis confirmed the negative effect that importance has upon satisfaction due to the overuse of social networks.

Nowadays, the young generation, which is permanently exposed to social media, has to distinguish between the positive and negative effects that this channel of communication may engender. Despite some mainstream opinions on this topic, the role and use of social networks is an endless subject of discussion and the literature is far from reaching a consensus.

Author Contributions: Conceptualization, D.M. and D.-A.S.-T.; methodology, D.M. and D.-A.S.-T.; validation, D.M.; formal analysis, D.M.; investigation, D.M., D.-A.S.-T., and I.-S.M.; resources, D.M., D.-A.S.-T., and I.-S.M.; data curation, D.M., D.-A.S.-T., and I.-S.M.; writing—Original draft preparation, D.M.; writing—Review and editing, D.M., D.-A.S.-T., and I.-S.M.; visualization, D.M.; supervision, D.-A.S.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Celi, G.; Guarascio, D.; Simonazzi, A. A fragile and divided European Union meets Covid-19: Further disintegration or 'Hamiltonian moment'? *J. Ind. Bus. Econ.* **2020**, *47*, 411–424. [CrossRef]
2. Pérez-Escoda, A.; Jiménez-Narros, C.; Perlado-Lamo-de-Espinosa, M.; Pedrero-Esteban, L.M. Social Networks' Engagement During the COVID-19 Pandemic in Spain: Health Media vs. Healthcare Professionals. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5261. [CrossRef] [PubMed]
3. Chen, B.; Chen, H.; Ning, D.; Zhu, M.; Ai, C.; Qiu, X.; Dai, W. A Two-Tier Partition Algorithm for the Optimization of the Large-Scale Simulation of Information Diffusion in Social Networks. *Symmetry* **2020**, *12*, 843. [CrossRef]
4. Yao, J.; Cao, X. The balancing mechanism of social networking overuse and rational usage. *Comput. Hum. Behav.* **2017**, *75*, 415–422. [CrossRef]
5. Lenhart, A.; Purcell, K.; Smith, A.; Zickuhr, K. Social Media & Mobile Internet Use among Teens and Young Adults. *Pew Internet Am. Life Proj.* **2010**. [CrossRef]
6. Ilakkuvan, V.; Johnson, A.; Villanti, A.C.; Evans, W.D.; Turner, M. Patterns of Social Media Use and Their Relationship to Health Risks Among Young Adults. *J. Adolesc. Health* **2019**, *64*, 158–164. [CrossRef]
7. Hruska, J.; Maresova, P. Use of Social Media Platforms among Adults in the United States—Behavior on Social Media. *Societies* **2020**, *10*, 27. [CrossRef]
8. Toma, C.L.; Hancock, J.T. Self-Affirmation Underlies Facebook Use. *Personal. Soc. Psychol. Bull.* **2013**, *39*, 321–331. [CrossRef]
9. Dhir, A.; Kaur, P.; Chen, S.; Lonka, K. Understanding online regret experience in Facebook use%Effects of brand participation, accessibility & problematic use. *Comput. Hum. Behav.* **2016**, *59*, 420–430. [CrossRef]
10. Liao, C.-H.; Chen, L.-X.; Yang, J.-C.; Yuan, S.-M. A Photo Post Recommendation System Based on Topic Model for Improving Facebook Fan Page Engagement. *Symmetry* **2020**, *12*, 1105. [CrossRef]
11. Paul, J.Z.; Friginal, E. The effects of symmetric and asymmetric social networks on second language communication. *Comput. Assist. Lang. Learn.* **2019**, *32*, 587–618. [CrossRef]
12. Conejero, J.; Sánchez-Figueroa, F.; Rodríguez-Echeverría, R.; Preciado, J. SCPL: A Social Cooperative Programming Language to Automate Cooperative Processes in (A)Symmetric Social Networks. *Symmetry* **2016**, *8*, 71. [CrossRef]
13. Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus. Horiz.* **2010**, *53*, 59–68. [CrossRef]

14. Kavianpour, S.; Tamimi, A.; Shanmugam, B. A privacy-preserving model to control social interaction behaviors in social network sites. *J. Inf. Secur. Appl.* **2019**, *49*, 102402. [CrossRef]
15. Schlosser, A.E. Self-disclosure versus self-presentation on social media. *Curr. Opin. Psychol.* **2020**, *31*, 1–6. [CrossRef]
16. Villanti, A.C.; Johnson, A.L.; Ilakkuvan, V.; Jacobs, M.A.; Graham, A.L.; Rath, J.M. Social media use and access to digital technology in US Young Adults in 2016. *J. Med. Internet Res.* **2017**, *19*, e196. [CrossRef]
17. Valkenburg, P.M.; Peter, J.; Walther, J.B. Media Effects: Theory and Research. *Annu. Rev. Psychol.* **2016**, *67*, 315–338. [CrossRef]
18. Boyd, D.M.; Ellison, N.B. Social network sites: Definition, history, and scholarship. *J. Comput. Commun.* **2007**, *13*, 210–230. [CrossRef]
19. Shane-Simpson, C.; Manago, A.; Gaggi, N.; Gillespie-Lynch, K. Why do college students prefer Facebook, Twitter, or Instagram? Site affordances, tensions between privacy and self-expression, and implications for social capital. *Comput. Hum. Behav.* **2018**, *86*, 276–288. [CrossRef]
20. Phua, J.; Jin, S.V.; Kim, J.J. Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of Facebook, Twitter, Instagram, and Snapchat. *Comput. Hum. Behav.* **2017**, *72*, 115–122. [CrossRef]
21. Kim, C.; Shen, C. Connecting activities on Social Network Sites and life satisfaction: A comparison of older and younger users. *Comput. Hum. Behav.* **2020**, *105*, 106222. [CrossRef]
22. Nabity-Grover, T.; Cheung, C.M.K.; Thatcher, J.B. Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media. *Int. J. Inf. Manag.* **2020**, 102188. [CrossRef] [PubMed]
23. Qi, M.; Zhou, S.J.; Guo, Z.C.; Zhang, L.G.; Min, H.J.; Li, X.M.; Chen, J.X. The Effect of Social Support on Mental Health in Chinese Adolescents During the Outbreak of COVID-19. *J. Adolesc. Health* **2020**, *67*, 514–518. [CrossRef]
24. Tang, C.S.; Koh, Y.Y.W. Online social networking addiction among college students in Singapore: Comorbidity with behavioral addiction and affective disorder. *Asian J. Psychiatr.* **2017**, *25*, 175–178. [CrossRef] [PubMed]
25. Tafesse, W. The effect of social networking site use on college students' academic performance: The mediating role of student engagement. *Educ. Inf. Technol.* **2020**. [CrossRef]
26. Abosag, I.; Ramadan, Z.B.; Baker, T.; Jin, Z. Customers' need for uniqueness theory versus brand congruence theory: The impact on satisfaction with social network sites. *J. Bus. Res.* **2020**, *117*, 862–872. [CrossRef]
27. Lee, Y. Motivations of employees' communicative behaviors on social media: Individual, interpersonal, and organizational factors. *Internet Res.* **2020**, *30*, 971–994. [CrossRef]
28. Moqbel, M.; Kock, N. Unveiling the dark side of social networking sites: Personal and work-related consequences of social networking site addiction. *Inf. Manag.* **2018**, *55*, 109–119. [CrossRef]
29. Khan, N.A.; Khan, A.N. What followers are saying about transformational leaders fostering employee innovation via organisational learning, knowledge sharing and social media use in public organisations? *Gov. Inf. Q.* **2019**, *36*, 101391. [CrossRef]
30. Verswijvel, K.; Heirman, W.; Hardies, K.; Walrave, M. Designing and validating the friendship quality on social network sites questionnaire. *Comput. Hum. Behav.* **2018**, *86*, 289–298. [CrossRef]
31. Abdul Malik, K.A.; Ahmad, A. The Effect of Use of Social Media on Prosocial Behavior. *Open J. Sci. Technol.* **2019**, *2*, 14–20. [CrossRef]
32. Sitar-Tăut, D.A.; Mican, D. MRS OZ: Managerial recommender system for electronic commerce based on Onicescu method and Zipf's law. *Inf. Technol. Manag.* **2020**, *21*, 131–143. [CrossRef]
33. Pollet, T.V.; Saxton, T.K. How Diverse Are the Samples Used in the Journals 'Evolution & Human Behavior' and 'Evolutionary Psychology'? *Evol. Psychol. Sci.* **2019**, *5*, 357–368. [CrossRef]
34. Cochran, W.G. *Sampling Techniques*, 3rd ed.; Wiley, Ed.; Wiley: Hoboken, NJ, USA, 1977; ISBN 978-0-471-16240-7.
35. Starnes, D.S.; Yates, D.; Moore, D.S. *The Practice of Statistics*, 4th ed.; W. H. Freeman: New York, NY, USA, 2010; ISBN 142924559X.
36. Bonett, D.G.; Wright, T.A. Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika* **2000**, *65*, 23–28. [CrossRef]
37. Hair, J.F.; Hult, G.T.M.; Ringle, C.M.; Sarstedt, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*; SAGE Publications: Thousand Oaks, CA, USA, 2016; ISBN 9781483377445.

38. Ringle, C.M.; Wende, S.; Becker, J.-M. SmartPLS 3. Available online: http://www.smartpls.com (accessed on 9 October 2020).
39. Hair, J.F.; Ringle, C.M.; Sarstedt, M. PLS-SEM: Indeed a Silver Bullet. *J. Mark. Theory Pract.* **2011**, *19*, 139–152. [CrossRef]
40. Hair, J.F.; Risher, J.J.; Sarstedt, M.; Ringle, C.M. When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **2019**, *31*, 2–24. [CrossRef]
41. Hair, J.F.; Anderson, R.E.; Tatham, R.L. *Multivariate Data Analysis with Readings*; Macmillan: New York, NY, USA, 1987; ISBN 0023489804.
42. Fornell, C.; Larcker, D.F. Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *J. Mark. Res.* **1981**, *18*, 39. [CrossRef]
43. Henseler, J.; Ringle, C.M.; Sarstedt, M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Mark. Sci.* **2015**, *43*, 115–135. [CrossRef]
44. Shmueli, G.; Ray, S.; Velasquez Estrada, J.M.; Chatla, S.B. The elephant in the room: Predictive performance of PLS models. *J. Bus. Res.* **2016**, *69*, 4552–4564. [CrossRef]
45. Zou, X.; Ingram, P.; Higgins, E.T. Social networks and life satisfaction: The interplay of network density and regulatory focus. *Motiv. Emot.* **2015**, *39*, 693–713. [CrossRef]
46. Reis, H.T.; Gable, S.L. Toward a Positive Psychology of Relationships. In *Flourishing: Positive Psychology and the Life Well-Lived*; American Psychological Association: Washington, DC, USA, 2004; pp. 129–159.
47. Baumeister, R.F.; Leary, M.R. The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychol. Bull.* **1995**, *117*, 497–529. [CrossRef]
48. Zhan, L.; Sun, Y.; Wang, N.; Zhang, X. Understanding the influence of social media on people's life satisfaction through two competing explanatory mechanisms. *Aslib J. Inf. Manag.* **2016**, *68*, 347–361. [CrossRef]
49. Diener, E.; Emmons, R.A.; Larsem, R.J.; Griffin, S. The Satisfaction with Life Scale. *J. Pers. Assess.* **1985**, *49*, 71–75. [CrossRef] [PubMed]
50. Arampatzi, E.; Burger, M.J.; Novik, N. Social Network Sites, Individual Social Capital and Happiness. *J. Happiness Stud.* **2018**, *19*, 99–122. [CrossRef]
51. Kross, E.; Verduyn, P.; Demiralp, E.; Park, J.; Lee, D.S.; Lin, N.; Shablack, H.; Jonides, J.; Ybarra, O. Facebook Use Predicts Declines in Subjective Well-Being in Young Adults. *PLoS ONE* **2013**, *8*, e69841. [CrossRef] [PubMed]
52. Sabatini, F.; Sarracino, F. Online Networks and Subjective Well-Being. *Kyklos* **2017**, *70*, 456–480. [CrossRef]
53. Burke, M.; Kraut, R.E. The Relationship Between Facebook Use and Well-Being Depends on Communication Type and Tie Strength. *J. Comput. Commun.* **2016**, *21*, 265–281. [CrossRef]
54. Seo, S.; Almanza, B.; Miao, L.; Behnke, C. The Effect of Social Media Comments on Consumers' Responses to Food Safety Information. *J. Foodserv. Bus. Res.* **2015**, *18*, 111–131. [CrossRef]
55. Alnjadat, R.; Hmaidi, M.M.; Samha, T.E.; Kilani, M.M.; Hasswan, A.M. Gender variations in social media usage and academic performance among the students of University of Sharjah. *J. Taibah Univ. Med. Sci.* **2019**, *14*, 390–394. [CrossRef]
56. Sahni, H.; Sharma, H. Role of social media during the COVID-19 pandemic: Beneficial, destructive, or reconstructive? *Int. J. Acad. Med.* **2020**, *6*, 70–75.
57. Bettmann, J.E.; Anstadt, G.; Casselman, B.; Ganesh, K. Young Adult Depression and Anxiety Linked to Social Media Use: Assessment and Treatment. *Clin. Soc. Work J.* **2020**. [CrossRef]
58. Valenzuela, S.; Park, N.; Kee, K.F. Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation. *J. Comput. Commun.* **2009**, *14*, 875–901. [CrossRef]

symmetry

MDPI

Article

Applied Identification of Industry Data Science Using an Advanced Multi-Componential Discretization Model

You-Shyang Chen [1,*], Arun Kumar Sangaiah [2], Su-Fen Chen [3,*] and Hsiu-Chen Huang [1]

[1] Department of Information Management, Hwa Hsia University of Technology, New Taipei City 23568, Taiwan; l0206113@go.hwh.edu.tw
[2] School of Computing Science and Engineering, VIT University, Vellore 632014, Tamil Nadu, India; sarunkumar@vit.ac.in
[3] National Museum of Marine Science & Technology, Keelung City 202010, Taiwan
* Correspondence: ys_chen@cc.hwh.edu.tw (Y.-S.C.); 10201a@webmail.nou.edu.tw (S.-F.C.)

Received: 11 August 2020; Accepted: 16 September 2020; Published: 30 September 2020

Abstract: Applied human large-scale data are collected from heterogeneous science or industry databases for the purposes of achieving data utilization in complex application environments, such as in financial applications. This has posed great opportunities and challenges to all kinds of scientific data researchers. Thus, finding an intelligent hybrid model that solves financial application problems of the stock market is an important issue for financial analysts. In practice, classification applications that focus on the earnings per share (EPS) with financial ratios from an industry database often demonstrate that the data meet the abovementioned standards and have particularly high application value. This study proposes several advanced multicomponential discretization models, named Models A–E, where each model identifies and presents a positive/negative diagnosis based on the experiences of the latest financial statements from six different industries. The varied components of the model test performance measurements comparatively by using data-preprocessing, data-discretization, feature-selection, two data split methods, machine learning, rule-based decision tree knowledge, time-lag effects, different times of running experiments, and two different class types. The experimental dataset had 24 condition features and a decision feature EPS that was used to classify the data into two and three classes for comparison. Empirically, the analytical results of this study showed that three main determinants were identified: total asset growth rate, operating income per share, and times interest earned. The core components of the following techniques are as follows: data-discretization and feature-selection, with some noted classifiers that had significantly better accuracy. Total solution results demonstrated the following key points: (1) The highest accuracy, 92.46%, occurred in Model C from the use of decision tree learning with a percentage-split method for two classes in one run; (2) the highest accuracy mean, 91.44%, occurred in Models D and E from the use of naïve Bayes learning for cross-validation and percentage-split methods for each class for 10 runs; (3) the highest average accuracy mean, 87.53%, occurred in Models D and E with a cross-validation method for each class; (4) the highest accuracy, 92.46%, occurred in Model C from the use of decision tree learning-C4.5 with the percentage-split method and no time-lag for each class. This study concludes that its contribution is regarded as managerial implication and technical direction for practical finance in which a multicomponential discretization model has limited use and is rarely seen as applied by scientific industry data due to various restrictions.

Keywords: data-mining techniques; data-discretization methods; feature-selection methods; industry data applications; advanced multicomponential discretization models

1. Introduction

Stock investments may earn profits but are often associated with inherent risks. That is, risks in terms of a stock are a part of financial investments. Stock investors typically face specific risks, such as systematic risk or undiversifiable risk (such as economic risks or inflation), unsystematic risk or idiosyncratic risk (such as market value risk, interest rate risk, and commodity risk), and others (e.g., opportunity risk and liquidity risk) in dealing with the major problems caused by economic and noneconomic factors. Thus, learning about the risk of making appropriate decisions to meet financial goals is an ongoing and important issue. In particular, escaping risks caused by economic uncertainty for stock investments has the following characteristics: dynamism and cyclicality, both of which form a focal and widely studied topic as they make a significant impact on the profitability of financial investments to the stock investors. Financial investments have attracted much attention because investors invest in the stock market, which results in various researchers using data to explore money markets [1] by achieving good data utilization in complex financial industry environments, such as through the use of smart/intelligent models or techniques. Therefore, this study is focused on financial applications with big data solutions that trigger advanced and intelligent componential models. These have posed great opportunities and challenges for data researchers because applied large-scale data are collected from diversified database services in a mixed-industry setting.

Within big data technology communities, large-scale data are collected from heterogeneous network services and integrated for good data utilization that addresses real-world problems. Technological advancement in financial management has been affected by the big data paradigm, such as the application of data mining to manage stock investments. Stocks with distinct characteristics in major markets have been a popular tool for analysis for investment purposes in Taiwan. Based on the data from the Taiwan Stock Exchange (TWSE), up to August 2015, the accumulated number of accounts opened was 17,771,994, and the number of accumulated investors with trading accounts was 9,715,629, while the population of Taiwan was just 23.51 million. Obviously, this shows active and busy market trends. Stock selection is a real-world problem because there are over 880 listed companies. It is a complex and risky issue to choose TWSE stocks. The measures used some main components that were unsystematic and had other risks involved in investing in stocks, such as commodity risk, currency risk, equity risk, headline risk, interest rate risk, obsolescence risk, or legislative risk, among which commodity risk is specific to the risk in connection with fluctuating commodity prices, such as copper, gold, or oil. Commodity risk is highly correlated with natural hazards; this allows it to be analyzed and managed with intelligent information from data mining [2,3]. Some advances in data mining technology (such as by Wang and Miao [4]) applied in stock analysis can definitely assist in escaping risk and achieving better investment performance. Thus, effective analytical tools or techniques are a priority to achieve the goals stated above. The principles of stock analysis have the following routes: technical analysis and fundamental analysis, which are used to prevent investors from searching for suitable research material on the stock market and are employed objectively and scientifically in modeling classification activities and functions for this study, as follows.

(1) Technical analysis: It is the knowledge and steps used to assess stocks to predict future trends by measuring the statistics collected from stock trading activities, such as price or volume changes. Technical analysts indicate that the past trading activity of a stock, in terms of price and volume changes, is a good sign for possible future price trends. First, market prices may offer rebates for various factors that affect the stock price. Second, market price trends or movements are not random but specific trends or movements of an identified pattern. Over the past decade, commonly used indicators include price trend indices, such as varied types of the moving average (MA) [5] or momentum indices, such as moving average convergence and divergence (MACD) [6]. MA has an average of closing prices of a specific number of time periods, such as a 7-period MA for seven days, which defines a flowing correlation between the time period and the money rate [5]. MACD uses exponential moving averages (EMAs) to calculate its value, equal to the value of the shorter time of EMAs and less than the value of the longer time of EMAs [7]. Furthermore, many varied and diverse perspectives for MA and MACD

were studied for these purposes, such as EMA, simple moving average (SMA), or weighted moving average (WMA), and variety of MACD-1–4 [7]. These are suitable for stock predictions, as shown in recent studies that demonstrated good performance and were involved in risk aversion.

(2) Fundamental analysis: It is the knowledge and steps of assessing a stock to estimate its intrinsic value by way of measuring related financial, economic, and other quantitative or qualitative attributes. Fundamental analysts study something that may affect stock value, such as the overall industry and economic qualifications of macroeconomic conditions. Its main purpose is highlighted to calculate a quantifiable value for an investor to measure its price at a specific time and to evaluate whether that stock is over- or undervalued. The fundamental analysis utilizes earnings, future growth, profit margins, return on equity, and revenues, among others, as evaluation instruments for stocks and equities because they identify a company's underlying value and future potential for growth. In particular, fundamental analysis reads a company's financial statements in terms of the company as a stock. A company's value is measured by its ability to realize cash flow under uncertainty; at the same time, it is the measure of modern finance to assess asset value, which is defined as being equal to the present value of the expected future cash flow since it is always discounted at the indispensable return. Three variables for its valuation models, including discounted cash flow models (DCFMs), dividend discount models (DDMs), and models depend on multiples (MDMs), are defined. First, DDMs make an elementary assumption, i.e., the value of a stock is decided by rebating the anticipated dividend as future cash. Therefore, the real value of a stock is decided by the present value of the stock cash dividend, which will be received by the shareholder. However, if the company chooses to reduce or stop dividends, the formula of DDMs becomes unworkable. Second, DCFMs were developed as an alternative. In DCFMs, analysts calculate the free cash flow of the company, consider the tax, depreciation, and amortization, change in working capital and capital expenditure, and discount the terminal value and free cash flow to gain the intrinsic value of a company. Third, compared to DDMs and DCFMs, more commonly used are the multiples that determine stock value. Such multiplier models (i.e., MDMs) assume that the company's value is several times the earnings, sales, cash flow, or book value. For example, referring to the earning multiplier model, the stock value is equal to multiple times the earnings. The information implies that the stock price is exactly several times its earnings per share (EPS) [8]. EPS is served as a financial indicator for the portion of earnings per outstanding share of allocated common stock for a company's profit over a fixed period of time [9,10]. The same logic applies to book value per share, cash flow per share, and sales per share to form variation of multiplier models. Among these models, company earnings are the most commonly used practical information. Estimating EPS is particularly crucial to analysts of fundamental analysis for complicated financial data. Moreover, as the market is developing and becoming more efficient, fundamental analysis is receiving more attention in the long-term. Given the above reasons, the EPS of fundamental analysis is emphasized in this study due to its preeminent advances and rich features.

Fundamental analysis is used to analyze data from financial reports, which assumes that the market price of a stock is centered on the company's intrinsic value. With further regard to financial reports, financial statements [11,12] should be addressed preferentially and contiguously to reflect a company's performance, such as an income statement, balance sheet, and a statement of cash flow. This study proposes advanced multicomponential discretization models that will analyze and mine data about financial ratios by predicting EPS. High EPS for fundamental analysis is the core selection used by investors to buy or hold a stock for a longer period of time. Classification applications for EPS from data from industry databases are of high practical value. Finding an intelligent hybrid classification model that develops out of financial ratio applications is considered an important issue for financial analysts. Many advanced componential data mining techniques used in this study are focused on some effective tools, such as data-discretization, feature-selection, and classification methods. Their performance is assessed and verified by looking at the application of financial ratios versus motivations by proposing an advanced multitechnological use for data processing analytics.

The main purpose of this study is as follows: (a) To build hybrid classification models using an advanced computing framework to diagnose and classify EPS from views of fundamental analysis and to evaluate their effectiveness; (b) to use the paths of literature review and relevant expert knowledge to select the essential financial ratios in advance; (c) to measure various multicomponent discretization performances of the applied hybrid models proposed; (d) to evaluate various classifiers' performance arising from big data recorded from the stock market to establish the identification of a specific target picture; the financial diagnosis application of comparative studies uses classifiers of advanced noted machine learning (ML) computing, such as the K-nearest neighbor (KNN) algorithm, decision tree (DT) learning, ensemble learning of stacking (STK), radial basis function network (RBFN), and naïve Bayes (NB) due to their past superior performance [11]; (e) to identify effective financial ratios useful to address EPS issues; (f) to discretize the selected condition attributes into corresponded linguistic terms of natural language; (g) to generate the knowledge of comprehensible decision tree-based rules extracted from the given datasets; (h) to identify suitable alternative work available for further research and use a difference-oriented exploration from the rich information.

As a whole, Section 2 introduces the literature to review relevant financial ratio issues. Section 3 presents the main framework of the study methods and materials used. Section 4 describes the data experiments and data analyses executed from the collected dataset, and Section 5 forms a complete discussion. Finally, Section 6 makes some conclusions and talks about future research.

2. Literature Review

The section discusses relevant issues for the identification of financial EPS diagnosis as well as some ML computing approaches, such as the applied financial ratio from scientific data, decision tree learning, the K-nearest neighbor algorithm, applied ensemble learning of stacking, the radial basis function network, and applied naïve Bayes.

2.1. Applied Financial Ratio from Scientific Data

From the perspective of fundamental analysis, financial ratios [13–16] are helpful to better understand the overall information of a specific company and its operating performance for various kinds of benefits. They are a premotivator to determine the company's financial condition from financial statements, and, in particular, they are used for comparisons for stakeholders, such as a company's owner and current and potential investors, in various object-oriented interpretations. Regarding the purposes of comparisons and analyses, financial ratios are measured between similar or competitive countries, industries, and companies, within different time ranges for one company or within a company with an industry-average and with benchmarked objects. Examples of financial ratios are often categorized into five specific groups, including debt ratios, market value ratios, profitability ratios, solvency ratios, and turnover ratios [14], and in most instances, quantify different aspects of a specific company from static and dynamic viewpoints. Mainly, (1) debt ratios are the measurement of the efficiency to refund previous long-term debt and the availability of using cash to return debt; (2) market value ratios value the investment return for shareholders and concern the response to return and the cost of issuing stock; (3) profitability ratios are used to assess the efficiency of using assets to manage operation performances for gaining an acceptable rate of earnings during a specific time period; (4) solvency ratios used are to evaluate the liquidity for uncashed assets converted into cash assets; (5) turnover ratios are used to measure operation ability by using owner assets to produce operating sales, and, in particular, they are valuable mediators in illustrating the operation performance of a company. Values of financial ratios used are calculated from various financial statements, such as the statements of balance sheet, income, cash flows, and the changes in equity [17]. As for most of these ratios, the value of the ratios improves with the performance that the company has compared with its competitors' ratio or a similar ratio with a specific time interval. Thus, financial ratios are of great interest to potential shareholders of a company, creditors, and financial managers to compare the SWOT (strengths, weaknesses, opportunities, and threats) analyses of various companies.

In practice, financial ratios are associated with predictive factors to EPS from the stock market. A study highlighted 20 key financial ratios that are known to be more relevant with EPS because they are considered to be highly correlated by financial experts and have had a limited literature review in earlier studies [9,11,12]. They are the accounts receivable turnover ratio, cash flow of each share, cash flow ratio, cash ratio, current liabilities, current ratio, debt ratio, fixed asset turnover ratio, interest expense, inventory turnover ratio, net asset value of each share, net worth turnover ratio, operating income per share, operating income margin, quick ratio, return on net asset, sales per share, times interest earned, total asset turnover ratio, and year-on-year percentage total assets, and they are grouped into five financial categories.

2.2. Decision Tree Learning

As to ML algorithms, DT learning [18] is a prominent classification method commonly used in data-mining domains that are specifically for multicriteria decision analysis, operations research in decision analysis, operations management, resource costs, and utility functions, which are faced in practical problems, and its main intent is to frame a tree-structured graph representation that predicts the value of a specific attribute from some input attributes. This tree-structured model, in which the specific attribute creates a possibly finite set of target values, is made as a classification decision-tree. The tree-like flowchart model in decision analysis involves three types of internal nodes that denote a test on a variable completely, branches that refer to the outcome of this test, and leaves that denote a decision-labeled class. The routes for root to leaf are linearized as decision-rules of classification, where the outcome is the contents of the leaf nodes. The decisional rules are generalized and formulated as a meaningful type of a comprehensive set like "IF condition 1 and condition 2 and condition 3 ... THEN outcome 1". Algorithms used for constructing DTs [19] usually work top-down in a flowchart-like construction by picking an attribute from each step to best break up the set of decision rules to help identify a better choice that is most likely to achieve a specific goal.

DTs have five major advantages, including (1) they are simple and easy to comprehend and definable in a concise explanation based on DT rules; (2) they use a white box model; (3) they are helpful to determine new possible scenarios; (4) they give values with little hard data; (5) they easily combine with other algorithms in ML techniques. Importantly, C4.5 [20] is used as a core algorithm to create a classification decision rule. It is a popular classifier with a better frequentation within numerous data-mining techniques than the other four DT algorithms, including C5.0, CART (classification and regression trees), CHAID (chi-square automatic interaction detector), and ID3 (iterative dichotomizer 3) [20].

2.3. K-Nearest Neighbors Algorithm

Regarding the field of pattern recognition for ML search processing, the instance where the specific class is projected to be a target class of the closest to training examples is made up as the nearest-neighbor algorithm; for deep definition, a K-nearest neighbor (abbreviated as KNN) algorithm [21] uses a nonparametric approach for important regression analysis and interested classification works. Within the classification stage, K is defined as a constant by the user, and an unlabeled vector is targeted by allocating the label, which has the most frequent use among these K training examples and is nearest to the query point. That is, an objective target is labeled by voting on the plurality of its neighbors and being allocated the target object to the most common class among the KNNs [22,23]. However, the majority voting mechanism occurs with a drawback when facing a skewed class distribution problem. To solve this problem, one way of weighting the classification [24] is by considering the distance calculated from the test light to each of the KNNs. Thus, for both cases of classification and regression tasks, KNN is helpful and useful [23] to set weights from the dedications of the neighbors, which results in the closest neighbor having more contributions compared to the average value than the more remote ones. Specifically, if K = 1, this object is directly dispatched to the class of the single nearest neighbor, and the selection of the best K is highly dependent on the given data [22,25];

although larger K-values lower the noise effect when processing the classification [26], they create less distinction for the boundaries from classes. Interestingly, good Ks are achieved through some heuristic learning approaches, and the performance evaluations (accuracies) of the KNN algorithm are significantly distorted and degraded by the irrelevant features and the presence of unexpected noise. Efforts for scaling and selecting features from many researchers have been improved with classification performance. Thus, an interesting and popular tool uses the reciprocal information [27] on the data with the classes trained and the evolutionary algorithms to optimize feature selecting and scaling. In particular, the output is a class membership in KNN's classification community.

Conversely speaking, KNNs have some disadvantages with larger sets of data as they are computationally intensive and tractable. In particular, KNNs are an instance-based heuristic learning method; the effects are locally approximative, and all mathematical computations are postponed until the classification work is complete [28]. The classification performance of KNNs is always significantly improved through supervised learning techniques, resulting in some strongly consistent results. Some algorithms for nearest neighbors have been intensively studied during the past decades, and this research has focused on seeking a decrease in all numbers of distance assessments actually fulfilled. Given the case, it is possible to make great improvements to the speed of KNNs by using proximity graphs [28].

KNN algorithms are the simplest type of algorithm for ML communities, where its algorithms are easier to perform by calculating the distances from all the test examples to the stored examples. Given these successful reasons for using the KNN algorithm in the context of various application data, the classification performance of KNNs will be increased effectively and significantly if the distance metric is studied and learned by a specialized algorithm like the large margin nearest neighbor (LMNN) classification algorithm [29]; the KNN algorithm is analyzed extensively [30] in this study to seek its performance in classification learning.

2.4. Applied Ensemble Learning of Stacking Classifier

Regarding the introduction of ML techniques, an ensemble method is important for conducting an empirical exploration. The ensemble method, which produces more accurate solutions, uses several learning algorithms to construct a new classifier [31,32] in unsupervised and supervised learning cases; it is called a meta-algorithm. Although this ensemble classification method tends to achieve empirically more predictive outcomes when a major variation between these models is faced and shows more flexibility functions [33], the prediction of this ensemble method needs more computational time than the prediction of a stand-alone model. Thus, ensemble methods are regarded as a way to perform extra computations to compensate for poor search-learning algorithms. There are various types of ensemble learning classification approaches [34] (that is, stacking, bagging, and boosting), and there is no single winner. First, bagging has a function to reduce the variation from the prediction by using a repetition method for adding data from the original dataset for training. Second, boosting adopts a two-step procedure to use subsets from the raw data to yield, on average, a series of carrying-out classification systems and boosts their classification performance by using a particular method of majority voting to unite them. Third, stacking [35] (STK) is a similar boosting method, and what is unique is that STK is able to determine the models constructed. In particular, it trains a learning model of such an ensemble method to make the predictions of combining other learning models to the STK classifier. Interestingly, a logistic regression algorithm is more used for the combiner in the STK method. STK [36] typically yields better performance than any of the stand-alone models trained, and it is well made for some supervised learning algorithms, such as regression analysis [37] and the classification method [38] and unsupervised learning algorithms such as chemical process fault diagnosis [39].

Although the performance that combines a variety of strong hybrid learning algorithms [40] to promote diversity is more effective than using single models, it is difficult to apply this method to find the best approach for overcoming practical problems. Theoretically, the STK algorithm [41] yields

better classification performance than any single one of the models trained. Thus, the performance of the STK algorithm is tested in practice in this study.

2.5. Radial Basis Function Network

Typically, the radial basis function network [42] (RBFN) is used for the construction of an algorithm of a type of artificial neural network (ANN), which gives radial basis algorithms the role of an activation function in forming a mathematical model that is helpful in defining linear problems of functioning inputs and neuron parameters. RBFNs have three layers of structure, including a pass-through input layer, a hidden layer, and an output layer. Regarding the formation of the layers' structure, the input layer is formed for a vector of realistic numbers and is linked to a number of hidden neurons, and the Euclidean distance for a center vector and a Gaussian function [42] is typically used in the norm formulation since the function is radially symmetric to this vector; thus, the radial basis algorithm is named. RBFNs consider different natures of the nonlinear hidden neurons versus the linear output neuron and are justified by weights. RBFNs are processed in a two-step algorithm corresponding to the training done. The first step involves performing unsupervised learning to choose and fix the center vectors, width, and weights in the hidden layers. The center vectors are defined by using a K-means clustering algorithm, or they are sampled and trained randomly from some sets of instances since no obvious way is effectively determined for the centers. Interestingly, RBFN is a type of nonparametric model, and its weights with parameters do not have a special intention associated with the problem to which it is used for estimating the values of a neural network in supervised learning [43,44]. The second step simply involves fitting a linear model with appropriate weights to create the hidden layer's output based on some objective function, such as the least-squares function that optimizes the accuracy of fit by the optimal choice of weights [45].

A main advantage of the RBFN is to keep the mathematics model simple. RBFN is relatively cheap in linear algebra and computational intelligence [43]. In particular, RBFN learns how to convert input data to a wanted response, and it has widely used for the functions of pattern recognition, classification, time series prediction, system control, and approximation [46–48]. The above rationality for the noticed well-being of the RBFN classifier in previous research is highlighted within this study; it is selected as one of the primary goals of complex comparison purposes to estimate the underlying function of neural networks due to their intrinsic meaning.

2.6. Applied Naïve Bayes

In applied views of ML domains, the classifier of naïve Bayes, NB, is a popular approach in a simple type of probabilistic process of classifier, with powerfully independent hypotheses between the attributes to process real-life problems according to Bayes' theorem [49,50]. NB assumes that the value of a specific attribute has independence from the value of any other attributes in the class variable given, and it effectively considers each of the attributes to be independently devoted to the possibility used. In spite of the naïve design and oversimplified hypotheses, the NB classifier [51,52] has done well, with considerable applications for complex real-world problems, particularly in financial prediction and medical diagnosis, with more advanced computing methods. From the limited literature on statistics and computer science, the NB classifier holds many capabilities that make it helpful enough to attract the attention of researchers. First, for completing the decoupling of the class distributions of conditional features, each distribution is estimated independently for a type of one-dimensional distribution, which serves to lessen a real problem derived from the effects of the curse of dimensionality on dataset features. Second, in many real applications, the method of maximum likelihood is used as an extensive parameter estimation for NB classification models; restated, it is unnecessary to accept Bayesian probability or to use any Bayesian method. Third, the NB classifier [53] of ML is highly concerned with scalable jobs in some linear parameters for the attributes used for solving real-life problems.

According to a smart advantage of NB [54], only small numbers of data training are needed to produce a good forecast of the linear parameter that is useful to classification works in many

applications in order to meet the requirement of a good classifier. Given this reason, the classifier is sound enough to neglect strict inadequacy in the fundamental naïve probabilities of models. Regarding some probability models, the NB classifier is efficiently trained in an ML technique. Thus, NB is used in this study as a comparative means to assess its performance due to its prominent results in past studies.

3. Methods and Materials

This section introduces the related methods and materials used in this study to address the applied financial application issues related to EPS to reach meaningful empirical research with a rich treasure trove of knowledge, as follows.

3.1. Background of the Applied Study Framework

In practice, benefiting the consideration of financial stock investment settled in the big data framework of the complicated stock market for interesting financial truths and options, accurately classifying the EPS of listed companies is an interesting issue attracting investors; however, unscientific decisions are unfortunately involved in the profit-making development, which may never be successful for the investing plan. Highly skilled investors have searched for an intelligent model to properly identify the potential positive/negative target from the vast sea of stocks for escaping possible losses and, conversely, maximizing profits. To keep making the right investor decisions, this study uses the advantages of past reviews of literature and the managerial experience of experts on financial ratios to propose a hybrid multicomponential discretization model with ML techniques to develop effective early-warning rules for the identification of positive/negative EPS. This study proposes a map of advanced multicomponential discretization models for identifying financial diagnoses and has the purpose of using 2009–2014 financial statements to research the EPS of companies on TWSE from six different industry online financial databases to assess the componential performance of the models, with effective comparative studies for getting rich features. The varied components of the models in the study test the performance measurements of differently organized data-preprocessing, data-discretization, feature-selection, two data split methods, machine learners, rule-based DT knowledge, time-lag effects, different times of running experiments, and two types of different classes.

Recently, a varied function of emerged linear and nonlinear ML techniques [55,56] in advanced soft computing algorithms such as DT-C4.5, the KNN algorithm, ensemble STK, RBFN, and NB, other than the nonlinear support vector machine, multilayer perceptron, and the linear logistic regression, has been used as an important research approach for both academicians and practitioners due to their superior past performance and has been well studied, with a wide application field with beneficial effects. Therefore, past prominent capability is very worthy of being a starting point to a study and overview of the linear or nonlinear comparative studies for further research work. In view of these interesting facts for modeling classification works [57,58], they are used as the basis for the complete construction of the study framework. There are nine components (stages), with 11 detailed steps for raising the advantages and rationalities of this study. (1) Data-preprocessing: This component is used to build a tangible benefit from the reviews of literature and experts from a specific database. (2) Data-discretization: The discretization facilitates the use of data from the view of natural language and improves classification accuracy. (3) Feature-selection: This core technique is used to lower data dimension and complexity to speed the benefits of operating experiments. (4) Two data-split methods: The different data-split methods are used for comparison of the review of practical datasets in the same environment. (5) Machine learners: Different classifiers are assessed and compared to discover the best suitable tool for EPS financial diagnosis. (6) Rule-based DT knowledge: The decisional rules generated by the DT-C4.5 algorithm are used for constructing the instructions of the "IF ... THEN ... " form used to help define a better choice to achieve a specific goal. (7) Different times of running experiments: It has the advantage of determining the performance of various classifiers with the same given data. (8) Time lag effects: The lag effects of different time periods are measured to understand whether the financial data consider a substantial time-lag effect demonstrated by the given data as

well as to identify the right time lag. (9) Types of different classes: The merit of measuring different classes is used to define the performance difference of various classifiers. Mainly, the nine major stages systematically detail the computing processes to support a further clear definition, which is described in Figure 1, to present a flowchart of the applied hybrid models proposed to represent data relativity, with the corresponding tools and steps for experiencing the empirical outcomes (results) of EPS.

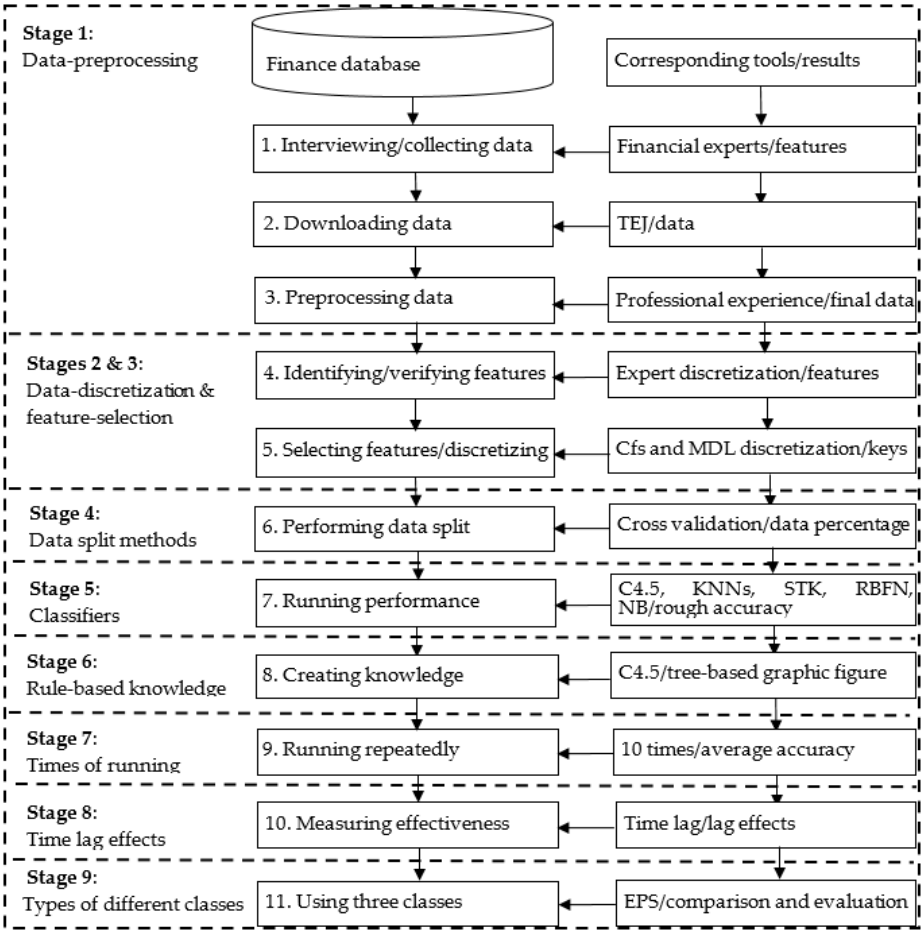


Figure 1. Flowchart of the applied hybrid models proposed, with corresponding tools in detailed steps.

These abovementioned components are selected based on the following three reasons. First, they had a high level of evaluation in past studies. Second, no one has tried to use them in predicting EPS of financial fields and then created rule-based decisional knowledge for the stock market to address this practical problem. Third, it is worthwhile to explore the interesting issue of multicomponential discretization models, and its application areas of research and its application advantage are accordingly expected and studied here.

From the restricted review, the study has a valuable contribution and directions in which the multicomponential discretization models are very limited and rarely seen for such an industry setting. Thus, the applied hybrid models proposed are of interest, with the uses of the above nine stages with 11 detailed steps. The proposed hybrid models are implemented and executed in a step-by-step mode;

some steps (e.g., five classifiers: DT-C4.5, KNNs, STK, RBFN, and NB) are respectively run for each model from a software package tool, and some other steps are coded or stored in the CSV format of Excel software (e.g., data-preprocessing is implemented in Microsoft Excel).

3.2. Algorithm of the Applied Hybrid Models Proposed

In the subsection, this study applies five main hybrid classification models, with nine stages, to address financial application issues related to EPS to find a meaningful empirical research method. To improve the readability of this study, the concepts and structures of these models are addressed in a tableau list report. The main stages of these proposed hybrid models (referred to as Models A–E) are data-preprocessing, data-discretization, feature-selection, data split methods, various classifiers, time-lag effects, and the measurement of two types of classes. Table 1 lists their information in detail below. Interestingly, the order performance for executing data-discretization and feature-selection is evaluated in this study.

Table 1. Component information of the applied hybrid five models proposed.

Stage (Component) \ Model	A	B	C	D	E
1. Data-preprocessing	√	√	√	√	√
2. Data-discretization		√		√(1)	√(2)
3. Feature-selection			√	√(2)	√(1)
4. Measurement of data-split methods	√	√	√	√	√
5. Classifiers	√	√	√	√	√
6. Measurement of rule-based knowledge	√	√	√	√	√
7.Measurement of times of running experiments	√	√	√	√	√
8. Measurement of time-lag effects	√	√	√	√	√
9. Measurement of types of classes	√	√	√	√	√

Note: The (1) and (2) denote the execution sequence for experiments to process data-discretization and feature-selection, respectively.

The algorithm for the above five hybrid models, with the experience of an empirical case study application extracted from a real financial database in Taiwan, is detailed and implemented in 11 steps systematically, as follows.

Stage 1. Data-preprocessing: Table 1 shows that this component is used in all the applied hybrid models proposed, Models A–E.

Step 1. Consulting field experts and gathering data: In this step, we first interviewed field experts interested in financial analysis and investment management and studied the online financial database of the noted Taiwan Economic Journal (TEJ) [59], which is targeted as the object of the experiments in order to identify and confirm essential features. The 25 essential features, including 24 condition features encoded as X1–X24 and one decision feature (X25) encoded as Class, were first defined with the help of financial experts and literature reviews. Financial experts added the first four extra features (X1–X4; Year, Season, Industrial Classification, and Total Capital) to the list of the 20 features (renamed, in order, as X5–X24) that were seen in Section 2.1. The classes of EPS were suggested by experts.

Step 2. Downloading the data used from the TEJ database: Accordingly, raw data for these features were downloaded as an experimental dataset, with 4702 instances from the TEJ database for a six-year period. For ease of presentation, the dataset is named the TEJ dataset.

Step 3. Preprocessing the data: In further analysis, this step filtered six industries. The six industries are electrical machinery, biotechnology and medical, semiconductor, optoelectronics, electronic components, and shipping industries, which are well-known and of high trading volume among Taiwanese investors. As this was the first attempt to find rules in the Taiwanese stock market, the study chose these industries to begin. To facilitate the operation of the experiment, this step cleared irrelevant columns with incomplete or inaccurate data, added and computed some new columns that were not present in the TEJ dataset (such as times interest earned), joined columns of special stock

and common stock as the feature of total capital, reconfirmed the information of the TEJ dataset with official reports reviewed from TWSE, and stored the dataset in Excel software format. Table 2 lists all the feature information with descriptive statistics from the TEJ dataset.

Table 2. Information of descriptive statistics for all features in the TEJ dataset.

Code	Feature	Type	Min.	Max.	Mean (μ)	S.D. (σ)
X1	Year	Symbolic	-	-	-	-
X2	Season	Symbolic	-	-	-	-
X3	Industrial classification	Symbolic	-	-	-	-
X4	Total capital	Numeric	200,000	259,291,239	7,721,909.23	23,565,808.62
X5	Current liabilities	Numeric	42,034	419,171,745	8,008,758.41	21,742,627.73
X6	Cash flow ratio	Numeric	−176.05	297.83	9.09	18.18
X7	Net asset value of each share	Numeric	0.09	249.42	21.48	14.48
X8	Cash flow of each share	Numeric	−23.34	37.03	0.88	1.89
X9	Sales per share	Numeric	−56.6	69.90	8.05	7.30
X10	Operating income per share	Numeric	−17.44	28.43	0.48	1.32
X11	Current ratio	Numeric	19.93	2247.51	224.54	175.44
X12	Quick ratio	Numeric	5.42	2127.72	165.93	152.63
X13	Debt ratio	Numeric	2.70	97.95	42.04	15.82
X14	Accounts receivable turnover ratio	Numeric	−2.63	305.33	1.80	7.90
X15	Inventory turnover ratio	Numeric	−1681.03	1274.55	4.59	75.30
X16	Fixed asset turnover ratio	Numeric	−6.38	91.33	1.31	4.97
X17	Operating income margin	Numeric	−640.32	167.52	17.66	23.38
X18	Return on net asset	Numeric	−84.40	45.29	4.21	10.17
X19	Cash ratio	Numeric	0	20.41	0.87	1.31
X20	Times interest earned	Numeric	−266,442	600,237	957.50	15,209.79
X21	Interest expense	Numeric	−77,927	2,173,528	38,994.51	130,730.96
X22	Year-on-year percentage total assets	Numeric	−58.02	349.27	5.62	22.05
X23	Total asset turnover ratio	Numeric	−0.55	1.49	0.19	0.10
X24	Net worth turnover ratio	Numeric	−1.06	4.37	0.37	0.27
X25	EPS (two classes and three classes)	Numeric	−11.95	25.38	0.40	1.19

Note: "S.D." refers to standard deviation, and "-" refers to a field with no answer given.

Stages 2 and 3. Data-discretization and feature-selection: Table 1 makes it is clear that the component of data-discretization is only used in Models B, D, and E, and the component of feature-selection is for Models C–E.

Step 4. Identifying and verifying and discretizing features: The decisional feature of this study is defined as EPS, which is divided into two types of two and three classes, and the conditional features are 20 financial ratios plus the related four financial variables of Year, Season, Industrial Classification, and Total Capital, which are listed in Table 2. In two classes, the decisional feature of EPS named as Class is firstly classified into P (>0 in NTD, i.e., positive profit) and N (≦0, negative profit), according to the opinion and selection of three experts.

Step 5. Selecting the features and/or discretizing the data: Subsequently, test five organized ways of data-mining the models using various components of data-preprocessing, data-discretization, and feature-selection for performance comparison, including (1) with data-preprocessing but without data-discretization and feature-selection, (2) with data-preprocessing and data-discretization but without feature-selection, (3) with data-preprocessing and feature-selection but without data-discretization, (4) conducting data-preprocessing and data-discretization before feature-selection, and (5) conducting data-preprocessing and feature-selection before data-discretization. The above five ways are mainly measured for component performances by using a Cfs subset evaluation algorithm of search method for feature-selection and a filter function of discretizing tool with minimum description length (MDL) of automatic data-discretization in professional software for data mining techniques from software-defined networks, followed by different classifiers that are executed by employing professional package services. As a result, there are a total of three core determinators, including times interest earned, operating income per share, and total assets growth rate, that are identified in

the feature-selection processes for influencing EPS in the type of two classes, which is significantly improved and associated with the stock price of a specific listed stock.

Stage 4. Data-split methods: This component is used for Models A–E.

Step 6. Performing cross-validation and percentage data-split: Furthermore, it is interesting to find out an appropriate model to overcome problems faced in real life by assessing the function of various components of models and then selecting a suitable model. It is desirable to avoid spending a long-time on training a model that is poor. Thus, this step uses two model-selection methods, cross-validation and percentage data-split, when training/testing the processing of the target dataset to make the right selection. On the one hand, based on a study by Džeroski and Zenko [60], the cross-validation method, a general approach for model-selection, is allocated and synthesized with "testing the models with the entire training dataset, and choosing the one that works best" when various models are used with a large set of real problems. In the reviews of model-selection, a basic form of cross-validation divides the dataset into two sets, with the bigger set used for training and the other smaller set used for testing if data are not scarce and are substantial enough to enable input–output measurements. On the contrary, another model-selection approach is to use percentage-split data; it is a popular approach in ML application algorithms from numerous studies [61,62]. The percentage-split for the dataset partitions the original set into different groups of training/testing sub-datasets, such as 90%/10% and 80%/20%. In practice, the training/testing sub-dataset is usually at a 2/1 ratio to achieve a good and reasonable result. To reverify their selection performance, the above two methods are adopted into all the proposed hybrid models, and, from past successful examples, the 10-fold cross-validation and 67%/33% percentage-split of data are commonly used in this step.

Stage 5. Classifiers: This component is used in Models A–E.

Step 7. Applying classifiers and making performance comparisons: Accordingly, run five classifiers (DT-C4.5, KNNs, STK, RBFN, and NB) for each model from the software package tool, with no changes predesigned on these learning classifiers; the five classifiers are selected from performance assessments based on their past high satisfactory results, and their classification performance in accuracy rate due to common use under the two classes of EPS is compared to see and judge their differences that may have some implied information for stock investors. The accuracy rate in one run for these classifiers is then achieved.

Stage 6. Rule-based knowledge: This component is used for Models A–E.

Step 8. Creating rule-based knowledge for interested parties: To generate and understand the hidden information of forming "IF … THEN … " in rule-based knowledge, this step employs a fundamental step of the DT-C4.5 algorithm to create a decisional-tree-based structure (in a graphic figure) to represent the formatted knowledge of the target TEJ dataset. The knowledge in the figure is crucial to a varied topic of EPS in financial investment. For easy presentation and reading, the created tree-based structure (in a graphic figure) and its explanations will be uniformly displayed in the next section.

Stage 7. Times of running experiments: This component is used in Models A–E.

Step 9. Running the experiments repeatedly: To further validate and test the classification accuracy of the five classifiers, the experiment is repeatedly run 10 times for the TEJ dataset, and the average accuracy is obtained for the difference analysis.

Stage 8. Time-lag effects: This component is used for Models A–E.

Step 10. Measuring the effectiveness of time lag: To consider whether time lag has an influence on the prediction of EPS, divide the dataset by seasons, and use the EPS of the current season (T+0), the next season (T+1), the third season (T+2), and the fourth season (T+3) as the decisional feature to test and find the best performing models, which are using the DT-C4.5 and RBFN classifiers and dividing EPS into two classes due to their suitability.

Stage 9. Types of different classes: Finally, this component is used for Models A–E.

Step 11. Using three classes of EPS: To further differentiate and create the class difference, use three classes (i.e., A (<0), B (0~3.75), and C (>3.75)) of EPS based on the automatic discretization

recommendation instead of two classes in all the applied hybrid models proposed, and the experiments from Steps 1–10 are run again.

4. Experimental Data Analysis and Research Finding

This section summarizes and concludes the above empirical results in the TEJ dataset on EPS and describes some rule-based knowledge construction, with some followed useful research findings and management implications, and the limitations of the study.

4.1. Empirical Results with Implication

Based on the applied hybrid models proposed with an empirical case study of financial ratios and financial variables, the experimental results are concluded uniformly in some tableau lists for the purposes of performance comparison. Moreover, regarding research-based conclusions, some meaningful information is referenced as a consultant datum to interested parties. First, the performance of models on the tests of cross-validation and percentage-split data was preferentially conducted in view of technical background and support. The cross-validation method selects 90% of the training data subset, and splits 67% of the training data for the other method. Accordingly, the performance of different periods of time-lag effects, different times of running experiments, and different classes were evaluated for the TEJ dataset. Conclusively, the analytical results from all experiments are respectively defined as the following three key points. First, Tables 3–6 show that the outputs of Steps 1–8 and 11 (i.e., Stages 1–6 and 9) in one run for two data-split methods and two types of classes; second, Tables 7–10 show the outputs of Steps 1–9 and 11 (i.e., Stages 1–7 and 9) in 10 runs; and third, Tables 11–14 show the outputs of Steps 1–8 and 10 (i.e., Stages 1–6 and 8) for measuring the time-lag effects.

Table 3. Testing results for the cross-validation method in two classes for the TEJ dataset.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
DT-C4.5	91.4292	91.4292	91.1740	91.4292	91.4292	91.3782
KNNs	74.6704	88.0476	86.2399	91.3441	91.3441	86.3292
STK	72.2033	72.2033	72.2033	72.2033	72.2033	72.2033
RBFN	74.3301	90.0893	85.8783	91.3016	91.3016	86.5802
NB	65.3126	88.9834	55.8911	91.3886	91.3886	78.5929
Avg. accuracy	75.5891	86.1506	78.3284	87.5334	87.5334	83.0270

Note: "Avg." refers to average, and the shading highlights the significance in the accuracy rate.

Table 4. Testing results of the percentage-split method in two classes for the TEJ dataset.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
DT-C4.5	92.1392	91.6881	92.4613	92.0103	92.0103	92.0618
KNNs	76.4175	88.7887	86.3402	91.9459	91.9459	87.0876
STK	72.6804	72.6804	72.6804	72.6804	72.6804	72.6804
RBFN	73.6469	91.3015	88.6598	92.0747	92.0747	87.5515
NB	69.4588	90.6572	59.1495	92.0103	92.0103	80.6572
Avg. accuracy	76.8686	87.0232	79.8582	88.1443	88.1443	84.0077

Table 5. Testing results of the cross-validation method in three classes for the TEJ dataset.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
DT-C4.5	90.8762	90.9188	90.8550	90.9188	90.9188	90.8975
KNNs	73.4794	88.2178	86.0910	90.8550	90.8550	85.8996
STK	71.2675	71.2675	71.2675	71.2675	71.2675	71.2675
RBFN	77.3926	88.7495	88.6431	90.5395	90.5395	87.1728
NB	66.9715	86.4738	63.6112	90.4509	90.4509	79.5917
Avg. accuracy	75.9974	85.1255	80.0936	86.8063	86.8063	82.9658

Table 6. Testing results of the percentage-split method in three classes for the TEJ dataset.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)	Total Avg. (%)
DT-C4.5	91.2371	91.5573	91.5593	91.5593	91.5593	91.4945	91.4708
KNNs	75.6443	88.2732	88.2732	91.4948	91.4948	87.0361	86.5881
STK	71.7784	71.7784	71.7784	71.7784	71.7784	71.7784	71.9824
RBFN	78.6727	90.3995	90.3995	91.5593	91.5593	88.5181	87.4557
NB	70.6186	89.1108	89.1778	91.1727	91.1727	86.2505	81.2731
Avg. accuracy	77.5902	86.2238	86.2376	87.5129	87.5129	85.0155	83.7540
Total avg.	76.5113	86.1308	81.1295	87.4992	87.4992	-	-

Table 7. Accuracy mean and standard deviation of the cross-validation method in two classes for the TEJ dataset.

Model	A (%)		B (%)		C (%)		D (%)		E (%)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
DT-C4.5	91.43	1.25	91.36	1.26	91.18	1.20	91.39	1.25	91.39	1.25
KNNs	74.70	1.74	88.28	1.31	86.20	1.47	91.28	1.23	91.28	1.23
STK	72.20	0.09	72.20	0.09	72.20	0.09	72.20	0.09	72.20	0.09
RBFN	74.09	2.61	90.23	1.35	86.10	1.89	91.35	1.25	91.35	1.25
NB	65.17	2.23	89.11	1.32	55.89	3.40	91.44	1.18	91.44	1.18
Avg. accuracy	75.52	1.58	86.24	1.07	78.31	1.61	87.53	1.00	87.53	1.00

Note: "S.D." refers to standard deviation, and the shading highlights the significance in the accuracy rate.

Table 8. Accuracy mean and standard deviation of the percentage-split method in two classes for the TEJ dataset.

Model	A (%)		B (%)		C (%)		D (%)		E (%)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
DT-C4.5	91.32	0.60	91.15	0.51	91.16	0.77	91.26	0.57	91.26	0.57
KNNs	74.58	1.13	88.48	0.79	85.86	0.63	91.09	0.66	91.09	0.66
STK	72.20	0.02	72.20	0.02	72.20	0.02	72.20	0.02	72.20	0.02
RBFN	74.58	1.81	90.49	0.67	86.64	1.53	91.30	0.62	91.30	0.62
NB	65.78	2.25	89.61	0.51	55.08	3.76	91.44	0.61	91.44	0.61
Avg. accuracy	75.69	1.16	86.39	0.50	78.19	1.34	87.46	0.50	87.46	0.50

Table 9. Accuracy mean and standard deviation of the cross-validation method in three classes for the TEJ dataset.

Model	A (%)		B (%)		C (%)		D (%)		E (%)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
DT-C4.5	91.13	1.21	91.14	1.21	91.21	1.16	91.15	1.20	91.15	1.20
KNNs	74.10	2.00	88.56	1.54	86.92	1.32	91.15	1.19	91.15	1.19
STK	71.74	0.08	71.74	0.08	71.74	0.08	71.74	0.08	71.74	0.08
RBFN	77.06	2.27	89.69	1.24	90.96	1.17	90.86	1.26	90.86	1.26
NB	67.42	2.46	87.80	1.42	62.27	4.85	90.57	1.20	90.57	1.20
Avg. accuracy	76.29	1.60	85.79	1.10	80.62	1.72	87.09	0.99	87.09	0.99

Table 10. Accuracy mean and standard deviation of the percentage-split method in three classes for the TEJ dataset.

Model	A (%)		B (%)		C (%)		D (%)		E (%)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
DT-C4.5	90.93	0.54	91.12	0.56	91.02	0.61	91.05	0.56	91.05	0.56
KNNs	74.08	1.11	88.55	0.70	86.85	0.62	91.15	0.58	91.15	0.58
STK	71.73	0.02	71.73	0.02	71.73	0.02	71.73	0.02	71.73	0.02
RBFN	77.39	1.11	90.09	0.49	90.76	0.63	90.70	0.44	90.70	0.44
NB	67.94	2.11	88.39	0.47	59.09	5.70	90.56	0.60	90.56	0.60
Avg. accuracy	76.41	0.98	85.98	0.45	79.89	1.52	87.04	0.44	87.04	0.44

Table 11. Test results of cross-validation with EPS of two classes, classifier C4.5, and time lag.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
T+0	91.4292	91.4292	91.1740	91.4292	91.4292	91.3782
T+1	83.2508	83.8536	84.9085	84.6932	84.6932	84.2799
T+2	78.8851	80.5424	80.1119	80.5424	80.5424	80.1248
T+3	76.7857	80.1282	79.8306	80.1282	80.1282	79.4002

Table 12. Test results of percentage-split with EPS of two classes, classifier C4.5, and time lag.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
T+0	92.1392	91.6881	92.4613	92.0103	92.0103	92.0618
T+1	82.9746	83.1050	84.5401	83.4964	83.4964	83.5225
T+2	77.8865	81.0176	79.9087	81.0176	81.0176	80.1696
T+3	76.6135	79.1117	79.1117	79.1117	79.1117	78.6121

Table 13. Test results of cross-validation with EPS of two classes, classifier RBFN, and time lag.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
T+0	74.3301	90.0893	85.8783	91.3016	91.3016	86.5802
T+1	71.0011	84.3272	81.0334	84.1550	54.1550	74.9343
T+2	72.7938	80.4348	80.3056	68.6612	68.6612	74.1713
T+3	72.4359	79.6016	79.0064	75.8700	75.8700	76.5568

Table 14. Test results of percentage-split with EPS of two classes, classifier RBFN, and time lag.

Model	A (%)	B (%)	C (%)	D (%)	E (%)	Avg. (%)
T+0	73.6469	91.3015	88.6598	92.0747	92.0747	87.5515
T+1	72.7984	84.2140	82.3875	83.4964	83.4964	81.2785
T+2	72.0809	79.8434	80.5610	77.9517	77.9517	77.6777
T+3	75.1561	78.1402	77.9320	76.6135	76.6135	76.8911

A total of 10 directions are simply defined from all the main empirical results of the above Tables 3–6 for Models A–E in the TEJ dataset; they are described as follows.

(1) Performance comparison of models and classifiers: As indicated in the above four tables, seven key results have been determined in terms of models and classifiers, and the shading highlights their significance in the accuracy rate. (a) The poorest accuracy is 75.59%: On average accuracy, Model A has the poorest performance, implying feature-selection and data-discretization could effectively and significantly enhance classification performance. (b) The highest accuracy is 88.14%: On average accuracy, Models D and E have the highest performance, providing valid proof that both feature-selection and data-discretization can improve classification accuracy. (c) The better performance: On total average accuracy, Model B (86.13%) outperforms Model C (81.13%), implying that by comparing the effects of feature-selection with data-discretization, the latter (data-discretization) has more impact on classification performance than on the former feature-selection method. (d) The same accuracy: Models D (87.50%) and E (87.50%) have the same performance in total average accuracy,

implying that the executing sequence of feature-selection with data-discretization does not affect model performance. (e) The better average accuracy: Among the five classifiers, DT-C4.5 (91.47%) and RBFN (87.46%) algorithms have better performance than the others in terms of accuracy rate or average accuracy. (f) The best model: Conclusively, Models D and E are the best models both in terms of average accuracy (87.51%) and total average accuracy (87.50%). (g) The best classifier: Conclusively, the DT-C4.5 algorithm wins as the best classifier in terms of both average accuracy (91.49%) and total average accuracy (91.47%).

(2) Assessment of data-discretization: When comparing Model A with Model B in Tables 3–6, it is clear that Model B has mostly better accuracy than Model A. This information implies that the data-discretization method is an effective tool to improve classification accuracy.

(3) Assessment of feature-selection: Similarly, when comparing Model A with Model C in the same tables, they show that Model C has better accuracy than Model A. This case implies that the feature-selection method benefits classification accuracy.

(4) Measurement of two data-split methods: The analytical results indicate that the cross-validation data approach has higher classification performance than the percentage-split data method in the TEJ dataset when Tables 3 and 4 are compared with Tables 5 and 6, respectively.

(5) Measurement of two types of classes of EPS: It is seen that dividing EPS into two classes yields better performance than dividing EPS into three classes. This phenomenon implies a converse management implication that the more classes a decision feature has, the less accuracy the classifier has.

(6) Measurement of two types of running times for experiments: Furthermore, to better justify these comparative results and conclusions, the study repeated the experiments 10 times to compute the mean as well as its standard deviation (S.D.) of the accuracy rate of each setting. Tables 7–10 show that there are five important viewpoints determined from the views of the mean and standard deviation of accuracy rate for the two data split methods and the two types of classes. (a) As the standard deviations (e.g., 1.00%) of accuracy rates are quite small compared to the mean values (e.g., 87.53%), it is known that the accuracy rates are densely distributed around the mean values. Therefore, the aforementioned conclusions based on the comparisons of average accuracy rates of each model setting are credible. (b) Regarding average accuracy, Models D and E have the same performance (i.e., 87.53%, 87.46%, 87.09%, and 87.04%) and win this accuracy competition regardless of two or three classes in 10 runs. (c) As for the standard deviation of average accuracy, Models D and E have a lower value (e.g., 1.00% < 1.07% < 1.58% < 1.61%, from Table 7) of relative stability, and it implies that the data-discretization and feature-selection methods not only improve classification accuracy but also enhance their high stability regardless of two or three classes in 10 runs. (d) It was found that the percentage-split data method in Models A (e.g., 75.69% > 75.52%) and B (e.g., 86.39% > 86.24%) has a better performance than the cross-validation method regardless of two and three classes. Conversely, the percentage-split data method in Models C–E (e.g., 78.19% < 78.31% and 87.46% < 87.53%) has less performance than the cross-validation method regardless of two and three classes. This interesting problematic issue should be further explored and examined in subsequent research. (e) More interestingly, the accuracy of 91.43% of DT-C4.5 in Model A without exogenous data-discretization and feature-selection has the second-highest accuracy in the 10 runs. A rational basis about the special case is that the DT-C4.5 algorithm already has an endogenous function of data-discretization and feature-selection, which is more suitable for the TEJ dataset than other methods.

(7) Measurement of time-lag effects: To test the influence that time lag has on EPS, the study ran the first two better-performing models found by using the classifiers of DT-C4.5 and RBFN in two classes. Tables 11–14 mainly show that there are four key directions useful to understanding the time effectiveness of rule-making for investors. (a) It is implied that the less the time lag, the higher the accuracy, and vice versa, regardless of data cross-validation or percentage-split methods and two or three classes. This analytical result indicates and implies that the models could not predict EPS in the far future and support the treatment of using the current season's EPS in most of the TEJ dataset. (b) Most of the rank of performance is followed by (T+0) ➔ (T+1) ➔ (T+2) ➔

(T+3) (e.g., 91.43% > 83.25% > 78.89% > 76.79% in Model A from Table 11), which shows in the cross-validation method in two classes for RBFN classifier. (c) From Model A, when Tables 11 and 12 are compared to Tables 13 and 14, it clearly indicates that DT-C4.5 (e.g., (91.38% and 92.06%) > (86.58% and 87.55%)) significantly outperforms RBFN in average accuracy, and it was found that the RBFN algorithm needs the support of some external techniques like feature-selection and data-discretization methods to improve its classification accuracy. (d) The above information implies that some classifiers in stand-alone models require some special methods to support its classification ability, and it implies hybrid models perform better than stand-alone models.

(8) Results of data-discretization: According to the estimation from Model E in two classes, Table 15 lists the results of the data-discretization method after implementing feature-selection. Table 15 shows that the three key condition features, X10, X20, and X22, are discretized into six, six, and five intervals based on the automatic discretization method, which correspond to linguistic terms A_1–A_6, B_1–B_6, and C_1–C_5, respectively. The decisional feature is discretized into two intervals (i.e., P and N), as suggested by three financial experts. These linguistic terms are referred to as a natural language, which is useful and helpful to good understanding and decision-making. For example, the linguistic terms A_1–A_6 are referenced as very low, low, medium, medium–high, high, and very high, respectively. Moreover, the data-discretization method for setting the linguistic terms improves the classification accuracy.

Table 15. Information of data-discretization for the four features in two classes in the TEJ dataset.

Feature	Cutoff Point	Interval	Linguistic Term	Natural Language	Corresponding Instances
X10	−0.195, −0.005, 0.135, 0.235, 0.385	6	A_1–A_6	Very low, Low, Medium, Medium high, High, and Very high	814, 472, 517, 352, 448, 2099
X20	−0.195, −0.005, 0.135, 0.235, 0.385	6	B_1–B_6	Very low, Low, Medium, Medium high, High, and Very high	20, 1080, 340, 282, 610, 2370
X22	−0.195, −0.005, 0.135, 0.235	5	C_1–C_5	Very low, Low, Medium, High, and Very high	110, 745, 1046, 471, 2330
X25	By expert suggestion	2	P and N	Positive and Negative	-

(9) Results of feature-selection: According to the estimation from Steps 4 and 5, three important factors, including total assets growth rate, times interest earned, and operating income per share, are specified in the feature-selection method for the two classes of EPS; simultaneously, two key features, times interest earned and operating income per share, are defined for the three classes. Obviously, the two key factors of times interest earned and operating income per share are identified in both the two and three classes.

(10) Additional measurement of hybrid models with stand-alone models: To further evaluate the capacity between the applied hybrid models proposed and the stand-alone models [18,30,34,45,54], the stand-alone models were run and their classification accuracy was achieved, respectively, in the two and three classes. Subsequently, the applied hybrid models proposed had the best of 92.46% compared to that of each stand-alone model, regardless of two or three classes. Clearly, the hybrid models did better than the stand-alone models used for the empirical case of this study.

4.2. Rule-based Knowledge Construction

The decision-rule-based trees created by the C4.5 algorithm in Step 8 of the applied hybrid models proposed, as well as three key features for influencing EPS, including times interest earned, total assets growth rate, and operating income per share, were determined. Figures 2 and 3 depict the framework of the tree-based rules in an interactive visualization form below. To illustrate the forming processes of the rules effectively, examples of some rules (highlighted in red in the type of two classes and three classes) are indicated in EPS object-oriented interpretations for explaining financial ratios and variables applications using advanced multicomponential discretization models toward a solution of big data benchmarks.

For example, the explanation of three rules in Figure 2 is described in detail, as follows:

(1) Rule 1: => IF X10 > −0.01 THEN Class = P.

This information of rule indicates that if only the feature of operating income per share (i.e., X10) is more than -0.01 (in NT$), then the EPS of a listed specific stock (i.e., a specific company) will be a positive profit. Thus, this study infers and forecasts that this listed stock will have an uptrend price associated with Rule 1 in the future.

(2) Rule 2: => IF X10 ≦ −0.01 and X10 ≦ −0.20 THEN Class = N.

This information of rule indicates that if the feature of operating income per share is less than or equal to -0.0 and -0.20, then the EPS of a listed specific stock will be a negative profit. Similarly, this study infers and forecasts that this listed stock will have a downtrend price associated with Rule 2 in the future.

(3) Rule 3: => IF X10 ≦ −0.01 and X10 > −0.20 and X10 ≦ −0.07 THEN Class = N.

This information of the rule indicates that if the feature of operating income per share is less than or equal to −0.01 and more than −0.20 and less than or equal to −0.07, then the EPS of a listed specific stock will be a negative profit. Additionally, this study infers and forecasts that this listed stock will have a downtrend price associated with Rule 3 in the future.

Certainly, following the same method, Rules 4–8 in Figure 2 are also formatted in the form "IF . . . THEN . . . " of rule-based knowledge construction.

Furthermore, their explanation for the four rules in Figure 3 is described, as follows:

(1) **Rule 1**: => IF X10 ≦ −0.01 THEN Class = A, or semantically low EPS.

This rule indicates that if only the feature of operating income per share (i.e., X10) is less than −0.01 (in NTD), then the EPS of a listed specific stock will have nonpositive benefits.

(2) **Rule 2**: => IF −0.01 > X10 ≦ 4.23 THEN Class = B, or semantically medium EPS.

This rule indicates that if only the feature of operating income per share is in this range (-0.01 > X10 ≦ 4.23), then the EPS of a listed specific stock will have a result between 0 and 3.75.

(3) **Rule 3**: => IF X10 > 4.23 and X20 ≦ 97.67 THEN Class = B, or semantically medium EPS.

This rule indicates that if the feature of operating income per share is greater than 4.23, and times interest earned (i.e., X20) is equal or less than 97.67, then the EPS of a listed specific stock will have a result between 0 and 3.75.

(4) **Rule 4**: => IF X10 > 4.23 and X20 > 97.67 THEN Class = C, or semantically high EPS.

This rule indicates that if the feature of operating income per share is greater than 4.23, and times interest earned is greater than 97.67, then the EPS of a listed specific stock will have a result higher than 3.75.

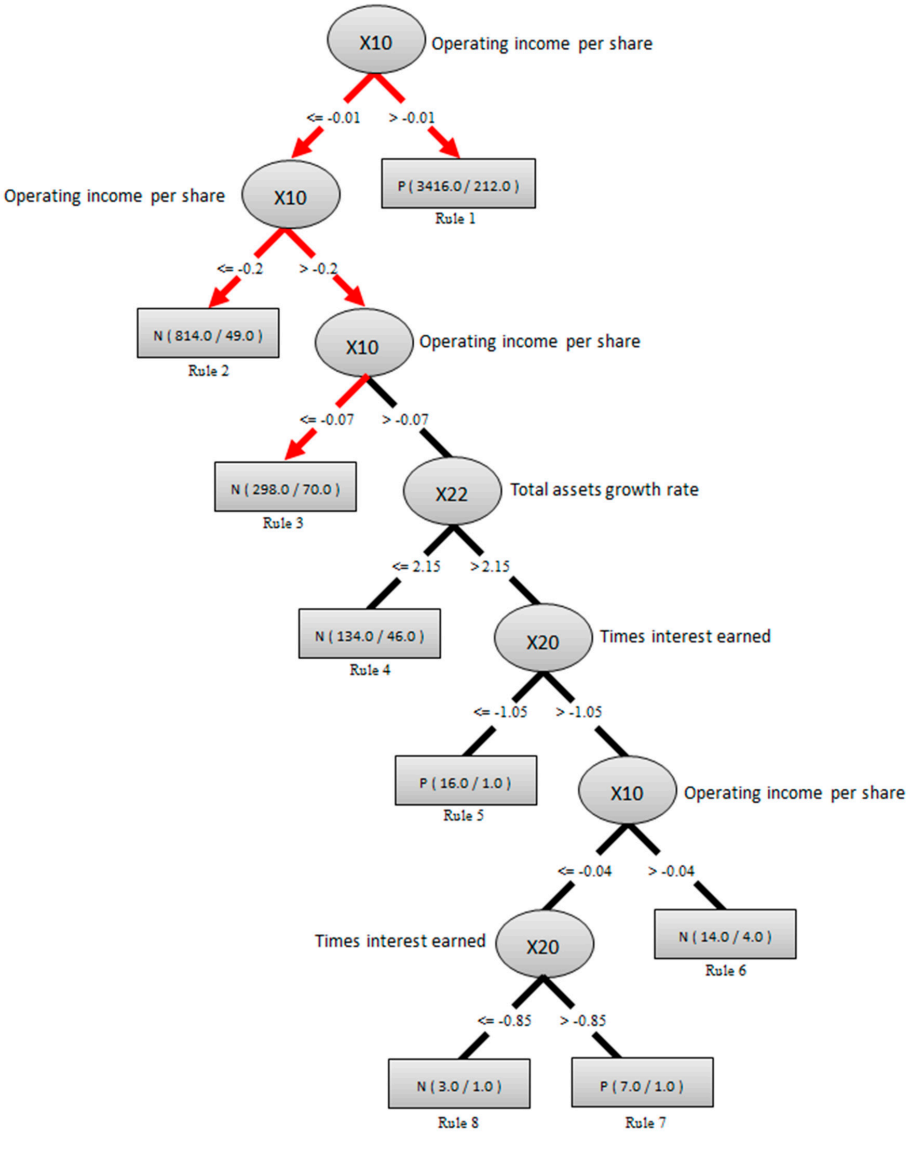


Figure 2. A tree-based rule set of two classes of EPS in the TEJ dataset.

Tables 3–6 illustrate the rule-based knowledge that could predict EPS with an accuracy rate of above 90%, which is a high accuracy to predict real targets. Following the rules in Figures 2 and 3, investors can easily consider and select positive/negative information of listed stock for their well-maintained investment portfolios with a positive affirmation.

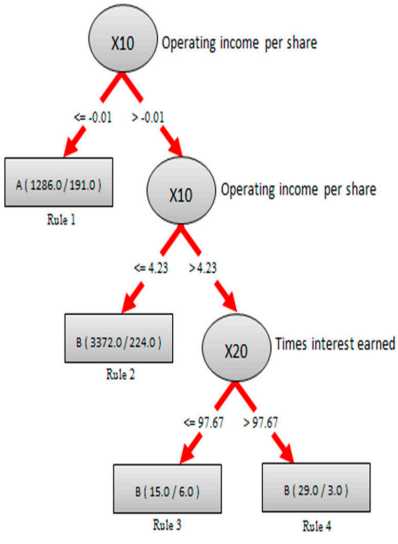


Figure 3. A tree-based rule set in three classes of EPS in the TEJ dataset.

4.3. Helpful Research Findings and Management Implications

To further mine the empirical information hidden in the TEJ dataset, the experimental results yield the following 11 useful research findings.

(1) Benefit of feature-selection: Through the Cfs algorithm for feature-selection in two classes, five features remained and 20 features were removed. In particular, the accuracy of the applied hybrid Models D and E proposed, with the five used features, still had significant performances when compared with the other listed models using the original features under the same settings. Based on Thangavel and Pethalakshmi [63], their postreduction performance of an experiment result showed that the hybrid model using the features after feature-selection had higher classification accuracy than the original features for the same conditions; the experimental results proved that the reasonably practicable solution of feature-selection techniques effectively reduces the features, yielding better accuracy than that of the original feature set.

(2) Feature-selection-based discretization method: After performing the feature-selection function, it was found, interestingly, that some key features were discretized into meaningful linguistic terms (e.g., small, medium, and large) of natural language by data-discretization methods to assist rule-based knowledge construction; concurrently, the functions of combining feature-selection with data-discretization works uncommonly improved classification accuracy.

(3) Important discovery from the functions of feature-selection and data-discretization works: It was found that the execution order for performing the works of feature-selection and data-discretization had clearly indicated no performance difference between the listed two methods of functional components when Tables 3–6 were compared to Tables 7–10, respectively.

(4) Interpretation of constructing the decision rule-based knowledge: This study generates some comprehensible decision-rule-based knowledge to explain how to use the graphic-based information visualization created by the C4.5 algorithm for investors to enable knowledge formation.

(5) Selection of good financial status EPS: From Figures 2 and 3, this study discovered important rules for the positive/negative status of listed stock, which can help investors to manage effective investment strategy.

(6) The best practice of enablers: From the study results, it was found that the best classifier is the decision tree C4.5 algorithm for the TEJ dataset, and the best model is Models D and E in terms of average accuracy.

(7) Interpretation of two data-split methods: It was found that the cross-validation data approach makes more appropriate use of the TEJ dataset than the percentage-split data approach.

(8) Interpretation of different running times of experiments: It was found that the rank of average accuracy for the Models D and E ding212 Model B ding212 Model C ding212 Model A is the same between one run and 10 runs of experiments, regardless of different data-split methods and two types of classes of EPS.

(9) Difference interpretation for different types of classes of EPS: It was found that a difference exists within the key features when different types of classes in the same decision feature of EPS are used.

(10) Effects of time lag for the financial industries: From Tables 11–14, the applied hybrid models proposed have the best classification capacity when the time lag reaches time T+0, and they do not have a deferred effect in the TEJ dataset on some public listed companies of Taiwan's stock market.

(11) Identification of the two most key determinants: From the above experimental results, two key features, operating income per share and times interest earned, are identified, regardless of two or three classes of EPS.

Three meaningful management implications from the empirical results of this study are found and catalogued.

(1) For individual investors: Three major attributes, total assets growth rate, times interest earned, and operating income per share, have been sorted out and deserve more attention from dozens of financial ratios when conducting fundamental analysis of a listed stock. The construction of knowledge-based decision rules provides effective instructs to select a proper target from the sea of considerable potential stocks.

(2) For investment-technology-developing companies: DT-C4.5 and RBFN are two good targets of classifiers suitable for predicting EPS in the TWSE with some empirical evidence. Investment decision support systems based on data-mining techniques deserve further valuable assessments and developments in the future.

(3) For practitioners and academicians: The experimental results are valuable information helpful to them by closing the knowledge gap between theoretical frameworks and evidence-based practices in future research.

4.4. Closing the Gap Report from the Applied Hybrid Models Proposed

Tables 3–14 clearly indicate that this study has four practical values to bridge the gap between past studies and this study, as follows: (1) The practical gap in field applications is filled with a small step for providing applicable remarks and practical experiences to control the performance of profit-making of the associated EPS classification for investors used in the financial environment. (2) The gap of technical services has been closed since the study has made an advanced appropriate hybrid model offered to the stock market application field. (3) The gap of knowledge creation to review applicable decision-rule-based diagnostic systems for identifying EPS has been closed with the successful results in financial ratio verification. (4) Moreover, the study bridges the gap of the lack of literature on EPS application support.

4.5. Research Limitations of the Study

Although the study is empirically valid, the following limitations must be noted and addressed in the future:

(1) A significant shortfall of this study is that the sample used only listed companies on the TWSE. Investor choice, such as companies traded over the counter and emerging the stock market, was

not considered in this study, and there are 969 such companies currently operating, which is quite large.

(2) Another limitation of this study is that the pool of possible predictors is confined to company financial statements, but, for the work of forecasting EPS, a broader range of information sources may be considered.

(3) The financial ratios and variables were predefined and calculated from financial statements of online databases, and researchers should have professional knowledge of this industry background and environment to know the financial topics better.

(4) The classifiers were limited to DT-C4.5, KNNs, STK, RBFN, and NB, which were used and validated. Future studies may take other classification algorithms in hybrid and stand-alone models into account for a generalized application in the identification of EPS for financial diagnosis.

5. Discussion

In the data-mining and machine-learning fields [64–66], so many challenges trigger the researchers' ability. Developing a model to overcome a dilemma in decision-making for financial analysts is necessary. This study proposed an advanced hybrid model [67] to fulfill a valuable research need for classifying EPS with good identification of financial ratios for various categories from large-scale data structures, with the empirical results as desired. However, regarding the key components of the applied hybrid models proposed, it was determined that three directions are required to ascertain the intrinsic needs of the experiences for this study.

First, financial data are continuous and of a big volume. Pal and Kar [68] suggested using the data-discretization technique to improve classification performance. Data-discretization could reduce the number of generated classifying rules, enhance the performance of classifiers, and ease the semantic representation. This study, through model performance comparison and the illustration of decision rule-based knowledge, confirmed Pal and Kar's study [68].

Second, feature-selection is a suggested technique to tackle big data by reducing the complexity of multifaceted data across disciplines [66,69]. Seeja [70] had mentioned several benefits of performing feature-selection functions, such as simplifying the model or accelerating model building and knowledge forming. The results of model performance comparisons and the generated simple DTs in this study were in line with Seeja's proposition [70]. This study has tested the effect of a sequence in applying data-discretization and feature-selection and found no difference between them. It was found that feature-selection had less impact on accuracy than data-discretization did. However, this study conjectured that the performance was determined by the fit between model and data because the reported performance from different model settings remained mixed.

Third, the five classifiers tested were all well-known and reported with good performance in previous studies. However, the performance of the learning classifier is strongly limited for some cases applied, such as the dependence of instance complexity [71], the objective of study [72], the subject experience of users [73], and the application field [74]. In particular, Tabassum [71] indicated that the performance of the classifiers was varied among data domains, such as the framework of the number of instances or attributes used in the experiences from the learning classifiers. It is problematic and difficult to mine suitable learning classifiers with the best performance from experimental instances. Interestingly, this study confirmed Tabassum's study [71]. With reference to the study results, DT-C4.5 and RBFN outperformed other classifiers significantly, indicating that there is no classifier universally good or poor in terms of classification performance. The reason for this is that DT-C4.5 and RBFN performed better than others, which may lie in the inherent nature of the Taiwanese stock market. Therefore, future studies require a greater variety of input data that might further test the proposition of fit between model and data, as well as explore the proper dimensions to describe the implicit nature of data to explain the fit.

6. Conclusions

Classifying EPS with financial ratios and variables has practical and researchable values in financial management and stock investment. This study contributed five hybrid models with advanced multicomponential discretization methods to use different classifiers of ML technologies to find key functions of data-discretization, feature-selection works, two data split methods, rule-based knowledge construction, effects of time lag, different running times of experiments, and two types of different classes in model building. This study also identified that DT-C4.5 and RBFN were relatively efficient classifiers for identifying positive EPS stocks from enormous financial ratio applications in the TEJ dataset for large-scale data analytics and solutions. Simultaneously, total assets growth rate, times interest earned, and operating income per share showed a specific feature-selection advantage in two classes of EPS. Particularly for Tables 7–10, some core results were concluded. (1) Most of Models D and E had higher accuracy rates with a lower standard deviation for the comparison of average accuracy. (2) Models D and E had the best accuracy and the same performance in average accuracy in 10 runs. (3) Feature-selection and data-discretization techniques had abilities for accuracy improvement and the stability of model implementation in the TEJ dataset. (4) Regarding the percentage-split data method, Models A and B had better performance than the cross-validation method; however, Models C–E for the percentage-split data method had worse performance than the cross-validation method. (5) Better accuracy, 91.43%, for DT-C4.5 in Model A was better suited than other classifiers for the cross-validation method. As for the better classification accuracy from all the experiments in Tables 3–14, four empirical results were identified for the TEJ dataset. (1) The highest accuracy of 92.46% occurred in Model C, using decision tree learning with the percentage-split method in two classes for one run. (2) The mean of highest accuracy, 91.44%, occurred in Models D and E using naïve Bayes learning, both with the cross-validation method and the percentage-split method in two classes for 10 runs. (3) The mean of highest average accuracy, 87.53%, occurred in Models D and E with the cross-validation method in two classes. (4) The highest accuracy of 92.46% occurred in Model C using decision tree learning-C4.5 with the percentage-split method and no time lag in two classes. It is feasible, due to the methodology used in this study, that the classifiers learned an acceptable classification accuracy rate and achieved an adequate ranking in the related financial characteristics applied. The proposed method provided a better possible way for a ranking, mentioned above, of the learning classifiers.

Stock selection is a real-world problem with a trade-off between risk and profit because returns from stock investments are risky. This study provides an alternate model to identify a good EPS stock associated with an expected high stock price, with risk aversion adequately and effectively addressed. Hence, in terms of research contribution, this study concludes that various performance measurements of multicomponents for comparison intentions, when searching for a suitable means and tool, are useful for financial analysts. Empirically, the analytical results showed that data-discretization and feature-selection with some classifiers had significantly better accuracy with a satisfactory result. There are five key findings identified and concluded from the empirical results. (1) Model A is the poorest one as it is without the key techniques of data-discretization and feature-selection; the abovementioned information implies that the hybrid models have better performance than that of stand-alone models. (2) Regarding model components used in this study, their performance degree is further assessed under various evaluation criteria for universal application in subsequent research. (3) From Pal and Kar [68], it was found that data first need to be discretized when putting it into a classification model. Thus, the MDL of automatic data-discretization to build thresholds in increasing performance provides just such a prominent example with good satisfaction to validate the importance of using rule-based models in this study. The above-detailed knowledge intensifies the requirement of discretizing data of the conditional features for financial applications. (4) The study results on data-discretization performance are matched and affirmed with the literature [68]. (5) Furthermore, the key data-discretization methods have been used as a useful tool for defining purposeful linguistic terms in natural language for manipulating rule-based knowledge representation. This study conclusively makes a valuable contribution to useful research findings, managerial implications, and technical directions in which multicomponential

discretization models are limited and rarely seen for such an industry setting, from the restricted reviews. This study is a stepping stone on the road towards an advanced data-mining application for financial data resolution [75,76], with some meaningful differentiation from past studies and having sentimental value. Furthermore, this study's originality is in the devising of multimodels with advanced settings to identify the best model and set for anticipating the EPS of Taiwanese companies.

In future studies, there is a need for expansion defined as follows: (1) For example, use the applied hybrid models proposed to analyze datasets from different industries and measure the model-selection performance. (2) It helps address good concerns to validate the same financial datasets, but collected from other countries, to develop different hybrid models, to use different years of financial datasets, and to evaluate the performance difference. (3) Future studies may take a variety of various companies into account for a well-identified application. Thus, other companies from different industries other than those used in this study may be studied for comparison. (4) In addition, different characteristics of features, such as market efficiency, might influence the model-selection of predictors or classifiers, which should be considered in the future. (5) Different discretization data techniques in identifying the EPS feature are worthy of future consideration in the applied hybrid models proposed; likewise, different feature-selection methods should be explored for the same purposes. (6) Various types of decisional features other than two and three classes of financial EPS status are necessary to retest the function of the hybrid models. (7) Extra conditional features differentiated from the study should be measured with the hybrid models. (8) Different time-lag effects should be further identified for measuring later events on the basis of the former ones in order to further validate the application function of the hybrid models. (9) In particular, the plan to add geo-information of firm locations to the TEJ dataset in Taiwan and to test a wider variety of classifiers to compare their performances is an interesting issue since stocks are a famous investment portfolio for global world investors. Taiwan is located over the Taiwan Strait, away from mainland China, with a 180-km distance. Regarding the stock market of mainland China, the three interests of Hong Kong Exchanges and Clearing Limited (HKEx), Shenzhen Stock Exchange (SZSE), and Shanghai Stock Exchange (SSE) are major stock markets with distinct characteristics from TWSE. Many geographically related interesting financial facts and rules about the companies reside in the big data of the four stock markets. Subsequent studies may research the EPS of companies in SSE, SZSE, and HKEx to construct a map of the Chinese stock market with rich features. (10) Exploring the topic is concerned with how these proposed hybrid models increase financial EPS to provide further verification. Additionally, differentiating and learning the function before/after feature-selection and data-discretization techniques in the study is needed in subsequent research.

Author Contributions: Conceptualization, H.-C.H.; methodology, H.-C.H. and Y.-S.C.; software, S.-F.C.; validation, Y.-S.C. and S.-F.C.; formal analysis, S.-F.C.; investigation, Y.-S.C. and A.K.S.; resources, Y.-S.C.; data curation, S.-F.C.; writing—original draft preparation, H.-C.H.; writing—review and editing, Y.-S.C. and A.K.S.; visualization, Y.-S.C. and A.K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of Republic of China, Taiwan, grant number MOST 108-2410-H-146-001.

Acknowledgments: The authors would like to cordially express many thanks for the above financial support of this paper. The authors also sincerely appreciate the editors and the four anonymous referees for their constructive comments and helpful suggestions to improve the quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Allen, K.D.; Winters, D.B. Auditor response to changing risk: Money market funds during the financial crisis. *Rev. Quant. Financ. Account.* **2020**. [CrossRef]
2. Cai, S.; Zhang, J. Exploration of credit risk of P2P platform based on data mining technology. *J. Comput. Appl. Math.* **2020**, *372*, 112718. [CrossRef]

3. Wang, Z.; Yin, J. Risk assessment of inland waterborne transportation using data mining. *Marit. Policy Manag.* **2020**, *47*, 633–648. [CrossRef]
4. Wang, G.; Miao, J. Design of data mining algorithm based on rough entropy for US stock market abnormality. *J. Intell. Fuzzy Syst.* **2020**, 1–9. [CrossRef]
5. Dimitrakopoulos, S.; Kolossiatis, M. Bayesian analysis of moving average stochastic volatility models: Modeling in-mean effects and leverage for financial time series. *Econ. Rev.* **2020**, *39*, 319–343. [CrossRef]
6. Muruganandan, S. Testing the profitability of technical trading rules across market cycles: Evidence from India. *Colombo Bus. J.* **2020**, *11*, 24–46. [CrossRef]
7. Hung, N.H. Various moving average convergence divergence trading strategies: A comparison. *Invest. Manag. Financ. Innov.* **2016**, *13*, 1–7. [CrossRef]
8. Chahine, S.; Malhotra, N.K. Impact of social media strategies on stock price: The case of Twitter. *Eur. J. Mark.* **2018**, *52*, 1526–1549. [CrossRef]
9. Cuestas, J.C.; Huang, Y.S.; Tang, B. Does internationalisation increase exchange rate exposure?—Evidence from Chinese financial firms. *Int. Rev. Financ. Anal.* **2018**, *56*, 253–263. [CrossRef]
10. Mehlawat, M.K.; Kumar, A.; Yadav, S.; Chen, W. Data envelopment analysis based fuzzy multi-objective portfolio selection model involving higher moments. *Inf. Sci.* **2018**, *460–461*, 128–150. [CrossRef]
11. Choi, H.; Son, H.; Kim, C. Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Syst. Appl.* **2018**, *110*, 1–10. [CrossRef]
12. Lu, R.; Wei, Y.C.; Chang, T.Y. The effects and applicability of financial media reports on corporation default ratings. *Int. Rev. Econ. Financ.* **2015**, *36*, 69–87. [CrossRef]
13. Kadim, A.; Sunardi, N.; Husain, T. The modeling firm's value based on financial ratios, intellectual capital and dividend policy. *Accounting* **2020**, *6*, 859–870. [CrossRef]
14. Bagina, R.W. Assessing the financial statement (ratios) of Anglogold-Ashanti Limited, Ghana. *Asian J. Econ. Bus. Account.* **2020**, *14*, 45–55. [CrossRef]
15. Sriram, M. Do firm specific characteristics and industry classification corroborate voluntary disclosure of financial ratios: An empirical investigation of S&P CNX 500 companies. *J. Manag. Gov.* **2020**, *24*, 431–448. [CrossRef]
16. Cengiz, H. The relationship between stock returns and financial ratios in Borsa Istanbul analysed by the classification tree method. *Int. J. Bus. Emerg. Markets* **2020**, *12*, 204–216. [CrossRef]
17. Mita, A.F.; Utama, S.; Fitriany, F.; Wulandari, E.R. The adoption of IFRS, comparability of financial statements and foreign investors' ownership. *Asian Rev. Account.* **2018**, *26*, 391–411. [CrossRef]
18. Rawal, B.; Agarwal, R. Improving accuracy of classification based on C4.5 decision tree algorithm using big data analytics. *Adv. Intell. Syst. Comput.* **2019**, *711*, 203–211.
19. Lee, C.-T.; Horng, S.-C. Abnormality detection of Cast-Resin transformers using the fuzzy logic clustering decision tree. *Energies* **2020**, *13*, 2546. [CrossRef]
20. Ghasemi, E.; Gholizadeh, H.; Adoko, A.C. Evaluation of rockburst occurrence and intensity in underground structures using decision tree approach. *Eng. Comput.* **2020**, *36*, 213–225. [CrossRef]
21. Saadatfar, H.; Khosravi, S.; Joloudari, J.H.; Mosavi, A.; Shamshirband, S. A new K-nearest neighbors classifier for big data based on efficient data pruning. *Mathematics* **2020**, *8*, 286. [CrossRef]
22. Gohari, M.; Eydi, A.M. Modelling of shaft unbalance: Modelling a multi discs rotor using K-Nearest Neighbor and Decision Tree Algorithms. *Measurement* **2020**, *151*, 107253. [CrossRef]
23. Qaddoura, R.; Faris, H.; Aljarah, I. An efficient clustering algorithm based on the k-nearest neighbors with an indexing ratio. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 675–714. [CrossRef]
24. Tran, H.Q.; Ha, C. High precision weighted optimum K-Nearest Neighbors algorithm for indoor visible light positioning applications. *IEEE Access* **2020**, *8*, 114597–114607. [CrossRef]
25. Tjahjadi, H.; Ramli, K. Noninvasive blood pressure classification based on Photoplethysmography using K-Nearest Neighbors algorithm: A feasibility study. *Information* **2020**, *11*, 93. [CrossRef]
26. Fiorentini, N.; Losa, M. Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures* **2020**, *5*, 61. [CrossRef]
27. Cai, W.; Pan, W.; Liu, J.; Chen, Z.; Ming, Z. k-Reciprocal nearest neighbors algorithm for one-class collaborative filtering. *Neurocomputing* **2020**, *381*, 207–216. [CrossRef]

28. Ala'raj, M.; Majdalawieh, M.; Abbod, M.F. Improving binary classification using filtering based on k-NN proximity graphs. *J. Big Data* **2020**, *7*, 15. [CrossRef]
29. Zhang, X.; Han, N.; Qiao, S.; Zhang, Y.; Huang, P.; Peng, J.; Zhou, K.; Yuan, C.-A. Balancing large margin nearest neighbours for imbalanced data. *J. Eng.* **2020**, *2020*, 316–321. [CrossRef]
30. Prajapati, B.P.; Kathiriya, D.R. A hybrid machine learning technique for fusing fast k-NN and training set reduction: Combining both improves the effectiveness of classification. *Adv. Intell. Syst. Comput.* **2019**, *714*, 229–240.
31. Jiang, M.; Liu, J.; Zhang, L.; Liu, C. An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Phys. A Stat. Mech. Appl.* **2020**, *541*, 122272. [CrossRef]
32. Pisula, T. An ensemble classifier-based scoring model for predicting bankruptcy of polish companies in the Podkarpackie Voivodeship. *J. Risk Financ. Manag.* **2020**, *13*, 37. [CrossRef]
33. Soui, M.; Smiti, S.; Mkaouer, M.W.; Ejbali, R. Bankruptcy prediction using stacked auto-encoders. *Appl. Artif. Intell.* **2020**, *34*, 80–100. [CrossRef]
34. García, V.; Marqués, A.I.; Sánchez, J.S. Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Inf. Fusion* **2019**, *47*, 88–101. [CrossRef]
35. Liang, D.; Tsai, C.F.; Lu, H.Y.R.; Chang, L.S. Combining corporate governance indicators with stacking ensembles for financial distress prediction. *J. Bus. Res.* **2020**, *120*, 137–146. [CrossRef]
36. Khan, W.; Ghazanfar, M.A.; Azam, M.A.; Karami, A.; Alyoubi, K.H.; Alfakeeh, A.S. Stock market prediction using machine learning classifiers and social media, news. *J. Ambient Intell. Hum. Comput.* **2020**. [CrossRef]
37. Saha, M.; Santara, A.; Mitra, P.; Chakraborty, A.; Nanjundiah, R.S. Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. *Int. J. Forecast.* **2020**. [CrossRef]
38. Patil, P.R.; Sivagami, M. Forest cover classification using stacking of ensemble learning and neural networks. In *Artificial Intelligence and Evolutionary Computations in Engineering Systems. Advances in Intelligent Systems and Computing*; Dash, S., Lakshmi, C., Das, S., Panigrahi, B., Eds.; Springer: Singapore, 2020; Volume 1056, pp. 89–102.
39. Zheng, S.; Zhao, J. A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis. *Comput. Chem. Eng.* **2020**, *135*, 106755. [CrossRef]
40. Liu, H.; Long, Z. An improved deep learning model for predicting stock market price time series. *Digital Signal Process.* **2020**, *102*, 102741. [CrossRef]
41. Ribeiro, M.H.D.M.; dos Santos Coelho, L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Appl. Soft Comput.* **2020**, *86*, 105837. [CrossRef]
42. Kanazawa, N. Radial basis functions neural networks for nonlinear time series analysis and time-varying effects of supply shocks. *J. Macroecon.* **2020**, *64*, 103210. [CrossRef]
43. Mansor, M.A.; Mohd Jamaludin, S.Z.; Mohd Kasihmuddin, M.S.; Alzaeemi, S.A.; Md Basir, M.F.; Sathasivam, S. Systematic boolean satisfiability programming in radial basis function neural network. *Processes* **2020**, *8*, 214. [CrossRef]
44. Teixeira Zavadzki de Pauli, S.; Kleina, M.; Bonat, W.H. Comparing artificial neural network architectures for Brazilian stock market prediction. *Ann. Data Sci.* **2020**. [CrossRef]
45. Mirjalili, S. Evolutionary radial basis function networks. *Stud. Comput. Intell.* **2019**, *780*, 105–139.
46. Buhmann, M.; Jäger, J. Multiply monotone functions for radial basis function interpolation: Extensions and new kernels. *J. Approx. Theory* **2020**, *256*, 105434. [CrossRef]
47. Karimi, N.; Kazem, S.; Ahmadian, D.; Adibi, H.; Ballestra, L.V. On a generalized Gaussian radial basis function: Analysis and applications. *Eng. Anal. Bound. Elem.* **2020**, *112*, 46–57. [CrossRef]
48. Soradi-Zeid, S. Efficient radial basis functions approaches for solving a class of fractional optimal control problems. *Comput. Appl. Math.* **2020**, *39*, 20. [CrossRef]
49. Nabipour, M.; Nayyeri, P.; Jabani, H.; Shahab, S.; Mosavi, A. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data: A comparative analysis. *IEEE Access* **2020**, *8*, 150199–150212. [CrossRef]
50. Vismayaa, V.; Pooja, K.R.; Alekhya, A.; Malavika, C.N.; Nair, B.B.; Kumar, P.N. Classifier based stock trading recommender systems for Indian stocks: An empirical evaluation. *Comput. Econ.* **2020**, *55*, 901–923. [CrossRef]

51. Bhandare, Y.; Bharsawade, S.; Nayyar, D.; Phadtare, O.; Gore, D. SMART: Stock Market Analyst Rating Technique Using Naive Bayes Classifier. In Proceedings of the 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 5–7 June 2020; pp. 1–4.
52. Rahul; Sarangi, S.; Kedia, P.; Monika. Analysis of various approaches for stock market prediction. *J. Stat. Manag. Syst.* **2020**, *23*, 285–293.
53. Ahmed, M.; Sriram, A.; Singh, S. Short term firm-specific stock forecasting with BDI framework. *Comput. Econ.* **2020**, *55*, 745–778. [CrossRef]
54. Chen, W.; Zhang, S.; Li, R.; Shahabi, H. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *Sci. Total Environ.* **2018**, *644*, 1006–1018. [CrossRef] [PubMed]
55. Nascimento, A.C.A.; Prudêncio, R.B.C.; Costa, I.G. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinf.* **2016**, *17*, 17–46. [CrossRef] [PubMed]
56. Tripathy, A.; Anand, A.; Rath, S.K. Document-level sentiment classification using hybrid machine learning approach. *Knowl. Inf. Syst.* **2017**, 1–27. [CrossRef]
57. Shon, H.S.; Batbaatar, E.; Kim, K.O.; Cha, E.J.; Kim, K.-A. Classification of kidney cancer data using cost-sensitive hybrid deep learning approach. *Symmetry* **2020**, *12*, 154. [CrossRef]
58. Liu, J.; Wang, Y.; Zhang, Y. A novel Isomap-SVR soft sensor model and its application in rotary kiln calcination zone temperature prediction. *Symmetry* **2020**, *12*, 167. [CrossRef]
59. Taiwan Economic Journal Website. Available online: http://www.tej.com.tw/twsite/Default.aspx?TabId=186 (accessed on 31 January 2020).
60. Džeroski, S.; Zenko, B. Is combining classifiers with stacking better than selecting the best one? *Mach. Learn.* **2004**, *54*, 255–273. [CrossRef]
61. Chen, Y.S. An empirical study of a hybrid imbalanced-class DT-RST classification procedure to elucidate therapeutic effects in uremia patients. *Med. Biol. Eng. Comput.* **2016**, *54*, 983–1001. [CrossRef]
62. Chen, Y.S. A comprehensive identification-evidence based alternative for HIV/AIDS treatment with HAART in the healthcare industries. *Comput. Methods Programs Biomed.* **2016**, *131*, 111–126. [CrossRef]
63. Thangavel, K.; Pethalakshmi, A. Dimensionality reduction based on rough set theory: A review. *Appl. Soft Comput.* **2009**, *9*, 1–12. [CrossRef]
64. Kuang, Y.; Wu, Q.; Shao, J.; Wu, J.; Wu, X. Extreme learning machine classification method for lower limb movement recognition. *Cluster Comput.* **2017**, *20*, 3051–3059. [CrossRef]
65. Ren, X.; Li, L.; Yu, Y.; Xiong, Z.; Yang, S.; Du, W.; Ren, M. A simplified climate change model and extreme weather model based on a machine learning method. *Symmetry* **2020**, *12*, 139. [CrossRef]
66. Alabdulwahab, S.; Moon, B. Feature selection methods simultaneously improve the detection accuracy and model building time of machine learning classifiers. *Symmetry* **2020**, *12*, 1424. [CrossRef]
67. Wu, Q.; Wang, L.; Zhu, Z. Research of pre-stack AVO elastic parameter inversion problem based on hybrid genetic algorithm. *Cluster Comput.* **2017**, *20*, 3173–3183. [CrossRef]
68. Pal, S.S.; Kar, S. Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory. *Math. Comput. Simul.* **2019**, *162*, 18–30. [CrossRef]
69. Balogun, A.O.; Basri, S.; Mahamad, S.; Abdulkadir, S.J.; Almomani, M.A.; Adeyemo, V.E.; Al-Tashi, Q.; Mojeed, H.A.; Imam, A.A.; Bajeh, A.O. Impact of feature selection methods on the predictive performance of software defect prediction models: An extensive empirical study. *Symmetry* **2020**, *12*, 1147. [CrossRef]
70. Seeja, K.R. Feature selection based on closed frequent itemset mining: A case study on SAGE data classification. *Neurocomputing* **2015**, *151*, 1027–1032. [CrossRef]
71. Tabassum, H. Enactment ranking of supervised algorithms dependence of data splitting algorithms: A case study of real datasets. *Int. J. Comput. Sci. Inf. Technol.* **2020**, *12*, 1–8. [CrossRef]
72. Fan, H.; Mark, A.E.; Zhu, J.; Honig, B. Comparative study of generalized born models: Protein dynamics. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6760–6764. [CrossRef]
73. Barber, S. Creating effective load models for performance testing with incomplete empirical data. In Proceedings of the Sixth IEEE International Workshop, Chicago, IL, USA, 11 September 2004; pp. 51–59.

74. Chen, C.C. A model for customer-focused objective-based performance evaluation of logistics service providers. *Asia Pac. J. Mark. Logist.* **2008**, *20*, 309–322. [CrossRef]
75. Li, Z.; Gan, S.; Jia, R.; Fang, J. Capture-removal model sampling estimation based on big data. *Cluster Comput.* **2017**, *20*, 949–957. [CrossRef]
76. Wu, Y.; Guo, Y.; Liu, L.; Huang, N.; Wang, L. Trend analysis of variations in carbon stock using stock big data. *Cluster Comput.* **2017**, *20*, 989–1005. [CrossRef]

symmetry

MDPI

Article

Telemedicine Acceptance during the COVID-19 Pandemic: An Empirical Example of Robust Consistent Partial Least Squares Path Modeling

Patricio Ramírez-Correa [1,*], Catalina Ramírez-Rivas [1], Jorge Alfaro-Pérez [1] and Ari Melo-Mariano [2]

[1] School of Engineering, Universidad Católica del Norte, Coquimbo 1781421, Chile; catalina.ramirez@alumnos.ucn.cl (C.R.-R.); jalfaro@ucn.cl (J.A.-P.)

[2] Department of Production Engineering, Universidade de Brasília, Campus Darcy Ribeiro Asa Norte, Brasília 04457, Brazil; arimariano@unb.br

* Correspondence: patricio.ramirez@ucn.cl

Received: 25 August 2020; Accepted: 19 September 2020; Published: 25 September 2020

Abstract: The explanation of behaviors concerning telemedicine acceptance is an evolving area of study. This topic is currently more critical than ever, given that the COVID-19 pandemic is making resources scarcer within the health industry. The objective of this study is to determine which model, the Theory of Planned Behavior or the Technology Acceptance Model, provides greater explanatory power for the adoption of telemedicine addressing outlier-associated bias. We carried out an online survey of patients. The data obtained through the survey were analyzed using both consistent partial least squares path modeling (PLSc) and robust PLSc. The latter used a robust estimator designed for elliptically symmetric unimodal distribution. Both estimation techniques led to similar results, without inconsistencies in interpretation. In short, the results indicate that the Theory of Planned Behavior Model provides a significant explanatory power. Furthermore, the findings show that attitude has the most substantial direct effect on behavioral intention to use telemedicine systems.

Keywords: telemedicine; technology acceptance; robust partial least squares path modeling

1. Introduction

Partial least squares path modeling (PLS) has been widely used to analyze data associated with complex phenomena [1]. The characteristics of PLS have managed to be seen by some social sciences researchers as a fundamental tool to try to explain causal relationships among concepts of the real world [2]. Many enhancements have been incorporated into PLS throughout the years. Among them, it is worth mentioning the following, multigroup analysis [3], identifying and treating unobserved heterogeneity [4], measures of model fit [5], predictive power assessment [6], and consistent PLS (PLSc) [7]. Despite the several enrichments of PLS [8], handling outliers in the context of PLS has been broadly ignored [9]. Johnson and Wichern [10] referred to an outlier as an observation in a dataset that appeared to be inconsistent with the rest of that dataset.

Commonly, two types of outliers are observed. Some outliers arise following no pattern, i.e., unsystematic outliers. Other outliers arise systematically, being part of a population different from the rest of the observations [11]. Considering that outliers are often found in empirical social sciences research, ignoring outliers is extraordinarily likely to lead to inaccurate results and debatable conclusions. Considering the above, robust PLS has recently been proposed to address this problem [9]. A highly robust estimator designed for elliptically symmetric unimodal distributions is central to this proposal. This option is considered to be a better approach for only identifying and manually removing outliers, which has two drawbacks. First, outliers may not be easily identified by visualization or

statistical methods. Second, even if this is possible, removing outliers would imply information lost and the sample size decreasing [9]. On the basis of the robust PLS proposal, this study is aimed at evaluating a social phenomenon where the analysis should be as free as possible from outlier-related bias. This research addresses a current social phenomenon, telemedicine acceptance during the COVID-19 pandemic, comparing the two known models, the Theory of Planned Behavior (TPB) and the Technology Acceptance Model (TAM). Therefore, in the following paragraphs, we develop both the telemedicine and technology acceptance concepts.

First, telemedicine refers to healthcare services provided by healthcare providers in a patient-centered manner, from a geographical distance, and using digital technologies [12]. Over the last decade, a technology shift has created a rise in the accessibility to technology and mobile services, including mobile health services [13]. However, although telemedicine technology has been in use for over five decades, it has still not moved past a pilot stage, with traditional in-person service preferred [14]. Global statistics back up this claim, as only ten percent of people have ever used telemedicine. Within this group of people, their approval level is positive, with two out of three individuals stating they would use the service again [15]. The usage of telemedicine is not the same across the globe. There is higher usage in developing countries within Asia and the Middle East (31% in Saudi Arabia, 27% in India, 24% in China, and 15% in Malaysia). However, in Europe, telemedicine is less common (2–4% in Belgium, Serbia, Russia, France, Spain, and Hungary) [15]. The current global COVID-19 crisis adds a new layer to the literature surrounding telemedicine and its usage. The onset of the virus has highlighted the ability of health providers to manage patient visits triaged to telemedicine services. It has also shown the importance of connectivity and how quickly the logistics behind these services could be put into place [16]. Telemedicine allows patients with mild conditions to obtain the attention that they need while minimizing their exposure to other patients with more severe conditions [17]. The ability to support healthcare workers during this time is a significant focus, as they are battling with pressure from the virus, which not only presents itself as a high rate of occupied resources but also a high rate of resources being removed due to exposure [18]. Concern regarding this quick spread of telemedicine is related to how long the measures in place will last past the pandemic. While the pre-pandemic adoption was not high, the telemedicine model greatly benefits both the patient and the provider from a business standpoint (e.g., [19]). Providers with better telemedicine services aim to gain a better competitive advantage, from which patients can significantly benefit [20]. This competitive advantage is more critical than ever at a time when governments are struggling to minimize both the death toll and the virus' economic impact [21].

Second, multiple authors have explored telemedicine acceptance using models rooted in technology acceptance theories or behavioral theories [14]. In general, these studies indicate that technology acceptance models perform better than behavioral models when it comes to telemedicine acceptance [14,22,23]. The TPB and the TAM are the two most popular models to explain the use of systems [24–26] and, in particular, within the adoption of telemedicine systems, their utilization has been highlighted separately [27–31] or in a complementary way [32]. Previous ideas led us to choose TPB and TAM in the present study as a research framework. The TPB originated from the Theory of Reasoned Action (TRA) [33]. The TRA proposed that attitude toward behavior and subjective norms surrounding that behavior directly affects the individual's behavioral intention. Attitude relates to how individuals perceive behavior. If the behavior is perceived as beneficial to themselves, they are more likely to partake in the behavior. Social norms are the way that individuals perceive others' beliefs regarding their behavior. If individuals see the behavior as viewed to be beneficial by those around them, then, they are more likely to partake in the behavior. Lastly, behavioral intention is how likely they are to participate in the observed behavior. The TPB adds a new concept to the TRA, i.e., perceived behavioral control [25]. Perceived behavioral control is the individual's perceived ability to perform the observed behavior. It considers if the individual believes that participating in this behavior is within their capabilities. If they believe that the behavior is within their reach, then they are likely to have a higher intention to take part in the behavior. Similar to the TPB, the TAM proposed by Fred

Davis [25] also had its roots in the TRA. The TAM looks at how users accept a technology through the same measure as the TRA and the TPB, i.e., behavioral intention. However, the TAM proposes different variables in order to predict behavioral intention. In the TAM, attitude, perceived use, and perceived ease-of-use are used to measure the individual's behavioral intention to use technology. Perceived use relates to how individuals perceive that the technology will be useful to them; perceived ease-of-use is how much effort the individual perceives that the technology requires to use it [25]. The TPB and the TAM both assume that once an individual develops an intention to partake in a behavior or use technology, they can carry out this behavior. This intention is the most significant predictor of this occurrence [25]. Since models are abstractions of a phenomenon within a context, their explanatory capabilities must be systematically tested to determine their usefulness in new settings. Therefore, comparing which model best explains telemedicine adoption in the current context emerges as a necessary action.

In this context, the objective of this study is to determine which model, TPB or TAM, provides greater explanatory power for the adoption of telemedicine addressing outlier-associated bias.

The main contributions of this study are three-fold. First, from a practical viewpoint, this research provides empirical evidence of the application of the robust PLS proposal to test the outlier bias effects in a PLS model based on primary data. Second, from an academic viewpoint, this study contributes by testing the technology acceptance theories' applicability in a new social context, validating the circumstances where these theories can be supported. Third, from a social perspective, this study gives an exploratory baseline to define public policies that support telemedicine implementation in a pandemic context.

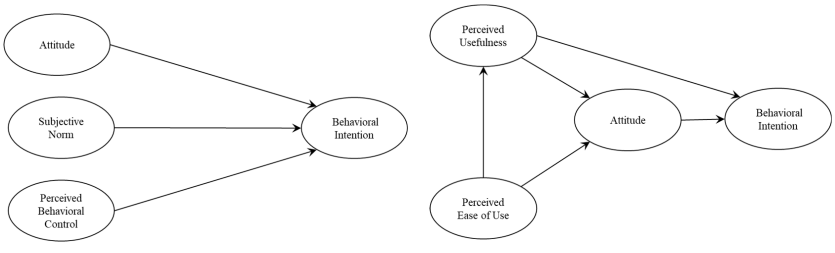
The organization of this paper is as follows: In Section 2, we explain the data collection process and methods used to analyze the data; in Section 3, We present the results of this data analysis; in Section 4, we offer a discussion of these results; and finally, in the last section, we provide a summary of the outcome of this study.

2. Methods

2.1. Data

A cross-sectional study was carried out between January and June 2020. A convenience sampling technique was used to collect data from Brazilian adults. The anonymity of the respondents was guaranteed in the data collection process. According to standard socioeconomic studies, no ethical concerns were involved other than preserving the participants' anonymity.

Specifically, the data was obtained through an online questionnaire for current and future adult telemedicine users in Brasilia. The scales were adapted from Jen and Hung [32]. A 7-point Likert scale was used with answers ranging from 1 (strongly disagree) to 7 (strongly agree). Table 1 shows the questions that were included in the online questionnaire. Figure 1A,B represents the variables and relationships associated with the models under study.



(**A**) TPB Model (**B**) TAM Model

Figure 1. (**A**) Theory of Planned Behavior (TPB) model; (**B**) Technology Acceptance Model (TAM) model.

Table 1. Questions included in the study questionnaire.

Latent Variable	Item	Description
Subjective norms	SN1	The experts who influence my behavior would think that I should use telemedicine services.
	SN2	My family would think that I should order the telemedicine service.
	SN3	My friends would think that I should order the telemedicine service.
Perceived behavioral control	PBC1	I have the knowledge and ability to operate the telemedicine service.
	PBC2	I think I can handle the telemedicine service.
	PBC3	Using the telemedicine service is entirely within my control.
Attitude	ATT1	Using the telemedicine service is a good idea.
	ATT2	The telemedicine service increases the healthcare service quality.
	ATT3	The adoption of telemedicine reduces the risks associated with health
	ATT4	The telemedicine service is valuable.
Perceived usefulness	PU1	The telemedicine service will be beneficial to the care of people.
	PU4	Using the telemedicine service will reduce the psychological burden of people.
	PU3	The advantages of the telemedicine service will outweigh the disadvantages.
Perceived ease of use	PEOU1	Instructions for using equipment in the telemedicine service will be easy to follow.
	PEOU2	It will be easy to learn how to use the telemedicine service.
	PEOU3	It will be easy for people to operate the equipment in the telemedicine service.
Behavioral intention	BI1	I am glad to present the telemedicine service to my close ones.
	BI2	I will adopt the telemedicine service.
	BI3	I will adopt the telemedicine service based on my close ones' necessities.

2.2. *Partial Least Squares Path Modeling and Robust Partial Least Squares Path Modeling*

Traditional and robust PLS were utilized to test the proposed research models. Two models define PLS, i.e., the measurement model and the structural model [34]. The first model examines the instrument's reliability and validity, and the second model evaluates the relationships among the latent variables. Figure 2 shows the PLS algorithm; a detailed description of the algorithm can be found in [35].

In the PLS procedure, a Pearson correlation matrix is a relevant input, even though Pearson estimates are highly sensitive to unsystematic outliers, which can finally conclude in distorted PLS results. To cope with this shortcoming, Schamberger et al. proposed using a robust correlation coefficient to define a robust PLS [9]. The minimum covariance determinant (MCD) was central to their approach [36]. The MCD estimator is a highly robust estimator of multivariate location and scatter, being the one with the highest asymptotic breakdown point (BP), see Figure 3. The MCD is designed for elliptically symmetric unimodal distributions. The MCD has been used to develop robust multivariate techniques, such as principal component analysis, factor analysis, and multiple regression [37]. In summary, the MCD coefficient estimates the variance-covariance matrix of a sample set based on a subsample of the total observations with the smallest positive determinant. The robust

PLS algorithm uses the MCD correlation as an input, maintaining unaltered the subsequent PLS steps, and therefore confronts the outlier issues without removing them from the sample set [9].

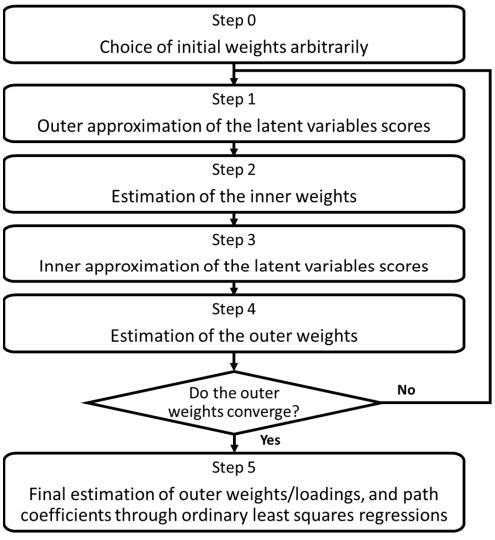


Figure 2. Flowchart of the partial least squares path modeling (PLS) algorithm.

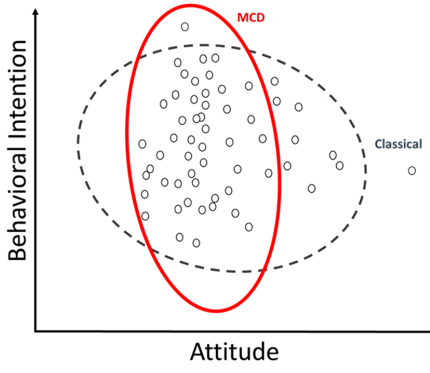


Figure 3. Compare classical and minimum covariance determinant (MCD) covariances.

All calculations were performed using the statistical programming environment R [38]. In particular, an ad hoc script was built based on the simplePLS function from the SEMinR package [39] to integrate the MCD correlation. MCD estimates were determined by the cov.rob function from the MASS package [40]; Figure 4 shows the MCD algorithm. These modifications affect Steps 2 and 4 of the PLS algorithm. Since the models contained common factors and followed the literature [41], the consistent PLS (PLSc) method was applied. The PLSc method applies a correction for attenuation to consistently estimate factor loadings and path coefficients among common factors [7,42]. The following section shows the results associated with these analyses for the empirical study.

$i \leftarrow 1$
repeat
Uniformly sample an initial subset
Compute T_i (mean matrix) and S_i (covariance matrix)
Compute the determinant of S_i. *det* (S_i)
until *det* (S_i) <> 0
repeat
$i \leftarrow i + 1$
Compute the Mahalanobis distance for all *n* points
Construct a new subset which contains samples with smaller distances
Update the estimates of T_i and S_i using the extracted samples
Compute *det* (S_i)
until *det* (S_i) = *det* (S_{i-1}) **or** *det* (S_i) = 0
Compute the mean and covariance only using select samples

Figure 4. Pseudocode of the MCD algorithm.

2.3. Statistical Analysis Plan

First, a primary analysis was carried out. This analysis consisted of the description of the characteristics of participants in the study and the preliminary evaluation of data using descriptive statistics. Next, a PLS analysis was carried out [34] which consisted of two broad phases. These phases applied to both the traditional [7] and the robust PLSc method [9]. The first phase was the measurement model analysis of TPB and TAM. Two analyses were carried out in this phase. First, the reliability analysis of the indicators and constructs associated with the models; second, we analyzed the convergent and discriminant validity of these same constructs. The second phase was the structural model analysis of TBP and TAM. This phase evaluated the relationships among the variables, considering the determination coefficients and the strength of the relationships. Finally, a resampling procedure evaluated the statistical significance of the estimates associated with the strength of the relationships.

3. Results

3.1. Primary Analysis

A total of 200 surveys were completed for the study. The majority of the completed surveys were from males (56%), and the average age was 39.9 years old. See Table 2 for more details of the distribution of the variables of interest.

Table 2. Distribution of the variables of interest.

Variable	N	%
Gender		
Male	111	56
Female	89	44
Total	200	100
Age	Mean 39.9 ± 16.65 Range 18–85 years	

Table 3 shows the descriptive statistics of the items that integrate the measurement models for TPB and TAM.

Table 3. Descriptive statistics.

Item	Average	SD	Asymmetry	Kurtosis
SN1	3.68	1.403	−0.055	0.020
SN2	3.52	1.378	−0.462	−0.590
SN3	3.61	1.421	−0.326	−0.526
PBC1	4.24	1.184	−0.174	−0.090
PBC2	4.46	1.267	−0.231	−0.064
PBC3	4.33	1.216	−0.148	−0.149
ATT1	5.00	1.315	−0.724	0.836
ATT2	4.66	1.358	−0.555	0.246
ATT3	4.21	1.286	−0.456	0.140
ATT4	4.98	1.260	−0.632	1.260
PU1	4.97	1.361	−0.618	0.591
PU4	3.96	1.256	−0.093	0.367
PU3	4.46	1.424	−0.314	−0.266
PEOU1	4.52	1.613	−0.530	−0.244
PEOU2	4.98	1.428	−0.675	0.347
PEOU3	4.64	1.698	−0.606	−0.211
BI1	4.23	1.448	−0.398	0.088
BI2	4.60	1.315	−0.356	0.241
BI3	3.99	1.470	−0.280	−0.141

3.2. Partial Least Squares Path Modeling (PLS) Analysis

3.2.1. Measurement Models Analysis

Table 4 indicates the assessment of the measurement models. The table shows the following two indicators: (1) Composite reliability which is a measure of internal consistency reliability that does not assume equal indicator loadings, in their place, it considers indicator loadings in its calculation and values greater than 0.7 are adequate [34] and (2) Average variance extracted (AVE) which is a measure of convergent validity, defined as the degree to which a construct explains the variance of its indicators, values exceeding 0.5 are acceptable [34]. In addition, the discriminant validity assessment using the Fornell–Larcker criterion indicates acceptable values [34].

Table 4. Assessment of the measurement models.

Model/Latent Variable	Traditional PLSc		Robust PLSc	
	Composite Reliability	AVE	Composite Reliability	AVE
TPB				
Behavioral intention	0.822	0.588	0.834	0.594
Attitude	0.908	0.693	0.910	0.694
Subjective norms	0.901	0.744	0.897	0.743
Perceived behavioral control	0.906	0.747	0.912	0.751
TAM				
Behavioral intention	0.824	0.589	0.834	0.594
Attitude	0.905	0.692	0.906	0.692
Perceived usefulness	0.869	0.646	0.914	0.652
Perceived ease of use	0.905	0.740	0.896	0.739

3.2.2. Structural Models Analysis

To indicate an intermediate result of the structural analysis, Figure 5 shows the plot of the score values for the two models, at the top with traditional PLSc and the bottom with robust PLSc.

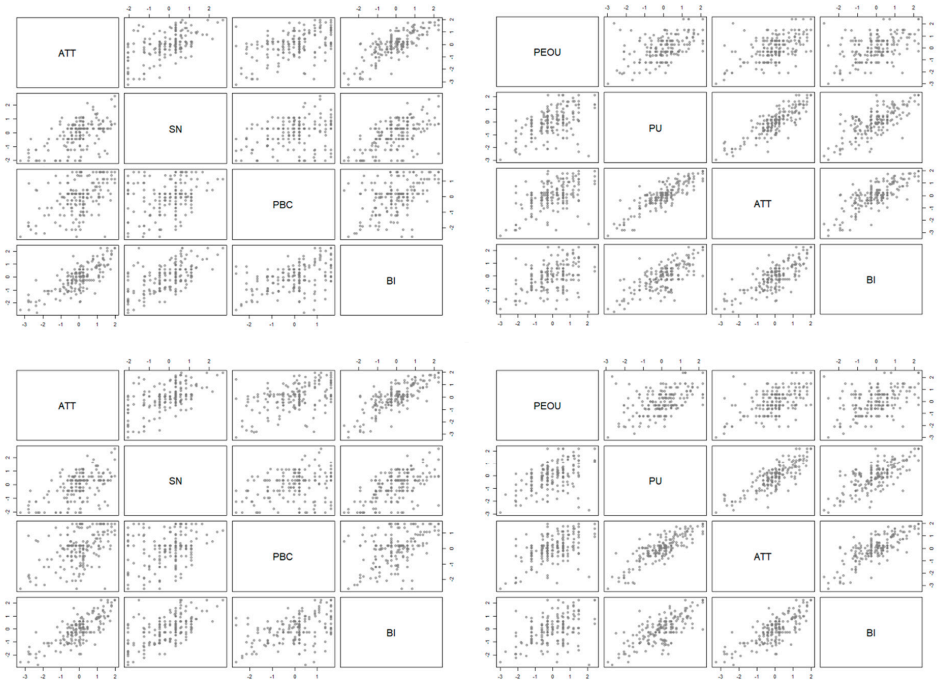


Figure 5. The plot of the score values for the two models. Traditional consistent partial least squares path modeling (PLSc) at the top and robust PLSc at the bottom. ATT, attitude; SN, subjective norms; PBC, perceived behavioral control; BI, behavioral intention; PEOU, perceived ease of use; PU, perceived usefulness.

The results concerning the analysis of the two research models are indicated in Table 5a,b. The coefficient of determination (R^2) indicates the amount of the variance of the dependent variables that is explained by the variables that predict it. The path coefficients (β) express the extent to which the independent variables contribute to the explained variance of the dependent variables. The significance of the β coefficients was calculated using bootstrapping. Bootstrapping is a resampling technique used to determine standard errors of coefficient estimates to evaluate the coefficient's statistical significance without relying on distributional assumptions. We used 999 bootstrap samples. Both estimation techniques lead to similar results without inconsistencies in interpretation. The analysis using traditional PLSc indicates that the TPB explains 85.8% of the intention to use telemedicine, whereas the TAM explains 81.5%. The analysis using robust PLSc indicates very similar results (84.5% versus 80.8% with TAM). In both methods, the TAM analysis indicates that none of the variables in that model which explain behavioral intention is statistically significant.

Table 5. (**a**) Structural results (coefficient of determination). (**b**) Structural results (path coefficients).

(**a**)

Model/Independent Variable	Traditional PLSc	Robust PLSc
	R^2	R^2
TPB		
Behavioral intention	0.858	0.845
TAM		
Perceived usefulness	0.336	0.310
Attitude	0.932	0.840
Behavioral intention	0.815	0.808

(**b**)

Model/Relationship	Traditional PLSc					Robust PLSc				
	Original	Boot Mean	Boot SD	Perc 0.025	Perc 0.975	Original	Boot Mean	Boot SD	Perc 0.025	Perc 0.975
TPB										
Attitude -> behavioral intention	0.713	0.707	0.075	0.546	0.838	0.712	0.709	0.079	0.534	0.851
Subjective norms -> behavioral intention	0.243	0.248	0.075	0.107	0.395	0.240	0.249	0.076	0.110	0.389
Perceived behavioral control -> behavioral intention	0.084	0.087	0.061	−0.034	0.206	0.080	0.084	0.064	−0.034	0.216
TAM										
Perceived ease of use -> perceived usefulness	0.579	0.576	0.084	0.403	0.732	0.557	0.580	0.084	0.397	0.729
Perceived ease of use -> attitude	0.031	0.025	0.075	−0.136	0.168	0.103	0.020	0.077	−0.134	0.169
Perceived usefulness -> attitude	0.947	0.953	0.055	0.843	1.063	0.855	0.956	0.058	0.844	1.071
Perceived usefulness -> behavioral intention	0.044	0.214	6.813	−2.819	2.563	0.121	−0.140	10.396	−2.215	2.743
Attitude -> behavioral intention	0.860	0.689	6.814	−1.634	3.691	0.787	1.043	10.395	−1.850	3.092

4. Discussion

This study was presented as the proper context to justify the use of robust PLSc. In this sense, we must highlight two elements. First, the differences between the results of both techniques are minimal, which rules out bias due to outliers in the models' estimations. Although all the estimates decrease when robust PLSc is applied instead of traditional PLSc, the variations are minimal. On the one hand, for the TPB model, the determination coefficient of the behavioral intention varies by 1.5%, and the maximum variation in the path coefficient is 1.2%. On the other hand, for the TAM model, the maximum variation in the path coefficients is 9.7%. Moreover, while this makes the following paragraphs of this discussion possible, we believe that this result is partly due to the TPB model's parsimony and broad application. Second, the robust PLSc approach has a significant challenge related to the extended computation time of the estimates, especially in the bootstrapping process. This problem calls into question the use of this technique beyond exploratory purposes if the sample size is large.

Our results show that the TPB model has significant explanatory power, while the TAM model does not. This outcome indicates that in the sample context, the TPB model is more parsimonious than the TAM model, meaning that we can have significant results with fewer measures. The TAM model does not explain the behavioral intention of using telemedicine. One possible explication is related to the fact that the current study is based on a concept to use technology rather than a demonstration of the technology itself. Since the TAM variables rely on the perceived usefulness and ease of use, the lack of specifications with respect to what the technology will look like could affect these results. Another possible explanation is related to telemedicine being a broader field than just the technology. There are more external variables that affect the behavioral intention to participate in telemedicine. Last but not least, the application of non-consistent PLS methods could be the cause of explanatory power lacking; the literature provides examples of the application of these methods [43,44].

The TPB-based results highlight four points. First, the determination coefficient of the behavioral intention variable (R^2 = 0.84), which results from applying robust PLSc, can be described as substantial [2]. This result implies that its predictor variables determine a high variability of the behavioral intention construct. This result must be supported by recent studies about the adoption of telemedicine in emerging countries, however, in general, the explanatory power of these studies has been moderate. This characteristic is evidenced in the following examples. On the basis of a sample of physicians and nurses in public hospitals in Malaysia, a TAM-based model explained 41.5% of the acceptance of telemedicine [45]. In Nigeria, using the data of physicians and nurses, a model based on the unified theory of acceptance and use of technology explained 49.7% of the variation in intention to use telemedicine [46]. On the basis of a sample of Pakistani patients, a TAM model explained a total of 62% of the variance of the intention to use telemedicine [43]. Second, the attitude variable was the most significant predictor of behavioral intention (β = 0.71, robust PLSc). This result is concordant with previous patient-based research [47]. Third, the subjective norms variable was a significative predictor of behavioral intention (β = 0.24, robust PLSc). In previous patient-based studies, the subjective norms factor had a significant effect on the intention to use [43,47], which contrasted with its effect on physician-based studies [48]. Fourth, the perceived behavioral control factor does not affect behavioral intention. This last result is in line with both previous patient-based and physician-based studies [48,49].

Telemedicine has been useful in crisis outbreaks in the past [50]. Today, telemedicine is displaying its potential in the COVID-19 pandemic, for example, e-triage, e-consultations, remote monitoring of the intensive care unit, and patients being attended to remotely by health personnel, including those currently in quarantine [51]. Unfortunately, telemedicine has not been promoted and scaled-up homogeneously in all countries [52]. For example, Italy did not include telemedicine at a fundamental level when the pandemic started. In comparison, France actively fostered the use of telemedicine [50]. COVID-19 is creating a great deal of learning about telemedicine's effectiveness in times of crisis. However, nation-wide telemedicine programs, especially in developing countries,

cannot be designed and implemented overnight [16,17]. According to this research's results, the attitude toward telemedicine is the most relevant variable to explain the intention of using these services by patients. A practical implication of this study is that communication strategies should focus on showing the benefits of these technologies, initially with vicarious experiences, and then stimulating engagement with these services. This promotion of engagement is associated with patients and also with family members or caregivers, as well as health service providers.

The outcomes of this study can serve as a good starting point for future research about telemedicine usage intention in developing countries. Future research could include larger sample sizes and different population samples. It would be noteworthy to see the difference between a population sample from a country that has many COVID-19 cases and a sample of a country with a low number of cases.

Some limitations must be considered in the present study. First, telemedicine adoption in developing countries, particularly at the COVID-19 pandemic, is an unexplored research area. Thus, the results of this investigation should not be lightly generalized to other settings. Second, this study used a convenience sampling technique appropriate for an initial exploratory research such as this one, but which limits the generalization of the findings. Third, this study used the more traditional versions of the TAM and the TPB. Although only the TPB model provides good explanatory power, these results indicate the necessity of considering other antecedent variables concerning developing countries, such as cultural values, hedonic motivation, self-efficacy, and habit. In this vein, future studies could make comparisons with extended models explicitly developed for telemedicine adoption.

5. Conclusions

This study was aimed at determining which model, TPB or TAM, provided greater explanatory power for the adoption of telemedicine addressing the outlier-associated bias. We carried out an empirical study on a sample of Brazilian adults. From the responses, we tested both the TPB and the TAM models to explain the behavioral intention to use telemedicine.

According to the results of both PLSc and robust PLSc analysis, the TPB provides significant explanatory power. Both estimation techniques lead to equivalent results without inconsistencies in interpretation. Additionally, the TPB structural results show that attitude has the strongest effect on behavioral intention to use telemedicine systems.

Our global findings suggest that statistical notions and methods associated with robustness can be effortlessly implemented in standard techniques used by social scientists. However, the community has not been readily receptive to these improvements. We hope that this study will be useful to advance in that sense.

Author Contributions: Conceptualization, C.R.-R., J.A.-P., and P.R.-C.; methodology, P.R.-C.; software, A.M.-M. and P.R.-C.; validation, C.R.-R. and J.A.-P.; formal analysis, C.R.-R. and J.A.-P.; investigation, C.R.-R. and J.A.-P.; resources, C.R.-R. and J.A.-P.; data curation, A.M.-M. and C.R.-R.; writing—original draft preparation, C.R.-R., J.A.-P., and P.R.-C.; writing—review and editing, C.R.-R. and J.A.-P.; visualization, C.R.-R. and J.A.-P.; supervision, C.R.-R.; project administration, C.R.-R.; funding acquisition, A.M.-M. and J.A.-P. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was partially funded by UCN.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khan, G.F.; Sarstedt, M.; Shiau, W.L.; Hair, J.F.; Ringle, C.M.; Fritze, M.P. Methodological research on partial least squares structural equation modeling (PLS-SEM): An analysis based on social network approaches. *Internet Res.* **2019**, *29*, 407–429. [CrossRef]
2. Hair, J.F.J.; Hult, G.T.M.; Ringle, C.; Sarstedt, M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed.; SAGE Publications: Thousand Oaks, CA, USA, 2016; ISBN 9781452217444.
3. Klesel, M.; Schuberth, F.; Henseler, J.; Niehaves, B. A test for multigroup comparison using partial least squares path modeling. *Internet Res.* **2019**, *29*, 464–477. [CrossRef]

4. Becker, J.M.; Rai, A.; Ringle, C.M.; Völckner, F. Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS Q. Manag. Inf. Syst.* **2013**, *37*, 665–694. [CrossRef]
5. Henseler, J.; Dijkstra, T.K.; Sarstedt, M.; Ringle, C.M.; Diamantopoulos, A.; Straub, D.W.; Ketchen, D.J.; Hair, J.F.; Hult, G.T.M.; Calantone, R.J. Common Beliefs and Reality About PLS: Comments on Rönkkö and Evermann (2013). *Organ. Res. Methods* **2014**, *17*, 182–209. [CrossRef]
6. Shmueli, G.; Sarstedt, M.; Hair, J.F.; Cheah, J.H.; Ting, H.; Vaithilingam, S.; Ringle, C.M. Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. *Eur. J. Mark.* **2019**, *53*, 2322–2347. [CrossRef]
7. Dijkstra, T.K.; Henseler, J. Consistent partial least squares path modeling. *MIS Q. Manag. Inf. Syst.* **2015**, *39*, 297–316. [CrossRef]
8. Henseler, J. Partial least squares path modeling: Quo vadis? *Qual. Quant.* **2018**, *52*, 1–8. [CrossRef]
9. Schamberger, T.; Schuberth, F.; Henseler, J.; Dijkstra, T.K. Robust partial least squares path modeling. *Behaviormetrika* **2020**, *47*, 307–334. [CrossRef]
10. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*; Pearson: Harlow, UK, 2018; ISBN 978-0134995397.
11. Niven, E.B.; Deutsch, C.V. Calculating a robust correlation coefficient and quantifying its uncertainty. *Comput. Geosci.* **2012**, *40*, 1–9. [CrossRef]
12. Sood, S.; Mbarika, V.; Jugoo, S.; Dookhy, R.; Doarn, C.R.; Prakash, N.; Merrell, R.C. What is telemedicine? A collection of 104 peer-reviewed perspectives and theoretical underpinnings. *Telemed. e-Health* **2007**, *13*, 573–590. [CrossRef]
13. Dick, S.; O'Connor, Y.; Thompson, M.J.; O'Donoghue, J.; Hardy, V.; Wu, T.-S.J.; O'Sullivan, T.; Chirambo, G.B.; Heavin, C. Considerations for Improved Mobile Health Evaluation: Retrospective Qualitative Investigation. *JMIR mHealth uHealth* **2020**, *8*, e12424. [CrossRef] [PubMed]
14. Harst, L.; Lantzsch, H.; Scheibe, M. Theories predicting end-user acceptance of telemedicine use: Systematic review. *J. Med. Internet Res.* **2019**, *21*, e13117. [CrossRef] [PubMed]
15. Ipsos Global Global Views On Healthcare—2018. Available online: https://www.ipsos.com/sites/default/files/Global%20Views%20on%20Healthcare%202018%20-%20Personel%20Health%20Perceptions.pdf (accessed on 20 March 2020).
16. Bashshur, R.; Doarn, C.R.; Frenk, J.M.; Kvedar, J.C.; Woolliscroft, J.O. Telemedicine and the COVID-19 pandemic, lessons for the future. *Telemed. e-Health* **2020**, *26*, 571–573. [CrossRef] [PubMed]
17. Portnoy, J.; Waller, M.; Elliott, T. Telemedicine in the Era of COVID-19. *J. Allergy Clin. Immunol. Pract.* **2020**, *8*, 1489–1491. [CrossRef]
18. Adams, J.G.; Walls, R.M. Supporting the Health Care Workforce During the COVID-19 Global Epidemic. *JAMA* **2020**, *323*, 1439–1440. [CrossRef]
19. Giudice, A.; Barone, S.; Muraca, D.; Averta, F.; Diodati, F.; Antonelli, A.; Fortunato, L. Can teledentistry improve the monitoring of patients during the Covid-19 dissemination? A descriptive pilot study. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3399. [CrossRef]
20. Duffy, S.; Lee, T.H. In-person health care as option B. *N. Engl. J. Med.* **2018**, *378*, 104–106. [CrossRef]
21. Anderson, R.M.; Heesterbeek, H.; Klinkenberg, D.; Hollingsworth, T.D. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **2020**, *395*, 931–934. [CrossRef]
22. Kim, J.; Park, H.A. Development of a health information technology acceptance model using consumers' health behavior intention. *J. Med. Internet Res.* **2012**. [CrossRef]
23. Vega-Barbas, M.; Seoane, F.; Pau, I. Characterization of user-centered security in telehealth services. *Int. J. Environ. Res. Public Health* **2019**, *16*, 693. [CrossRef]
24. Xie, Q.; Song, W.; Peng, X.; Shabbir, M. Predictors for e-government adoption: Integrating TAM, TPB, trust and perceived risk. *Electron. Libr.* **2017**, *35*, 2–20. [CrossRef]
25. Rondan-Cataluña, F.J.; Arenas-Gaitán, J.; Ramírez-Correa, P.E. A comparison of the different versions of popular technology acceptance models a non-linear perspective. *Kybernetes* **2015**, *44*, 788–805. [CrossRef]
26. Ramírez-Correa, P.; Rondán-Cataluña, F.J.; Moulaz, M.T.; Arenas-Gaitán, J. Purchase intention of specialty coffee. *Sustainability* **2020**, *12*, 1329. [CrossRef]
27. Lin, S.P.; Yang, H.Y. Exploring key factors in the choice of e-health using an asthma care mobile service model. *Telemed. e-Health* **2009**, *15*, 884–890. [CrossRef] [PubMed]

28. Zhang, X.; Han, X.; Dang, Y.; Meng, F.; Guo, X.; Lin, J. User acceptance of mobile health services from users' perspectives: The role of self-efficacy and response-efficacy in technology acceptance. *Inform. Health Soc. Care* **2017**, *42*, 194–206. [CrossRef] [PubMed]
29. Saigi-Rubió, F.; Jiménez-Zarco, A.; Torrent-Sellens, J. Determinants of the intention to use telemedicine: Evidence from Primary Care Physicians. *Int. J. Technol. Assess. Health Care* **2016**, *32*, 29–36. [CrossRef] [PubMed]
30. Vidal-Alaball, J.; Mateo, G.F.; Domingo, J.L.G.; Gomez, X.M.; Valmaña, G.S.; Ruiz-Comellas, A.; Seguí, F.L.; Cuyàs, F.G. Validation of a short questionnaire to assess healthcare professionals' perceptions of asynchronous telemedicine services: The Catalan version of the health optimum telemedicine acceptance questionnaire. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2202. [CrossRef]
31. Saigí-Rubió, F.; Torrent-Sellens, J.; Jiménez-Zarco, A.I. Drivers of telemedicine use: International evidence from three samples of physicians. *IN3 Work. Pap. Ser.* **2014**. [CrossRef]
32. Jen, W.Y.; Hung, M.C. An empirical study of adopting mobile healthcare service: The family's perspective on the healthcare needs of their elderly members. *Telemed. e-Health* **2010**, *16*, 41–48. [CrossRef]
33. Hill, R.J.; Fishbein, M.; Ajzen, I. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research.*; Addison-Wesley: Reading, MA, USA, 1977; ISBN 9780201020892.
34. Henseler, J.; Hubona, G.; Ray, P.A. Using PLS path modeling in new technology research: Updated guidelines. *Ind. Manag. Data Syst.* **2016**, *116*, 2–20. [CrossRef]
35. Sarstedt, M.; Ringle, C.M.; Hair, J.F. Partial Least Squares Structural Equation Modeling. In *Handbook of Market Research*; Homburg, C., Klarmann, M., Vomberg, A., Eds.; Springer: Cham, Switzerland, 2017; pp. 1–40.
36. Hubert, M.; Debruyne, M.; Rousseeuw, P.J. Minimum covariance determinant and extensions. *Wiley Interdiscip. Rev. Comput. Stat.* **2018**, *10*, e1421. [CrossRef]
37. Hubert, M.; Debruyne, M. Minimum covariance determinant. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 36–43. [CrossRef]
38. R Core Team R: A Language and Environment for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 9 September 2019).
39. Ray, S.; Danks, N.P.; Velasquez Estrada, J.M.; Uanhoro, J.; Bejar, A.H.C. Package "SEMinR". Domain-Specific Language for Building and Estimating Structural Equation Models. Available online: https://CRAN.R-project.org/package=seminr (accessed on 20 December 2019).
40. Ripley, B.; Venables, B.; Bates, D.; Hornik, K.; Gebhardt, A.; Firth, D. Package "MASS". Support Functions and Datasets for Venables and Ripley's MASS. Available online: https://CRAN.R-project.org/package=MASS (accessed on 10 April 2020).
41. Cepeda-Carrion, G.; Cegarra-Navarro, J.G.; Cillo, V. Tips to use partial least squares structural equation modelling (PLS-SEM) in knowledge management. *J. Knowl. Manag.* **2019**, *23*, 67–89. [CrossRef]
42. Dijkstra, T.K.; Schermelleh-Engel, K. Consistent Partial Least Squares for Nonlinear Structural Equation Models. *Psychometrika* **2014**, *79*, 585–604. [CrossRef]
43. Kamal, S.A.; Shafiq, M.; Kakria, P. Investigating acceptance of telemedicine services through an extended technology acceptance model (TAM). *Technol. Soc.* **2020**, *60*, 101212. [CrossRef]
44. Dünnebeil, S.; Sunyaev, A.; Blohm, I.; Leimeister, J.M.; Krcmar, H. Determinants of physicians' technology acceptance for e-health in ambulatory care. *Int. J. Med. Inform.* **2012**, *81*, 746–760. [CrossRef]
45. Zailani, S.; Gilani, M.S.; Nikbin, D.; Iranmanesh, M. Determinants of telemedicine acceptance in selected public hospitals in Malaysia: Clinical perspective. *J. Med. Syst.* **2014**, *38*, 111. [CrossRef]
46. Adenuga, K.I.; Iahad, N.A.; Miskon, S. Towards reinforcing telemedicine adoption amongst clinicians in Nigeria. *Int. J. Med. Inform.* **2017**, *104*, 84–96. [CrossRef]
47. Tao, D.; Wang, T.; Wang, T.; Zhang, T.; Zhang, X.; Qu, X. A systematic review and meta-analysis of user acceptance of consumer-oriented health information technologies. *Comput. Hum. Behav.* **2020**, *104*, 106147. [CrossRef]
48. Chau, P.Y.K.; Hu, P.J.H. Investigating healthcare professionals' decisions to accept telemedicine technology: An empirical test of competing theories. *Inf. Manag.* **2002**, *39*, 297–311. [CrossRef]
49. Kim, J.; DelliFraine, J.L.; Dansky, K.H.; McCleary, K.J. Physicians' acceptance of telemedicine technology: An empirical test of competing theories. *Int. J. Inf. Syst. Change Manag.* **2010**, *4*, 210–225. [CrossRef]

50. Ohannessian, R.; Duong, T.A.; Odone, A. Global Telemedicine Implementation and Integration Within Health Systems to Fight the COVID-19 Pandemic: A Call to Action. *JMIR Public Health Surveill.* **2020**, *6*, e18810. [CrossRef] [PubMed]
51. Hollander, J.E.; Carr, B.G. Virtually Perfect? Telemedicine for Covid-19. *N. Engl. J. Med.* **2020**, *382*, 1679–1681. [CrossRef] [PubMed]
52. Smith, A.C.; Thomas, E.; Snoswell, C.L.; Haydon, H.; Mehrotra, A.; Clemensen, J.; Caffery, L.J. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). *J. Telemed. Telecare* **2020**, *26*, 309–313. [CrossRef] [PubMed]

symmetry

MDPI

Article

Tele-Education under the COVID-19 Crisis: Asymmetries in Romanian Education

Cristina Nicolau [1], Ramona Henter [2,*], Nadinne Roman [3], Andrea Neculau [3] and Roxana Miclaus [3]

[1] Faculty of Economic Sciences and Business Administration, Transilvania University of Brașov, 500368 Brasov, Romania; cristina.nicolau@unitbv.ro
[2] Faculty of Psychology and Education Sciences, Transilvania University of Brașov, Str. N. Bălcescu 56, 500368 Brașov, Romania
[3] Faculty of Medicine, Transilvania University of Brașov, 500368 Brasov, Romania; nadinneroman@unitbv.ro (N.R.); andrea.neculau@unitbv.ro (A.N.); roxicum@unitbv.ro (R.M.)
* Correspondence: ramona.henter@unitbv.ro

Received: 13 August 2020; Accepted: 10 September 2020; Published: 12 September 2020

Abstract: The COVID-19 pandemic has deepened social and educational asymmetries in some developing countries, such as Romania. Tele-education failed to replace face-to-face education due to the lack of symmetrical policy, connectivity, infrastructure, digitalized educational materials and digital competences. Was this issue predictable and, hence, the stakeholders' mission failed? Our qualitative research aims at analyzing, in depth, these digitalization asymmetries, with a sample formed of Information and Communication Technology (ICT) specialists working for/with Romanian 1–4 International Standard Classification of Education (ISCED) schools. The collected primary data were processed with Atlas.ti 8. The results emphasize major key areas to be addressed with future public symmetrical policy and change management strategies: equal access to infrastructure, as well as development of compulsory and complementary digital skills (for teachers and students). The necessity to support school management in accessing funding is also required to enhance digitalization.

Keywords: asymmetry; tele-education; digitalization; ICT infrastructure; digital teacher training; replace face-to-face education

1. Introduction

The current coronavirus pandemic, generated by the outbreak of the novel virus SARS-CoV2 [1,2], determined at a global level certain social, economic, medical and educational issues, forcing professionals in all fields to adapt their activities to the new governmental measures and health policies. A lack of symmetry in education, together with the increase in unemployment [3] and inflation rates, generate an educational crisis with outcomes that cannot be neglected as education is the driver of any economic system in terms of knowledge transfer and competence development, including the digital teaching competences [4].

Digital technologies have been introduced in educational institutions, where they are considered to have had a positive impact [5]; but, in a developing country like Romania, this process was a slow one (digitalization was not perceived as a need to be rapidly addressed). Face-to-face education seemed to be comfortable (not requiring changes) and successful (meeting settled standards and indicators) whereas the pedagogical methods and techniques seemed to be sufficient and efficient, at least from the management's perspective, although less creative and lacking attractiveness in today's students' eyes. In time, incorporating multimedia was not enough for teaching and training. Computer-aided instruction and intelligent tutoring systems were developed as interactive learning environments [6], which with the support of the growth of the World Wide Web and of the extension

of Internet connections, gave birth to online learning. The main consequence was a re-shaping of the educational profession, influencing both teachers and trainers, including the most resisting ones, and included networks, television and computers into the educational process, as they enabled long-distance communication and use of multimedia [7].

Under the circumstances generated by the COVID-19 crisis, the main aim of this paper was to determine, in a developing country like Romania, the capacity of its educational system to replace face-to-face education with any form of tele-education because of the lockdown caused by this pandemic. Its general objective (GO) was identifying the level of digitalization that supports Romanian tele-education under the COVID-19 crisis whereas the specific objectives (SO) were as follows:

SO1—Evaluating the asymmetries of Romanian education.

SO2—Examining the resources used for supporting tele-education.

SO3—Identifying key-areas for future development of tele-education in Romania.

2. Literature Review

2.1. The Importance of Tele-Education in Developing Countries during the COVID-19 Crisis

Research shows that children have an important role as transmitters of COVID-19 [8–10]. Under such circumstances, one of the most important measures of public health was to interrupt face-to-face education activities placed in public and private educational organizations and to ban the organization of any event which would gather a large number of people, education-oriented events included.

Before the first months of 2020 when the COVID-19 pandemic started spreading worldwide, digitalization of education in a developing country like Romania seemed to be a regular, smooth process (ICT was introduced when needed). Hence, it has rapidly started to show its lack of symmetry when government policies started banning participation of face-to-face education. Students in all the levels of education, whether public or private, irrespective of their ages (children or adults) or generations, needed to adapt to tele-education as quickly as possible, regardless of whether they had Internet connectivity and/or electric power (Challenge 1), proper ICT infrastructure at home (Challenge 2), or possessed digital competences or had an available adult to provide ICT support (Challenge 3). Teaching and training staff (Challenge 4) faced legal issues and a lack of public policy regarding remote and online job tasks (Challenge 5). In the developing countries where low funding has been used for training public teaching staff [11] and digitalization in general, schools may not have implemented a secure and telecommunication-based educational environment, exposing students (mostly children) to cyber risks (Challenge 6). Hence, the major challenge of tele-education during the COVID-19 pandemic was to properly and timely respond to these challenges in terms of public policy, funding, infrastructure and digital competence training.

2.2. The Symmetry with the Old and New Tele-Education

Tele-education is the teaching–learning process that uses any device merging with telecommunications and computer science for education-purposes. Tele-education does not refer only to a geographical distance between students and teachers as opposed to face-to-face education, but also to the distance between these two actors of education and the resources used for educational purposes, which may be virtual and computer-assisted, or physical like books, textbooks, copybooks and handouts, but used in the absence of an in-person teacher.

The electronic tutor was firstly envisioned by Arthur Clarke in the 1970s [12]. Two-decade experiments showed the efficiency of both wide- and narrow-band techniques used in responding to global education's needs in developed and developing countries [13]. This brought interactivity in the classrooms as well as other opportunities of distant learning (which at first mainly consisted of pen-and-paper-based tasks exchanged by post-mail). Distance education, which exploits ICT, provides the right means of communication needed so as to avoid any physical contact between students/trainees and teachers/trainers.

The flipped learning pedagogical approach transfers the direct instruction from the group to the individual, offering an active and cooperating learning space, with the educator as facilitator guiding students to apply concepts in a creative manner [14]. On the other hand, tele-education is presumably named by the European Commission (EC) as digital and online learning (DOL), a concept consisting of two main constitutive parts [15]: digital learning (ICT-supported teaching and learning) and online learning (desktop, mobile devices, Internet and web services support learning from a distance and creating a personalized learning experience, reducing time and place as educational constraints).

Tele-education brought about the need for a new didactics, and hence e-Didactics [16] emerged and marked a change of focus from simply teaching to learning engineering, promoting blended and online education; moreover, the classroom is moved into virtual environments, including social networks, using various learning systems and benefitting from unlimited ICT resources whereas teachers' role changed from delivering information to engineering the learning of their students who have thus become actively involved learners [16]. In Romania, teachers make use of ICT and multimedia to enhance learning, but this piece of research shall analyze the level of their use of MOOC platforms and directories such as Moodle, Open edX, Canvas, NovoEd, Udemy, Miriada Coursera, etc., within the processes of e-learning and e-teaching. Still, the real challenge in any type of network-based education is to use the most appropriate pedagogical models that support the teaching–studying–learning process [17].

We investigated the development of this concept as a whole (not including the development of its constitutive parts namely tele-teaching and tele-learning) and we identified that, except from medicine where it received a lot of attention and the concept of tele-medicine is largely used, previous research may be divided according to the two major fields this concept has received special attention from: computer science and engineering, and social sciences and language acquisition. Hence, there are two forms of tele-education [18]: (1) asynchronized—the Internet is used to publish hyperlinked multimedia content whereas offering the opportunity of reaching a large audience with digitalized material (self-learning is highly encouraged as assessment may be delivered as well, not needing the assistance of a professor); and (2) synchronized—the major form being real-time interactive virtual classrooms supported by a multitude of applications developed to meet all students' needs.

The results of tele-education are also influenced by various factors, which reveal a lack of symmetries, starting with demographic ones (unequal educational levels, education isolation, information isolation, regional disparity, rural versus urban gaps and gender issues, with females registering a more serious lack of education [19]). There are also disadvantages in this type of learning: students' low motivation for learning, social isolation, technical incompatibilities among the learning systems available, technology dependency and higher costs for the institutions [20].

2.3. Digital Intelligence of Schools, Classrooms and Students

Intelligent schools refer to the use of computers, multimedia and teacher droids/roboprofessors for educational purposes. In the 1990s, the Program on Educational Building of the Organization for Economic Cooperation and Development (OECD) aimed at encouraging the design of school architecture and environments that serve and foster learning [21]. It also advanced some important ideas to be implemented in all the schools, irrespective of whether the school is located in an OECD country or not: (1) locating ICT resources throughout the school (not in dedicated computer rooms); and (2) reducing the gap between privileged and disadvantaged schools with regard to ICT resources (bringing the educational process outside the school providing more availability to learning).

Additionally, the micro-unit within intelligent schools is the intelligent classroom, a platform or application providing teaching and training staff with written and oral services like digital handwriting, speech and gestures delivery, making annotations on courseware, pointing to objects displayed on media boards and saying predefined commands [21]. Hi-tech component technologies used in the mid-2000s were interactive interfaces, which could be run with virtual assistants able to speak, digital pens and laser points (hardware) as well as structure providing similarly structured modules or the possibility

to create and integrate new ones: architecture and runtime structure, inter-agent communication, runtime environment management and debugging support as well as agent development interface (software) [18]. Irrespective of how much technology is developed, intelligence did not refer to the classroom, but to instructor's ability to intelligently use the room by concentrating on the lecture and not on the technology [22]. Under such circumstances, we highlight that education success depends on the teachers' intermediate to advanced levels of digital competences.

To conclude, tele-education can be performed via various media [23]:

- audio (transmission of the spoken word between learners and instructors, either synchronously, e.g., videoconferencing and short-wave audio, or asynchronously, e.g., audiotape or audiocassette);
- video (either synchronously, e.g., videoconferencing and interactive television, or asynchronously, e.g., slow-scan video, interactive videodiscs and videotapes);
- computer-assisted (Internet, www, email, applications and multimedia applications—interactive or on CD-ROM.

More and more technologies, functionalities and educational benefits of the virtual reality approach [24] have been added. Nevertheless, the technology itself is not enough. The human resource must be able to use it properly, to have the necessary digital competencies. The European Centre for the Development of Vocational Training (Cedefop) [25] considers that a competence is the capacity of using knowledge, skills and abilities from the personal, social and activity fields in various situations. Therefore, the digital competence (or digital literacy), a key competence for lifelong learning in the European Union (EU), covers ICT knowledge and the ability to use this knowledge for problem-solving in any domain.

In this view, this paper explores the asymmetries of Romanian education with regard to digitalization as a response to the COVID-19 crisis; this challenges education in terms of learning acquisition and educational needs, teachers' digital competences, the ICT infrastructure available and cybersecurity.

3. Materials and Methods

3.1. Research Localization

Our study will provide Romania as an example of a developing country. Romania is a member of the European Union (EU). In order to understand the context of using tele-education during the COVID-19 pandemic, we analyzed three types of descriptive data: about the COVID-19 pandemic, about expenditure in education and about national digital performance.

At present (22 August 2020), in Romania, 76,355 cases of COVID-19-infested persons were reported, with 3196 (4.18%) deaths and 34,523 (45.21%) recovered patients. The first case of COVID-19 was reported on 26 February 2020. Since then, Romania has passed through two periods: a state of emergency from 16 March to 15 May 2020, (lockdown for all the residents [26–29]) and a state of alert from 16 May 2020, and still continuing (measures of economic and social relaxation were progressively installed [27,30,31]; but, a second wave of COVID-19 infections set in with a peak of infections on 13 August 2020, with 1454 new cases [26], as presented in Figure 1.

Starting from 11 March 2020, face-to-face learning was banned in schools and on 15 June 2020, the summer vacation started (the new school year is forecast to start on 14 September 2020). However, it is not officially announced how to conduct courses; there are three possibilities, depending on the evolution of the number of cases of COVID-19 on the territory of Romania. There are three scenarios in which the school will normally start: in case the number of infections decreases, the variant in which the pupils and students will be divided in half and to alternate the physical courses with the online ones and the variant in which all the education will take place online, if there is an increased epidemiological risk [32].

Under such circumstances, expenditure on education became a very important indicator, at the government and family level, too. In 2018, Romania spent on education 3.2% of its gross domestic

product (GDP); it totaled LEI 30,100 million [33] (approximately Euro 6468 million given the annual average exchange rate of Lei 4.6535 for 1 Euro [34]). This expenditure (see Figure 2) ranked Romania the last within the EU in 2018 [33].

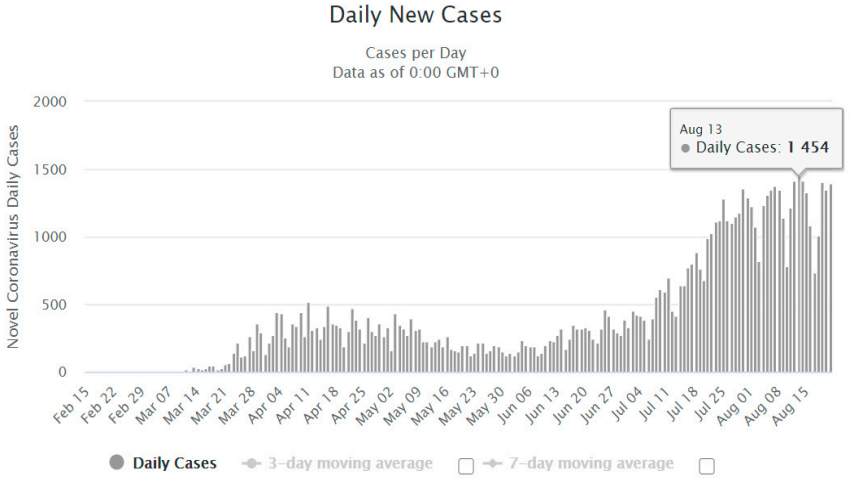


Figure 1. Daily new cases of novel coronavirus infections [26].

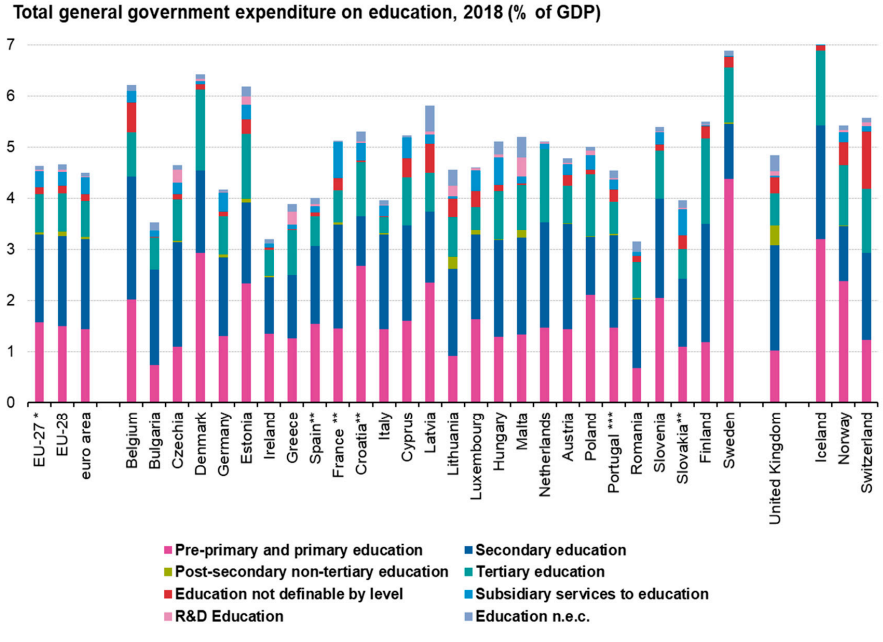


Figure 2. Romania's expenditure on education in 2018 [33].

Other national indicators showing Romania's asymmetries with impact on digitalization are presented in Table 1. Descriptive data show that, in 2018, a household's average revenue was approximately Euro 340 (approximately Euro 260 in rural areas) and spent only 0.32% on education (0.16% in rural areas), whereas at the national level, approximately Euro 3.04 were spent on phoning, audio, video, photo, information processing equipment and its accessories and Euro 0.40 on information

processing equipment and accessories. These data clearly show that the individual investment in ICT is very low and there is a critical need for funding in digitalization in general.

Table 1. Romania's asymmetries with effects on education [35].

Indicator	2017 *	2018 **	Observation
Total monthly income per household.	LEI 1290.9 = Euro 283 per person (LEI 1005 = Euro 220 in rural areas)	LEI 1613 = Euro 346 per person (LEI 1226.84 = Euro 264 in rural areas)	+Euro 63 per person (+Euro 44 in rural areas)
Total monthly expenditure on education per household.	LEI 3.72 = Euro 0.81 (LEI 2.04 = Euro 0.45 in rural areas)	LEI 5.15 = Euro 1.11 (LEI 2.01 = Euro 0.43 in rural areas)	+Euro 0.3 (-Euro 0.02 in rural areas)
Monthly average expenditure on phoning, audio, video, photo, information processing equipment and its accessories.	LEI 12.08 = Euro 2.65	LEI 14.13 = Euro 3.04	+Euro 0.39
Monthly average expenditure on information processing equipment and accessories.	LEI 2.3 = Euro 0.50	LEI 1.87 = Euro 0.40	-Euro 0.10
Share of households having access to computer at home.	65.60% (51.9% in rural areas)	n/a	-/-
Share of households having access to Internet at home.	68.60% (56.9% in rural areas)	72.40% (61.5% in rural areas)	+3.8% (+4.6% in rural areas)
Total number of the population aged 0–17.	3,704,601 persons (1,857,626 in rural areas)	3,680,850 persons (1,825,322 in rural areas)	−23,751 persons (−32,304 persons)
At-risk-of-poverty or social exclusion rate in the population aged 0–17 years.	41.70%	38.10%	-3.60%
Total number of matriculated students.	3,578,561 persons (1,022,507 persons in rural units, of which 77.84% in ISCED 1–3 units)	3,547,301 persons (994,183 persons in rural units, of which 77.50% in ISCED 1–3 units)	−31,260 persons (−28,324 persons in rural units, of which −0.34% in ISCED 1–3 units)
Total number of PC's managed by schools.	387,786 pieces (99,750 pieces in rural areas, of which 93.25% in ISCED 1–3 units)	396,614 pieces (99,275 in rural areas, of which 92.91% in ISCED 1–3 units)	+8828 pieces (−475 pieces in rural areas, of which −0.34% in ISCED 1–3 units)

* Annual average exchange rate was Lei 4.5681 for 1 Euro (source: www.bnr.ro) [34]; ** Annual average exchange rate was Lei 4.6535 for 1 Euro (source: www.bncr.ro) [34].

Moreover, there are still households that do not have a computer or an Internet connection at home (in 2017, more than 30% for the whole country and more than 40% in rural areas for both indicators); so, children belonging to these households have no opportunity for online education if public funding is not used to help them. The number of children with a high poverty or social exclusion risk was still high in 2018 at the national level, being about 1,402,404 children aged 0–17 years. However, we identified severe infrastructure deprivation in Romanian schools, not only in households. In 2018, there were 8.94 students that matriculated with ISCED 1–3 using one personal computer (PC), whereas in rural areas, this number reached 24.40, slightly decreasing from 2017 by 0.17% [35].

In 2019, Romania's index of Digital Economy and Society (DESI), measuring digital performance [36], positioned Romania as 27th within the 28 member states (see Table 2), with a score of 36.5 (increasing by approximately 14% since 2017), whereas the EU average was 52.5 in 2019. It is obvious that the lack of symmetry identified show an urgent need for investment in digitalization of all the Romanian public services, including education.

Table 2. The Digital Economy and Society Index (DESI), 2019 [36].

Index	DESI 2019	Connectivity	Human Capital	Use of Internet	Integration of Digital Technology	Digital Public Service
The EU's score	52.5	53.5	48.0	53.4	41.1	62.9
Romania's score	36.5	59.3	31.1	31.9	20.5	43.2
Romania's rank in the UE	27th	22nd	27th	28th	27th	28th

3.2. Outcome Measures

We designed a semi-structured in-depth research instrument presented in Table 3 consisting of a list of topics and sub-topics to discuss with participants. To reduce fatigue and generate ideas, Sub-topic 2.3. was designed as an "energizer", stimulating the respondents' creativity to think and present examples of any digital item, instrument, device, tool, application, solution or platform used as learning aid in Romanian school that was impressive/surprising enough to say "Wow!".

Table 3. The qualitative research instrument.

Analysis of Digitalization Supporting Romanian Tele-Education Under the COVID-19 Crisis	
Topic	**Sub-Topic**
1. The asymmetries within Romanian education in terms of digitalization.	1.1. The present asymmetries of the Romanian education system. 1.2. Main educational activities using ICT.
2. Digital resources needed for tele-education.	2.1. Existing infrastructure and teaching staff's hardware and software competences. 2.2. IT support for students and teachers. 2.3. Digital "Wow" in education.
3. Future digitalization of schools.	3.1. Fully digital schools and online teachers. 3.2. Costs needed to digitalize a school.

Because of the isolation imposed by the COVID-19 pandemic, interviews were carried out over the phone or by using online applications and every interview lasted about 60 min, out of which 10 min were given in the end to add the participants' personal insights and concerns about the topics discussed.

3.3. Participants

The research design aimed at collecting primary data about the ISCED 1–4 levels of Romanian education (primary, secondary and high school levels) from subjects who may provide the necessary data from within the Romanian public schools. The ICT specialists' opinions brought more objective and accurate data as they are not biased by the desire to distort the truth. The ideal participant has worked with/for public Romanian schools in the ICT field for more than 3 years. We used the snowball sampling method as it is convenient when accessing subjects without target characteristics [37] and we formed a representative non-probabilistic sample with social significance (no statistical logic is needed in qualitative research), which reproduces all the characteristics of the researched population.

This study was conducted at mid-March–April 2020 at the very beginning of the COVID-19 pandemic in Romania and fourteen respondents positively answered our invitation to participate in the study. The research methodology stipulated, according to research ethics, was to obtain informed consent from every participant. The interviews first collected information on the sample characteristics: 100% were male, with an average age of 44.5 years; average IT experience was 18 years, of which 12.5 years in the Romanian public education; and 85.71% holds a bachelor's degree.

3.4. Procedures and Data Analysis

The primary data collected was processed with Atlas.ti 8. We imported the interview notes into this qualitative software and the data analysis consisted of two processes: one inductive, which meant describing seven themes and organizing the participants' words and sentences according to these themes, and one iterative, in which we assigned a label or code to the meanings. Coding of the qualitative data is made on every objective. For the first objective, we grouped the respondents' answers on two paradigms: (1) the perception of *management's and teaching staff's intentions, planning and implementation of digitalization* in education received the following codes, (+/− management support) and (+/− teacher digitalization), used for the following pieces of information: knowledge, skills, intentions, motivations, competitions and funding; and (2) *infrastructure* received the following codes, (+/− infrastructure), used to mark the existence, use or lack of Internet, PC's, phones, labs, YouTube, smart phones, digital textbooks. Moreover, our respondents underlined the need to observe similarities and differences between urban and rural educational environments, which received the following codes (+ urban digital development) and (+ rural digital environment), referring to urban digitalization, rural digitalization, funding and mayor houses' support.

In order to highlight the present condition of the digitalization of the Romanian education, we identified in the subjects' response data three major paradigms, presenting the resources existing in terms of hardware and software, which were divided into the following:

1. *Logistics*: (+ old hardware infrastructure, + new hardware infrastructure), where "old" means both "used" and "old-generation technology" and "new" means both "newly acquired" and "the latest technology"; (+ old software infrastructure, + new software infrastructure), where "old" refers to "outdated" and "new" refers to "the latest generation" and a focus is put on expired and needed licenses coded with (+/− software license); (− enough infrastructure) presents opinions regarding the need of investment in infrastructure, an idea also suggested by (+ teacher's own infrastructure), which shows that teachers use their personal devices in labor interest; and ICT security was simply coded (+/− ICT security).
2. *Human resources* were divided into teaching personnel (herein teachers of ICT and computer science are included) and administrative staff who should support the educational process in terms of ICT infrastructure and secure use:

    - teaching personnel: (+/− teacher hardware competences), (+/− teacher software competences), (+/− teacher solution) expressing efforts made for digitalizing teaching and learning, (+/− teacher attitude) measuring intentions to digitalize, and (+/− teacher IT training) coding all ICT training;
    - support ICT specialists: (+ ICT service externalization, + Own ICT employee), showing the managerial solution for supporting a school's digital processes.

3. *Financial resources*: As education is or shall be student-centered, we considered the financial resources needed for the three constitutive parts of digitalization: to form (+ students' digital competences) and (+ teaching staff training) and acquire (+ appropriate infrastructure), which shall consist of (+ interactive contents) and (+ ICT security), everything resulting in a time-, quality- and cost-constrained digitalization process. With regard to costs, the last sub-topic of discussion was designed to result in an average cost of endowing an average Romanian school a schooling capacity with the necessary technology and specific training so as to deliver qualitative education, but also to have the ability of easily switching from face-to-face education to tele-education if needed.

4. Results

4.1. Digitalization of Romanian Education at the Beginning of the COVID-19 Crisis

Qualitative analysis showed that the digitalization of Romanian schools was very low, as stated by participants: "schools have no infrastructure, no Internet", "teachers' digital competences are obsolete", "some school have no digital infrastructure because of the management" and "no acquisitions for the ICT infrastructure were made in the last 10 years". The implementation of modern technology in learning had been a main objective of education [38] and of all its educational stakeholders, but in fact, ICT had not reached all students. The interviewed IT specialists considered that school management was mainly responsible for the digitalization or the lack of it, indicating that there were schools with no Internet connectivity and as far as they knew, in some remote villages there was not even a working computer in a school: "in schools everything is desired to be digitized and very well developed, but actually the ICT infrastructure is missing" and "each school manages its public funding the way it wants". There are very well-equipped schools with Computer Science laboratories and/or laptops, printers and video-projectors in every classroom and "smartboards in some classrooms". We identified two gaps between schools: rural versus urban (Asymmetry 1) and poor versus rich (Asymmetry 2), "as funding comes from the local council, in rural areas there is no money for schools", "usually only in central urban and high rated schools it is a priority", "teachers bring their own ICT equipment if they have", "people in rural areas are so poor that children don't have any electronic devices" and even in urban areas, there were poor and rich schools. Therefore, school management (Figure 3) emerges both as main factor of progress and obstacle of technological advancement in schools (Asymmetry 3): "each school decides what to buy", "teachers' interest in ICT helps with the introduction of new equipment" and "the school management is influenced by the local management", being the decision factor in accessing funding for infrastructure acquisition and development of human resources and implementing projects of development.

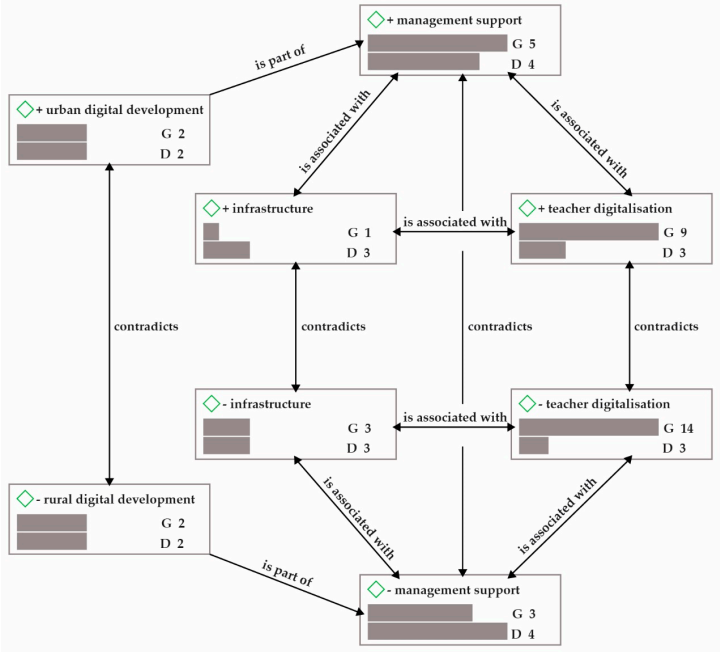


Figure 3. Asymmetries in the Romanian public education's digitalization.

The lack of ICT infrastructure (G = 3) is associated with the poor development of the ICT competences of teachers (G = 14). There are teachers with well-developed digital skills (Asymmetry 4), mainly as a result of their interest in this field (Asymmetry 5): "some teachers are really interested in digitalization" and "they enter their students in various e-contests". However, "many teachers can and use PowerPoint presentations, YouTube films and digital textbooks"; though some participants stated that "these activities are not enough to consider Romanian education digitalized".

The necessary infrastructure relies on the public funding of the school and on the principals' ability to access and attract funding: "the acquisition of ICT equipment depends on the managers' plan and vision", "all acquisitions depend on the local policy makers" and "it depends on the manager to attract funding" (Asymmetry 6). However well-endowed the school is, it comes down to the teachers' desire to use ICT in teaching and, more specifically, to their ability to do so.

In this regard, school management's openness and focus on digitalization shall be cultivated; as two subjects say, sometimes, computers used in some schools were donated by large companies, which renewed their infrastructure, but "students cannot benefit from their use as they might be outdated or out of service" (Asymmetry 7). To conclude, public policy in terms of funding infrastructure is needed after "school management would have done an inventory of ICT-assisted activities to respond to students' and teachers' needs", to increase digital intelligence at the national level.

4.2. Resources Needed to Enhance Digitalization in Romanian Public Schools

From this piece of research, a Romanian school would need three types of resources in order to become digital: logistics (infrastructure in terms of tangible and intangible assets), human resources and financial resources so as to support tele-education (this shows how oversimplified online education is). According to data collected from our subjects, Romanian schools were endowed with infrastructure for ICT activities, "PCs, laptops, video-projectors, smartboards, scanners and printers", especially the primary education levels, but they seemed not to be "enough" (Asymmetry 8).

As presented in Figure 4, infrastructure in Romanian schools is quantitatively centered on hardware: "we have computers but no all are functioning", "donated equipment is mostly outdated" and "old computers are kept"; little focus is also given to software and users' security (students' and teachers' security): "teachers do not know how to use an anti-virus software", "teachers use the Internet without any fear of being hacked", "school management does not support cyber-security software", "20% of the teachers know that they can protect information through antivirus software, but only 1% know the concept of firewall which, in terms of security, provides protection against external and internal attacks on the computer network (I consider the following phenomena very dangerous for students: pornography, theft of personal data, fraudulent use of online catalogs with addition/modification/deletion of information, access and modification of data in work folders) and internal (if a teacher presents online on a smart board a material stored on his phone, a student in the class can attack the network and copy, delete or modify the data of the two devices)" and "most teachers are unfamiliar with services such as software updates and virus scanning" (Asymmetry 9).

Most schools had old infrastructure (G = 6): "not new", "donated" and "second-hand", without the necessary software (G = 4); respondents stated that neither the hardware nor software was "enough" (G = 15). In general, infrastructure was located in Computer Science labs where access was given "only during specialized classes of Computer literacy", as a rule (Asymmetry 10). The lack of licensed software was even more strenuous (G = 5); for example, "the infrastructure was unable to support the security elements, the recommendations were not to make software updates because they would lead to the malfunction of the equipment that does not allow later versions of those initially installed", and the subjects complained about the severe lack of IT security (G = 11) for their work, confidential information and students. Hence, insufficient and/or inappropriate infrastructure, lack of access to infrastructure (kept only for the ICT Lab) and lack of cybersecurity determined teachers "use their own infrastructure" in teaching (G = 6), which clearly underlines the need to implement a national program to massively digitalize Romanian schools.

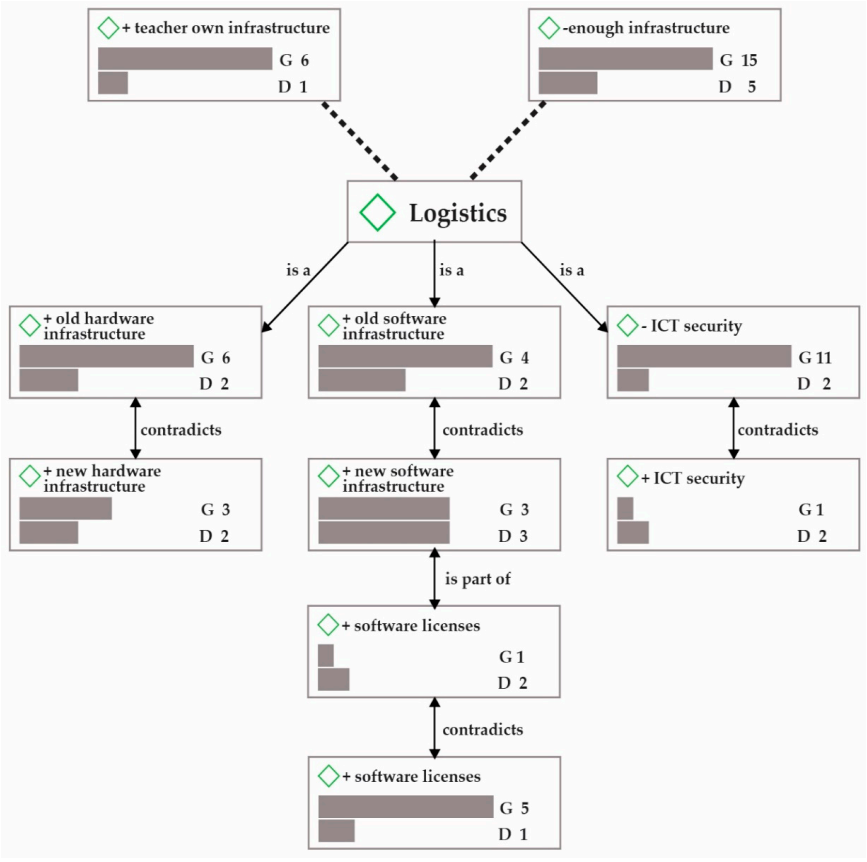


Figure 4. Asymmetries of the logistic resources in Romanian public schools.

Furthermore, a participant underlined that Romanian education lacks the equipment specialized for learning specific subjects (Asymmetry 11), such as science, technology, engineering and math (STEM). Large ICT providers sell a lot of physical devices and online synchronized and asynchronized applications and activities for education use at very affordable costs [39], which may also be purchased within a national program.

With regard to the human resources implied in the digitalization of education (Figure 5), apart from the use of infrastructure that was previously discussed, the respondents offered valuable information on the digital competences of teaching personnel and ICT specialists in order to properly support tele-education: "not all teachers have digital competences or interest towards developing them (especially for those who graduate more than 15 years ago, when ICT was not a compulsory subject matter)" and "teachers lack basic skills—how to save data, send students data by email, copy data to memory sticks or other storage media, such as the cloud (if I talk about the cloud, teachers don't know what it is, what it uses, they have no information about this concept)". Our participants referred to both teachers with very low digital competences and to "very few" teachers who were "highly interested in ICT and STEM" and who "changed whole schools" in terms of digitalization. One subject gave the example of a primary school teacher "who was involved in projects aiming at digitalizing education such as eTwinning or STEM projects" (Asymmetry 12) and "changed lives with innovative education methods offered with devices especially purchased to support them".

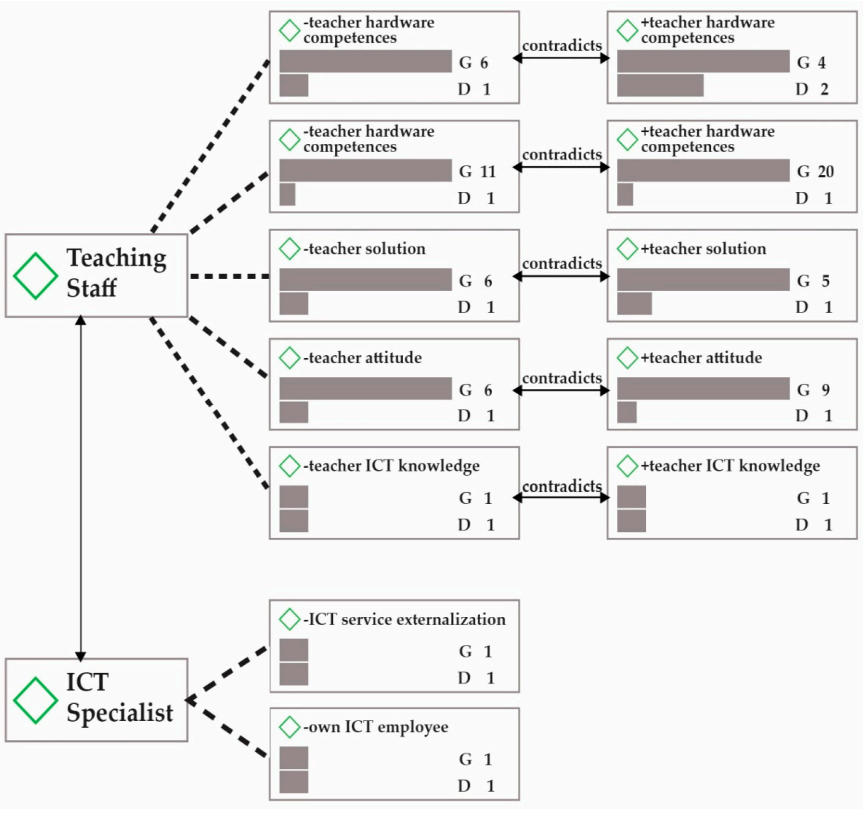


Figure 5. Asymmetries of human resources in Romanian public schools.

Moreover, a subject named three categories of teachers representative for Romania in terms of digitalization: those who have ICT competencies and "do everything themselves"; those "under the age of 35" who do not know and ask for specialized help (they "firstly try by themselves to solve the problem and ask for help only if they do not succeed, they are polite, accept online or remote assistance, offer a period for solving"); and those "over the age of 35 who do not have hardware or software competencies and are not interested from the beginning to have ICT knowledge or to develop their competences" (they are "authoritarian and show a superiority attitude, demanding solutions for their problems, they do not admit improper use of equipment"). Regardless of their age, "50% of teachers are totally disinterested in using digital education".

Teachers are not the only category of staff involved in the digitalization of education. Respondents also spoke about ICT specialists who should "support students and teachers undertaking digitalized activities". Our research showed that IT support (which "could not exist in many schools") was provided in two ways: by having an employee in charge of all the digital aspects of a school or by externalizing serviced to an IT company. Participants highlighted though the need of creating more ICT specialized jobs in public education that shall build a strong education-centered "cooperation between ICT specialist, teachers and students" resulting in an increase in students' learning outcomes (Asymmetry 13).

Furthermore, another major result of this research is that the main obstacle in the digitalization of Romanian schools is the lack of financial resources to acquire the needed infrastructure, increase cybersecurity, make learning interactive and properly develop students' and teachers' digital skills (Figure 6). Investment in digitalizing education shall be realistic in terms of (+time), (+costs) and

(+quality): "massive" and "short-term" investment in "appropriate", "sustainable" and "fair priced" products and services to "match the existent needs of the education system" is recommended by the sample. Two participants even made a forecast of costs to provide immediate access to digital education to a student:

(a) For a school with 25 classes of 25 students, the needs are as follows:

- basic digital infrastructure located in the school: one router and a firewall (Euro 3000), annual licensing (Euro 3900), a switch with management (Euro 1000), one server to keep an e-library, secretarial, accounting work, etc., (Euro 4000), two large printers (Euro 4000), and two small multifunctional printers (Euro 1200). It totals Euro 17,100 (Euro 27.36 per student);
- individual devices for every student (to be used at school or remote): one all-in-one (AIO) unit with licensed software and a mobile Internet device to use both at school and at home (Euro 600). It totals Euro 375,000.

(b) Basic infrastructure of approximately 1500 euro/classroom, "affordable costs for face-to-face education".

Figure 6. Qualitative analysis of financial resources in Romanian public schools.

4.3. Key Areas for Future Development of Tele-Education in Romania

The three major key areas of development highlighted in this research are the improvement of teachers' digital skills; support for education about applications, software and devices to animate and stimulate learning; and provisioning of fully online education.

The range of competences for further ICT training for the teachers within the Romanian education, which may help them reach intermediate to advance levels of digital competences, were, according to our respondents:

(a) Hardware:

- how to use a specific device: PC, laptop, printer, scanner, smart board and multifunctional projector;
- how to interconnect devices.

(b) Software:

- how to operate Windows, Microsoft Office (Excel, Office, PowerPoint), Adobe Reader and other operation systems and software (general and specific to a specific subject taught);
- how to work with archives, folders, files, etc., on a desktop, on physical devices (CDs, memory sticks, etc.) and in the cloud;
- how to synchronously teach by using online platforms and applications (Skype, Zoom, Google Meet, Microsoft Teams, Discord, WebEx, YouNow, etc.);
- how to use Internet, online platforms and online resources available to provide and produce interactive material, such as videos, music, presentations, digital textbooks, digital handouts, questionnaires, cards, etc.;
- how to update software, clean unnecessary contents, scan computer for viruses and work with cybersecurity;
- how to adjust settings when interconnecting devices (e.g., the resolution).

The teachers' digital skills can be inferred from the kind of support they ask from the ICT specialist. All subjects declared that some teachers could not even connect two devices whereas others had no difficulties in using state-of-the-art devices, these being indicators for the importance of personal involvement in self-development. Some teachers are pro-active in finding new ways to interact with students and to present new materials, they are creative and find solutions to introduce digital learning in the traditional school and, as revealed by these interviews, they bring their own devices to be used in teaching. Their positive attitude towards digitalization makes them find solutions and innovate school systems. However, not all teachers are interested in keeping their teaching updated with the evolution of technology, lacking in interest or having a negative attitude towards it.

Our interviews also revealed that there were surprising facts about ICT development in Romanian schools, one of the subjects calling it the "wow" factor: there are scattered situations of teachers using state-of-the-art technology and bringing their schools to the upper end of the educational offer in schools from all over the world. One subject considered that the real "wow" of Romanian education was represented by teaching staff who were considered hard to be replaced by technology. Participants were all in favour of tele-education, but they said it would never replace face-to-face education and, as one subject pointed out, it should not, as children "need to socialize without ICT and work in teams and groups".

5. Discussion

Our study proposed an innovative research methodology focused on analyzing the asymmetries in digitalization of developing countries. We applied it to Romanian public education in which the shift from face-to-face learning to tele-education forced by the COVID-19 pandemic has brought

major challenges. The results showed that digitalization depended on, one side, on the necessary infrastructure, on the teachers' and students' levels of digital competences and on capabilities of accessing and attracting funding to support digitalization (school management's and ministerial bodies' capabilities); on the other side, it depended on the stakeholders' motivations and needs to keep up with state-of-the art technology and to use and develop the latest digital skills.

For the identified challenges (1—students' lack of Internet connectivity and/or electric power; 2—students' lack of proper ICT infrastructure; 3—students' lack of digital competences or ICT support; 4—teaching staff's lack of Internet connectivity, proper ICT infrastructure and digital competences; 5—legal issues with remote and online job tasks; and 6—cyber risks), there appears to be a need for national infrastructure development (for Challenges 1 to 4), changes in the education curricula (for Challenges 3, 4 and 6) and new legal support (for Challenges 5 and 6).

In our view, these challenges led to an educational crisis. In Romania, the present educational crisis has three major players:

- tele-learners who are or should be/become skilled users of ICT devices;
- educational institutions that are improperly digitalized in terms of infrastructure;
- tele-teaching staff that are low skilled (according to the E-Didactics framework [13]).

Considering that today's students are digital natives, whereas most teachers are digital immigrants [40], the asymmetry is even deeper. However, the concepts of digital natives and digital immigrants imply that digital skills are innate [41] and, therefore, they should or could not be trained, making the differences between these two generations seem too deep to overcome. This is not the reality [41], as in both generations there are people with high and low digital skills. This approach better highlights our findings that in both younger and older generations there are skilled and unskilled users of digital instruments and the difference lies in the individual's digital literacy. We also identified asymmetries in the use of digital skills and infrastructure in learning (Table 4) that need to be addressed at the national policy level as the socio-economic and technical disparities lead to educational disparities and a lack of equality of chances for all students in Romania. The necessary infrastructure is still not enough as tele-teaching requires a completely different set of pedagogical competences than traditional teaching (Comenius's [42]), as well as specific personality traits [43].

Table 4. Asymmetries in the Romanian public education's digitalization.

Asymmetry	Consequences
Asymmetry 1—Rural versus urban (student residence and school location).	**Socio-economic and technical disparities**
Asymmetry 2—Poor versus rich (student wealth and school patrimony).	
Asymmetry 3—Managerial funding versus no funding.	
Asymmetry 4—Developed versus undeveloped digital skills.	
Asymmetry 5—Teachers' high versus no motivation to digitalize.	↓
Asymmetry 6—School management's high versus low capability of accessing/attracting funding.	
Asymmetry 7—High versus low sustainability.	
Asymmetry 8—Educational gaps between primary, secondary and tertiary levels of education (ISCED 1–4).	
Asymmetry 9—Hardware versus software infrastructure and security.	**Educational Disparities**
Asymmetry 10—Full versus no access to infrastructure.	
Asymmetry 11—Gaps between digitalization between taught subjects.	
Asymmetry 12—Extra-work (project management) versus ordinary tasks.	

Although many specialists may call the education required in times of a pandemic online education, in developing countries, it cannot be named so as online education requires not only certain equipment, but also a specific pedagogy. Schooling during COVID-19 pandemic has created a precedent and it will not be so utterly new and unknown to turn to tele-education when there is no longer the possibility of meeting in classrooms. How teachers started developing ICT skills during this period is a topic that needs to be addressed in further research as well as its impact on students and their parents.

The need for a digital school is also triggered by an exponential development of devices that can help or induce learning. Students learn best when they use virtual environments that bring the real world into the classroom. Technology must be embedded in teaching and the key words should be interactive (during a tele-class, real-time dialogue can emerge between the educational actors, teachers getting the students' feedback and responses instantly and responding to their questions, too, using the digital tools offered) and personalized (teaching/learning tailored to each student's personal interests, while the learning situations are offered to all students using the same contents and methods).

Tele-education development, regardless of the name of the concept used, will further advance the research of ICT product design, development, production and promotion to address the educational needs of specific individuals and to reduce the lack of symmetry.

6. Conclusions

To conclude, the teaching and learning process, accompanied by assessments, have undergone a profound change in Romania during the latest pandemic, challenging the core of the education system, which, like in many developing countries, used to be mainly face-to-face education. Teachers and all the other education stakeholders had to face the students' readiness to learn in virtual environments as opposed to many teachers' reluctance to teach in virtual environments. The entire teaching and training system should be rethought so as to include e-didactics—a new topic absolutely necessary for educating future generations.

Author Contributions: Conceptualization, C.N. and R.H.; formal analysis, C.N. and R.H.; funding acquisition, C.N.; investigation, C.N. and R.H.; methodology, C.N. and R.H.; software, R.H.; supervision, C.N.; validation, N.R., A.N. and R.M.; writing—original draft, C.N., R.H. and A.N.; writing—review and editing, N.R. and R.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research was conducted and the article was prepared and published with funding from "Transilvania University Scholarship 2016" that C.N. benefitted from.

Acknowledgments: We are highly grateful to our participants who responded positively to our kind request as well as provided valuable information and to Mr Adrian Dascăl from Dascal Visual who improved the quality of our figures by processing them with Adobe Illustrator CC 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sohrabi, C.; Alsafi, Z.; O'neill, N.; Khan, M.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, R. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surg.* **2020**, *76*, 71–76. [CrossRef] [PubMed]
2. Karabag, S.F. An unprecedented global crisis! the global, regional, national, political, economic and commercial impact of the Coronavirus pandemic. *J. Appl. Econ. Bus. Res.* **2020**, *10*, 1–6.
3. Watson, I. The youth labour market: From education to work before and after the global financial crisis. *J. Ind. Relat.* **2020**, *62*, 33–57. [CrossRef]
4. Guillen-Gamez, F.D.; Mayorga-Fernandez, M.J. Quantitative-comparative research on digital competence in students, graduates and professors of faculty education: An analysis with ANOVA. *Educ. Inf. Technol.* **2020**. [CrossRef]
5. Al-Huneini, H.; Walker, S.A.; Badger, R. Introducing tablet computers to a rural primary school: An activity-theory case study. *Comput. Educ.* **2020**, *143*, 103648. [CrossRef]

6. Jacak, W.; Jahn, G.; Rozenblit, J. System theoretical approach to control and synchronization of teleeducation in open distributed environment. In *Computer Aided Systems Theory—EUROCAST'97*; Lecture Notes in Computer Science; Pichler, F., Moreno-Díaz, R., Eds.; Springer: Berlin/Heidelberg, Germany, 1997; Volume 1333.
7. Guérit, P. Tele-education: Benefits of space and integration of multimedia technologies. In *Space and the Global Village: Tele-Services for the 21st Century*; Haskell, G., Rycroft, M., Eds.; Springer: Dordrecht, The Netherlands, 1999; Volume 3.
8. Shen, K.; Yang, Y.; Wang, T.; Zhao, D.; Jiang, Y.; Jin, R.; Zheng, Y.; Xu, B.; Xie, Z.; Lin, L.; et al. Diagnosis, treatment, and prevention of 2019 novel coronavirus infection in children: Experts' consensus statement. *World J. Pediatr.* **2020**, *16*, 223–231. [CrossRef] [PubMed]
9. Li, Y.; Guo, F.F.; Cao, Y.; Li, L.F.; Guo, Y.J. Insight into COVID-2019 for paediatricians. *Pediatr. Pulmonol.* **2020**, *55*, E1–E4. [CrossRef] [PubMed]
10. Hong, H.; Wang, Y.; Chung, H.T.; Chen, C.J. Clinical characteristics of novel coronavirus disease 2019 (COVID-19) in new-borns, infants and children. *Pediatr. Neonatol.* **2020**, *61*, 131–132. [CrossRef] [PubMed]
11. Tabira, Y.; Otieno, F.X. Integration and implementation of sustainable ICT-based education in developing countries: Low-cost, en masse methodology in Kenya. *Sustain. Sci.* **2017**, *12*, 221–234. [CrossRef]
12. Pelton, J. A satellite based global education system: The Knowledge Network of the World. *Acta Astronaut.* **1992**, *26*, 835–839. [CrossRef]
13. Pelton, J.N.; Quigley, J. New opportunities in satellite tele-education. *Space Commun.* **1992**, *9*, 253–261.
14. Association of Flipped Learning Network. What Is Flipped Learning? Available online: https://flippedlearning.org/wp-content/uploads/2016/07/FLIP_handout_FNL_Web.pdf (accessed on 27 April 2020).
15. Brolpito, A. *Digital Skills and Competence, and Digital and Online Learning*; European Training Foundation: Turin, Italy, 2019; Available online: https://www.etf.europa.eu/sites/default/files/2018-10/DSC%20and%20DOL_0.pdf (accessed on 26 May 2020).
16. Tchoshanov, M. *Engineering of Learning: Conceptualizing e-Didactics*; UNESCO Institute for Information Technologies in Education: Moscow, Russia, 2013; pp. 21–24.
17. Vahtivuori-Hanninen, S. Pedagogical Models in Network-based Education. In *E-Training Practices for Professional Organizations*; Nicholson, P., Thompson, J.B., Ruohonen, M., Multisilta, J.E., Eds.; The International Federation for Information Processing: New York, NY, USA, 2005.
18. Shi, Y.C.; Xie, W.K.; Xu, G.Y.; Shi, R.; Chen, E.; Mao, Y.H.; Liu, F. The smart classroom: Merging technologies for seamless tele-education. *IEEE Pervasive Comput.* **2003**, *2*, 47–55.
19. Wang, L.; Han, D. Tele-education technology eliminating Chinese knowledge poverty based on information technology. In *Advances in Computer Science, Intelligent System and Environment*; Jin, D., Lin, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 1, pp. 327–332.
20. Srivastava, D. Advantages & disadvantages of e-education & e-learning. *J. Retail. Mark. Distrib. Manag.* **2019**, *2*, 22–27.
21. Xie, W.K.; Shi, Y.C.; Xu, G.Y.; Xie, D. Smart classroom—An intelligent environment for tele-education. In *Advances in Multimedia Information Processing—PCM 2001 Proceedings*; Shum, H.Y., Liao, M., Chang, S.F., Eds.; Springer: Berlin, Germany, 2001; Volume 2195, pp. 662–668.
22. Winer, L.; Coopenstock, J. The "intelligent classroom": Changing teaching and learning with an evolving technological environment. *Comput. Educ.* **2002**, *38*, 253–266. [CrossRef]
23. Curran, V.R. Tele-education. *J. Telemed. Telecare* **2006**, *12*, 57–63. [CrossRef] [PubMed]
24. Vernet, M.P.; Schilling, K. *Virtual Reality for Tele-Education Experiments with Remote Mobile Robot Hardware*; Pergamon Press: Oxford, UK, 2002.
25. *Terminology of European Education and Training Policy*; The European Centre for the Development of Vocational Training; Publications office of the European Union: Luxembourg, 2014; pp. 47–48. Available online: www.cedefop.europa.eu/it/publications-and-resources/publications/4117 (accessed on 17 April 2020).
26. Worldometer. Available online: https://www.worldometers.info/coronavirus/country/romania/ (accessed on 22 August 2020).
27. CNSCBT—Institutul National de Sanatate Publica. Available online: https://www.cnscbt.ro/index.php/lex/1748-ordonanta-de-urgenta-nr-70-2020-privind-reglementarea-unor-masuri-incepand-cu-data-de-15-mai-2020-in-contextul-situatiei-epidemiologice-determinate-de-raspandirea-coronavirusului-sars-cov-2-pentru-pre/file (accessed on 27 May 2020).

28. Portal Legislative. Available online: http://legislatie.just.ro/Public/DetaliiDocumentAfis/223831 (accessed on 27 May 2020).
29. Portal Legislative. Available online: http://legislatie.just.ro/Public/DetaliiDocumentAfis/224849 (accessed on 27 May 2020).
30. CNSCBT—Institutul National de Sanatate Publica. Available online: https://www.cnscbt.ro/index.php/lex/1805-hotararea-nr-476-2020-privind-prelungirea-starii-de-alerta/file (accessed on 27 July 2020).
31. Portal Legislative. Available online: http://legislatie.just.ro/Public/DetaliiDocument/229151 (accessed on 27 May 2020).
32. Ministerul Educației și Cercetării. Available online: https://www.edu.ro/structura-anului-%C8%99colar-2020-2021-fost-lansat%C4%83-%C3%AEn-dezbatere-public%C4%83 (accessed on 27 May 2020).
33. European Commission. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php/Government_expenditure_on_education#Large_differences_between_countries_in_the_importance_of_expenditure_on_education (accessed on 20 August 2020).
34. BCR. Available online: www.bncr.ro (accessed on 20 August 2020).
35. Nation Institute of Statistics. Available online: www.insse.ro (accessed on 20 August 2020).
36. European Commission. Digital Economy and Society Index. Available online: https://digital-agenda-data.eu/datasets/desi/indicators#desi-individual-indicators (accessed on 20 August 2020).
37. Naderifar, M.; Goli, H.; Ghaljaie, F. Snowball sampling: A purposeful method of sampling in qualitative research. *Strides Dev. Med. Educ.* **2017**, *14*. [CrossRef]
38. Law of National Education in Romania. Available online: https://edu.ro/sites/default/files/_fi%C8%99iere/Legislatie/2019/Legea%20nr%201%20Educatiei%20Nationale%20actualizata%202019.pdf (accessed on 20 July 2020).
39. Welcome to the Hacking STEM Library. Available online: https://www.microsoft.com/en-us/education/education-workshop/activity-library.aspx (accessed on 17 April 2020).
40. Prensky, M. *From Digital Natives to Digital Wisdom. Hopeful Essays for 21st Century Education*; Corwin. A SAGE Company: Thousand Oaks, CA, USA, 2012.
41. Boyd, D. *It's Complicated: The Social Lives of Networked Teens*; Yale University Press: London, UK, 2014.
42. Comenius, J.A. *Didactica Magna*; EDP: Bucharest, Romania, 1970.
43. Maican, C.I.; Cazan, A.M.; Lixandroiu, R.C.; Dovleac, L. A study on academic staff personality and technology acceptance: The case of communication and collaboration applications. *Comput. Educ.* **2019**, *128*, 113–131. [CrossRef]

symmetry

MDPI

Article

Predicting Business Innovation Intention Based on Perceived Barriers: A Machine Learning Approach

Carolina Rojas-Córdova [1,*], Boris Heredia-Rojas [2] and Patricio Ramírez-Correa [3]

[1] Department of Industrial Engineering, Universidad Católica del Norte, Antofagasta 1240000, Chile
[2] Department of Construction Management, Universidad Católica del Norte, Antofagasta 1240000, Chile; bheredia@ucn.cl
[3] School of Engineering, Universidad Católica del Norte, Coquimbo 1781421, Chile; patricio.ramirez@ucn.cl
* Correspondence: carojas@ucn.cl; Tel.: +56-55-2355000

Received: 24 June 2020; Accepted: 19 July 2020; Published: 19 August 2020

Abstract: In the Industry 4.0 scenario, innovation emerges as a clear driver for the economic development of societies. This effect is particularly true for the least developed countries. Nevertheless, there is a lack of studies that analyze this phenomenon in these nations. In this context, this study aims to examine the impact of perceived barriers to innovation to predict companies' innovative intentions in an emerging economy. This study is a preliminary effort to use data mining and symmetry-based learning concepts, especially classification, to assist the identification of strategies to incentivize intention to innovate in companies. Using the decision tree classification technique, we analyzed a sample of Chilean companies (N = 5876). The sample was divided into large enterprises (LEs) and small and medium enterprises (SMEs). In the group of large companies, the barriers that most impact the intention to innovate are innovation cost, lack of demand innovations, and lack of qualified personnel. Alternatively, in the group of small-medium companies, the barriers that most impact the intention to innovate are lack of own funds, lack of demand innovations, and lack of information about technology. These results show how the perceptions of barriers are significant to predict the intentions of innovation in Chilean companies. Furthermore, the perceptions of these barriers are contingent on the organizational sizes. These findings contribute to understanding the effect of contingencies on innovative intention in an emerging economy.

Keywords: innovation; business; machine learning; decision tree; predictive analytics; social data science; contingencies

1. Introduction

The fourth Industrial Revolution refers to a change in manufacturing logic in the direction of an increasingly decentralized and self-regulating approach to value creation; a process supported by concepts and new technologies that help companies meet future production requirements [1]. Innovation is key to this process of adopting concepts and new technologies, as exemplified by the adoption of the Internet of Things as an enabler of service innovation [2]. Additionally, both application-pull and technology-push directions for Industry 4.0 are related to innovation capabilities [3]. In this scenario, innovation emerges as a driver for the economic development of societies. This effect is particularly true for the least developed countries. Given that developed nations have a high intensity of knowledge, Industry 4.0 is used in these territories as a tool for further development of the knowledge economy. In contrast, in developing countries, Industry 4.0 is seen as a self-goal [4]. Besides, the development prospects of Industry 4.0 in the global economy indicate that in some future scenarios, this phenomenon may become a competitive advantage for developing countries compared to developed nations or at least a source of competitive parity. However, in other future scenarios, emerging economies will not be winners [5]. Despite the importance of understanding

innovation as crucial for Industry 4.0, as far as we know, there is a lack of studies that analyze business innovation in developing countries.

In this context, it is crucial to understand the barriers that prevent the fostering of innovation. Researchers have used different theoretical lenses to explain the obstacles to innovation. For example, scholars have explained these barriers from the perspective of absorptive capacity [6], dynamic capabilities [7], organizational learning [8], organizational routines [9], and social capital [10]. Most of these studies agree that the availability of resources can help explain the perception of barriers to innovation, since resources may shape capabilities and motivations. For instance, authors claim that resource-rich organizations are more capable of innovating, so they will encounter fewer barriers to innovation [11,12]. This perception of higher capabilities also promotes an aggregate sense of self- enhancement that motivates to initiate the exploration of new opportunities. Nevertheless, the availability of resources depends on the munificence of the countries. Munificence describes "the extent to which an environment supports the sustained growth of a firm" [13]. Developed countries are more munificent than developing countries. This means that in developed countries there are more opportunities to pursue and more availability of resources to pursue them [14]. In developing countries, this situation differs because organizations tend to be "deprived of superior technology and the supporting infrastructure often found in developed countries" [15]. Organizations are open systems that interact with the environment and depend on the availability of environmental resources [16]. Thus, we can presume that organizations in developed countries will distinctly perceive the barriers as organizations in developed countries. However, current studies do not account for the contingent dynamics that affect emerging countries because more of the empirical studies have assessed the barriers in developed countries c.f. [17,18].

As such, we asked how organizations in developing countries perceived barriers to innovation. To address this question, we examined the perception of barriers in organizations from Chile, a developing country. Consequently, this study aims to analyze the impact of perceived barriers to innovation to predict the innovative intention of companies in an emerging economy. We drew on a contingency theoretical lens that argues that mechanisms and barriers change according to internal and external contingencies [19]. Thus, we compared how perceptions change towards the barriers for innovation between large organizations that possess more resources, and small and medium organizations that possess fewer resources. We used a machine learning approach to assess this contingent model, in particular, the decision tree classification technique. This technique is a hierarchical structure consisting of nodes and directed edges that "solves a classification problem by asking a series of carefully crafted questions about the attributes of the test record" [20]. In general, this study procedure is an instance of the use of data mining and symmetry-based learning concepts for particular classification and subsequent prediction.

This study contributes to understanding and comparing the effects of the critical perceived barriers to the intention to innovate between different realities and environments, such as those that happen in large, and small-medium enterprises in an emerging country like Chile. These findings will help policymakers and managers make adequate decisions to avoid or eliminate these obstacles in a way that achieves better business performance and sustainable results.

The organization of this paper is as follows. In Section 2, we describe the secondary data gathered from the Chilean Innovation Survey and the method used to analyze the data about perceived barriers of innovation. The results of applying the decision tree method to predict the innovating perception in large and SMEs organizations are presented in Section 3. Section 4 includes the discussion, implications for practitioners and policymakers, and future research. The last section presents the main conclusions of this study.

2. Methods

2.1. Data

We used archival data as secondary sources of information [21]. The source of information was acquired from the Innovation Survey (EI-10) from the Ministry of Economy [22]. This survey was based in the Oslo Manual 3rd edition [23]. The data collected by this survey is part of the information included in the business innovation statistics and indicators database of the OCDE developed by the working party of National Experts on Science and Technology Indicators (NESTI) [24,25]. The Innovation questions cover two years of 2015–2016. The survey was self-applied using the web page of the Innovation Division of the Ministry of Economy from the Ministry of Economy. The analysts checked whether the online surveys were answered. In the case of surveys without answers, the analysts sent a reminder by e-mail. Alternatively, the analysts contacted the organizations by phone, offering help for answering the survey when needed. The mean number of contacts between the analyst and the informant was three. However, reaching the organizations could take up to seven tries [25]. The sample comprises 5876 Chilean organizations that had sales over US$90,000 per year. The demographic characteristics of the organizations in our sample are shown in Table 1. Sales per year defined the size of the organization. The small organizations comprise sales over US$90M and under US$880M and represent 41% of the sample. The medium organizations include sales over US$880M and under US$3500M and represent 23% of the sample. Finally, large organizations comprise sales of over US$3500M and represent 36% of the sample. The average organization age was 18 years (s.d. = 15.3). The organizations encompassed a broad range of industries, covering natural resources (13%), manufacturing (19%), electricity and water (2%), construction (10%), trade (16%), transport (6%), telecommunication (3%), and services (31%).

Table 1. Demographic characteristics of the sample.

Variables	Total	Percentage
Number of Organizations	5876	
Small Organizations	2406	41%
Medium Organizations	1369	23%
Larger Organizations	2101	36%
Organizations Age	18	
Location		
Capital	2012	34%
Regions	3864	66%
Industry		
Natural Resources	735	13%
Manufacturing	1094	19%
Electricity and Water	143	2%
Construction	577	10%
Trade	960	16%
Transport	336	6%
Telecommunication	203	3%
Services	1828	31%

According to the innovation type, Table 2 shows the intention of innovating in the next two years of these companies. For the study, we established a company as intending to innovate if it wanted to perform at least one innovation type.

This study considered the obstacles or barriers perceived by the companies of the sample as attributes to predict the intention to innovate. These obstacles were based on the guidelines of the Oslo Manual [23]. Table 3 shows that the analysis finds a dozen obstacles. Furthermore, given the possible differences between the company sizes, we included the size as an attribute in SME analysis. The scores detail the perceived importance of each obstacle by company size and in total. We categorized the

importance of the barriers as null, low, medium, and high, and for the calculation, we coded it as 0, 1, 2, and 3, respectively.

Table 2. Innovation intention of the sample.

Type of Innovation	In the Next Two Years, We Plan to Carry out This Type of Innovation (%)	
	Large	SME
Product	33.5	32.0
Process	41.8	28.3
Marketing	23.2	24.5
Management	33.3	25.5
Social	15.9	12.5
Product or Process or Marketing or Management or Social	*56.1*	*49.4*

Table 3. Perceptions of the importance of obstacles to innovation.

Obstacles	Large (%)				SME (%)			
	N	L	M	H	N	L	M	H
The very high cost of innovation	23.1	15.1	26.7	35.1	21.6	10.2	22.2	46.0
Lack of own funds	26.6	20.7	24.6	28.1	21.1	12.3	25.0	41.6
A market dominated by established companies	28.0	23.4	25.9	22.7	24.6	15.8	25.5	34.1
Uncertainty of demand for innovative goods or services	26.4	20.9	30.1	22.6	23.3	15.0	28.0	33.7
Lack of external financing to the company	31.3	22.9	25.9	20.0	26.3	14.9	26.7	32.2
Difficulty finding cooperation from partners for innovation	29.6	21.8	28.4	20.2	26.5	16.0	25.4	32.1
Lack of qualified personnel	25.4	22.0	30.6	21.9	23.8	15.5	30.2	30.5
Lack of information about technology	26.3	25.6	31.6	16.5	24.4	18.0	31.0	26.6
Lack of market information	29.2	26.8	29.6	14.4	25.1	19.4	30.5	25.0
Regulatory difficulty	37.9	28.1	21.4	12.6	36.5	24.4	21.8	17.3
Not necessary due to lack of demand for innovations	35.4	29.8	22.4	12.4	35.0	24.9	23.6	16.5
Not necessary due to previous innovations	40.0	30.1	20.3	9.6	42.6	24.5	20.6	12.3

N: Null; L: Low; M: Medium; H: High

2.2. Machine Learning Approach

We used a machine learning approach to achieve the objectives of this study. In particular, the prediction of innovation intention was conducted using decision trees. A decision tree is an inverted tree-shaped model made up of a set of nodes intended to decide on values affiliated to a class. This learning algorithm identifies a model that best fits the relationship between the attribute set and the class label of the input data. This research chooses this nonparametric technique due to several reasons following Tan, Steinbach and Kumar [20]. First, the decision trees do not require prior assumptions about the type of probability distributions satisfied by the class and other attributes. Second, the computational construction of decision trees is inexpensive and fast, even when the training set size is considerable. Third, the interpretation of decision trees, especially smaller-sized trees, requires less effort. Fourth, decision tree algorithms are quite robust to the presence of noise, especially when methods for avoiding overfitting are employed. Fifth, the accuracy of decision trees is not impacted by the presence of strongly correlated attributes and irrelevant attributes during preprocessing. Ultimately, the technique is useful for predictive modeling, i.e., anticipating the class label of unknown records, in this case, the perceived barriers to business innovation intention in small and medium-sized (SME) and large-sized organizations.

In the decision trees method, an algorithm is used to divide a dataset into categories belonging to the response variable. For the implementation we used RapidMiner. Specifically, we employed the C4.5 algorithm, which builds decision trees from a collection of training data using the notion of information entropy [26] see Figure 1.

We used a grid optimization strategy as a procedure for the setting of the parameters. In particular, the algorithm was optimized based on the split and stop criteria. The division criteria evaluated were

gain ratio, information gain, Gini index, and accuracy. Regarding the detention criteria, the maximum depth was assessed with possible values ranging from 2 to 25. In the case of SME companies, the procedure result indicates information gain as the division criterion and value three as the maximum depth. In regards to large companies, the procedure result indicates accuracy as the division criterion and value five as the maximum depth.

Input values: *D*: Dataset, *Tree*: Tree
Tree ← {}
if *D* is pure OR other stopping criteria met **then**
    break
end if
for all attribute a ϵ *D* **do**
    Calculate information-theoretic criteria if we split on a
end for
a* ← Best attribute giving to above calculated criteria
*Tree** ← Create decision node for finding a* in the root
*D** ← Sub-datasets form *D* based on a*
for all *D** **do**
    *Tree** ← C4.5(*D**)
    Attach *Tree** to the corresponding branch of *Tree*
end for
return *Tree*

Figure 1. Pseudocode of the C4.5 algorithm.

Additionally, to avoid overfitting, all analyses were performed using 10-fold cross-validation. In general, the cross-validation procedure consists of two phases. The first phase trains a model, and after that, the second phase applies the trained model and measures its performance. In the case of 10-fold cross-validation, the procedure divides the data sample into ten subsets of equal size. Of the ten subsets, the method preserves a single subgroup as test data, and the remaining nine subsets are used as training data. This process is repeated ten times, and each of the ten subsets is used once as test data. Lastly, the procedure averages the results of the ten iterations to produce a single estimation.

Finally, the two-class criteria measured the performance prediction: sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), precision = TP/(TP + FP), and accuracy = (TP + TN)/(TP + FP + FN + TN); where TP is truly positive, TN is a true negative, FP is false positive, and FN is a false negative.

3. Results

In the case of SME companies, prediction results in Table 4 indicate that the method performs well concerning the ability to select the cases that need to be chosen (innovation intention of SME companies) with a sensitivity of 81.90% ± 3.25%. The values of the other calculated criteria are specificity 46.45% ± 4.71%, precision 59.91% ± 2.11%, and accuracy 63.95% ± 2.44%.

Table 4. Decision tree confusion matrix for small and medium enterprises (SMEs) companies.

Accuracy: 63.95%	No Innovation Intention (True)	Innovation Intention (True)	Class Precision
No Innovation Intention (pred.)	816	310	72.47%
Innovation Intention (pred.)	941	1403	59.85%
Class recall	46.44%	81.90%	

According to the results, the following points guide the intention of innovation for SME companies. Firstly, when the perception of lack of own funds as an obstacle to innovation is nil, the perception of lack of information on technology arises as a discriminating variable; in this group, companies that perceive this obstacle as null do not declare their intention to innovate. Secondly, when there is a perception of a lack of own funds as an obstacle to innovation, the perception of lack of demand

predicts their intention to innovate. The companies in this group that perceive it as a critical barrier "not necessarily due to lack of demand for innovations" indicate that they have no intention to innovate. In contrast, companies that perceive this barrier as medium or less declare a higher purpose to innovate. Figure 2 shows these results.

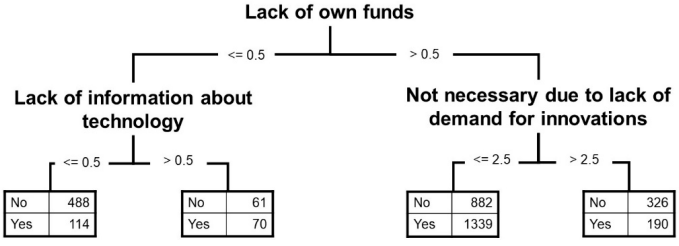


Figure 2. Decision tree graph results for innovation intention for SME companies.

In the case of large companies, prediction results in Table 5 indicate that the method performs properly concerning the ability to select the cases that need to be chosen (innovation intention of large companies) with a sensitivity of 87.11% ± 2.39%. The values of the other calculated criteria are specificity 40.62% ± 3.86%, precision 65.24% ± 1.71%, and accuracy 66.71% ± 2.39%.

Table 5. Decision tree confusion matrix for large companies.

Accuracy: 66.71%	No Innovation Intention (True)	Innovation Intention (True)	Class Precision
No Innovation Intention (pred.)	429	174	71.14%
Innovation Intention (pred.)	627	1176	65.22%
Class recall	40.62%	87.11%	

Following the results, the subsequent perceptions guide the intention of innovation for large companies. Firstly, when the perception of the very high cost of innovation as an obstacle to innovation is nil, various understandings of barriers emerge as essential to predict, for the most part, the lack of intention to innovate. Secondly, when there is a perception that the very high cost of innovation is an obstacle to innovation, but the perceived importance of the barrier "not necessary due to lack of demand for innovations" is not high, companies have a greater intention to innovate. Figure 3 shows these results.

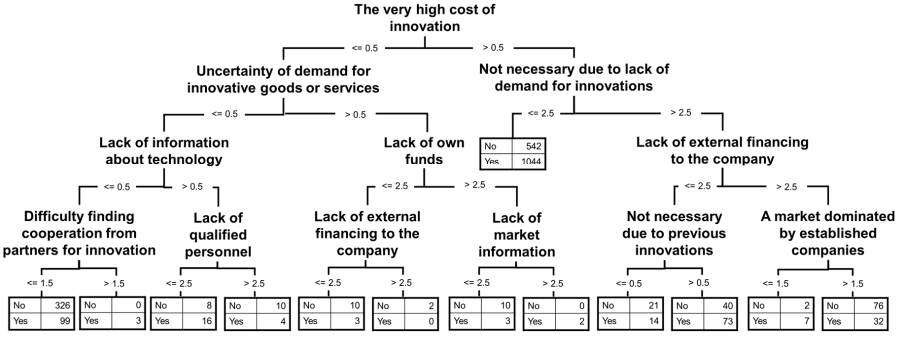


Figure 3. Decision tree graph results for innovation intention for large companies.

4. Discussions

Data science is rising as an interdisciplinary field that mixes statistics, data mining, machine learning, and analytics to understand and explain how to generate prediction models. Consequently,

the value and effectiveness of problems related to social data and data science are being recognized in social disciplines. Particularly in the management discipline, since data science methods allow scholars, among other benefits, to get better answers to existing questions, more immediate and accurate results are expected to evaluate existing theories [27]. In this light, this study has used a machine learning approach applied to business innovation data as a way to explore the barriers that exist for this behavior, and their results confirm the usefulness of data science to evaluating conceptual propositions in business. In this vein, we believe that the analysis findings could reveal the nonlinear effects of independent variables in the intention to innovate, as stated [28].

Our findings show how organizational resources moderate the perception of barriers to innovate. Small and medium organizations (SMEs) perceive these barriers differently than large organizations. For SMEs, the awareness of lack of funds appeared as a relevant discriminant variable for the intention to innovate. Nevertheless, in both cases (when organizations do or do not perceive the lack of own funds as a barrier), the absorptive capacity of these organizations explained their intention to innovate. Absorptive capacity describes the organizational ability to scan the change of environment. In our study, we can recognize this capacity in organizations that can identify the lack of information about technology or demand. In other words, although SMEs perceive the lack of resources as a barrier, yet they can recognize the environmental needs, SMEs will show a higher intention to innovate.

Alternatively, in the case of large organizations, the cost of innovation is a more relevant discriminant variable than the perception of own funds since these organizations count with their resources. When large organizations perceived the high cost of innovation as nil, several barriers relative to technology and market uncertainty emerged as predictors of the intention to innovate. On the contrary, when these organizations perceived a high cost of organizations yet recognized that there was a demand for innovation, they also showed an interest to innovate. Thus, as the same as SMEs, we can suggest that organizations that can scan the environmental needs or have higher absorptive capacity will show a higher intention to innovate. This last result is in agreement with previous studies [10,29,30].

These results are especially interesting for both government and managers. The findings could be used in the development of public policy for supporting and encouraging innovation in large, medium, and small-sized companies. These government policies can help countries remain competitive in a global market and improve firms' competitiveness and sustainability through direct implications for employment and a country's economic viability. The results may also provide insights for managers who are attempting to encourage innovation, especially when the Industry 4.0 era becomes more widespread in developed countries and world-class industries.

Understanding these perceived barriers can help decision-makers foster an innovative culture by supporting new technology strategies or avoiding an attitude of resistance to new ideas. The rules of business are changing with the inclusion of Industry 4.0, where the consumer market is looking for smart products and services more personalized to satisfy unique needs. However, many traditional industries continue to operate under marketing strategies that have demonstrated their ineffectiveness. Companies that are transitioning into Industry 4.0 need to plan new marketing actions [27] if predicting the business innovation intention based on perceived barriers is the first step to advance.

This study is not without limitations, which suggests avenues for future research. The main limitation is the heterogeneity of our sample, reducing the accuracy of the study results. Our sample includes a wide variety of industries that moderate the intention to innovate of the organizations. Thus, future research could examine which are the barriers to innovation within specific industries. Another limitation is that the organizations of our study are in Chile. As such, the results can be contingent on the country's characteristics. However, the data of the sample of this study is based on an OCDE survey. Thus, future studies could benefit from geographically diverse research settings. Finally, future studies could explore other machine learning techniques to improve the predictive capacity of the models, such as random forest or support vector machine, or adding different attributes to the learning process.

5. Conclusions

Innovation emerges as a fundamental driver for the economic development of societies and more in the current Industry 4.0 era. However, there are a lack of studies that analyze business innovation in developing countries compared to empirical research available in developed countries. Thus, this study investigates the impact of perceived barriers for business innovation to predict companies' innovative intentions in an emerging economy like Chile. Drawing on the contingent theory and applying a machine learning approach, the research concludes that organizational resources moderate the perception of barriers to innovate. SMEs perceive these barriers differently than large organizations. Notably, the knowledge of lack of own funds appears as a relevant discriminant variable for the intention to innovate for SMEs. At the same time, for larger companies, the cost of innovation is the most pertinent discriminant variable since these organizations count with their resources. Additionally, large firms and SMEs that can scan the environment, recognize the market needs, and have higher absorptive capacity will show a higher intention to innovate.

These findings can benefit organizations in other emerging economies to predict the innovation intention of organizations based on perceived barriers. Understanding these perceived barriers can help decision-makers foster an innovative culture by supporting new technology strategies or avoiding an attitude of resistance to new ideas. For example, countries and companies that are transitioning into Industry 4.0 could establish new actions, and policies about innovation intention referred to the perceived barriers.

The study's results from Chile cannot be generalized entirely. However, since these barriers are based on an international standard, Oslo Manual, the methodology applied in this study can be used to predict the business innovation intentions of large, medium, and small enterprises in other developing countries and verify the moderator effect of organizational resources.

Author Contributions: Conceptualization, C.R.-C., B.H.-R. and P.R.-C.; methodology, P.R.-C.; software, P.R.-C.; validation, C.R.-C. and B.H.-R.; formal analysis, C.R.-C. and B.H.-R.; investigation, C.R.-C. and B.H.-R.; resources, C.R.-C. and B.H.-R.; data curation, C.R.-C.; writing—original draft preparation, C.R.-C., B.H.-R. and P.R.-C.; writing—review and editing, C.R.-C. and B.H.-R.; visualization, C.R.-C. and B.H.-R.; supervision, C.R.-C.; project administration, C.R.-C.; funding acquisition, C.R.-C. All authors have read and agreed to the published version of the manuscript.

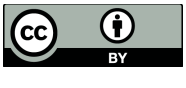
Funding: The APC was partially funded by UCN.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hofmann, E.; Rüsch, M. Industry 4.0 and the current status as well as future prospects on logistics. *Comput. Ind.* **2017**, *89*, 23–34. [CrossRef]
2. Andersson, P.; Mattsson, L.G. Service innovations enabled by the "internet of things". *IMP J.* **2015**, *9*, 85–106. [CrossRef]
3. Lasi, H.; Fettke, P.; Kemper, H.G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [CrossRef]
4. Bogoviz, A.V.; Osipov, V.S.; Chistyakova, M.K.; Borisov, M.Y. Comparative analysis of formation of industry 4.0 in developed and developing countries. *Stud. Syst. Decis. Control* **2019**, *169*, 155–164.
5. Sozinova, A.A. Causal connections of formation of industry 4.0 from the positions of the global economy. In *Industry 4.0: Industrial Revolution of the 21st Century*; Springer: Cham, Switzerland, 2019; Volume 169, pp. 131–143.
6. Zahra, A.; George, G. Absorptive Capacity: A Review, Reconceptualization, and Extension. *Acad. Manag. Rev.* **2002**, *27*, 185–203. [CrossRef]
7. Teece, D.J.; Pisano, G.; Shuen, A. Dynamic Capabilities and Strategic Management. *Strateg. Manag. J.* **1997**, *18*, 509–533. [CrossRef]
8. Argote, L.; Miron-Spektor, E. Organizational: From Experience to Knowledge. *Organ. Sci.* **2011**, *22*, 1123–1137. [CrossRef]
9. Feldman, M.S.; Pentland, B.T. Reconceptualizing Organizational Routines as a Source of Flexibility and Change. *Adm. Sci. Q.* **2003**, *48*, 94–118. [CrossRef]

10. Kwon, S.W.W.; Adler, P.S. Social Capital: Maturation of a field of Research. *Acad. Manag. Rev.* **2014**, *39*, 412–422. [CrossRef]
11. Cao, Q.; Gedajlovic, E.; Zhang, H. Unpacking Organizational Ambidexterity: Dimensions, Contingencies, and Synergistic Effects. *Organ. Sci.* **2009**, *20*, 781–796. [CrossRef]
12. Voss, G.B. The effects of slack resourcs and environmental threat on product exploration and exploitation. *Acad. Manag. J.* **2008**, *51*, 147–164. [CrossRef]
13. Dess, G.G.; Robinson, R.B. Measuring organizational performance in the absence of objective measures: The case of the privately-held firm and conglomerate business unit. *Strateg. Manag. J.* **1984**, *5*, 265–273. [CrossRef]
14. Lavie, D.; Stettner, U.; Tushman, M.L. Exploration and exploitation within and across organizations. *Acad. Manag. Ann.* **2010**, *4*, 109–155. [CrossRef]
15. Mesquita, L.F.; Lazzarini, S.G. Horizontal and Vertical Relationships in Developing Economies: Implications for SMEs' Access to Global Markets. *Acad. Manag. J.* **2008**, *51*, 359–380. [CrossRef]
16. Anand, G.; Ward, P.T.; Tatikonda, M.V.; Schilling, D.A. Dynamic capabilities through continuous improvement infrastructure. *J. Oper. Manag.* **2009**, *27*, 444–461. [CrossRef]
17. Zapata-Cantu, L. Boosting innovation in emerging markets: The moderating role of human capital. *Int. J. Emerg. Mark.* **2020**. Available online: https://www.emerald.com/insight/1746-8809.htm (accessed on 24 June 2020). [CrossRef]
18. Belausteguigoitia Rius, I.; De Clercq, D. Knowledge sharing and unethical pro-organizational behavior in a Mexican organization: Moderating effects of dispositional resistance to change and perceived organizational politics. *Manag. Res.* **2018**, *16*, 248–269. [CrossRef]
19. Flynn, B.B.; Huo, B.; Zhao, X. The impact of supply chain integration on performance: A contingency and configuration approach. *J. Oper. Manag.* **2010**, *28*, 58–71. [CrossRef]
20. Tan, P.N.; Steinbach, M.; Kumar, M. *Introduction to Data Mining*; Pearson Education: London, UK, 2016.
21. Keats, B.W.; Hitt, M.A. A Causal Model of Linkages Among Environmental Dimensions, Macro Organizational Characteristics, and Performance. *Acad. Manag. J.* **1988**, *31*, 570–598.
22. Ministerio de Economía Fomento y Turismo. *Décima Encuesta de Innovación en Empresas, 2015–2016*; Ministerio de Economía Fomento y Turismo: Santiago de Chile, Chile, 2017.
23. OECD/Eurostat. *OSLO Manual: Guidelines for Collecting and Interpreting Innovation Data*, 3rd ed.; OECD: Paris, France, 2005.
24. OECD. Business Innovation Statistics and Indicators. Available online: https://www.oecd.org/sti/inno/inno-stats.htm#indicators (accessed on 6 July 2020).
25. Bello, M. *Innovation Survey Metadata Wave 2014–2016*; OECD: Paris, France, 2019.
26. Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16*, 235–240. [CrossRef]
27. George, G.; Osinga, E.; Lavie, D.; Scott, B. Big data and data science methods for management research. *Acad. Manag. J.* **2016**, *59*, 1493–1507. [CrossRef]
28. Varian, H.R. Big data: New tricks for econometrics. *J. Econ. Perspect.* **2014**, *28*, 3–28. [CrossRef]
29. Jansen, J.; Van Den Bosch, F.; Volberda, H.; Van den Ven, F. Explorative Innovation, Exploitative Innovation and Performance: Effects of Organizational and Environmental Moderators Antecedents. *Manag. Sci.* **2006**, *52*, 1661–1674. [CrossRef]
30. Uotila, J.; Maula, M.; Keil, T.; Zahra, S.A. Exploration, exploitation, and financial performance: Analysis of S&P 500 corporations. *Strateg. Manag. J.* **2009**, *30*, 221–231.



symmetry

MDPI

Article

Blockchain Paradigm for Healthcare: Performance Evaluation

Leila Ismail * and Huned Materwala

Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain, Abu Dhabi 15551, UAE; huned.m@uaeu.ac.ae
* Correspondence: leila@uaeu.ac.ae

Received: 2 June 2020; Accepted: 17 July 2020; Published: 22 July 2020

Abstract: Electronic health records (EHRs) have become a popular method to store and manage patients' data in hospitals. Sharing these records makes the current healthcare data management system more accurate and cost-efficient. Currently, EHRs are stored using the client/server architecture by which each hospital retains the stewardship of the patients' data. The records of a patient are scattered among different hospitals using heterogeneous database servers. These limitations constitute a burden towards a personalized healthcare, when it comes to offering a cohesive view and a shared, secure and private access to patients' health history for multiple allied professionals and the patients. The data availability, privacy and security characteristics of the blockchain have a propitious future in the healthcare presenting solutions to the complexity, confidentiality, integrity, interoperability and privacy issues of the current client/server architecture-based EHR management system. This paper analyzes and compares the performance of the blockchain and the client/server paradigms. The results reveal that notable performance can be achieved using blockchain in a patient-centric approach. In addition, the immutable and valid patients' data in the blockchain can aid allied health professionals in better prognosis and diagnosis support through machine learning and artificial intelligence.

Keywords: artificial intelligence; blockchain; client/server; electronic health records; health information management; privacy; security

1. Introduction

Healthcare data management is the process of storing and analyzing patients' health records to improve patient treatment, track the causes of diseases efficiently, manufacture effective medicines and establish an accurate prevention agenda. The early form of data management includes documenting patient's complaints, diagnosis and the corresponding treatment manually introduced in a health record. Later, with the development of digital data, electronic health records (EHRs) came into existence [1]. For accurate health care, EHRs are often required to be shared among different healthcare organizations, medical drug manufacturers, pharmacists, medical insurance providers, researchers and patients. This poses a serious challenge in keeping the patients' sensitive data secure and up to date. A patient may visit/get transferred to different hospitals during the treatment lifecycle. A patient in such a situation owns the right over their own medical data and may require defining access control limits, says the U.S Department of Health and Human Services [2] and the European Parliament and the Council of the European Union [3]. The patient needs to sign a consent form stating what data will be shared and with whom and for how long. The consent form and the data are then posted to the recipient instead of sending them over the Internet due to security concerns [2]. Consequently, the process of sharing a patient's data between multiple hospitals becomes complex, time-consuming and difficult to coordinate. With the trend toward a personalized healthcare for better diagnosis

and prognosis, current EHR systems using the client/server approach are limited when it comes to providing a cohesive view and secure shared-access to patients' health history to multiple stakeholders, including the patients. In a centralized client/server approach, electronic health records are vulnerable to mistakes that can lead to life-threatening situations. Moreover, the patients need to trust the service provider and their exists privacy and security concerns.

Blockchain [4,5] is a very promising technology that enables a secure, private and distributed environment among peers without any trusted third-party using consensus. It is based on a shared, distributed and immutable ledger. Each transaction in the blockchain network is processed and validated by the majority of the network participants that eliminates the need of a trusted third-party. The validated transactions after verification are packed in blocks. Each block is linked to the preceding block by hashing the block's data along with the previous block's hash, providing immutability. The blocks in the network are considered valid when more than 50% of the participants reach an agreement using a consensus algorithm [6]. The immutable and replicated blockchain ledger is capable of solving the issues of scattered data, delayed sharing, lack of audit trail, privacy and security that prevail in the client/server model [7]. In addition, blockchain has an important characteristic of enforcing smart contracts—pieces of codes that are executed automatically once certain conditions are met. Blockchain-based EHRs have a tremendous potential in healthcare to enable allied health professionals to manage and share, not only clinical data, but also important patient-reported social and contextual data. Moreover, implementing artificial intelligence approaches on the ledger data that includes patients' health data from all the hospital in the network can aid allied health professionals in better prognosis and diagnosis support. While many researchers investigated the development blockchain-based system for healthcare data management, most of the works focus on the comparison of different blockchain development platforms [8–10]. As far as we know, this is the first work to evaluate and compare the blockchain technology with the current client/server model.

The major contributions of the paper are as follows:

- While the blockchain characteristics are suitable for implementing a healthcare system, these mechanisms are still costly considering execution time and amount of data transferred for ledger update.
- In spite of these costly mechanisms, notable performance can be achieved thanks to the blockchain model, especially in a patient-centric approach. In this approach, the patients and/or the physicians are constantly visiting the health records to construct a cohesive view from different hospitals for a better diagnosis or prognosis of diseases using artificial intelligence.

The rest of the paper is organized as follows. We overview the related work in Section 2. In Section 3, we introduce the principles of blockchain and its benefits to healthcare. We discuss the system model for the developed blockchain-based healthcare platform in Section 4. We present the experiments, and comparative analysis of client/server model versus blockchain using several application scenarios in Section 5. Section 6 concludes this paper.

2. Related Work

Traditionally, EHRs are employed by the healthcare organizations using a client/server architecture where the hospitals retain primary stewardship [11]. The medical data of a patient receiving treatment from multiple hospitals are scattered among different databases. To address this issue, several cloud-based eHealth applications are proposed [12–17]. However, privacy and security are the major concerns in the client/server and cloud-based models. Several research efforts address the privacy concern by managing the health data in cloud storage and recording the hash of the data in a special blockchain network [18–25]. Wang et al. [26–33] propose mechanisms for data integrity and authenticity of health records in a blockchain.

Several works [8–10,34–43] propose a blockchain-based healthcare data management system involving multiple hospitals. Azaria et al. [35] propose an interoperable blockchain-based application

that enables allied health professionals to share patients' health records. Dagher et al. [8] and Li et al. [36] propose an Ethereum-based blockchain framework for smart contract enabled medical data access. However, the blockchain networks in [8,35,36] use the energy-hungry and non-scalable Proof of Work (PoW) consensus mechanism [44–46]. Fan et al. [9] and Zghaibeh et al. [34] develop an EHR blockchain-based framework using Practical Byzantine Fault Tolerance (PBFT) consensus which is more energy-efficient and has better performance than PoW [47]. The healthcare data management systems proposed by [37,38,42,43] allow patients to retain the primary stewardship of the medical data because only the patients have the right to upload their data to the network. In these systems, the allied health professionals can only access the patients' health record. Uddin et al. [10,39–41] propose a blockchain-based healthcare system with both allied health professionals and patients having data update authority. As far as we know, there is no work that compares the client/server and blockchain-based healthcare data management systems. In this paper, we implement a minimal blockchain-based healthcare platform and compare its execution time and amount of data transferred with the client/server system model for health records update and query. This is with increasing number of records and hospitals.

3. Blockchain Paradigm

Figure 1 shows the blockchain architecture. The architecture consists of the following layers [6]:

1. *Infrastructure layer:* It includes the network nodes (known as participants), network modules and storage provisions. There are three types of participants: (1) simple which only performs the transactions, (2) validating which performs and validates transactions, and has a copy of the ledger and (3) mining which generates a new block and has a copy of the ledger.
2. *Platform layer:* It includes modules for communication between the blockchain participants.
3. *Computing layer:* It includes the underlying blockchain mechanisms for immutability, availability, finality, provenance, privacy and security.
4. *Application layer:* It enables the blockchain participants to communicate with the application.

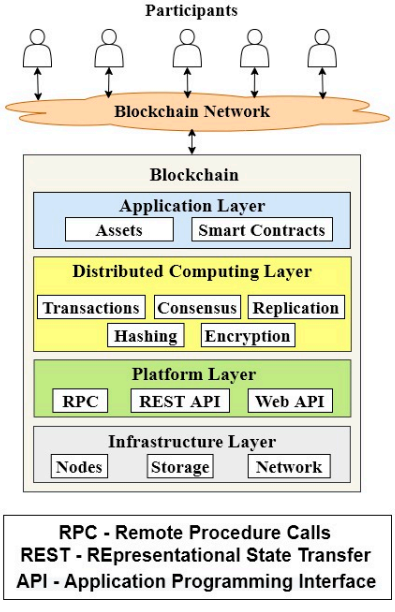


Figure 1. Blockchain architecture.

A blockchain network can be classified as either permission-less or permissioned [48]. The permission-less network (also called public) is open for anyone to join using a pseudo-name [49]. The users are encouraged to join the network for enhanced security by offering incentives. The ledger data in this network are visible to everyone. The permissioned network (also known as private) is an invitation-based network that allows only authorized participation. The visibility of the data is subject to access control rights as defined by the certificate authority. Considering the restricted network participation, the permissioned blockchain is more scalable and has a higher throughput compared to the permission-less.

3.1. Features of Blockchain

Blockchain is a peer-to-peer network that enables the stakeholders to share transactions with each other ensuring decentralization. It has the following features:

- *Decentralization:* A centralized third-party is not required as the ledger is updated after the majority of the participants in the network reaches a consensus.
- *Immutability:* A block in the ledger is hashed using its contents and the hash of the previous block. Consequently, any modification in a block will modify all the following blocks in the ledger. This makes the modification of a block in blockchain computationally difficult because the ledger is replicated among peers. In case data are entered by error, these data are corrected by issuing a new transaction.
- *Transparency:* Any change in the network is recorded as a transaction and can be viewed by all the participants maintaining a copy of the ledger.
- *Traceability:* The replication of any event in the network enables convenient tracing and audit trail.
- *Trustless:* Participants unknown to each other can perform transactions among each other as the consensus mechanism maintains the trust in the network.

3.2. Transaction Execution Mechanism

A transaction is an action that alters the blockchain ledger. It is application dependent and can be the transfer of monetary assets or execution of a smart contract. The transaction execution flow in blockchain is as follows:

- *Transaction proposal:* The user hashes the transaction using a hashing algorithm. The user's private key is then used to encrypt this hashed value. The result is known as the digital signature. The digital signature along with the data is broadcasted to the network.
- *Transaction validation:* The transaction is validated by each validating node. This is by authenticating the user identity and ensuring the data integrity. The identity is authenticated by decrypting the signature and the integrity is ensured by hashing the transaction and comparing it with the decrypted result. The valid transaction is sent to the mining node.
- *Block generation:* The mining node (selected based on the consensus protocol used) verifies the valid transactions and groups them in a block in a way that the block size does not exceed a predetermined threshold. It hashes the transactions data, block version, timestamp and previous block's hash value, and then hashes this hash value to obtain the hash of the block. The miner broadcasts the block to the network.
- *Replication:* The validating and mining nodes verify the validity of the block as part of the consensus protocol. Once valid, each node updates its copy of the ledger by appending the block.

3.3. Benefits to Healthcare

A blockchain-based healthcare platform provides the following benefits compared to the client/server approach:

- *Fault tolerance:* In a client/server-based system the patients' health data are managed in a centralized database. Once the data are lost, they cannot be recovered. The replication characteristic of blockchain aids in fault tolerance.
- *Data sharing:* In the current client/server systems, a patient's data are scattered over multiple hospitals' databases. The sharing of data among different hospitals and medical organizations is a complex process. However, in a blockchain-based platform, the patients' data recorded in the ledger is replicated among all the hospitals in the network.
- *Interoperability:* In a client/server-based system, each hospital stores the patients' data in a different database using heterogeneous data formats and structures resulting in interoperability challenges. The synchronized and replicated ledger in the blockchain solves this issue.
- *Avoidance of tests repetition:* Currently the patients' data are scattered across different healthcare providers, a patient often needs to repeat various laboratory and pathological tests. This not only incurs huge medical bills but also has adverse effects on the human body. The replicated blockchain ledger aids in avoiding medical tests.
- *Security:* The existing client/server-based system is prone to different cyber-attacks such as phishing and hacking. The stolen health records can be used to buy medical equipment by creating a fake ID or combining a patient number with a false provider to claim medical insurance. Table 1 shows the number of health data records breached in America based on a report by the Health Insurance Portability and Accountability Act (HIPAA) [50], and the cost per breached health record based on a report by the Federal Bureau [51] between 2009–2019. This cost of health record breach includes the expenses for forensic experts, outsourcing hotline support, the value of customer loss and free subscriptions and discounts for future services [52]. The table shows a spike in the number of health records breached in 2015. This is due to the largest health records breach encountered so far by the health insurance company, Anthem, with almost 78.8 million individuals affected as the patients' records were not encrypted [53]. The immutability feature of blockchain ensures data security.

Table 1. Number of health record breaches and cost per breached health record between 2009–2019.

Year	Number of Health Records Breached	Cost per Breached Record (USD)
2009	0	204
2010	6,006,063	214
2011	13,407,992	194
2012	2,808,042	233
2013	7,401,928	255
2014	12,946,972	308
2015	113,270,000	363
2016	27,300,000	355
2017	5,138,179	380
2018	13,947,909	408
2019	41,335,889	429

4. A Blockchain-Based Healthcare System Model

We develop a basic blockchain platform for healthcare which provides minimal functionality to program healthcare transactions. This is using a permissioned blockchain network due to its advantages over the permission-less. The blockchain-based healthcare system model consists of participants such as patients, allied health professionals (doctors, nurses and pharmacists) and administrators; assets such as medical test data; and transactions such as health record update and query. Figure 2 shows the developed blockchain-based healthcare platform overview. It shows that the platform involves several hospitals and participants, including patients, connected to the blockchain network. A hospital includes different departments such as radiology and pathology laboratories

as well as doctors, nurses, pharmacists and administrators. Let N represent total hospitals, K total participants and M total patients ($M < K$) in the network. Each hospital i, $i \in \{1, \ldots, N\}$, maintains a copy of the ledger. A participant k, $k \in \{1, \ldots, K\}$ updates/queries the health record of a patient j, $j \in \{1, \ldots, M\}$. For every healthcare transaction, the doctors and pharmacists act as full nodes and the administrator acts as a mining node. The developed minimal blockchain-based healthcare platform supports two types of transactions: (1) data update and (2) data query.

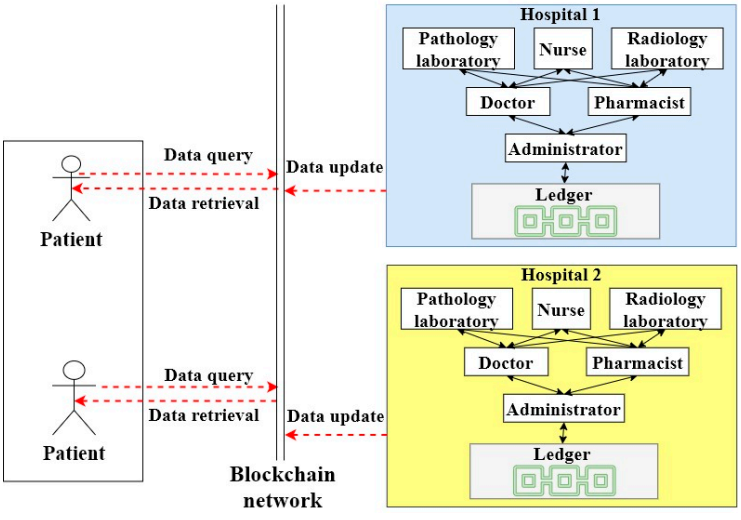


Figure 2. Blockchain-based healthcare data management platform.

In the developed platform, in order to perform a healthcare transaction, the patient and/or allied health professionals send the transaction to the blockchain network. The doctors and pharmacists validate the transaction, and the validated transaction is sent to the administrator that acts as a miner. The administrator will generate the block for the transaction and broadcast it to all other hospitals' administrators for replication. The execution flow for the healthcare transaction in the developed blockchain-based healthcare platform is shown in Figure 3. The selection of the miner and the ledger update is done using a consensus protocol.

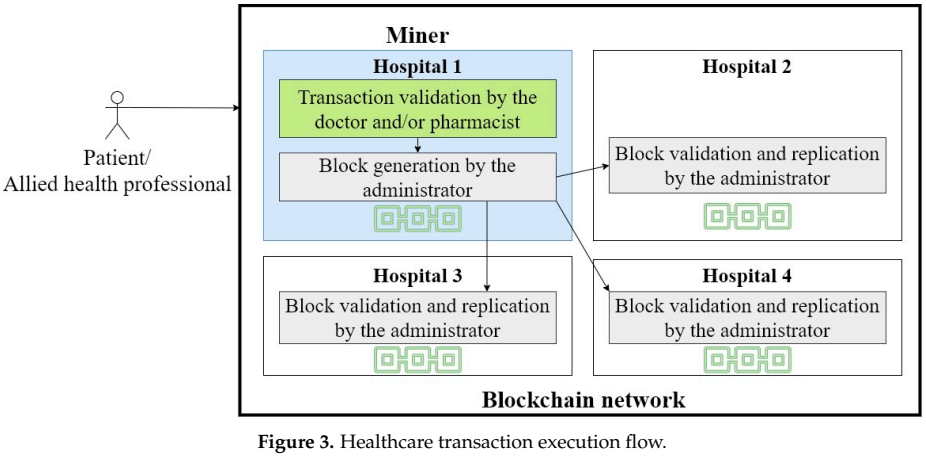


Figure 3. Healthcare transaction execution flow.

The consensus protocol used in the developed blockchain-based healthcare platform is the PBFT [54]. The PBFT consensus protocol is selected over the mostly used Proof of Work (PoW) because the latter consumes more energy [44,45] and less throughput [46] than the former. In PBFT, the mining nodes are known as peer nodes. A leader node is selected from the peers in a round-robin manner. The leader receives all the transactions from the network participants. The transactions are validated and a block is generated. The block is then broadcasted to all the peers. Each peer node verifies the transactions in the block, recalculates the block's hash and broadcasts the block's hash to all other peer nodes in the network. Each peer node then updates its ledger with the block if it receives the same block's hash value from 2/3 of the total peer nodes in the network.

A block in the ledger corresponding to a health record of a patient j primarily consists of the transaction with the workload $W_t(j)$ and the block's hash with the workload $W_{bh}(j)$. The total workload of the block is represented by $W_b(j)$, which is the summation of $W_t(j)$ and $W_{bh}(j)$. When a participant k initiates an update/query transaction to a health record of patient j, the execution time (ET_j) is given by Equation (1).

$$ET_j = TC_j + TP_j \tag{1}$$

where TC_j represents the communication time for transferring the total workload between the communicating nodes. It is calculated by dividing the total workload by the network bandwidth between the communicating nodes. TP_j indicates the processing time which is calculated as the summation of the time taken to generate the digital signature of the transaction j, the time taken to generate the block by the leader node and the time taken to validate the hash of the block by the peer nodes. In this paper, we consider the communication time and neglect processing time. The communication time (TC_j) for the blockchain ledger update/query is calculated as stated in Equation (2).

$$TC_j = TC_{j(k,miner_i)} + TC_{j(validator_i,leader_i)} + TC_{j(leader_i,peer)} + TC_{j(peer)} + TC_{j(miner_i,k)} \tag{2}$$

where $TC_{j(k,miner_i)}$ represents the communication time for transferring the health record transaction of patient j between the participant k and the blockchain mining node i (hospital), $TC_{j(validator_i,leader_i)}$ indicates the transaction communication time between the validator and the miner of the mining node i, $TC_{j(leader_i,peer)}$ denotes the block communication time between the leader of the node i and the peer nodes, $TC_{j(peer)}$ represents the block's hash communication time between the peer nodes and $TC_{j(miner_i,k)}$ denotes the communication time for the acknowledgement between the mining node i and the participant k. Equations (3)–(7) show the calculation of these terms.

$$TC_{j(k,miner_i)} = \frac{W_t(j)}{C_{k,i}} \tag{3}$$

$$TC_{j(validator_i,leader_i)} = \frac{W_t(j)}{C_{validator_i,leader_i}} \tag{4}$$

$$TC_{j(leader_i,peer)} = \sum_{i \in \{N\} - \{leader\}} \frac{W_b(j)}{C_{leader,i}} \tag{5}$$

$$TC_{j(peer)} = \sum_{h \in \{N\} - \{leader\}} \sum_{i \in \{N\} - \{leader,h\}} \frac{W_{bh}(j)}{C_{h,i}} \tag{6}$$

$$TC_{j(miner_i,k)} = \frac{W_{ack}(j)}{C_{i,k}} \tag{7}$$

Similarly, the amount of data transferred ($W_{total}(j)$) to update/query the health record of a patient j can be calculated using Equation (8).

$$W_{total}(j) = 2 * W_t(j) + W'_b(j) + W'_{bh}(j) + W_{ack}(j) \tag{8}$$

where, $W_t(j)$ represents the amount of data transferred while broadcasting the transaction from the participant to the mining node. It is similar to the amount of data transferred between the validating and the mining node. Consequently, the term $W_t(j)$ is multiplied by 2 in Equation (8). It can be calculated using the size of the transaction. The term $W_{ack}(j)$ denotes the amount of data transferred between the mining node and the participant for acknowledgment and it is same as the size of the acknowledgment signal. $W'_b(j)$ and $W'_{bh}(j)$ in Equation (8) are calculated using Equations (9) and (10), respectively.

$$W'_b(j) = \sum_{i \in \{N\} - \{leader\}} W_b(j) \tag{9}$$

$$W'_{bh}(j) = \sum_{h \in \{N\} - \{leader\}} \sum_{i \in \{N\} - \{leader, h\}} W_{bh}(j) \tag{10}$$

where $W'_b(j)$ indicates the amount of data transferred while broadcasting the block having workload $W_b(j)$ from the leader node to the peer nodes and $W'_{bh}(j)$ represents the amount of data transferred between the peer nodes for consensus.

The blockchain-based healthcare system model enables the sharing of health records to provide more accurate and timely patient care. However, protecting the privacy, confidentiality and security of the records is crucial to effective data sharing [55]. Privacy refers to the rights of a patient to control their own data. Confidentiality refers to the obligations of allied health professionals and administrators who use patient's data to maintain the privacy of patient identity. According to the title 42 of the Code of the Federal Regulations part 2 in the USA, the healthcare providers are required to obtain the patient's written consent in order to share the health records with other medical organizations, even for treatment [56]. The Data Protection Act 1998 [57] and the Human Rights Act 1998 [58] provide a framework that governs a confidential usage and sharing of patients' health records. These privacy and confidentiality laws can be reinforced using smart contracts and access control mechanisms. Furthermore, according to the HIPAA security rule [59], the integrity of health records should be ensured by employing proper encryption and authentication methods. In blockchain, security is established by signing every health transaction digitally using encryption mechanisms [60] and hash-chaining of transactions to reinforce data integrity.

In order to maintain data privacy and security, various blockchain participants have different role-based access rights to patients' health records. A primary care health provider should a have full access to all patient's health history, including patient's identification such as name, contact details, photographic image, biometric details and medical record number; biological data such as height, weight and waist circumference; medical data such as body temperature, blood pressure, sugar level, diagnosis, treatment and allergies; laboratory and pathological results such as x-rays, magnetic resonance imaging, electrocardiography and computed tomography scans; and social data such as smoking habits, sleeping patterns, physical activity and diet plans. This provides the primary care with a cohesive view of the patient's health records for a personalized care and artificial intelligence-based prognosis/diagnosis. For biomedical research purposes, another level of accessibility to patients' data is defined. Biomedical researchers have access to anonymous data. Anonymity is reinforced via consent rules executed by the blockchain smart contracts.

5. Performance Evaluation

In this section, we analyze in which conditions the blockchain platform outperforms the client/server model by using two application scenarios. We evaluate and compare the execution time and amount of data transferred of these models for health records update and query with increasing number of health records and hospitals in the network.

5.1. Methods

5.1.1. Application Scenarios

We perform two experimental application scenarios to evaluate the performance of client/server and blockchain models: (1) health records update and (2) health records query.

In the first scenario, an allied health professional updates patients' health records. We develop this scenario using the client/server model where the allied health professional updates the hospital local database that acts as the server, as shown in Figure 4. After successful data update, the server sends an acknowledgement to the allied health professional. We then measure the total execution time and the amount of data transferred for this process. We also developed the application using our minimal blockchain platform, in which case the allied health professional sends a health record to the blockchain network (Figure 4). The administrator of the mining node elected as the leader creates a block for that record and sends it to the administrators of the peer nodes for verification. Each administrator will verify the block and send the block's hash to all other administrators in the network. An administrator updates the hospital's copy of the ledger once it receives the same hash value from the majority of the administrators. Once the block containing the health record is added to the chain, an acknowledgement is sent back to the allied health professional notifying successful data update. The execution time and the amount of data transferred for this process are then measured.

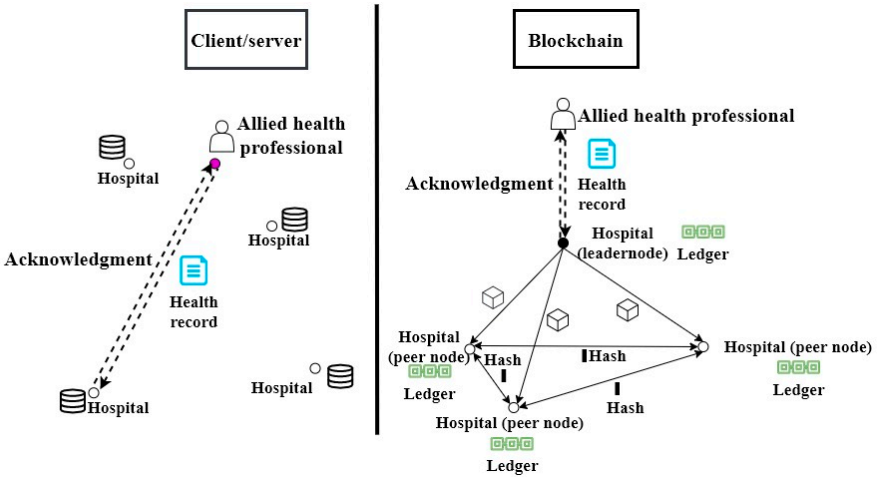


Figure 4. Application scenario for health record update using client/server and blockchain models.

In the second scenario, a patient and/or an allied health professional queries a health record. We develop this scenario using the client/server system model where the patient and/or allied health professional sends a health record query request to the hospital that has that record as shown in Figure 5. In response to the request, the hospital sends back the health records. We measure the total execution time and the amount of data transferred for this process. We also developed the application scenario using the minimal blockchain-based healthcare platform in which the patient and/or allied health professional sends the query request to the blockchain network (Figure 5). The administrator of the mining node elected as the leader creates a block for the query transaction and sends it to the other hospitals' administrators in the network. Each administrator will verify the block and send the block's hash to all other administrators in the network. An administrator updates the hospital's copy of the ledger once it receives the same hash value from the majority of the administrators. Upon ledger update, the health record is forwarded to the patient or allied health professional. The allied health

professional will retrieve the health record from its local copy of the ledger whereas the patient will retrieve it from the nearest hospital having a copy of the ledger.

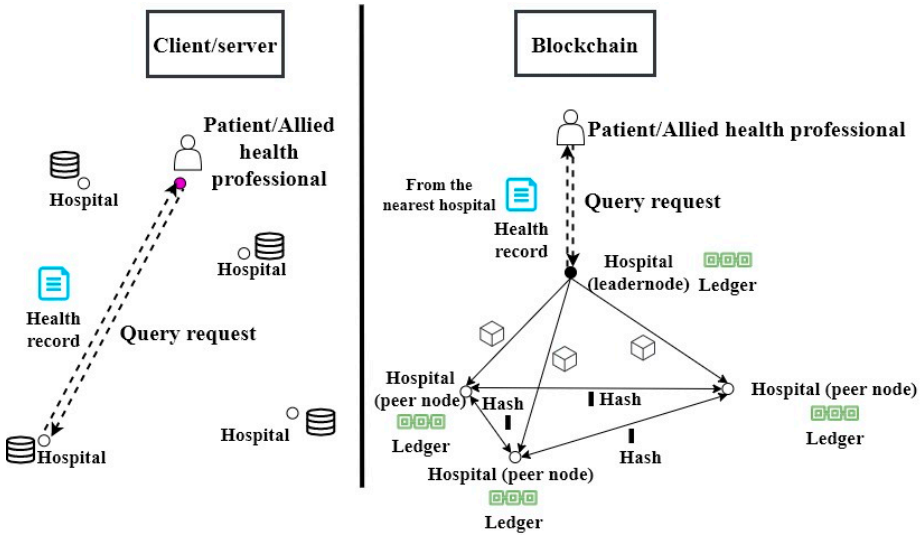


Figure 5. Application scenario for health record query using client/server and blockchain models.

5.1.2. Experimental Environment

To evaluate the client/server model and the blockchain-based healthcare platform, we implemented both network models. In the blockchain system model, we consider a block consisting of a single health record. This is to avoid the delay in the processing of health records in a situation when several records are grouped into a block that is limited by the block size. We use a health record of the size of 25.86 MB in the experiments. This is by considering that a record consists of images and texts of standard sizes, such as intraoral photography (1.64 MB), dental panoramic X-ray (0.85 MB), orthodontic cephalogram (1.20 MB) and skin lesion photography (22.17 MB) [61]. For the query application scenario, we used the block size of 1MB that represents a health record query. A block in blockchain consists of header and body. The body includes a health record in our experiments and the header consists of metadata information such as previous block's hash, timestamp, Merkle root hash value, block number and version [6]. In our experiments, we use the standard block header size of 80 Bytes. The hashing algorithm used in the experiments is SHA-256 that generates a unique 256-bit output for a given input [62]. The selection of SHA-256 is due to its popularity among the blockchain implementations. To evaluate the impact of a dynamic healthcare data management system, we perform all the experiments for the client/server and blockchain system models with increasing number of health records (4000, 5000, 6000, 7000, 8000 and 9000) and increasing number of hospitals (10, 20, 30, 40, 50 and 100). We increase the number of records while keeping the number of hospitals constant at 10, and we increase the number of hospitals while keeping the number of records constant at 4000. The selection of the minimum number of records, i.e., 4000 represents the average patients' records, visiting 10 hospitals (minimum number of hospitals in the experiments) per day, based on the Center for Health Statistics report [63]. We use Network Simulator NS3 [64] to develop the experiments.

5.2. Results Analysis

Figure 6 shows the execution time for updating health records using client/server and the developed minimal blockchain-based healthcare models with increasing number of health records.

It shows that the execution time for client/server and blockchain models increase linearly with increasing health records. However, the execution time for the client/server model is less than that of blockchain. This is because of the consensus mechanism used in the blockchain for data validation and replication. The health record that has to be updated to the ledger is transmitted to all the hospitals having the copy of the ledger for validation. In addition, each peer node will transmit the block's hash to the other peers in the network for consensus before appending it to the ledger. On the other hand, in the client/server approach, the data update request is transmitted to the hospital where the patient data exists. Consequently, the execution time of the client/server is less compared to blockchain approach for health records update. On average, the client/server approach takes 8.5 times less time for updating health records compared to the blockchain platform. Figure 7 shows the performance of client/server and blockchain models in terms of the amount of data transferred to update health records versus increasing number of records. It shows that the amount of data transferred in the blockchain platform is more compared to the client/server network. This is because each health record update request is broadcasted to all the peer nodes in the network by the leader node generating more data transfer. In addition, each peer nodes broadcasts the block's hash to all other peers in the network increasing the data transfer. On average, blockchain model transfers 10 times more data compared to the client/server approach.

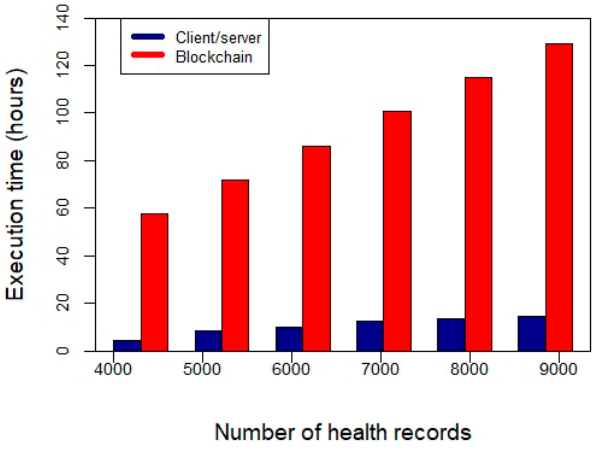


Figure 6. Execution time for health records update using client/server and blockchain with increasing number of health records.

Figure 8 shows the execution time of client/server and the developed minimal blockchain-based healthcare models for querying health records from the databases with an increasing number of health records. It shows the execution time of blockchain is significantly less than the client/server approach. This is because in client/server, the data are retrieved from the server where the record exists, while in blockchain the data are retrieved from the local copy of the ledger. The execution time in the blockchain is only due to transmission of data query request to all the nodes in the network and to add the request as a transaction in a block upon consensus. On average, blockchain is 11.7 times faster compared to the client/server approach for querying health records. Figure 9 shows the amount of data transferred by client/server and blockchain models for querying health records from the databases with an increasing number of health records. It shows the amount of data transferred by the blockchain platform is more compared to the client/server approach. This is because of the PBFT consensus protocol used by blockchain. On average, blockchain transfers 1.1 times more data compared to the client/server

approach for querying health records. However, the amount of data transferred by blockchain for ledger query (Figure 9) is less compared to the one for ledger update (Figure 7). This is because, for ledger update, a block includes a single health record to be updated, whereas, for ledger query, it includes a query request which is comparatively small in size.

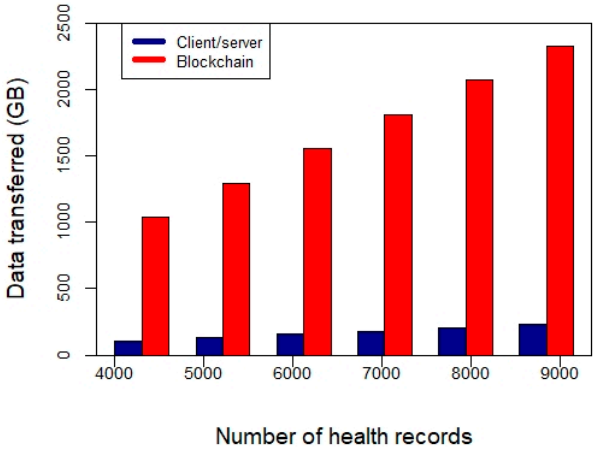


Figure 7. Amount of data transferred for health records update using client/server and blockchain with increasing number of health records.

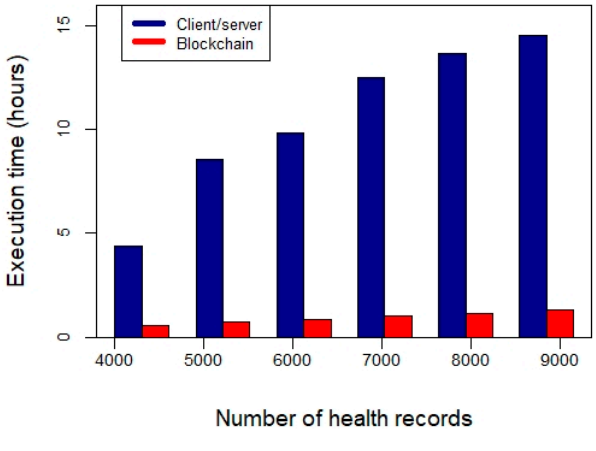


Figure 8. Execution time for health records query using client/server and blockchain with increasing number of health records.

Figure 10 shows the execution time of client/server and blockchain models for updating health records with an increasing number of hospitals. The relative performance is similar to that with an increasing number of health records (Figure 6). It shows that the blockchain model has more execution time than the client/server. On average, the client/server approach takes 13 times less time for updating health records compared to blockchain. Figure 11 shows the amount of data transferred

by client/server and blockchain approaches for updating health records with an increasing number of hospitals. It shows the execution time of blockchain is more than the client/server approach due to the consensus protocol used by the former. On average, the amount of data transferred by blockchain increases 10 times compared to the client/server approach with every 10 hospitals increased.

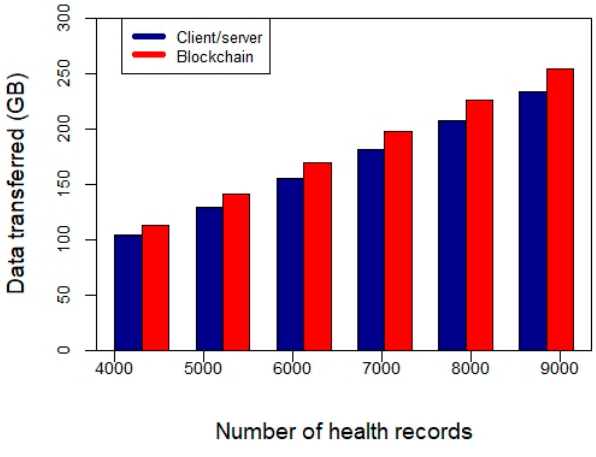


Figure 9. Amount of data transferred for health records query using client/server and blockchain with increasing number of health records.

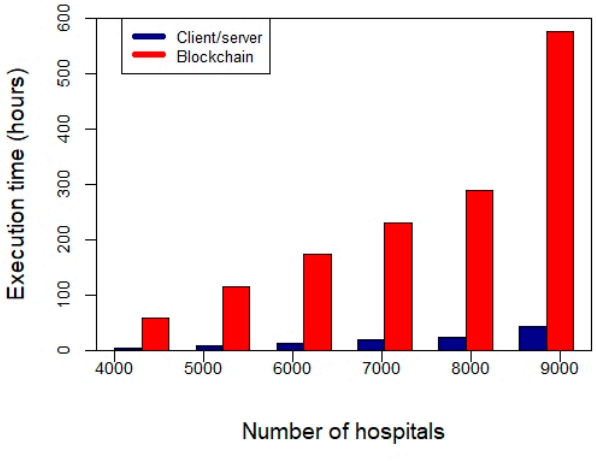


Figure 10. Execution time for health records update using client/server and blockchain with increasing number of hospitals.

Figure 12 shows the execution time for querying health records from the databases with an increasing number of hospitals. The relative performance is same when we increase the number of health records (Figure 8). It shows the execution time of blockchain is significantly less compared to that of the client/server approach. On average, blockchain is 8 times faster compared to the client/server approach for health records query. Figure 13 shows the amount of data transferred by

client/server and blockchain models for querying health records from the databases with an increasing number of hospitals. It shows that the amount of data transferred by the client/server model is constant. This is because the number of health records queried is constant irrespective of the number of hospitals. The query will be performed to the hospital having the required record. However, the amount of data transferred by the blockchain platform for querying health records increases with the number of hospitals. This is because of the exchange of messages between the hospitals due to PBFT consensus. On average, the amount of data transferred by blockchain is 1.1 times more compared to the client/server model for 10 hospitals.

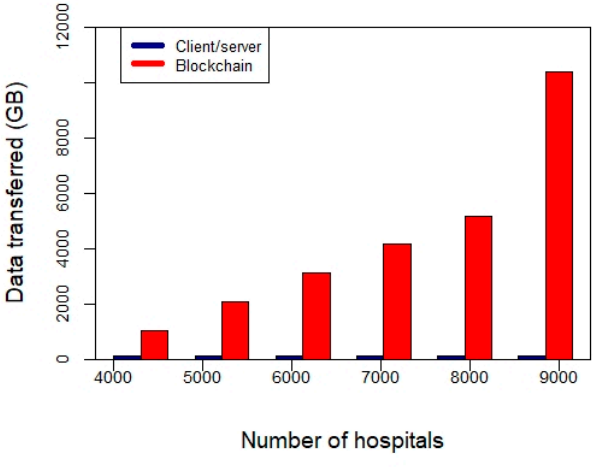


Figure 11. Amount of data transferred for health records update using client/server and blockchain with increasing number of hospitals.

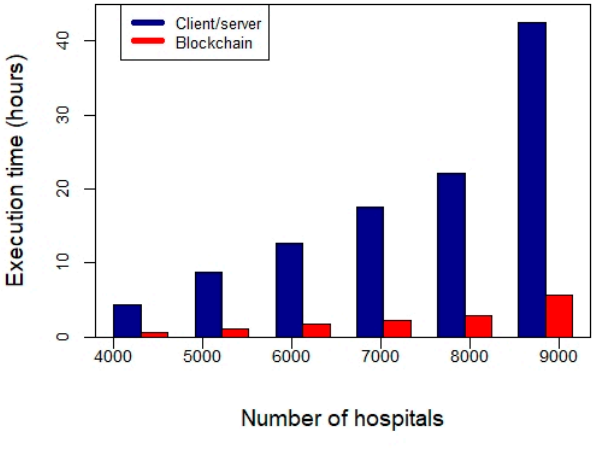


Figure 12. Execution time for health records query using client/server and blockchain with increasing number of hospitals.

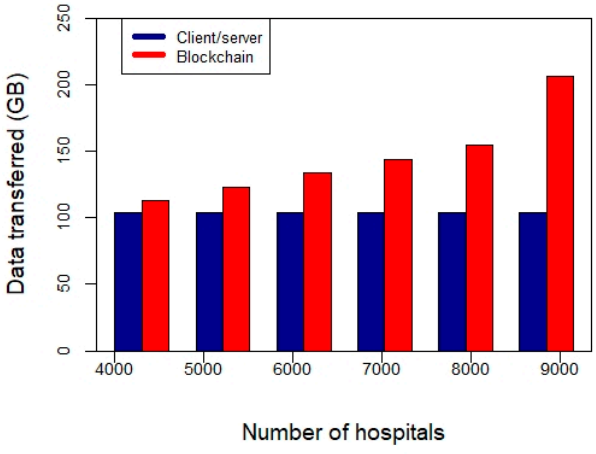


Figure 13. Amount of data transferred for health records query using client/server and blockchain with increasing number of hospitals.

Table 2 shows the performance of client/server and blockchain approaches in terms of execution time for ledger update and query with increasing number of health records and hospitals. It shows that blockchain outperforms client/server for ledger query with both increasing number of health records and hospitals. Table 3 shows the amount of data transferred of client/server and blockchain approaches for ledger update and query with increasing number of health records and hospitals. The client/server approach outperforms blockchain in all application scenarios.

Table 2. Execution time for client/server and blockchain with increasing health records and hospitals.

		Average Increase in Execution Time (Hours)			
		Health Records Update		**Health Records Query**	
Increasing variable	Variable increasing factor	Client/server	Blockchain	Client/server	Blockchain
Number of health records	+1000	2.03	14.36	2.03	0.14
Number of hospitals	+10	4.42	57.44	4.42	0.58

Table 3. Amount of data transferred for client/server and blockchain with increasing health records and hospitals.

		Average Increase in Data Transfer (GB)			
		Health Records Update		**Health Records Query**	
Increasing variable	Variable increasing factor	Client/server	Blockchain	Client/server	Blockchain
Number of health records	+1000	25.86	258.61	25.86	28.18
Number of hospitals	+10	0	1034.98	0	10.34

6. Conclusions

While many researchers investigated the application of blockchain for healthcare data management, however to our knowledge there is no evaluation of this new paradigm with the traditional client/server model. The client/server model suffers from the issue of data stewardship, data fragmentation, vulnerability, security and privacy. Blockchain paradigm has a strong potential to enhance health records management due to its immutability, security, privacy and data replication

features. In this paper, we presented a comparative analysis of the two models for healthcare data management.

To analyze the performance of the models under study for updating and querying health records, we developed a basic healthcare platform based on blockchain paradigm. The results of the experiments show that the blockchain paradigm can lead to significant improvements. This is in particular in a patient-centric approach where the patients and/or the physicians are constantly visiting the health records to construct a cohesive view from different hospitals for a better diagnosis or prognosis using machine learning and artificial intelligence algorithms. Our results show that the health records query for the blockchain platform is 11.7 times faster compared with the client/server model with increasing number of health records. However, the blockchain-based system model is more costly than the client/server system model considering the execution time and the amount of data transferred whenever the transaction involves an update of the health record. This is due to the consensus mechanism involved in the ledger update. One of the future research directions is to implement the developed platform to evaluate its privacy and security.

Author Contributions: Conceptualization, L.I.; methodology, L.I.; investigation, L.I. and H.M.; writing—original draft preparation, L.I. and H.M.; writing—review and editing, L.I.; supervision, L.I.; project administration, L.I. All authors have read and agreed to the published version of the manuscript.

Funding: Thanks to the Emirates Center for Energy and Environment Research of the United Arab Emirates University for supporting this work (Grant G00003304).

Acknowledgments: We would like to thank the anonymous reviewers for their invaluable feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jamoom, E.; Yang, N.; Hing, E. *Adoption of Certified Electronic Health Record Systems and Electronic Information Sharing in Physician Offices: United States, 2013 and 2014*; Technical Report 236, NCHS Data Brief; U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics: Hyattsville, MD, USA, 2016.
2. The HIPAA Privacy Rule. Available online: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (accessed on 5 March 2020).
3. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31995L0046 (accessed on 5 March 2020).
4. Nakamoto, S. *Bitcoin: A Peer-to-Peer Electronic Cash System.* Available online: https://bitcoin.org/bitcoin.pdf (accessed on 20 July 2020).
5. Ismail, L.; Hameed, H.; AlShamsi, M.; AlHammadi, M.; AlDhanhani, N. Towards a Blockchain Deployment at UAE University: Performance Evaluation and Blockchain Taxonomy. In Proceedings of the 2019 International Conference on Blockchain Technology, Honolulu, HI, USA, 15–18 March 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 30–38.
6. Ismail, L.; Materwala, H. A Review of Blockchain Architecture and Consensus Protocols: Use Cases, Challenges, and Solutions. *Symmetry* **2019**, *11*, 1198. [CrossRef]
7. Hölbl, M.; Kompara, M.; Kamišalić, A.; Nemec Zlatolas, L. A systematic review of the use of blockchain in healthcare. *Symmetry* **2018**, *10*, 470. [CrossRef]
8. Dagher, G.G.; Mohler, J.; Milojkovic, M.; Marella, P.B. Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. *Sustain. Cities Soc.* **2018**, *39*, 283–297. [CrossRef]
9. Fan, K.; Wang, S.; Ren, Y.; Li, H.; Yang, Y. Medblock: Efficient and secure medical data sharing via blockchain. *J. Med. Syst.* **2018**, *42*, 136. [CrossRef] [PubMed]
10. Uddin, M.A.; Stranieri, A.; Gondal, I.; Balasubramanian, V. Continuous patient monitoring with a patient centric agent: A block architecture. *IEEE Access* **2018**, *6*, 32700–32726. [CrossRef]

11. Who Owns Medical Records: 50 State Comparison. Available online: http://www.healthinfolaw.org/comparative-analysis/who-owns-medical-records-50-state-comparison (accessed on 5 March 2020).
12. Introduction-HealthVault Development. Available online: https://docs.microsoft.com/en-us/healthvault/introduction/introduction (accessed on 5 March 2020).
13. MTBC PHR: Personal Health Records for Patients. Available online: https://phr.mtbc.com/phrdefault.aspx (accessed on 5 March 2020).
14. OpenClinical e-Health Applications: MyPHR. Available online: http://www.openclinical.org/publicApp_myPHR.html (accessed on 5 March 2020).
15. Capzule PHR: Your Family Health Data in One App. (Personal Medical/Health Records). Available online: https://www.capzule.com/ (accessed on 5 March 2020).
16. My Medical—The Personal Medical Record for You, The Patient. Available online: http://mymedicalapp.com/ (accessed on 5 March 2020).
17. Individual Electronic Healthrecord-GenexEHR. Available online: https://www.genexehr.com/individual-electronic-healthrecord (accessed on 5 March 2020).
18. Saravanan, M.; Shubha, R.; Marks, A.M.; Iyer, V. SMEAD: A secured mobile enabled assisting device for diabetics monitoring. In Proceedings of the 2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Bhubaneswar, India, 17–20 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
19. Liang, X.; Zhao, J.; Shetty, S.; Liu, J.; Li, D. Integrating blockchain for data sharing and collaboration in mobile healthcare applications. In Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, 8–13 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
20. Patel, V. A framework for secure and decentralized sharing of medical imaging data via blockchain consensus. *Health Inf. J.* **2019**, *25*, 1398–1411. [CrossRef]
21. Juneja, A.; Marefat, M. Leveraging blockchain for retraining deep learning architecture in patient-specific arrhythmia classification. In Proceedings of the 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, USA, 4–7 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 393–397.
22. Griggs, K.N.; Ossipova, O.; Kohlios, C.P.; Baccarini, A.N.; Howson, E.A.; Hayajneh, T. Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. *J. Med. Syst.* **2018**, *42*, 130. [CrossRef]
23. Kleinaki, A.S.; Mytis-Gkometh, P.; Drosatos, G.; Efraimidis, P.S.; Kaldoudi, E. A blockchain-based notarization service for biomedical knowledge retrieval. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 288–297. [CrossRef]
24. Mytis-Gkometh, P.; Drosatos, G.; Efraimidis, P.; Kaldoudi, E. Notarization of knowledge retrieval from biomedical repositories using blockchain technology. In *Precision Medicine Powered by pHealth and Connected Health*; Springer: Singapore, 2018; pp. 69–73.
25. Wu, H.; Shang, Y.; Wang, L.; Shi, L.; Jiang, K.; Dong, J. A Patient-Centric Interoperable Framework for Health Information Exchange via Blockchain. In Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications, Xi'an, China, 9–11 December 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 76–80.
26. Wang, H.; Song, Y. Secure cloud-based EHR system using attribute-based cryptosystem and blockchain. *J. Med. Syst.* **2018**, *42*, 152. [CrossRef]
27. Zhang, X.; Poslad, S. Blockchain support for flexible queries with granular access control to electronic medical records (EMR). In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
28. Badr, S.; Gomaa, I.; Abd-Elrahman, E. Multi-tier blockchain framework for IoT-EHRs systems. *Procedia Comput. Sci.* **2018**, *141*, 159–166. [CrossRef]
29. Guo, R.; Shi, H.; Zhao, Q.; Zheng, D. Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems. *IEEE Access* **2018**, *6*, 11676–11686. [CrossRef]
30. Zhang, J.; Xue, N.; Huang, X. A secure system for pervasive social network-based healthcare. *IEEE Access* **2016**, *4*, 9239–9250. [CrossRef]

31. Brogan, J.; Baskaran, I.; Ramachandran, N. Authenticating health activity data using distributed ledger technologies. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 257–266. [CrossRef]
32. Hussein, A.F.; ArunKumar, N.; Ramirez-Gonzalez, G.; Abdulhay, E.; Tavares, J.M.R.; de Albuquerque, V.H.C. A medical records managing and securing blockchain based system supported by a genetic algorithm and discrete wavelet transform. *Cogn. Syst. Res.* **2018**, *52*, 1–11. [CrossRef]
33. Chen, L.; Lee, W.K.; Chang, C.C.; Choo, K.K.R.; Zhang, N. Blockchain based searchable encryption for electronic health record sharing. *Future Gener. Comput. Syst.* **2019**, *95*, 420–429. [CrossRef]
34. Zghaibeh, M.; Farooq, U.; Hasan, N.U.; Baig, I. SHealth: A Blockchain-Based Health System With Smart Contracts Capabilities. *IEEE Access* **2020**, *8*, 70030–70043. [CrossRef]
35. Azaria, A.; Ekblaw, A.; Vieira, T.; Lippman, A. Medrec: Using blockchain for medical data access and permission management. In Proceedings of the 2016 2nd International Conference on Open and Big Data (OBD), Vienna, Austria, 22–24 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 25–30.
36. Li, H.; Zhu, L.; Shen, M.; Gao, F.; Tao, X.; Liu, S. Blockchain-based data preservation system for medical data. *J. Med. Syst.* **2018**, *42*, 141. [CrossRef] [PubMed]
37. Dey, T.; Jaiswal, S.; Sunderkrishnan, S.; Katre, N. HealthSense: A medical use case of Internet of Things and blockchain. In Proceedings of the 2017 International conference on intelligent sustainable systems (ICISS), Palladam, India, 7–8 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 486–491.
38. Yue, X.; Wang, H.; Jin, D.; Li, M.; Jiang, W. Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control. *J. Med. Syst.* **2016**, *40*, 218. [CrossRef]
39. Wang, S.; Wang, J.; Wang, X.; Qiu, T.; Yuan, Y.; Ouyang, L.; Guo, Y.; Wang, F.Y. Blockchain-powered parallel healthcare systems based on the ACP approach. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 942–950. [CrossRef]
40. Roehrs, A.; da Costa, C.A.; da Rosa Righi, R. OmniPHR: A distributed architecture model to integrate personal health records. *J. Biomed. Inf.* **2017**, *71*, 70–81. [CrossRef]
41. Kaur, H.; Alam, M.A.; Jameel, R.; Mourya, A.K.; Chang, V. A proposed solution and future direction for blockchain-based heterogeneous medicare data in cloud environment. *J. Med. Syst.* **2018**, *42*, 156. [CrossRef] [PubMed]
42. Aswin, A.; Basil, K.; Viswan, V.P.; Reji, B.; Kuriakose, B. Design of AYUSH: A Blockchain-Based Health Record Management System. In *Inventive Communication and Computational Technologies*; Springer: Singapore, 2020; pp. 665–672.
43. Tanwar, S.; Parekh, K.; Evans, R. Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *J. Inf. Secur. Appl.* **2020**, *50*, 102407. [CrossRef]
44. Bitcoin Mining Consumes More Electricity a Year Than Ireland. Available online: https://www.theguardian.com/technology/2017/nov/27/bitcoin-mining-consumes-electricity-ireland (accessed on 3 May 2020).
45. Bitcoin Energy Consumption Index. Available online: https://digiconomist.net/bitcoin-energy-consumption (accessed on 3 May 2020).
46. Scherer, M. Performance and Scalability of Blockchain Networks and Smart Contracts. Ph.D. Thesis, Umeå University, Faculty of Science and Technology, Department of Computing Science, Umeå, Sweden, 2017.
47. What is Practical Byzantine Fault Tolerance (pBFT)? Available online: https://crushcrypto.com/what-is-practical-byzantine-fault-tolerance/ (accessed on 3 May 2020).
48. Zheng, Z.; Xie, S.; Dai, H.N.; Chen, X.; Wang, H. Blockchain challenges and opportunities: A survey. *Int. J. Web Grid Serv.* **2018**, *14*, 352–375. [CrossRef]
49. Pseudonymity. Available online: https://en.wikipedia.org/wiki/Pseudonymity (accessed on 3 May 2020).
50. Healthcare Data Breach Statistics. Available online: https://www.hipaajournal.com/healthcare-data-breach-statistics/ (accessed on 3 May 2020).
51. Health Care Systems and Medical Devices at Risk for Increased Cyber Intrusions for Financial Gain. Available online: http://www.calhospital.org/sites/main/files/file-attachments/dp___attachment_fbi_alert.pdf (accessed on 3 May 2020).
52. Health Sector Cybersecurity Coordination Center (HC3). A Cost Analysis of Healthcare Sector Data Breaches. Available online: https://content.govdelivery.com/attachments/USDHSCIKR/2019/04/16/file_attachments/1193648/HC3%20-%20HPH%20Breach%20Cost%20whitepaper.pdf (accessed on 28 May 2020).

53. The Breach of Anthem Health—The Largest Healthcare Breach in History. Available online: https://resources.infosecinstitute.com/category/healthcare-information-security/healthcare-attack-statistics-and-case-studies/case-study-health-insurer-anthem/#gref (accessed on 24 June 2020).
54. Castro, M.; Liskov, B. Practical Byzantine fault tolerance. In Proceedings of the 3rd Symposium on Operating System Design and Implementation (OSDI), New Orleans, LA, USA, 22–25 February 1999; Unisex Association: Berkeley, CA, USA, 1999; pp. 173–186.
55. Hodge, J.G.; Kaufman, T.; Jaques, C. Legal Issues Concerning Identifiable Health Data Sharing Between State/Local Public Health Authorities and Tribal Epidemiology Centers in Selected US Jurisdiction. 2011. Available online: https://cdn.ymaws.com/www.cste.org/resource/resmgr/PDFs/LegalIssuesTribalJuris.pdf (accessed on 21 June 2020).
56. Health Information Privacy Law and Policy | HealthIT.gov. Available online: https://www.healthit.gov/topic/health-information-privacy-law-and-policy (accessed on 21 June 2020).
57. Data Protection Act 1998. Available online: http://www.legislation.gov.uk/ukpga/1998/29/contents (accessed on 21 June 2020).
58. The Human Rights Act 1998 | Department of Health. Available online: https://www.health-ni.gov.uk/articles/human-rights-act-1998 (accessed on 21 June 2020).
59. The HIPAA Security Rule. Available online: https://www.hhs.gov/hipaa/for-professionals/security/index.html (accessed on 21 June 2020).
60. Merkle, R.C. A digital signature based on a conventional encryption function. In *Conference on the Theory and Application of Cryptographic Techniques*; Springer: Berlin/Heidelberg, Germany, 1987; pp. 369–378.
61. Ruiz, M.G.; Chaves, A.G.; Ibañez, C.R.; Mazo, J.M.G.; Giraldo, J.C.R.; Echavarria, A.P.; Diaz, E.V.; Restrepo, G.P.; Munera, E.N.M.; Loaiza, B.G.; et al. mantisGRID: A grid platform for DICOM medical images management in Colombia and Latin America. *J. Digit. Imaging* **2011**, *24*, 271–283. [CrossRef] [PubMed]
62. SHA-256 Cryptographic Hash Algorithm. Available online: https://www.movable-type.co.uk/scripts/sha256.html (accessed on 5 March 2020).
63. National Hospital Ambulatory Medical Care Survey: 2017 Emergency Department Summary Tables. Available online: https://www.cdc.gov/nchs/data/nhamcs/web_tables/2017_ed_web_tables-508.pdf (accessed on 5 March 2020).
64. ns-3 | A Discrete-Event Network Simulator for Internet Systems. Available online: https://www.nsnam.org/ (accessed on 5 March 2020).

symmetry

MDPI

Article

A Photo Post Recommendation System Based on Topic Model for Improving Facebook Fan Page Engagement

Chia-Hung Liao [1], Li-Xian Chen [2,*], Jhih-Cheng Yang [1] and Shyan-Ming Yuan [1,*]

[1] Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan; aiallen.cs07g@nctu.edu.tw (C.-H.L.); qqoo12342001.pcs05g@nctu.edu.tw (J.-C.Y.)

[2] School of Technology, Fuzhou University of International Studies and Trade, Fuzhou 350202, China

* Correspondence: lixian.cs98g@g2.nctu.edu.tw (L.-X.C.); smyuan@cs.nctu.edu.tw (S.-M.Y.)

Received: 9 June 2020; Accepted: 30 June 2020; Published: 2 July 2020

Abstract: Digital advertising on social media officially surpassed traditional advertising and became the largest marketing media in many countries. However, how to maximize the value of the overall marketing budget is one of the most concerning issues of all enterprises. The content of the Facebook photo post needs to be analyzed effectively so that the social media companies and managers can concentrate on handling their fan pages. This research aimed to use text mining techniques to find the audience accurately. Therefore, we built a topic model recommendation system (TMRS) to analyze Facebook posts by sorting the target posts according to the recommended scores. The TMRS includes six stages, such as data preprocessing, Chinese word segmentation, word refinement, TF-IDF word vector conversion, creating model via Latent Semantic Indexing (LSI), or Latent Dirichlet Allocation (LDA), and calculating the recommendation score. In addition to automatically selecting posts to create advertisements, this model is more effective in using marketing budgets and getting more engagements. Based on the recommendation results, it is verified that the TMRS can increase the engagement rate compared to the traditional engagement rate recommended method (ERRM). Ultimately, advertisers can have the chance to create ads for the post with potentially high engagements under a limited budget.

Keywords: Facebook advertising post; social media marketing; text mining; recommendation system; topic model; post engagement

1. Introduction

Web activity data, as in e-commerce, e-learning, e-government, social networks, and so on, represent diverse information that can provide useful data for particular users. Several studies have proposed a variety of recommendation systems to solve the problem of information retrieval and filtering. General used recommendation methods are content-based (CB), knowledge-based (KB) and collaborative filtering (CF) techniques [1]. However, CB and KB require a lot of domain knowledge and have limited expanded ability problems. CF has data sparseness, synonymous and shilling attacked problems. Many improved methods are proposed to solve these problems, such as social relationship-based recommendation systems [2–5] and context-awareness-based recommendation systems [1,6,7]. Recommendation systems have been widely regarded as an effective mechanism that contributes to social media companies' (i.e., Facebook, Instagram, LinkedIn, and Twitter) digital advertising aims and strategy.

Precise digital advertising brings greater business benefits to enterprises and customers. In 2017, Taiwan Media White Paper pointed out two important interpretations. First, digital advertising has accelerated growth and traditional media encounters are suffering the decline. Second, the growth and decline have changed faster in Taiwanese tradition and digital media [8]. Digital advertising volume surpassed magazines in 2009 and the newspaper in 2012. In 2016, Taiwan's overall advertising volume

reached 60.46 billion, of which digital ads were NT$25.87 billion, surpassing NT$22.53 billion of TV (including $19.16 for cable TV and $3.37 for wireless TV) ads for the first time and then digital media became the largest media. Therefore, how to effectively use the largest media in the advertising market is our goal.

Most social media managers have a heavy workload. In addition to spending a lot of time writing postscripts, adjusting photos, and even making videos, the fan page managers have to squeeze time to manipulate the ads. For example, Taiwan Apple's Daily fan page team needs to process more than 120 posts in a day. It is a difficult and time-consuming task for the fan page manager to pick out high-quality posts to create ads. Furthermore, managers painstakingly operating the fan pages have not received a relative return. The organic reach rate of Facebook posts all over the world declines year by year due to constant changes in the news feed algorithm on Facebook. According to Buzzsum's statistics of 880 million posts, the analysis of the engagement rate dropped by 20% from 2016 to 2017 [9]. This is viewed as Facebook's alternative claim for advertisers to improve the quality of their material or to spend more money on advertising to maintain the discussion of the fan pages.

This research aims to help advertisers or social media managers to concentrate on the content of their fan pages. Therefore, the advertising part is handed over to our topic model recommendation system (TMRS). We use text mining technology to automate the selection of posts with a high engagement rate. Thus, this system can help advertisers to get the most benefit within the same advertising budget. In response to the above issues, (a) we choose posted photo posts to be the training data. (b) Then, input the texts of the target post into the trained topic model, (c) find similar ad posts in the training set, (d) sort these similar ad posts in the order of cosine similarity, and (e) take the appropriate number of ad post samples. (f) Then, use the advertising insight data of these similar ad posts, such as positive feedback filed, to make the weight for the recommendation score. The positive feedback field has three levels, which have been verified to be highly correlated with the cost per post engagement (CPE). i.e., each target post can use the topic model to find their own similar ad posts, and then combine the similarities with positive feedback levels to calculate the recommendation scores to make recommendations for these target posts.

The TMRS includes six stages. First, we preprocess the data. Second, the Chinese word segmentation. Third, we do the word refinement, which means the words that would not be the topic of the post will be removed after the word segmentation. Fourth, the words are converted into TF-IDF vectors. Fifth, we use Latent Semantic Indexing (LSI) or Latent Dirichlet Allocation (LDA) to create a model to identify potential topics or features of the ad post texts. Finally, after feeding the target post texts into the trained topic model, the similarity calculation is performed and the similar post texts are output. Use the positive feedback levels and the similarities of the similar ad posts to calculate the recommendation score for the target post.

The rest of this paper is organized as follows. Section 2 talks about the background knowledge and the related work of recommendation systems, recommendation techniques, and topic modeling. Section 3 presents how we analyze important advertising insight data. Section 4 introduces the procedure of building the model structure. Section 5 describes the experiment scenarios and dataset. Section 6 shows the way we decide the model hyperparameters such as the number of topics and the number of sampling. Section 7 illustrates the idea of how we evaluate the effectiveness of TMRS. Section 8 discusses the summary of the results. Finally, Section 9 gives the conclusion of this study.

2. Related Work

2.1. Recommendation Systems

Recommendation systems (RSs) attempt to recommend the most effective items (advertisements, products, or services) to particular users (individuals, social media managers, or advertising companies). Those use some relevant item information and the interaction between users and items to predict a user's interest [6,7]. These systems are really critical in specific industries as they can generate a large

amount of profit when they are efficient or also be a way to transcend significantly from competitors. RSs methods have been developed by academic researchers and applied in a variety of different social media applications, including marketing, movie box-office, information dissemination, elections, macroeconomic, and many others [10].

Our research mainly focuses on social media marketing, especially on Facebook advertising. The methodologies may include text mining, topic model, document similarity, and recommendation system. Therefore, existing recommendation systems relevant to this study can be roughly categorized into two different groups: content awareness-based and social relationship-based recommendation systems.

2.1.1. Context-Awareness Recommendation Systems

Traditional content-based recommendation systems use a lot of information generated by a large number of user activities to analyze group preferences. Content-based recommendation system refers to the description of the product and the configuration file that matches the active users' interests to suggest products that are similar to those that the active user used to like [11,12]. News content was analyzed by using Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) method to make personalized recommendations [13]. They explored the internal relationships between news articles, and the different characteristics of news items. Effective clustering of newly published news articles, as well as high-level recommendations. Moreover, a new method for Facebook fan page ranking that the ranks of pages estimated by this new method are close to the ranks estimated by an engagement-based method [14]. The traditional ranking methods rely on user engagement including the number of posts, comments, and "likes". However, the polarity of each comment is ignored in these methods, which can be positive, neutral, or negative. It has developed a content-based ranking method that takes into account users' engagement and comment polarity. In addition, the new page ranking method concerning the comment polarity is more accurate regarding users' opinions.

Social media identify individuals shared a connection with others and view their connected information within a public or semi-public profile system. Facebook, a giant social media in the world, always refined their recommendation systems. Content analysis in the official Facebook pages of 70 global brands was used to explore the companies' marketing and advertising strategy in social media [15]. Different fan page types will post in different ways. Interestingly, it also found that a large number of fans on a fan page cannot clearly measure sales figures or purchase intentions. This gives us a perspective on how many fans may not be the focus, but the quality of the post or the product itself. Moreover, Facebook's popularity and effectiveness are largely related to the content and semantic of the posts. The popularity of the post and the engagement rate were used to dynamically adjust the model parameters [16]. The purpose was to increase the exposure of the fan page and apply it to the political fan page. There is also a multifactor model that shows how time, the number of people and their genders, and how the media type contributes to the popularity and effectiveness of the post. It is also found that a fan page with more fans does not necessarily lead to more popular posts or higher engagements. For analyzing the text and photo posts, we need to consider the semantic and context of the posts.

2.1.2. Social Relationship-Based Recommendation Systems

The users' trust relationship, direct trust, and indirect trust are established according to whether or not users directly give trust value [3]. People can receive friends' recommendations through social media such as Facebook or Twitter. The trust relationship between users in a social network can be inferred based on past user interaction records or users' specified items. Users can also use indirect data like rating information to compute and infer their trust relationships. Text mining was an effective method to explore business value from a large amount of available data. The value of social media competitive analysis was demonstrated by analyzing the text content on Facebook and Twitter of the three largest pizza chains [17]. Its recommendation system provides help to companies to develop their

social media strategy. It is also found that references to other competitors' posts with high engagements published in their own social media with the same concept had good engagement with customers.

Many social relationship-based recommendation techniques were analyzed according to implicit trust-based information like user trust relation, interaction, product popularity, and user credibility [3,18]. Users' implicit trust relationship and corresponding degrees can be inferred by their common items with coratings and social networks' role importance. Some studies using machine learning techniques, Connectionist Inductive Learning, to generate recommendations in Web communities or supporting Web navigation [19]. Social relationship-based recommendation systems discriminate against the users' commonalities according to their ratings and generate new recommendations considering the comparisons between different users. Our previous work analyzed the separation degree problem from two aspects: (a) between two normal-persons or famous-persons and (b) two individuals with special characteristics [20]. The six degrees of separation theory was re-evaluated and extended by using the real Facebook tool "We T So Close." The result pointed out that the average acquaintance number was 3.9 regardless of two normal individuals or two persons with rare features. This study aims to select posts with a high interaction rate automatically, so the designed recommendation system needs to consider implicit trust-based information.

2.2. Topic Modeling

In order to investigate the representation of documents, generative topic models, such as Latent Semantic Indexing (LSI) [21,22] and Latent Dirichlet Allocation (LDA) [23,24] are widely adopted methods. Using knowledge representation form established by term frequency-inverse document frequency (TF-IDF) or bag-of-words (BOW) can improve LSI and LDA methods' effects. Topic models refer to a statistical model used to discover abstract topics in a series of documents [25,26]. Intuitively, if an article has a central idea, then some specific words will appear more frequently. For example, if an article is about a dog, the words "dog" and "bone" appear more frequently. If an article is about cats, the words "cat" and "fish" appear more frequently. However, the real situation is that an article usually contains multiple topics, and each topic has a different proportion. Therefore, if an article is 10% related to cats and 90% is related to dogs, the number of keywords associated with dogs will be approximately nine times the number of keywords associated with cats. A topic model uses a mathematical framework to implement the document feature and it decides what topics are included in the current document by analyzing each document, the word counts in the document and other statistical document information.

LSI refers to how to find the relationship between words through massive literature [21,22]. When two words or a group of words appear in the same document in numbers, they can be considered semantically related. LSI uses singular value decomposition (SVD) to decompose the word-document matrix. SVD maps the original data into the semantic space by finding irrelevant index variables so that two dissimilar documents in the word-document matrix may resemble in the semantic space. In addition, LDA is a document-generation model that considers each topic in an article corresponding to different words [23,24]. In the process of constructing an article, it will select a topic with a certain probability, then selects a word with some specific probability under this topic, and finally generates the first word for this article. Repeat this process and generate an entire article. LDA assumes that there is no order between words. At the same time, it is an unsupervised learning algorithm and it does not need to manually label the training set. Using LDA only requires the document set and specifying the number of topics. Moreover, LDA can find some words for each topic to describe it. In order to select a topic model suitable for analyzing social media copywriting, this study uses LSI and LDA topic models for analysis.

Using the topic model toolkit, Gensim, to process corpus data created by LSI and LDA topic models can analyze paradigmatic and syntagmatic relations between lemma within topics [27,28]. The topical similarity can be queried between plain text documents and other documents when the semantic topics were found. Gensim is a free python module dedicated to working with raw, unstructured

text that automatically extracts semantic topics from documents [29]. Modules are developed from three concepts: corpus, vector, and model. According to the topic model toolkit Gensim official document, the latent semantic index (LSI) converts documents from word bag or TF-IDF weighted space to lower-dimensional potential space. The 200–500 topic dimension is recommended as the "gold standard." However, this standard is suitable for long articles. We investigate the social media managers' community copies. The findings show the average number of words is about 50–150 words and the number of concepts to be expressed is about five or so. Therefore, in our research, we will not use the number of topics in the general article, but use the number of topics from 1 to 15 to build a topic model experiment on the TMRS. Moreover, the computer cuts the Chinese characters into units of "meaning" that are important [30]. Without special treatment, the computer will treat each Chinese character separately, but this is meaningless for analyzing semantics and potential topics. To process the Chinese word segmentation correctly, Jieba library was needed to import in this study. Jieba is an open-source project. This Chinese word segmentation program is written by a developer of Baidu in China [30]. Its core is actually Simplified Chinese. However, since it is an open-source project, there are already enthusiastic developers on the Internet plus a traditional Chinese dictionary.

3. Advertising Insight Analysis

3.1. System Overview

Facebook posts can be composed of a variety of attributes, such as texts, images, photos, videos, and call to action [31]. The text is an attribute that can be found on every fan page. If the post texts are well written, it will resonate with the user. The TMRS obtains a weighted score by calculating the cosine similarity [32] of current texts for the target post and past texts for the ad posts. The higher the score, the higher the post engagement that the system predicts will be, and the lower the cost per post engagement (CPE). When each target post is fed into the TMRS, it can form a recommendation order according to the weight score, and provide a priority reference for the social media manager to post advertising.

3.2. Advertising Performance Indicator

In order to evaluate our system, we should compare the results of TMRS with the existing system, engagement rate recommended method (ERRM). Its recommended ranking of the post is based on the level of post engagement rate (PER), and the PER is calculated as follows:

$$PER = \frac{post\_engaged\_users}{post\_impressions\_unique} \quad (1)$$

post_engaged_users is the number of post engagements that users interact with the post, after posting the ad. post_impressions_unique is the number of exposures that the post appears on the users' screen. Both post_engaged_users and post_impressions_unique are collected from Facebook Graph API.

According to Facebook's official document, Facebook's advertising insights API provides a variety of advertising insights for developers [31]. Post engagements refer to all actions taken by the user for the advertisement during the delivery of the advertisement; for example: convey a mood, leave a message or share, request for a discount, view photos or videos, or click a link. In the case of a limited marketing budget, the lower the CPE is, and the more user engagements the ad post gets. The public API can collect lots of the average cost of post engagements data. Our goal is how to effectively use the largest media resources in the advertising market so that we are most concerned about the CPE. Therefore, we hope to find the field of Facebook associated with the post engagement most.

This indicator is calculated by dividing the total cost by the number of post engagements, which is shown as Equation (2):

$$CPE(Cost\ per\ Post\ Engagement) = \frac{total\_ad\_spending}{num\_post\_engage} \quad (2)$$

where total_ad_spending is the total amount of ad post spending, and num_post_engage is the number of post engagements.

In addition, Facebook ads will have different benchmarks for calculating post engagement depending on the type of post. For example, the movie has three seconds, 10 s of views. The photo has photo clicks, etc. In order to avoid the difference of the benchmark, we have chosen the most popular type of post, photo post for experiment, and analysis.

3.3. Ad Insights Select

3.3.1. Relevance Score

The role of the relevance score is to allow the advertiser to evaluate how much the ad resonates with the user he or she wants to reach. The higher the relevance score of an ad post, the better the performance of the ad. This score is based on a comparison between the ad posted by social media manager and other ads that lock the same customer. The factors also include positive feedback (ex: clicks, app installs, video views) and negative feedback (ex: someone clicks "I don't want to see this" on your ad). The relevance score is scored on a scale of 1–10.

3.3.2. Observation

We first analyze the relevance scores in the advertising insights to see if the relevance score can be used to judge the quality of the post texts. In addition, the relevance scores have extended fields, which are positive feedback and negative feedback. The feedback level of the advertisement may be low, medium, or high.

First, we want to find out how the relevance score of the post is related to post engagement, and we will calculate the correlation coefficient between them. In addition, Facebook fan page types are divided into 10 categories, and each category is subdivided into different items. In every experiment, we pick out the same category of related fan pages and calculate the correlation coefficient between their CPEs and relevance scores.

3.3.3. Correlation Coefficient

After the analysis, the correlation coefficient of the relevance score is not very high. Then, we take the extended two fields at the same time, the effect after the analysis is not good too. It may be because the negative feedback does not reflect on the CPE. Finally, we pick positive feedback for analysis, as shown in Table 1. The correlation coefficient between positive feedback and cost per post engagement has come to −0.65, which is strongly correlated. Table 2 shows the degree of correlation strength. It means that the higher the positive feedback level is, the lower the CPE will be. This trend can also be seen in Figure 1.

Table 1. The correlation coefficient between ad insights and the cost per post engagement.

Ad Insights Fields	Correlation Coefficient
Relevance Score Field	−0.41
Negative and Positive Feedback Field	−0.32
Positive Feedback Field	−0.65

Table 2. Pearson product-moment correlation coefficient table.

Degree of Relationship	Negative	Positive
No relationship	−0.09 to 0.0	0.0 to 0.09
Weakly correlated	−0.3 to −0.1	0.1 to 0.3
Moderately correlated	−0.5 to −0.3	0.3 to 0.5
Highly correlated	−1.0 to −0.5	0.5 to 1.0

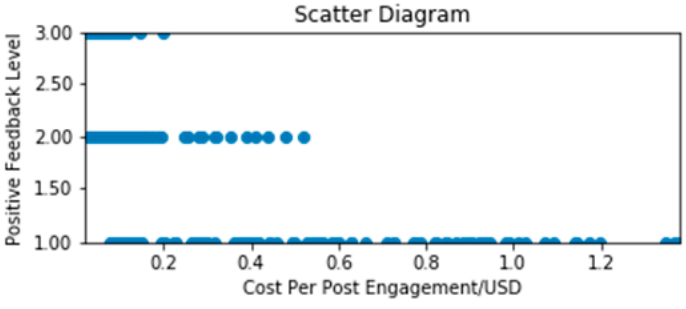


Figure 1. Scatter diagram for positive feedback and cost per post engagement.

Therefore, in our study, we will use a positive feedback level to be the main weighted factor for calculating the recommendation score.

4. System Architecture and Implementation

In order to provide a recommended post list with high engagement potential for social media managers, we design a system for computing a recommendation score by comparing the target post and ad posts. We use Facebook Graph API to get the post data which we need, then input them to model and get the score. Finally, sort the score from high to low. There are six stages for the system: preprocessing, word segmentation, word refinement, TF-IDF vector conversion, creating the LSI/LDA model, and calculating the recommendation score. The system structure is shown in Figure 2.

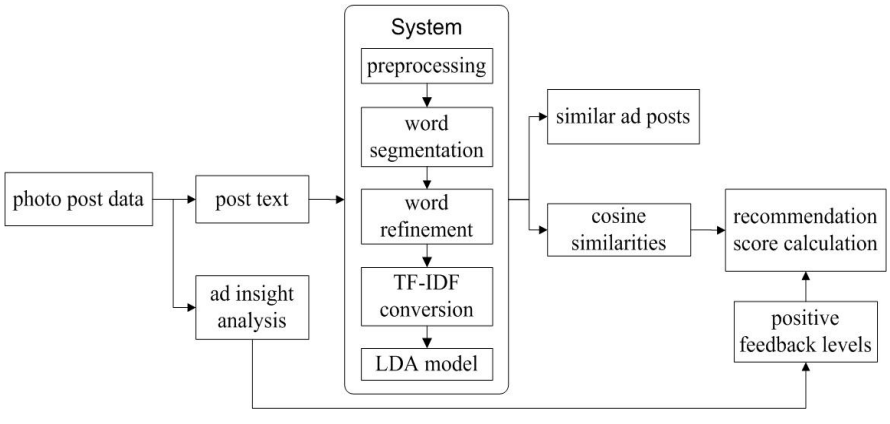


Figure 2. Schematic for system structure.

Procedure

Stage 1: Preprocessing

The actual data will be affected by different factors, so there may exist extreme value. In order to prevent the extreme value from affecting the result accuracy, and avoiding the influences on analyzing posts, we preprocess the data. First, the model removes the extreme value. If there exists some extreme value in the data that has a big difference with others, the credibility of the overall data may be reduced. Therefore, we remove the data in which the CPEs fall outside the two plus and minus standard deviations from the mean. Through this step, we can prevent the overall data from being affected by values that are too large or too small. Then, the model removes special characters. After we got the post from Facebook Graph API, it may contain emoji and special characters, for example, ♡ or line break symbol. It is relatively irrelevant to the quality of the post content. We hope to retain only the story or artistic concept of the post, therefore, we use the program to remove these special characters from the post. Additionally, there are some URLs in the post. These URLs may be an official website or event registration page, but the URL has nothing to do with the quality of the post and will be removed from the post texts here.

Stage 2: Word Segmentation

Chinese word segmentation [33,34] is the most important preprocess in Chinese. If the Chinese word segmentation correctly identifies the words with the smallest unit of meaning, we may have a way to conduct higher-level natural language analysis. This study used Jieba participle (an open-source project) to do Chinese word segmentation. After doing participles, a sequence of words is regrouped into a sequence according to certain specifications. Therefore, the correctness of the Chinese word segmentation has affected the success or failure of many natural language processing applications.

Stage 3: Word Refinement

We remove the words that should not be the topic of the post after the word segmentation. For example, if words such as "it, is, that", are not removed and appear many times in the post, it will be misunderstood for the post topic. Therefore, before training the model, it is necessary to remove such words from the bag of words after the word segmentation. There are three steps to do for word refinement. First, synonym replacement replaces words with the same or similar meaning, such as wine and spirit. In the post texts, it would be better if the words with the same meaning are expressed by the same word, to ensure better performance when calculating the similarity of the posts [35]. Second, removing the brand or product name from the words bag makes this recommended system common to any fan page copywriting. If the brand or product name do not be removed from the post texts, the model will misjudge them to be kinds of topics when doing text analysis. Moreover, the brand and the product name will disturb the similarity and will make the post texts too similar to each other. Finally, removing the hashtag lets irrelevant text be deleted. The hashtag is composed of # with a word or a sentence without spaces. Users can link to the same platform with the same hashtag. The reason for the removal of the hashtag is the same as the brand name.

Stage 4: TF-IDF Conversion

This study uses the Gensim module in the topic model to convert words into vectors and feed them to the TF-IDF model. TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and text mining [36,37]. TF-IDF is a statistical method that is used to evaluate the importance of a word for a document in a group of documents or corpus. The importance of a word is proportional to the number of times it appears in the document, but the word importance also decreases inversely with the frequency it appears in the corpus. After TF-IDF conversion, the meaningful words' weights will be increased.

Stage 5: Create the LSI/LDA Model

After the TF-IDF vector conversion, each word has its own weighted vector. Then, use these weighted vectors and specify the number of topics via Gensim library, the LSI/LDA model is generated separately for the cosine similarity of the subsequent target post.

Cosine similarity is commonly used for file comparison in text mining, and the similarity between them is measured by the cosine of the angle between the two vectors [32]. Cosine similarity is usually used in positive spaces and the value is between 0 and 1. For example, cosine similarity is one when two vectors have the same orientation and the value is 0 when two vectors angle is 90°.

Stage 6: Recommendation Score Calculation

After the LSI and LDA are established, the target post can be fed into these trained topic models to calculate the similarity between the target post and each ad post in the training set. Then, according to the similarity order, output those indices of similar ad posts. The indices here are the index numbers of the ad posts in the training set. Then, return the advertising data of the ad posts and observe its positive feedback levels. Use these levels to calculate the recommendation score of the target post, which is calculated by Equation (3).

$$\hat{R}_i = \frac{\sum \mathrm{sim}\left(a_i, t_j\right) \times p_i}{\sum \mathrm{sim}\left(a_i, t_j\right)} \tag{3}$$

$\hat{R}_i$: The target post recommendation score predicted by the similar ad posts. $\sum \mathrm{sim}\left(a_i, t_j\right)$: Cosine similarity of the target post the ad posts. p_i: Positive feedback level of ad post (high = 3, medium = 2, low = 1).

For example, assuming that the target post texts are fed into the system, the system takes the positive feedback rating of the first 10 most similar posts, like Table 3. We put the "high" level for three points, "medium" for two points, and "low" for one point. The level scores are multiplied by the similarity then added, and finally divided by the total score. This is the final recommendation score.

Example: (0.99123 × 3 + 0.97456 × 2 + 0.96111 × 2 + … … + 0.86666 × 1) / (0.99123 + 0.97456 + … … + 0.86666)

Table 3. Similar ad post index, similarity and its positive feedback level (example).

Ad Post Index	Positive Feedback Level	Cosine Similarity
3	high	0.99123
9	medium	0.97456
10	medium	0.96111
90	high	0.96000
100	low	0.95444
200	low	0.93214
305	medium	0.93001
446	medium	0.88888
555	medium	0.87777
666	low	0.86666

5. Scenarios and Dataset

5.1. Experiment Scenarios

To select a topic model suitable for analyzing social media copywriting, this paper designs three scenarios for experimentation which are shown by Table 4. In the first scenario, we consider the recommendation effectiveness of the post texts of the wine fan page under the LSI and LDA models. The second scenario is to use the ad post screened by marketing experts from the wine fan pages, then re-generate LDA models and check the recommendation effectiveness. The third scenario will be based on the above two experiments, to see which way will win the most. Then, we select the best way to do

experiments on different types of fan pages. In scenario3, we choose makeup/skincare fan pages to test our TMRS.

Table 4. Three kinds of scenarios.

Scenario1	All the advertising post texts of the Wine/Spirits fan pages.
Scenario2	Selected ad post texts of the Wine/Spirits fan pages by three marketing experts.
Scenario3	Apply the better solution from 1 and 2 to the Makeup/Skincare fan pages.

5.2. Marketing Expert Screening

In the experiment of a photo post, which is composed of photos and texts, our experiments focus on analyzing the post texts. Therefore, in order to reduce the variation factor of the photo, we invited three marketing experts to vote for each ad post to see whether the positive feedback of each post was influenced by the texts or the photo or the half. When an ad post is more than 1.5 points after the experts' vote, it will be selected into our data set. The voting score rules are shown in Table 5.

Table 5. The voting score rule for marketing experts.

Score	Voting of Rule
0	The positive feedback level of this ad post is mainly caused by the photo.
0.5	The positive feedback level of this post is mainly caused by the photo and the texts.
1	The positive feedback level of this ad post is mainly caused by the texts.

5.3. Dataset

If one wants to do a text analysis of ad posts, one must use the past post data. This study used Facebook's Graph API to access individuals' information without requiring their passwords. After accessing fan pages' tokens, this API collected the post data including manage_pages and ads_management to do the following analysis. Instead of the common 80:20 rule, we use older data as the training data, and later data as test data. Thus, it can be in line with the actual advertisement created by social media managers. In this case, we can also know whether our TMRS will get more post engagements (PEs). We use the ad post and ad data that implement the promotion of PE ads from Mar. 2015 to Mar. 2018 as training data for scenario1 and scenario2. The training data of scenario3 come from Oct. 2016 to Mar. 2018. Test data were all obtained from Apr. 2018 to Jun. 2018. Table 6 shows the ad post data for different scenarios. Taking scenario1 as an example, in the training data that actually has the post for creating ads, there are 688 posts with delivery data, from 18 wine-related fan pages. Test data has 92 posts for 11 fan pages to create ads. Table 7 shows the important ad data for different scenarios, such as their total advertising spending (AS), post engagements (PE), cost per post engagement (CPE), total post moods, and exposures. Taking the training data from scenario1 as an example, its total AS is 14,253528 New Taiwan dollars (NTD), total PE is 3,039,045 times, and CPE is 4.69 NTD which is calculated by using AS, PE, and Equation (2). It has 2,353,152 post moods and 141,386,675 exposures. Scenario2 and scenario3 are similar, and so on. PE includes all actions taken by the user for the ad during the delivery. PE includes the following actions: conveying moods, messages or sharing, requesting offers, viewing photos or videos, or clicking on a link. Post mood is the amount of mood the ad receives. The mood button of the ad post allows the user to express different moods for the post, such as "like", "big heart", "ha", "wow", "cry" or "angry". Exposures are the number of times an ad appeared on the user's screen. When the ad is first displayed on the user's screen, it is counted as one exposure. (For example, if a user scrolls down after seeing an ad and then scrolls back up to the same ad, it counts as one exposure. If the user sees the same ad two times a day, it counts as two exposures.)

Table 6. Ad post data for different scenarios.

Experiment	Use	Time Interval	Total Fan Pages	Total Posts
Scenario1	Training data	March 2015 to March 2018	18	688
	Test data	April 2018 to June 2018	11	92
Scenario2	Training data	March 2015 to March 2018	18	411
	Test data	April 2018 to June 2018	11	77
Scenario3	Training data	October 2016 to March 2018	20	590
	Test data	April 2018 to June 2018	8	104

Table 7. Ad data for different scenarios.

Experiment	Use	AS (NTD)	PE	CPE (NTD)	Post Moods	Exposures
Scenario1	Training data	14,253,528	3,039,045	4.69	2,353,152	141,386,675
	Test data	688,398	241,564	2.84	155,212	3,225,919
Scenario2	Training data	6,825,854	1,571,306	4.34	1,116,576	60,549,680
	Test data	419,032	164,317	2.55	101,236	2,041,252
Scenario3	Training data	6,926,666	2,365,664	2.93	1,313,777	51,991,448
	Test data	862,344	364,772	2.36	111,879	9,726,080

6. Model Hyperparameter Selection

6.1. Number of Topics

The topic model refers to a set of methods for extracting hidden topics from a document [26]. When training the model, we need to set the number of topics in advance, manually adjust the parameters according to the results of the training, optimize the number of topics, and then optimize the text classification results. The length of the post texts in social advertising is generally not too long, so the experiment will set the number of topics to 1–15 and use the training data to obtain the best number of topics for the TMRS.

6.2. Number of Samples

When calculating the recommendation score, how many most similar post samples are needed to be taken from the training set? Through the experiment, we will test by sampling 1% to 10% of the total number of training data to obtain the most suitable number of samples for the TMRS.

6.3. Settings and Methods

First, after the training data are preprocessed, the ad post texts are sent to the LSI and LDA models respectively. The difference between each model is the number of topics. Then, compare the CPE of each monthly ad post list of each fan page of the training data, and then decide the best number of topics for each scenario. While experimenting with the best number of topics, we also experiment with the optimal number of samples required for TMRS. We took 1% to 10% of the total number of training samples. For example, there are 411 posts in the training data of scenario2. We will take 4, 8, 12, ... , ad posts to be the similar numbers of samples, and use these numbers of samples to calculate the recommendation score.

Then, we segment the training data of scenario2 according to the fan page and the month, to form a total of 46 cases. Note that we have removed the case where there are only one or two ad posts for the month. We use these segmented cases to compare the recommendation effectiveness of ERRM and TMRS. If the CPE from the TMRS is relatively low, the number of topics and the number of samples of the model are recorded, and the LSI-based TMRS number of wins table is constructed as shown in Table 8. Then, use the table lookup method to find the combination of the number of topics and the number of samples, that this best combination means TMRS has the most wins. If the best

combination has more than one, choose a smaller number of topics and the number of samples as the optimal combination to reduce the time to build the model.

Table 8. Wins count table for the LSI-based topic model recommendation system (TMRS) in scenario2.

Sampling Topics	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
1	22	22	22	22	22	22	22	22	22	22
2	20	21	22	20	20	20	20	20	20	20
3	22	22	22	20	20	20	21	20	20	20
4	21	23	21	21	21	21	20	20	20	21
5	23	23	22	22	22	22	22	22	22	21
6	21	22	23	23	23	22	22	22	21	21
7	23	24	22	23	22	22	22	22	21	21
8	24	24	23	24	23	23	21	22	22	21
9	21	22	21	21	21	21	21	21	22	22
10	24	24	22	21	21	21	21	21	21	20
11	21	23	23	23	21	21	22	21	22	21
12	22	24	23	22	21	21	21	21	21	22
13	23	24	22	21	22	21	20	20	19	19
14	23	23	22	21	20	21	21	21	21	21
15	24	22	23	24	22	22	22	22	21	20

Construct the tables for the LSI and LDA of each scenario in the same way, and take the most wins combination of the number of topics and the number of samples. The obtained (sampling number, topic number) of each scenario is shown in Table 9. Taking the LSI of scenario2 as an example, the training data can be segmented into 46 cases, and the maximum number of wins of TMRS is 24. Therefore, the number of topics suitable for the wine fan page is set to seven for the LSI model, and the words that make up a certain topic are, for example, "Activities, flavors, classics, messages, absolutes, fans, first time, double barrels, time, friends ", and the top 2% of samples for the ad posts are used to evaluate the recommendation scores. Others and so on.

Table 9. The results of model hyperparameter (sampling and topics) after experiments.

Experiment	Model	Cases/Max Wins	Sampling	Topics	Representative Words for One of the Topics (Example)
Scenario1	LSI	48/26	7%	2	Activities, classics, events, tastes, fans, messages, flavors, first time, sharing, original intentions
	LDA	48/28	4%	12	Absolute, travel bag, taste, flavor, aroma, oak barrel, limited edition, one bite, departure, greet
Scenario2	LSI	46/24	2%	7	Activities, flavors, classics, messages, absolutes, fans, first time, double barrels, time, friends
	LDA	46/24	2%	15	Cherry blossoms, appearance, flower season, cans, faces, couples, friends, aftertaste, rogue, lobster
Scenario3	LSI	58/34	1%	13	Official website, consumption, limited gift, reward, purchase, discount, full, forgive, exclusive, gift
	LDA	58/34	8%	13	Official website, exclusive, essence, purchase, repair, moisturizing, activities, skin, discount, reservation

The test data uses LSI- or LDA- based TMRS to compare their CPE with traditional ERRM to see how the recommendation effectiveness works. Test data is also segmented according to the fan page and month. The test data for each scenario form 20 cases. If the fan page has three target posts in the month, take the first one for creating an ad, and take the first two to do so if the fan page has six target posts, and calculate the average CPE of these first posts, and so on.

7. Evaluation Method for Recommendation

After building the model by using training data and setting the model hyperparameters, we use the test data and go through the following steps to evaluate our TMRS. Here, we show an example that

the results are shown in Table 10, to illustrate the idea of the evaluation method. Table 10 shows the effectiveness of LSI-based TMRS by using test data in scenario1. In accordance with the usual habits of the social media managers, we do the sorting of the posts in cases with monthly units. (Let $Apr\_P_i^A$ denotes the i^{th} post of fan page A in April. In this case, i ranges from one to five.)

Step 1 Start from the fan page A in April, and this case has five target posts.

Step 2 Sort these five posts by traditional ERRM and TMRS, respectively.

Step 3 Take the first two posts respectively from ERRM and TMRS, and calculate their average cost per post engagement (ACPE). (Take the same fan page month as the case unit, choose the first one if there are three posts, or choose the first two if there are six posts, etc.)

Step 4 Compare which ACPE is lower and decide whether the ERRM wins or TMRS wins. For the fan page A in April, the ERRM-ACPE is 2.52 NTD, and it is lower than the TMRS-ACPE 2.85 NTD. Therefore, the TMRS loses this round.

Step 5 Calculate the CPE gain (CPEG). Here, the CPEG is −13%, which is calculated by Equation (4)

$$\text{CPEG} = \frac{\text{ERRM\_ACPE} - \text{TMRS\_ACPE}}{\text{ERRM\_ACPE}} \tag{4}$$

Step 6 Recursively implement the above steps to the other cases in test data, until all cases belonging to this test data have been done. (According to the rule for sorting the posts in cases with monthly units, each test data for each scenario are divided into 20 cases.)

Table 10. The effectiveness of LSI-based TMRS by using test data in scenario1.

Fan Page/Post Month	Posts/Selected Posts	ERRM-ACPE (NTD)	TMRS-ACPE (NTD)	Win Lose	CPEG
A/Apr	5/2	2.52	2.85	lose	−13%
A/May	4/1	2.97	2.52	win	15%
B/Jun	5/2	1.47	1.29	win	12%
C/May	5/2	3.63	3.18	win	12%
C/Jun	4/1	2.94	2.97	lose	−1%
D/Apr	7/2	1.95	1.92	win	2%
D/May	8/3	2.37	2.46	lose	−4%
D/Jun	6/2	2.25	2.55	lose	−13%
E/May	2/1	4.8	4.8	tie	0%
F/Apr	3/1	5.97	1.26	win	79%
F/May	2/1	2.61	2.70	lose	−3%
F/Jun	3/1	2.28	2.28	tie	0%
G/May	3/1	3.24	3.33	lose	−3%
H/Apr	8/3	2.34	1.98	win	15%
H/May	7/2	2.49	2.19	win	12%
H/Jun	6/2	2.04	2.16	lose	−6%
I/Apr	4/1	2.88	2.88	tie	0%
I/May	3/1	2.91	2.91	tie	0%
J/Jun	5/2	2.13	3.24	lose	−52%
K/May	2/1	33.09	32.37	win	2%

Note that there are two possible situations in the tie: In situation1, assume the first two posts of ERRM and TMRS in the same case are the same. This means both ERRM and TMRS obtain the same best post list to get the same value of ACPE. In situation2, assume the first two posts of ERRM and TMRS in the same case are different. For example, the traditional ERRM obtains $Apr\_P_1^A$, $Apr\_P_2^A$, and TMRS obtains $Apr\_P_3^A$, $Apr\_P_4^A$. This implies that although it is in the tie, our TMRS can still obtain better ad posts for earning more engagements under the same budget. The reason is that the TMRS is prerecommended and does not need to be publicized first, it will be considered to have won the ERRM.

(ERRM needs to publish the post on the community for a while to calculate the engagement rate.) That is, in situation1, TMRS and ERRM are in a true tie. However, when it comes to situation2, TMRS will be recognized to win over the ERRM.

8. Results

After recursively implementing the steps of the evaluation method, there are effectiveness tables that are similar to Table 10 for LSI- and LDA-based TMRS by using test data in the three scenarios. Then, we count the numbers of win, lose and tie, and calculate the win rate, lose rate and tie rate, which are defined by the following equations:

$$Win\ rate = \frac{number\ of\ win}{total\ number\ of\ win\ and\ lose} \tag{5}$$

$$Lose\ rate = \frac{number\ of\ lose}{total\ number\ of\ win\ and\ lose} \tag{6}$$

$$Tie\ rate = \frac{number\ of\ tie\ in\ situation1}{total\ number\ of\ tie} \tag{7}$$

Then, taking Table 10 as an example, there are a total of 20 cases in scenario1, and we can find that the numbers of win, lose, and tie are eight, eight, and four, respectively. Furthermore, we calculate the average CPEG to see how much gain percentage of the post engagements under the same ad budget. When calculating the average CPE increasing gain (ACPE-IG), we only consider and add the cases where CPEGs are larger than 0%, and take the average. When it comes to the average CPE decreasing gain (ACPE-DG), we only consider the ones lower than 0%. Repeating the above procedure, we can obtain the results of LSI and LDA for the three scenarios, which are summarized in Table 11.

Table 11. Summary of results.

	Scenario1		**Scenario2**		**Scenario3**	
	LSI	**LDA**	**LSI**	**LDA**	**LSI**	**LDA**
Win	8	6	10	7	11	9
Lose	8	9	5	4	3	5
Tie	4	5	5	9	6	6
cases	20	20	20	20	20	20
Win rate	50%	40%	67%	64%	79%	64%
ACPE-IG	18.6%	11.5%	21.9%	15.4%	22.5%	20.3%
Lose rate	50%	60%	33%	36%	21%	36%
ACPE-DG	11.9%	18.9%	13.6%	8%	14%	11.6%
Tie rate	75%	60%	100%	78%	17%	33%

In scenario1, we directly use the photo post texts of the wine fan page and compare the recommendation effectiveness by the LSI- and LDA-based TMRS. LSI-based TMRS achieves a 50% win rate and increases the ACPE-IG by 18.6%, while it reduces the ACPE-DG by 11.9% in the lost part. LDA-based TMRS only achieves a 40% win rate and increases the ACPE-IG by 11.5%, while it reduces the ACPE-DG by 18.9% in the lost part. The tie rates for LSI and LDA are 75% and 60%, respectively. Additionally, we can see the results of scenario2 and scenario3, which are shown in Table 11.

An advertising post example from a wine fan page recommended by LSI-based TMRS was shown in Figure 3. The slogans of figure were "Burn your passion, win your Bud beer." "Login invoices and win the prize," and "No drunk driving. Don't drive after drinking let you safe and secure." The number of likes was 1750 times, and this recommended post received about 150 comments and 180 shared times. The TMRS analysis results showed that the representative words included prize, share it, limited gifts, invoice, and so on. This result was similar to the result of manual inspection.

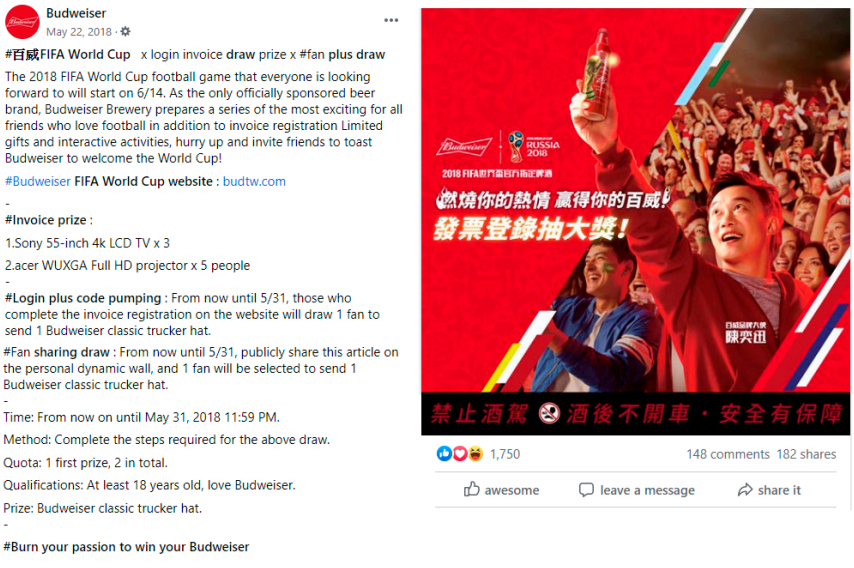


Figure 3. A recommended post example of TMRS.

According to the results of the above experiments, the engagement effect of LSI is better than that of LDA. Take scenario2 in Table 9 as an example, the representative words for one of the topics extracted by LSI are "activities, flavors, classics, messages, absolutes, fans, first time, double barrels, time, friends", among which "activities, messages, fans, friends" and "Classics, first time, time" have a certain correlation with each other. Those representative words for one of the topics extracted by LDA in scenario2 are "cherry blossoms, appearance, flower season, cans, faces, couples, friends, aftertaste, rogue, lobster", among them, only "cherry blossoms, flower season" are related to each other, and other words are less relevant. That is to say, the topic formed by the words obtained by LSI is more obvious than the topic of LDA. This is due to the weak correlation between the components of the random vector of the Dirichlet distribution (The reason why there is some "relevance" is that the sum of the weights must be 1), making the potential topics of the LDA hypothesis almost irrelevant. Therefore, from the results of each scenario, it can be inferred that in the fan page posts, if the topics extracted by LSI or LDA are not completely independent, it will affect the recommended effectiveness of TMRS.

Then, from the comparison of the results of scenario2 and scenario1, it indicates that photo post data identified by marketing experts and then used in TMRS is significantly better than ERRM. Therefore, one can gain more post engagements under the same marketing budget. Finally, we apply the best setting and method for TMRS from scenario1 and scenario2 to scenario3 to verify whether the TMRS is still as effective. From Table 11, it can be seen that the experimental results of scenario3 are in line with expectations, and LSI has a 79% win rate, which is higher than the LDA model. ACPE-IG is also as high as 22.5%.

9. Conclusions

In this paper, we successfully propose a Facebook photo post recommendation system based on the topic model that can increase the fan page post engagement rate, and develop an automated method to select posts to create ads to replace the manual selection by social media managers, and reduce the managers' daily workload. The text mining method we proposed here, LSI is more suitable for the TMRS than LDA from the experiment results, and effectively improves the traditional ERRM of the existing system. These results confirm that LSI and LDA techniques are useful in context-awareness-based recommendation systems [13]. In the recommendation results from the

experimental fan page, we have helped more than half of the fan pages to effectively increase the post engagement rate or achieve the effect of saving the budget. TMRS can also provide social media managers with popular keywords referring to the previous Facebook ad posts. The need of considering using implicit trust-based information to select fan page posts with a high interaction rate automatically is also verified [20]. In addition, the photo post datasets of the wine fan page identified by marketing experts are more effective in improving the effectiveness of the TMRS, and we have proved the effectiveness of the TMRS by applying it to other types of fan pages, such as makeup/skincare fan pages. Furthermore, even in the tie situation of TMRS and ERRM, our TMRS is still better than ERRM, since it is not necessary to publish posts or create post ads first to help the social media managers to prerecommend. All the above results prove that the advertising budget can be saved and more engagements can be achieved than the existing recommendation methods.

In the future, there are still several points that can be improved. For example, designing an automatic classifier to replace the experts' identification for improving the winning rate of the recommendation system. This requires many times to communicate with experts to learn and analyze their identification knowledge. Furthermore, how to determine the number of model topics for different fan page types is difficult. Although we can decide a value based on past advertising data, whether this value will cause overfitting or underfitting remains to be evaluated. In addition, Facebook posts have a comment mechanism, so that users can leave their feelings under the related post. Therefore, we can consider the sentiment analysis of the comments under the post, which can be used as another reference indicator to provide a more accurate recommended post order. Finally, TMRS is constructed using the text content of the photo post selected by experts, but the photo is another important factor. In the future, we will also think about how to include the advertising features of photos to the recommendation system, so as to enhance the recommendation effectiveness of the entire model and provide more reference value for the social media managers.

Author Contributions: Conceptualization, S.-M.Y.; methodology, S.-M.Y. and C.-H.L.; software, J.-C.Y.; validation, C.-H.L. and J.-C.Y.; formal analysis, J.-C.Y.; investigation, C.-H.L. and L.-X.C.; resources, S.-M.Y.; data curation, J.-C.Y.; writing—original draft preparation, J.-C.Y.; writing—review and editing, C.-H.L. and L.-X.C.; visualization, C.-H.L. and L.-X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Ministry of Science and Technology of Taiwan under the grant no. 108-2511-H-009 -009 -MY3.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lu, J.; Wu, D.; Mao, M.; Wang, W.; Zhang, G. Recommender system application developments: A survey. *Decis. Support Syst.* **2015**, *74*, 12–32. [CrossRef]
2. Kumar, P.; Reddy, G.R.M. Friendship recommendation system using topological structure of social networks. In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 237–246.
3. Lai, C.-H.; Lee, S.-J.; Huang, H.-L. A social recommendation method based on the integration of social relationship and product popularity. *Int. J. Hum. Comput. Stud.* **2019**, *121*, 42–57. [CrossRef]
4. Lee, D.; Brusilovsky, P. Recommendations based on social links. In *Social Information Access*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 391–440.
5. Ma, X.; Ma, J.; Li, H.; Jiang, Q.; Gao, S. ARMOR: A trust-based privacy-preserving framework for decentralized friend recommendation in online social networks. *Future Gener. Comput. Syst.* **2018**, *79*, 82–94. [CrossRef]
6. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl. Based Syst.* **2013**, *46*, 109–132. [CrossRef]
7. Ricci, F.; Rokach, L.; Shapira, B. Recommender systems: Context-aware recommender systems. In *Recommender Systems Handbook*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 191–221.
8. MAA. Taipei Media Agency Association (MAA). Available online: https://maataipei.org/ (accessed on 10 May 2018).

9. Rayson, S. Facebook Engagement for Brands and Publishers Falls 20% in 2017. Available online: https://buzzsumo.com/blog/facebook-engagement-brands-publishers-falls-20-2017/ (accessed on 20 February 2018).
10. Yu, S.; Kak, S. A survey of prediction using social media. *arXiv preprint* arXiv:1203.1647 2012.
11. Yera, R.; Martinez, L. Fuzzy tools in recommender systems: A survey. *Int. J. Comput. Intell. Syst.* **2017**, *10*, 776–803. [CrossRef]
12. Rosaci, D. Finding semantic associations in hierarchically structured groups of Web data. *Form. Asp. Comput.* **2015**, *27*, 867–884. [CrossRef]
13. Li, L.; Wang, D.; Li, T.; Knox, D.; Padmanabhan, B. SCENE: A Scalable Two-Stage Personalized News Recommendation System. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 25–29 July 2011; ACM: New York, NY, USA, 2011; pp. 125–134.
14. Ngoc, P.T.; Yoo, M. The Lexicon-based Sentiment Analysis for Fan Page Ranking in Facebook. In Proceedings of the 2014 International Conference on Information Networking (ICOIN), Phuket, Thailand, 10–12 February 2014; IEEE: Piscataway, NJ, USA; pp. 444–448.
15. Parsons, A. Using Social Media to Reach Consumers: A Content Analysis of Official Facebook Pages. *Acad. Mark. Stud. J.* **2013**, *17*, 27.
16. Goncalves, J.; Liu, Y.; Xiao, B.; Chaudhry, S.; Hosio, S.; Kostakos, V. Increasing the Reach of Government Social Media: A Case Study in Modeling Government–Citizen Interaction on Facebook. *Policy Internet* **2015**, *7*, 80–102. [CrossRef]
17. He, W.; Zha, S.; Li, L. Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry. *Int. J. Inf. Manag.* **2013**, *33*, 464–472. [CrossRef]
18. Poongodi, M.; Vijayakumar, V.; Rawal, B.; Bhardwaj, V.; Agarwal, T.; Jain, A.; Ramanathan, L.; Sriram, V. Recommendation model based on trust relations & user credibility. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4057–4064.
19. Rosaci, D. CILIOS: Connectionist inductive learning and inter-ontology similarities for recommending information agents. *Inf. Syst.* **2007**, *32*, 793–825. [CrossRef]
20. Daraghmi, E.Y.; Yuan, S.-M. We are so close, less than 4 degrees separating you and me! *Comput. Hum. Behav.* **2014**, *30*, 273–285. [CrossRef]
21. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
22. Hofmann, T. Probabilistic Latent Semantic Indexing. *Acm Sigir Forum* **2017**, *51*, 211–218. [CrossRef]
23. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
24. Hoffman, M.; Bach, F.R.; Blei, D.M. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: Vancouver, BC, Canada, 2010; pp. 856–864.
25. Blei, D.M.; Lafferty, J.D. A Correlated Topic Model of Science. *Ann. Appl. Stat.* **2007**, *1*, 17–35. [CrossRef]
26. Steyvers, M.; Griffiths, T. Probabilistic topic models. *Handb. Latent Semant. Anal.* **2007**, *427*, 424–440.
27. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1608. [CrossRef]
28. Mitrofanova, O. Probabilistic Topic Modeling of the Russian Text Corpus on Musicology. In Proceedings of the International Workshop on Language, Music, and Computing, St. Petersburg, Russia, 20–22 April 2015; pp. 69–76.
29. Řehůřek, R. Gensim Tutorial. Available online: https://radimrehurek.com/gensim/tut2.html#id6 (accessed on 20 March 2018).
30. Fxsjy. Jieba. Available online: https://github.com/fxsjy/jieba (accessed on 25 April 2018).
31. Facebook. Facebook Ad Insights. Available online: https://developers.facebook.com/docs/marketing-api/insights/?locale=en_US (accessed on 28 February 2018).
32. Nguyen, H.V.; Bai, L. *Cosine Similarity Metric Learning for Face Verification. Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 709–720.
33. Xue, N. Chinese Word Segmentation as Character Tagging. *Comput. Linguist. Chin. Lang. Process.* **2003**, *8*, 29–48.
34. Sproat, R.; Emerson, T. The First International Chinese Word Segmentation Bakeoff. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing—Volume 17; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 133–143.

35. Keskisärkkä, R. Automatic Text Simplification via Synonym Replacement. Master's Thesis. Available online: http://www.diva-portal.org/smash/get/diva2:560901/FULLTEXT01.pdf (accessed on 10 April 2018).
36. Ramos, J. Using TF-IDF to Determine Word Relevance In Document Queries. *Proc. First Instr. Conf. Mach. Learn.* **2003**, *242*, 133–142.
37. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]

symmetry

MDPI

Article

Micro-Distortion Detection of Lidar Scanning Signals Based on Geometric Analysis

Shuai Liu [1,2,3], Xiang Chen [1], Ying Li [4] and Xiaochun Cheng [5,*]

1 State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang 471000, China; cs.liu.shuai@gmail.com (S.L.); cemee_xchen@163.com (X.C.)
2 College of Computer Science, Inner Mongolia University, Hohhot 010012, China
3 College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
4 College of information and communication engineering, Harbin Engineering University, Harbin 150000, China; 3188279500@hrbeu.edu.cn
5 College of Computer Science, Middlesex University, London NW4 4BT, UK
* Correspondence: x.cheng@mdx.ac.uk

Received: 20 October 2019; Accepted: 29 November 2019; Published: 3 December 2019

Abstract: When detecting micro-distortion of lidar scanning signals, current hardwires and algorithms have low compatibility, resulting in slow detection speed, high energy consumption, and poor performance against interference. A geometric statistics-based micro-distortion detection technology for lidar scanning signals was proposed. The proposed method built the overall framework of the technology, used TCD1209DG (made by TOSHIBA, Tokyo, Japan) to implement a linear array CCD (charge-coupled device) module for photoelectric conversion, signal charge storage, and transfer. Chip FPGA was used as the core component of the signal processing module for signal preprocessing of TCD1209DG output. Signal transmission units were designed with chip C8051, FT232, and RS-485 to perform lossless signal transmission between the host and any slave. The signal distortion feature matching algorithm based on geometric statistics was adopted. Micro-distortion detection of lidar scanning signals was achieved by extracting, counting, and matching the distorted signals. The correction of distorted signals was implemented with the proposed method. Experimental results showed that the proposed method had faster detection speed, lower detection energy consumption, and stronger anti-interference ability, which effectively improved micro-distortion correction.

Keywords: geometric analysis; lidar scanning signal; micro-distortion; detection technology; TCD1209DG; lossless signal transmission

1. Introduction

The recent development of wireless information has promoted the maturity of laser lidar mapping technology. As a key technology in the field of surveying and mapping, lidar scanning has received more and more attention from relevant experts and scholars [1,2]. Lidar mapping technology quickly and accurately acquires three-dimensional information of objects, making it widely used in production and life [3]. However, due to factors, such as the outdoor light, the lidar scanning signal is slightly distorted, which affects the accuracy of the acquired information [4]. Detection for micro-distortion of lidar scanning signals improves the quality of the acquired lidar scanning signal, which is of great significance for ensuring the utilization efficiency of lidar scanning signals [5]. However, existing distortion detection technologies of lidar scanning signals only detect the region where the target's distortion of three-dimensional information is severe. With the popularity of lidar scanning applications, accuracy requirements for lidar scanning are getting higher [6]. Since current complicated micro-distortion detection technologies of lidar scanning signals have poor anti-interference, micro-distortion detection technology of scanning signals has become the focus of

research in this area. With the deepening of the research content, some mature theories and applications have been produced [7].

Different experts and scholars realized the detection of micro-distortion for scanning signals by different methods in years. However, there are still some shortcomings in this research domain and need follow-up experts and scholars to study. A micro-distortion detection technology based on fiber optic gyroscope was proposed by Zheng et al. [8]. The calibrated signal structure was tested linearly by achieving a level gauge of horizontal and vertical signal distortion calibration. Micro-distortion detection was further realized by the calibration result. However, this method did not pre-process scanning signals, so its distortion detection was easily interfered by factors, such as illumination change. Lupi et al. [9] proposed a distortion detection technique for lidar scanning signals based on active panoramic vision. Lidar scanning signal information was obtained by acquiring three-dimensional coordinates of point clouds for lidar scanning signals. Then, lidar scanning signals were preprocessed, and their three-dimensional coordinates of the feature points for lidar scanning signals were determined to realize quantitative analysis. According to the analysis results, a three-dimensional model was constructed to realize distortion detection for lidar scanning signals. However, the detection process of this method was complicated, which affected high time-consuming detection. Xu et al. [10] proposed a distortion detection technique based on ASODVS (Active Stereo Omni-Directional Vision Sensing) for lidar scanning signals. The midpoint of the lidar scanning signal was determined by a Gaussian curve. The waveform of the lidar signal was smoothed by the Bezier curve. Lidar scanning signal was calibrated by ASODVS, and its distortion detection was realized by qualitative analysis. However, in the process of detecting the distortion signal designed by this method, more energy was consumed, and its detection cost was higher. In [10], it was proposed to use FPGA (Stratix, made by ALTERA, San Jose, California, USA) as the main control equipment, design a general echo signal acquisition card with 20 MHz sampling rate, filtering and hardware accumulation functions, and configurable parameters. For the design of the analog-to-digital conversion circuit and its peripheral circuit, signal conditioning circuit, level conversion circuit, RS232 interface circuit, and power circuit in the analog board card, the logic design of FPGA was given in that circuit too. For the radar scanning signal detection, the basic parameter configuration function and accumulation function were implemented, but the method had poor detection accuracy for long-distance signal detection.

Therefore, a micro-distortion detection method for lidar scanning signals with geometric statistical characteristics was proposed in this paper. The experimental study of lidar with a frequency of 500 MHz was carried out. The proposed method used TCD1209DG to model the linear array CCD (charge-coupled device), reverse optoelectronics, and store and transmit the signal charge. It improved the time-consuming and energy consumption of [9,10], respectively. The linear CCD module was designed according to LS series lidar. FPGA was used for signal preprocessing of TCD1209DG output. C8051, FT232, and RS-485 were used to reduce the loss of signal transmission. The signal TRAM publishing unit sent the distortion information of the radar scanning signal to the control host computer for micro-distortion detection of the radar scanning signal. The anti-interference ability of the method in [8] was improved. Through the experiment of micro-distortion correction, compared with the existing methods, this method had faster detection speed, lower detection energy consumption, and stronger anti-interference ability.

2. Hardware Support for the Technology

2.1. Overall Framework of the Technology

In order to realize micro-distortion detection of lidar scanning signals, technical modules, such as linear array CCD module, signal processing module, signal transmission unit, and PC control program, were designed according to LS series laser lidar. The overall framework of technology is shown in Figure 1.

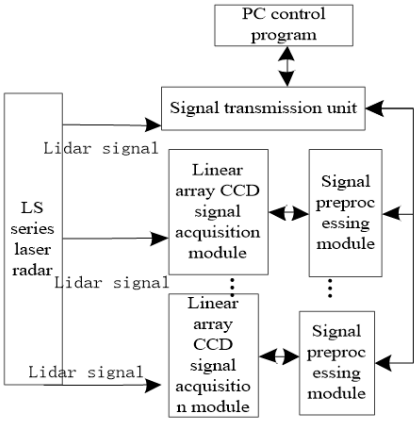


Figure 1. The overall framework of the proposed method.

According to Figure 1, each linear CCD module in the method received a word line laser emitted by the horizontal line lidar and outputted a corresponding signal through the lidar signal. When the lidar scanning signal changed slightly, its signal output from the line CCD module changed. The signal acquisition module collected the output signal and then digitized the acquired signal to determine micro-distortion information for the scanning signal. The signal transmission unit transmitted distortion information of the lidar scanning signal to the control host to realize micro-distortion detection for the lidar scanning signal. The control host could implement the setting of technical parameters and receive and process signals.

2.2. Development of Linear Array CCD Module

A charge-coupled device (CCD) [11] is a sensor that uses charge to realize signal transmission, enabling photoelectric conversion and signal charge storage and transfer. According to the working content of the linear array CCD module, chip TCD1209DG [12] with high sensitivity and high resolution is used as a chip of the linear array CCD module. The size of each photosensitive unit of tcd1209dg was 14 μm × 14 μm × 14 μm, the total length of the photosensitive array was 28.6 mm, the best working frequency was 1 MHz, and the maximum working frequency could reach 201 MHz. Because of its anti-interference advantages compared with the traditional chip, this paper chose this chip. The design of this hardware was to detect the micro-distortion signal effectively, so as to realize the detection of radar scanning signal. Using photodiode as the used pixel, the size of each pixel was set to 9.33 μm × 9.33 μm. The line CCD is shown in Figure 2.

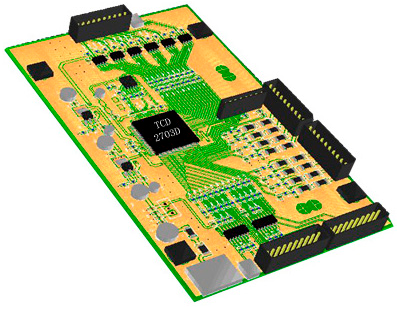
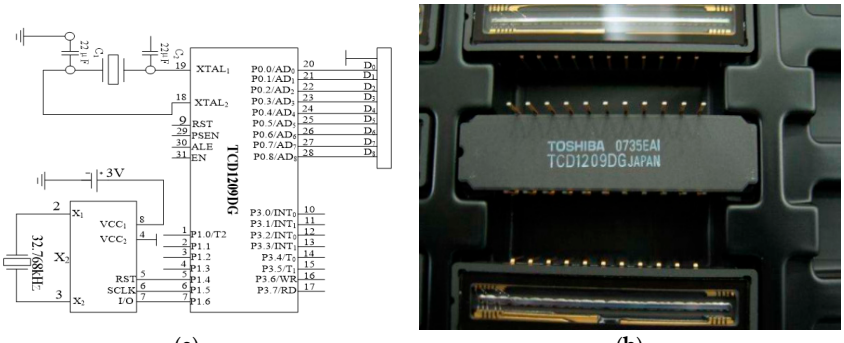


Figure 2. A general linear array CCD (charge-coupled device).

2.3. Signal Preprocessing Module

In order to realize the output of red, green, and blue light, chip TCD1209DG realized two channels of six frames of output, and its output signal size of each frame was 3984.2 × 16 *bits*. In the process of using TCD1209DG, the light entering the linear array CCD was made red by providing a red filter lens in front of it. It only processed the red light output and reduced the effect of external stray light on signal processing.

Chip TCD1209DG had five driving signals, including two-phase clock signals φ1A and φ2A, a charge conversion signal SH, a reset signal RS, and a clamp signal CP. There was a strict timing and phase relationship between the signals in TCD1209DG. Drive signal circuit diagram of TCD1209DG is shown in Figure 3a. When TCD1209DG was operating normally, the two-phase clock signal contained a high-level clamp signal and a reset signal, and the clamp signal lagged behind the reset signal. When the scanning signal was distorted, the clamp signal and the reset signal were low.



(**a**) (**b**)

Figure 3. TCD1209DG driver and its circuit. (**a**) TCD1209DG drive signal circuit, (**b**) A real TCD1209DG driver.

TCD1209DG drive signal was generated by the internal logic of FPGA and was implemented by VHDL programming [13]. Drive resulting from TCD1209DG is shown in Figure 3b.

The signal processing module mainly adjusted the signal output by TCD1209DG, digitized the processed signal, and stored the processed signal.

For signal output features of chip TCD1209DG, it was necessary to preprocess the output signal of each frame. The signal threshold processing was outputted for each frame, and the processed signal was transmitted to the signal acquisition module. The high sensitivity of TCD1209DG made the light have a greater impact on the output signal. By processing the threshold, interference of the light could be effectively reduced, and the quality of the output signal with the linear array CCD module was improved.

2.4. Signal Transmission Unit

In order to realize the lossless signal transmission for the lidar scanning signals of the control host and other slaves, a signal transmission unit was designed. The unit used FPGA as the core device of the signal processing module and adopted EP1C6Q144 [14] as the chip FPGA. After processing by FPGA, the micro-distortion information of the lidar scanning signal was stored in the off-chip SRAM.

The signal transmission unit was mainly composed of chip C8051 [15], chip FT232, and chip RS-485 [16]. In technology, the signal transmission unit host could be connected to any slave to realize communication. However, communication between slaves was impossible.

In order to realize the signal transmission unit, the connection between the USB interface with RS-232 and RS-485 interfaces was completed by using chip FT232. RS-232 interface was connected to

the RS-485 interface through chip RS-485. Using C8051F (made by Silicon Labs, Austin, Texas, USA) as an MCU, the timing between the chips was controlled to avoid communication interruption caused by bus conflicts. The specific implementation process of the signal transmission unit is shown in Figure 4.

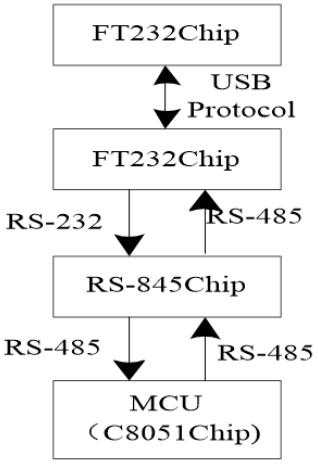


Figure 4. Composition of the signal transmission unit.

Through the above discussion, based on the requirements of the micro-distortion detection of the lidar scanning signals, the overall framework of the technology was analyzed. Designs of the linear array CCD module, the signal processing module, and the signal transmission unit for implementation determined the detection process of the micro-distortion detection technology for the lidar scanning signal.

3. Feature Matching Algorithm for Micro-Distortion Signal Based on Geometric Statistics

3.1. Frequency Matching of Micro-Distortion Based on Geometric Statistical Algorithm

The premise of realizing the detection of micro-distortion for the lidar scanning signals was the need to describe the lidar scanning signals. By accurately scanning the local properties of a signal, the frequency of the distortion signal was matched. According to the principle of lidar scanning and mapping, combined with the geometric statistical algorithm, a feature frequency band of the distorted signal was matched.

The distortion feature F_P of the scanning point band P on the segment S was introduced. A coordinate system was constructed by taking the scanning point band P as the origin and the normal [17] direction α_p as the abscissa. A tangent along the point P was equally divided into two sides. The length of S was set to L. Each micro-segment was recorded as S_i by projecting the points on S into S_i in sequence until the end of S. Using S_i statistics for a continuous segment on S, the distortion feature F_P of the scanning point band P thus obtained is:

$$F_P = \left(L, U_P^B(l,r), \left\{F'_{S_i} | i = \lfloor l/L \rfloor, \cdots, \lceil r/L \rceil\right\}\right) \tag{1}$$

In the above Equation (1), $U_P^B(l,r)$ represents the boundary farthest from the origin P projection, $\left\{F'_{S_i} | i = \lfloor l/L \rfloor, \cdots, \lceil r/L \rceil\right\}$ represents the set of all S_i distortion feature points, and $\lceil r/L \rceil - \lfloor l/L \rfloor$ represents the number of S_i associated with the scan point band.

Micro-segment S_i associated with the scanning point band was used as a segment with coordinate system information, and its distortion feature F'_{S_i} could be expressed as:

$$F'_{S_i} = (L_{S_i}, U^{\alpha}_{S_i}(\mu^{\alpha}_{S_i} - \sqrt{3}\delta^{\alpha}_{S_i}, \mu^{\alpha}_{S_i} + \sqrt{3}\delta^{\alpha}_{S_i}), C_{S_i}, U^{H}_{S_i}(t, d)) \tag{2}$$

In the above Equation (2), L_{S_i} and C_{S_i}, respectively, represent the length and overall unevenness of each micro-segment S_i after division. $U^{H}_{S_i}$ represents the vertical distance distribution range of the projected scan point band and point P [18]. $\mu^{\alpha}_{S_i}$ represents the average off-angle of the normal direction for the scanning point band with α_p in each micro-segment S_i. $\delta^{\alpha}_{S_i}$ represents the off-angle of the normal direction for the scanning point band with α_p in each micro-segment S_i. The normal direction of the micro-segment was set to satisfy consistent distribution. The associated microsegment was simplified into a distorted feature segment or arc. The range of the central angle was represented by $U^{H}_{S_i}$. S_i and $U^{H}_{S_i}$ were used to determine the position in the coordinate system. The direction of the opening was determined by $\mu^{\alpha}_{S_i}$ and C_{S_i}.

The segmentation distortion feature F_P could be seen as a special form of the associated segment distortion feature F'_{S_i}:

$$F_{S_i} = (L_{S_i}, \delta^{\alpha}_{S_i}, C_{S_i}) \Leftrightarrow F'_{S_i} = (L_{S_i}, U^{\alpha}_{S_i}(-\sqrt{3}\delta^{\alpha}_{S_i}, \sqrt{3}\delta^{\alpha}_{S_i}), C_{S_i}, \phi) \tag{3}$$

The similarity between different micro-segments S_1 and S_2 was calculated by micro-segment similarity S'_F, which is:

$$S'_F = R_L \cdot R_\delta \cdot R_H \frac{1}{1 + \left| C_{S_1}\delta^{\alpha}_{S_1} - C_{S_2}\delta^{\alpha}_{S_2} \right|} \tag{4}$$

In the above Equation (4), R_L represents the length ratio of each micro-segment in a lidar scanning signal. R_δ represents the overlap ratio of the central angle range between micro-segments. R_H represents the projection distribution ratio. According to the above equation, $S'_F \in [0, 1]$, and S'_F increases as the similarity increases.

In order to ensure the accuracy of distortion detection for the lidar scanning signal, it was necessary to accurately match the scanning point's frequency bands. For any of the scanning point bands P_1 and P_2, their coordinate systems were overlapped. Then, the matching degree of the scanning point band M_P could be expressed as:

$$M_P = (\overline{S}_F, O_B, \overline{S}'_F, O'_B) \tag{5}$$

In the above Equation (5), $\overline{S}_F$ and $\overline{S}'_F$, respectively, describe the average similarity of the corresponding micro-segment and the matching micro-segment. O_B represents overlap ratio of lidar scanning signal distortion feature span U^{B}_{P}. O'_B indicates the proportion of matching micro-segments. The average match case was described using $\overline{S}_F$ and O_B. $\overline{S}'_F$ and O'_B were used to improve the matching of scanning point bands. The highest matching degree was set to $M_{best} = (1, 1, 1, 1)$, and the matching degree could be expressed as:

$$V_{M_P} = \cos(\frac{\pi}{2} \cdot \sum_{i=1}^{4} w_i(\eta_i - 1)^2) \tag{6}$$

In the above Equation (6), $\sum_{i=1}^{4} w_i = 1$, w_i represents the weights of the distortion feature matching degree vector. η_i indicates an influencing factor of the matching. $V_{M_P} \in [0, 1]$, the larger V_{M_P}, the higher the matching. The geometrical statistical algorithm was used to match frequency bands of micro-distortion signals, and the geometrical statistical algorithm was combined according to frequency bands.

3.2. Nonlinear Micro-Distortion Signal Detection

During the process of signal scanning, the scanning signal is distorted due to the influence of illumination and scanning target chromatic aberration. Some obvious distortions can be easily detected by linear detection algorithms [3]. However, some distortion signals are nonlinear due to their distortion features, which are often overlooked by monitoring algorithms. Therefore, this paper adopted a nonlinear micro-distortion detection algorithm to detect micro-distortion signals based on the matched distortion signal frequency band. Figure 5 shows the nonlinear detection processing in this paper.

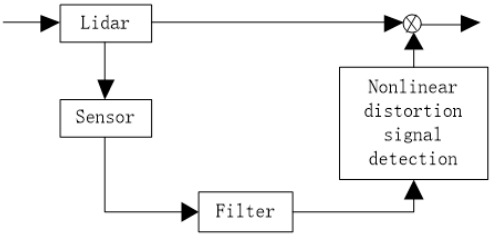


Figure 5. Nonlinear detection process of the micro-distortion signal.

The state of laser lidar scanning was fully considered. An output source of laser and the scanning signal were used as a basis for distortion detection. The analysis of the band structure for the distorted optical signal was as follows:

$$\begin{cases} q(k+1) = Aq(k) + Bh(k) + Cf(k) \\ p(k) = Dq(k) + Eh(k) \end{cases} \tag{7}$$

where $q(k)$ is the normal lidar signal scanning shape, $p(k)$ is the scanning result output, $h(k)$ is the disturbance output, $f(k)$ is the micro-distortion signal, and A, B, C, D, E are distortion dimension constant matrixes.

A lidar would have distortion problems during the scanning process. Therefore, a nonlinear state detection method was needed to detect the signal scanning with the sensor reaching the filter. Let the micro-distortion signal be $\alpha(k)$; then, its initial output signal could be expressed as:

$$p'(k) = Dq(k - \alpha(k)) + Eh(k - \alpha(k)) + Ff(k) \tag{8}$$

In the above Equation (8), $p'(k)$ is the output of dividing the micro-distortion signal set. $0 < \underline{\alpha} < \alpha(k) < \overline{\alpha}$, where $\underline{\alpha}$ is the lower bound of the micro-distortion signal, $\overline{\alpha}$ is the upper bound of the micro-distortion signal, and F is the matrix of the distortion constant for the distortion signal.

Common micro-distortion signal structures, such as the nonlinear micro-distortion state detection mechanism, could be expressed as Equation (9):

$$\begin{cases} q'(k+1) = Aq'(k) + Bg(k) + L\left(\overline{p'}(k) - p'(k)\right) \\ p'(k) = Dq'(k) \\ r(k) = M(\overline{p}(k) - p'(k)) \end{cases} \tag{9}$$

where $q'(k)$ is the control host status detection situation. $p'(k)$ is a micro-distortion output with a micro-distortion signal. $g(k)$ is the detection output. $r(k)$ is the residual signal with a micro-distortion signal. L and M are a gain matrix of the micro-distortion detection and the gain matrix of the residual signal, respectively [4].

3.3. Distortion Correction of Lidar Scanning Micro

The detected micro-distortion signal existing in the process of nonlinear micro-distortion detection was fully considered. According to the distortion distribution sequence, statistical signal features were:

$$\begin{array}{l} p(\beta(k) = 0) = R\{\beta(k)\} = \overline{\beta} \\ p(\beta(k) = 1) = 1 - R\{\beta(k)\} = 1 - \overline{\beta} \end{array} \tag{10}$$

Detection signal in the scanning and the random micro-distortion signal of lidar scanning signal were represented by $\alpha(k)$ in Equation (8) and $\beta(k)$ in Equation (10), respectively. Where $\alpha(k) = 1$ indicates the detection signal received by the control host, and $\beta(k) = 1$ indicates the lidar signal received by the control host. At this time, $\overline{p}(k) = p(k-1)$ and $g(k) = g'(k-1)$ indicate that the lidar signal had slight distortion. $\alpha(k) = 0$ and $\beta(k) = 0$ indicate that the host did not receive the detection signal and the scanning signal, respectively. At this time, $\overline{p}(k) = p(k)$ and $g(k) = g'(k)$, indicating that there was no micro-distortion signal in this frequency band.

The nonlinear state detection method was used to detect signal scanning with the sensor reaching the detection filter. According to this situation, a nonlinear micro-distortion state detection flow was designed. The micro-distortion state detection mechanism was set, the micro-distortion signal condition was analyzed, and the signal features were statistically reported to detect the micro-distortion [5].

In order to make an analysis of lidar problem more precise, the following distortion correction rules should be set:

(1) In order to improve the performance of the control host, the sensor design of the entire lidar needs to use the clock as the corrective drive, and the controller uses the event as the corrective drive.
(2) Data is scanned in a single package.
(3) The local scanning state of the micro-distortion signal is controllable.

When the lidar signal showed random micro-distortion, its output was:

$$y(k') = \alpha'(k')E'\mathrm{x}(k') \tag{11}$$

where $y(k')$ is the output of the micro-distortion detection, and E' is the matrix of the micro-distortion dimension constant.

According to the micro-distortion structure of the lidar signal, a nonlinear micro-distortion state correction mechanism, such as Equation (12), was constructed:

$$\begin{cases} x'(k'+1) = A'x'(k') + B'\overline{g}(k') + L'\left(\overline{y'}(k') - y'(k')\right) \\ y'(k') = D'g'(k') \end{cases} \tag{12}$$

where $x'(k')$ is the nonlinear control master state. $\overline{g}(k')$ is the micro-distortion signal input of the lidar. $g'(k')$ is a micro-distortion signal input without a lidar scanning signal. L', M' are the observer gain matrix and controller gain matrix for minor distortion correction, respectively [9].

The combined variable $\tau(k')$ obeyed Bernoulli distribution, and the statistical signal features were:

$$\begin{array}{l} \mathrm{Prob}(\tau(k') = 1) = R'\{\tau(k')\}\tau \\ \mathrm{Prob}(\tau(k') = 0) = 1 - R'\{\tau(k')\}1 - \tau \\ Var(\tau(k')) = R'(\tau(k') - \tau)^2 = (1-\tau)\tau = \tau^2 \end{array} \tag{13}$$

where $\alpha'(k')$ in Equation (11) and $\tau(k')$ in Equation (13) that obey Bernoulli distribution represent the distortion that occurs when the sensor transmits to the controller and the controller transmits to the actuator. $\alpha'(k') = 1$ indicates that the sensor successfully scans the signal set to the controller, and $\alpha'(k') = 0$ indicates that the sensor is distorted when transmitting the signal set to the controller.

$\tau(k') = 1$ indicates that the controller successfully scans the signal set to the actuator, and $\tau(k') = 0$ indicates that the controller is distorted when scanning the signal set to the actuator.

According to the block diagram of the micro-distortion structure containing lidar, the nonlinear state correction method was used to describe the micro-distortion signal structure of the lidar scanning signal. According to stated problems of micro-distortion signal structure, the reasonable correction rule was analyzed, and the output result when the micro-distortion had a random distortion phenomenon was calculated. Signal features under the distortion phenomenon were statistically combined with the variable $\tau(k')$ obeying Bernoulli distribution to correct the small lidar distortion.

When there was no micro-distortion of a lidar signal, the error caused by other interference factors of the control host could be ignored. At this time, the control host output result was 0, and the micro-distortion observation error was also 0. When micro-distortion of the lidar signal occurred, the output of the control host at this time was not 0, and the observation error of the micro-distortion was not 0. The magnitude of the error changed as the output was processed. Using the nonlinear feedback control law, when the micro-distortion control error increased, the micro-distortion observation error needed to be used as the residual. The entire scanning state of the micro-distortion was analyzed by observing the change in the residual. It completed the correction of micro-distortion for the lidar scanning signal.

According to the above description, these correction principles should be followed. If the absolute value of the error was less than the output result threshold, then the lidar was in a normal scanning state. If the absolute value of the error was greater than or equal to the output result threshold, then the scan signal was slightly distorted. According to this principle, the distortion correction mechanism could be designed to complete the fault-tolerant control for micro-distortion laser lidar, and the fault-tolerant phenomenon in the distortion correction process was divided and compensated.

4. Results and Analysis of Simulation Experiments

In order to test the detection time and the detection degree of the proposed geometric statistics-based micro-distortion detection technology for lidar scanning signals, the experiment was compared with methods in [8,9]. The experimental platform was built by the control host of Intel B360 i7-8700. Using Windows 2010 as the operating system, MATLAB was used to simulate the process. The detection interface of distortion signals is shown in Figure 6.

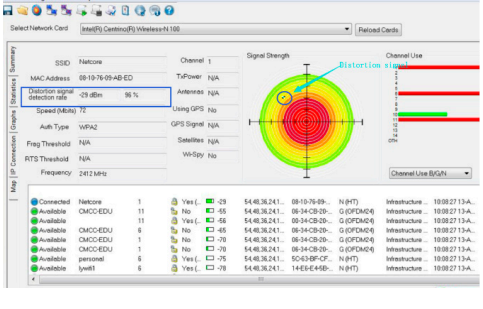


Figure 6. Detection interface of lidar signal distortion.

The proposed method and methods in [8,9] were used to detect the micro-distortion of 3000 sets for lidar scanning signals. Through experiments, time-consuming results of the three methods were recorded, as shown in Table 1.

Table 1. Comparison of time-consuming for micro-distortion detection by different methods.

Number of Experiments/Time	Proposed Method/s	Literature [8] Method/s	Literature [9] Method/s
1	2.13	2.58	3.03
2	2.16	2.64	3.09
3	2.08	2.52	2.94
4	2.09	2.53	2.95
5	2.12	2.56	3.02
6	2.15	2.63	3.07
7	2.10	2.54	3.96
8	2.14	2.60	3.05
9	2.13	2.59	3.04
10	2.11	2.26	2.99

It could be seen from Table 1 that it took less time to detect micro-distortion for the lidar scanning signal by the proposed method. It showed that its detection speed was faster, and the detection efficiency was higher. The proposed method directly matched scanning points in the process of detecting micro-distortion for lidar scanning signals. Calculation steps of the micro-distortion detection for lidar scanning signals were reduced, and its detection speed was improved. Therefore, the detection of micro-distortion for lidar scanning signals took a short time.

In order to ensure the reliability of the research and development technology, a noise signal was added through MATLAB and then connected to software Original Pro 7.0 (Website: www.Originlab.com). Using the proposed method and methods in [8,9], the micro-distortion of the lidar scanning signal was detected in a strong interference environment. The three-dimensional output of the three methods is shown in Figure 7.

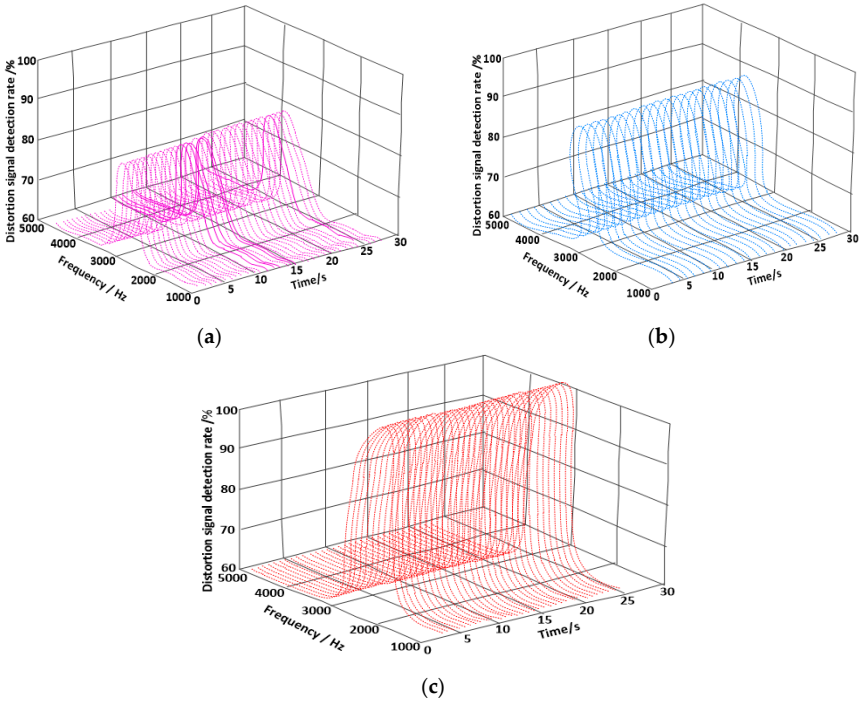


Figure 7. Comparison of detection rates for distortion signals. (**a**) Distortion detection rate of the method in [8], (**b**) Distortion detection rate of the method in [9], (**c**) Distortion detection rate of the proposed method.

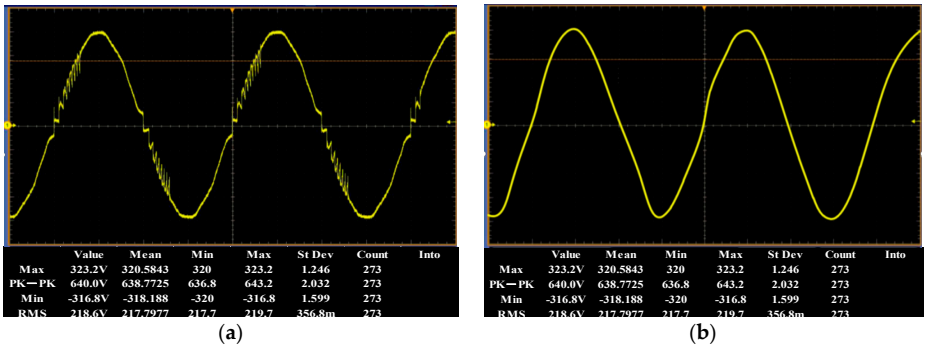
It could be seen from Figure 7 that the proposed method had a high detection rate of the distortion signal in the process of monitoring micro-distortion for the lidar scanning signal. It showed that the external factor had the least influence on the detection rate of the distortion signal, and the proposed detection method had the strongest anti-interference. In the process of detecting micro-distortion of the lidar scanning signal, the proposed method effectively reduced interference of the light and improved the anti-interference of the signal.

An experiment used the proposed method and the classical methods to compare the energy consumption of micro-distortion detection for lidar scanning signals. During the experiment, the results obtained are shown in Table 2.

Table 2. Comparison of energy consumption (nJ) for distortion signals detected by different methods.

Number of Experiments/Time	Proposed Method/nJ	Literature [8] Method/nJ	Literature [9] Method/nJ
1	235	342	318
2	241	350	326
3	237	345	321
4	240	352	328
5	236	343	317
6	235	342	315
7	238	347	323
mean			

It could be seen from Table 2 that the lidar scanning signal detected by the proposed method had the least energy consumption. It showed that the proposed method was less costly for distortion signal detection. In the process of energy consumption detection, the proposed method had fewer modules, which required less energy consumption. In order to ensure the correct performance of the distortion signal for the R&D technology, MATLAB was used to connect an oscilloscope software. The detected micro-distortion signal was corrected by the proposed method. Its correction result is shown in Figure 8.



(**a**) (**b**)

Figure 8. Comparison of micro-distortion signals before and after correction. (**a**) Minor distortion signal before correction, (**b**) Corrected micro-distortion signal.

It could be seen from Figure 8 that there were five micro-distortion signal bands in the lidar scanning signal samples selected before the correction. After the lidar scanning micro-distortion signal was corrected by the proposed method, the signal waveform before the correction was smoother, and there was no distortion phenomenon. It was proved that the proposed method had strong feasibility in the function of micro-distortion correction.

5. Conclusions

Micro-distortion detection of the lidar scanning signal could be used to improve the lidar systems. Existing micro-distortion detection technologies of lidar scanning signal have the problems of long

detection time, high energy consumption, and poor performance against interference [19,20]. To deal with these problems, a technique based on geometric statistics for micro-distortion detection of the lidar signal was proposed. This project built an overall framework for the micro-distortion detection using TCD1209DG in linear array CCD module for photoelectric conversion, signal charge storage, and transfer. FPGA chip was used for the signal preprocessing of TCD1209DG output. C8051, FT232, and RS-485 were used for signal transmission. The signal distortion features were analyzed by geometric statistics for micro-distortion detection. Experimental results showed the effectiveness of the proposed method. The following conclusions were drawn:

(1) In the process of micro-distortion monitoring of radar scanning signal, this method had a high detection rate of distortion signal compared with the methods in [8,9].
(2) The energy consumption of the radar scanning signal detected by this method was the least compared with the methods in [8,9].
(3) This method could effectively correct the distorted signal using the frequency difference formula "f = −2v/Lamda" for the change of the Doppler principle compared with the methods in [8,9].

Author Contributions: Conceptualization, S.L. and X.C. (Xiaochun Cheng); methodology, Y.L.; software, X.C. (Xiang Chen); validation, X.C. (Xiang Chen) and Y.L.; formal analysis, S.L.; investigation, X.C. (Xiaochun Cheng); resources, X.C. (Xiang Chen); data curation, X.C. (Xiang Chen); writing—original draft preparation, X.C. (Xiang Chen) and S.L.; writing—review and editing, X.C. (Xiaochun Cheng); visualization, X.C. (Xiang Chen); supervision, S.L.; project administration, S.L.; funding acquisition, S.L.

Funding: This work is supported by the Open Project Program of the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System under Grant 2019K0104B. National Natural Science Foundation of China project under Grant 61502254, Program for Yong Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT-18-B10.

Acknowledgments: We want to give our sincere gratitude for the effective work of the editorial board of journal "Symmerty", as well as the guest editors of the special section "Recent Advances in Social Data and Artificial Intelligence 2019".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aubry, A.; Maio, A.D.; Pallotta, L.A. Geometric Approach to Covariance Matrix Estimation and its Applications to Radar Problems. *IEEE Trans. Signal Process.* **2018**, *66*, 907–922. [CrossRef]
2. Muqaibel, A.H.; Abdalla, A.T.; Alkhodary, M.T.; Alawsh, S.A. Through-the-wall radar imaging exploiting Pythagorean apertures with sparse reconstruction. *Digit. Signal Process.* **2017**, *61*, 86–96. [CrossRef]
3. Lin, Y.; Wang, C.; Wan, J.; Dou, Z. A Novel Dynamic Spectrum Access Framework Based on Reinforcement Learning for Cognitive Radio Sensor Networks. *Sensors* **2016**, *16*, 86–96.
4. Rakesh, P.R.; Narayanan, G. Investigation on Zero-Sequence Signal Injection for Improved Harmonic Performance in Split-Phase Induction Motor Drives. *IEEE Trans. Ind. Electron.* **2017**, *64*, 2732–2741. [CrossRef]
5. Lin, Y.; Zhu, X.; Zheng, Z.; Dou, Z.; Zhou, R. The individual identification method of wireless device based on dimensionality reduction and machine learning. *J. Supercomput.* **2017**, *75*, 3010–3027. [CrossRef]
6. Liu, S.; Bai, W.; Zeng, N.; Wang, S. A Fast Fractal Based Compression for MRI Images. *IEEE Access* **2019**, *7*, 62412–62420. [CrossRef]
7. Piazzo, L.; Raguso, M.C.; Seu, R.; Mastrogiuseppe, M. Signal enhancement for planetary radar sounders. *Electron. Lett.* **2019**, *55*, 153–155. [CrossRef]
8. Zheng, Y.; Zhang, C.; Li, L. Influences of optical-spectrum errors on excess relative intensity noise in a fiber-optic gyroscope. *Opt. Commun.* **2018**, *410*, 504–513. [CrossRef]
9. Lupi, S.M.; Galinetto, P.; Cislaghi, M.; y Baena, A.R.; Scribante, A.; y Baena, R.R. Geometric distortion of panoramic reconstruction in third molar tilting assessments: A comprehensive evaluation. *Dentomaxillofacial Radiol.* **2018**, *47*, 20170467. [CrossRef] [PubMed]
10. Wu, T.; Lu, S.; Tang, Y. Research on panoramic point cloud data acquisition technology based on ASODVS. *J. Comput. Meas. Control* **2014**, *22*, 2284–2287.

11. Hua, X.; Cheng, Y.; Wang, H.; Qin, Y.; Li, Y. Geometric means and medians with applications to target detection. *IET Signal Process.* **2017**, *11*, 711–720. [CrossRef]
12. Gui, R.; Wang, W.Q.; Cui, C.; So, H.C. Coherent Pulsed-FDA Radar Receiver Design with Time-Variance Consideration: SINR and CRB Analysis. *IEEE Trans. Signal Process* **2017**, *66*, 200–214. [CrossRef]
13. Le, Z.; Wang, X. Super-Resolution Delay-Doppler Estimation for OFDM Passive Radar. *IEEE Trans. Signal Process.* **2017**, *65*, 2197–2210.
14. Zhang, Y.; Pan, S. Broadband Microwave Signal Processing Enabled by Polarization-Based Photonic Microwave Phase Shifters. *IEEE J. Quantum Electron.* **2018**, *54*, 1–12. [CrossRef]
15. Engels, F.; Heidenreich, P.; Zoubir, A.M.; Jondral, F.K.; Wintermantel, M. Advances in Automotive Radar: A framework on computationally efficient high-resolution frequency estimation. *IEEE Signal Process. Mag.* **2017**, *34*, 36–46. [CrossRef]
16. Wanchun, L.; Qiu, T.; Chengfeng, H.; Yingxiang, L. Location algorithms for moving target in non-coherent distributed multiple-input multiple-output radar systems. *IET Signal Process.* **2017**, *11*, 503–514. [CrossRef]
17. Pan, Z.; Liu, S.; Sangaiah, A.K.; Muhammad, K. Visual attention feature (VAF): A novel strategy for visual tracking based on cloud platform in intelligent surveillance systems. *J. Parallel Distrib. Comput.* **2018**, *120*, 182–194. [CrossRef]
18. Cheng, Z.; Liao, B.; He, Z.; Li, Y.; Li, J. Spectrally Compatible Waveform Design for MIMO Radar in the Presence of Multiple Targets. *IEEE Trans. Signal Process.* **2018**, *66*, 3543–3555. [CrossRef]
19. Duan, K.; Wang, Z.; Xie, W.; Chen, H.; Wang, Y. Sparsity-based STAP algorithm with multiple measurement vectors via sparse Bayesian learning strategy for airborne radar. *IET Signal Process.* **2017**, *11*, 544–553. [CrossRef]
20. Zhang, W.; Fu, Y.; Nie, L.; Zhao, G.; Yang, W.; Yang, J. Parameter estimation of micro-motion targets for high-range-resolution radar using high-order difference sequence. *IET Signal Process.* **2018**, *12*, 1–11. [CrossRef]

symmetry

MDPI

Article

A Two-Tier Partition Algorithm for the Optimization of the Large-Scale Simulation of Information Diffusion in Social Networks

Bin Chen [1], Hailiang Chen [1,*], Dandan Ning [1], Mengna Zhu [1], Chuan Ai [1], Xiaogang Qiu [1] and Weihui Dai [2]

[1] College of Systems Engineering, National University of Defense Technology, Changsha 410073, China; chenbin06@nudt.edu.cn (B.C.); lizhen08@nudt.edu.cn (D.N.); zhumengna16@nudt.edu.cn (M.Z.); aichuan@nudt.edu.cn (C.A.); xgqiu@nudt.edu.cn (X.Q.)

[2] Department of Information Management and Information Systems, School of Management, Fudan University, Shanghai 200433, China; whdai@fudan.edu.cn

* Correspondence: chenhailiang@nudt.edu.cn; Tel.: +86-18681203974

Received: 26 April 2020; Accepted: 14 May 2020; Published: 21 May 2020

Abstract: As online social networks play a more and more important role in public opinion, the large-scale simulation of social networks has been focused on by many scientists from sociology, communication, informatics, and so on. It is a good way to study real information diffusion in a symmetrical simulation world by agent-based modeling and simulation (ABMS), which is considered an effective solution by scholars from computational sociology. However, on the one hand, classical ABMS tools such as NetLogo cannot support the simulation of more than thousands of agents. On the other hand, big data platforms such as Hadoop and Spark used to study big datasets do not provide optimization for the simulation of large-scale social networks. A two-tier partition algorithm for the optimization of large-scale simulation of social networks is proposed in this paper. First, the simulation kernel of ABMS for information diffusion is implemented based on the Spark platform. Both the data structure and the scheduling mechanism are implemented by Resilient Distributed Data (RDD) to simulate the millions of agents. Second, a two-tier partition algorithm is implemented by community detection and graph cut. Community detection is used to find the partition of high interactions in the social network. A graph cut is used to achieve the goal of load balance. Finally, with the support of the dataset recorded from Twitter, a series of experiments are used to testify the performance of the two-tier partition algorithm in both the communication cost and load balance.

Keywords: social network simulation; ABMS; Spark; two-tier partition algorithm

1. Introduction

With the development of Internet technology, Facebook, Twitter, WeChat, Weibo, and other social network applications have developed rapidly. As of 31 December 2018, Facebook had 2.32 billion monthly active users, with an average of 1.52 billion daily active users in December 2018. According to the annual data report of WeChat in 2018 [1], up to September of this year, there were 1.0825 billion active online users every month, and the daily information delivery volume of WeChat reached 45 billion times. The rapid development of all kinds of social media makes the use of the Internet has had a profound change, from the simple information search and browsing to the establishment and maintenance of online social relations, information creation, communication, and sharing based on social relationships. Social networks have penetrated every aspect of our life [2]. The role of social relationships in information diffusion, guidance for individuals, media influence, and promoting attitude or behavior change are influenced by social networks [3]. Research on social networks can be

summarized into three types: the structural characteristics and evolution mechanism of social networks, the formation and interaction of group behaviors in social networks, and the law and evolution of information diffusion in social networks [4]. However, the coverage of current mainstream media is widespread, and corresponding social network nodes are massive, complex in structure, and the calculation of whether nodes transmit information is complicated. Relevant theories in the fields of psychology, communication, and complex networks are difficult to be directly applied in quantitative researches. As a result, agent-based modeling and simulation (ABMS) [5] is one of the most popular methods to solve these problems in a symmetry simulation world. In ABMS, every node in the social network is mapped into an agent. As all the agents follow human behavioral mode, a social network evolves during the running of the simulation. Under certain conditions, a specific emergence may occur, reflecting the evolution of real society. In this symmetrical way, real-world data can be used to revise the information diffusion in the symmetrical simulation world, and at the same time, through the study of information diffusion in the symmetrical simulation world, we can intervene in real-world information diffusion.

The applications of ABMS tools, such as NetLogo, Repast, and Swarm, are very popular in the fields of sociology, communication, and psychology [6–9]. However, these simulation tools are designed with more consideration of generalization, and the flexibility and scalability are limited; in particular, the performance is not acceptable in large-scale simulation. Although Repast has an HPC(High Performance Computing) version, which is for high performance computing, it is not optimized for social network simulation. Repast HPC is suitable for modeling general agent-based complex systems, but complex networks themselves have characteristics that are different from general complex systems, such as small-world characteristics, scale-free characteristics, etc. At the same time, the propagation calculations in complex networks have special fast calculation requirements related to networks such as node degrees, clustering coefficients, and community structure. Therefore, as a generic ABMS tool, Repast HPC does not directly support the fast calculation of these characteristics of complex networks. Although Repast HPC is well known in the implementation of large scale ABMS, it hadn't been proven to perform well in the simulation of large-scale complex network propagation. Thus, large-scale social network simulation is a problem worth studying.

Large-scale social network simulation is used to study information diffusion and topology changes in the network through the modeling and simulation of large-scale social networks. The amount of data involved is huge, and various difficulties are faced in the actual studies.

- Difficulties in the storage, management, and analysis of big data. For large-scale social networks, such as WeChat and Weibo, the scale of data is very large—usually hundreds of millions or even billions. It is difficult to manage such unstructured data based on traditional file storage or relational databases. When the network scale is too large, the time consumption of connectivity is not acceptable in the traditional relational database, let alone in the analysis and calculation of network indexes such as node degree, cascades, and so on.
- Difficulties with the implementation of the simulation of large-scale social networks. Existing high-performance simulation engines, including MPI-based(Message Passing Interface) simulation tools, GPU-based(Graphic Processing Unit) simulation engines [10], and so on, cannot obtain high performance in the simulation of large-scale social networks because of the strong correlations in the network. Too many interactions among nodes decrease the performance greatly, especially when the nodes are distributed in different machines.

With the advent of the era of big data [11], many big data platforms such as Hadoop and Spark have been spawned. With the support of these technologies, it is easy to analyze big data by the efficient management and application of clusters, cloud environments, etc. [12,13]. These platforms can also support the simulation of the social networks obtained from big data. This paper proposes a method to implement a large-scale social network simulation based on Spark, and the optimization based on the network structure is also given to improve the performance of the simulation. The rest of

this paper is organized as follows. Section 2 gives the current research on large-scale social network simulation; Section 3 introduces how to build the large-scale social network simulation using the Spark platform; Section 4 presents the optimization of the social network simulation based on the network structure; the effectiveness of the optimization method is testified by the experiments in Section 5; the conclusion and areas for future work are given in Section 6.

2. Related Works

There are many studies on agent modeling based on existing tools, such as NetLogo, Swarm, MASON, and Repast [14,15]. NetLogo, with its heritage as an educational tool, stands out for its ease of use and excellent documentation. However, the performance of large models is hard to bear. Usually, only hundreds of agents are supported. For example, Anuj and Caroline et al. proposed an agent-based model to predict the performance of different residential distributed solar models with respect to the stakeholders' objectives [16]. Swarm is relatively small and well-organized while providing a fairly complete set of tools. However, it cannot support a large-scale simulation either. It is fast for simple models but slow for complex ones. MASON could be a good choice for experienced programmers working on computationally intensive models. However, it is nonstandard and sometimes confusing terminology is given. Repast is certainly the most complete Java platform, but it is not suitable for large-scale simulation [9]. The scale of agents developed in Repast is usually less than 1 million because of performance. For example, in reference [17], Griffin and Stanish developed an agent-based model using the Repast agent-based modeling toolkit. They took 500,000 agents for this model in total. In reference [18], Mock and Testa developed an agent-based model of predator–prey relationships between transient killer whales and threatened marine mammal species in Alaska based on Repast for about 200,000 agents. Existing agent-based modeling tools cannot support the large-scale simulation of information diffusion well.

Based on Hadoop, Kangsun Lee et al. designed ARLS (After-action Reviewer for Large-scale Simulations), a tool for analyzing simulation results, and used MapReduce to batch process simulation log files to accelerate the analysis of the simulation results [19]. Reference [20] proposed an agent-based modeling method based on Hadoop; the MapReduce programming model was used to implement the simulation kernel. Hadoop automatically realized the underlying load balance and fault tolerance. Compared with the traditional simulation tools, this framework greatly improves the performance and scalability of the simulation. References [21,22] applied Hadoop in the simulation of molecular dynamics and power systems. He Liang et al. proposed a large-scale Online Social Network (OSN) worm simulation based on MapReduce. The OSN worm propagation process was divided into different stages and MapReduce was used to construct the corresponding map and reduce algorithms [23]. However, because the programming model of Hadoop is relatively simple, and the hard disk needs to be read and written frequently during the operations, the above large-scale simulation research based on Hadoop still has many limitations in performance.

With the improvement of the Spark ecosystem, Spark has gradually become the most popular big data framework. Chuan Ai et al. [24] implemented the propagation simulation of large-scale social networks based on Spark's graph computing library GraphX and the Pregel algorithm. The performance of the simulation kernel was increased greatly GraphX is Apache Spark's API (Application Programming Interface) for graphs and graph-parallel computation, with a built-in library of common algorithms. The networks are organized and managed in the form of nodes and edges. Pregel is a bulk synchronous parallel computing model (BSP model). The simulation of propagation is abstracted as a series of supersteps. Nodes in each superstep pass messages to each other and the nodes update the state according to the messages received previously and their state update mechanism. The large-scale social network simulation method based on Pregel is simple and fast. However, because Pregel is tightly packaged, users can analyze the simulation process and results only after the simulation is completed, and state transformations cannot visually be observed during the simulation. As a result,

the difficulties faced by actual large-scale social network simulations lie in the whole process from data management to the analysis of the simulation results.

Based on the problems discussed above, this paper proposes a large-scale social network simulation framework based on Spark. The built-in components in the Spark ecosystem are used to support data management, data analysis, and simulation experiments in a large-scale social network simulation.

3. The Spark-Based Large-Scale Simulation of Information Siffusion in Social Networks

3.1. The Infrastructure of Simulation

The infrastructure of the Spark-based large-scale simulation of information diffusion in social networks designed in this paper is shown in Figure 1. The simulation framework is composed of three layers. From the bottom up, they are the hibernate layer, the scheduling layer, and the execution layer. The hibernate layer is used for the big data storage of large-scale social networks in HDFS (Hadoop Distributed File System). The scheduling layer is responsible for job scheduling and distribution, distributing large-scale social network simulation and analysis jobs to different nodes in the cluster. The execution layer is the top layer, which is composed of a data analysis module and a simulation kernel. The simulation kernel is the core part of the large-scale social network simulation framework based on Spark. Based on the data interaction characteristics and the characteristics of large-scale social networks, the optimization of the simulation kernel is proposed to further accelerate the execution of a large-scale social network simulation.

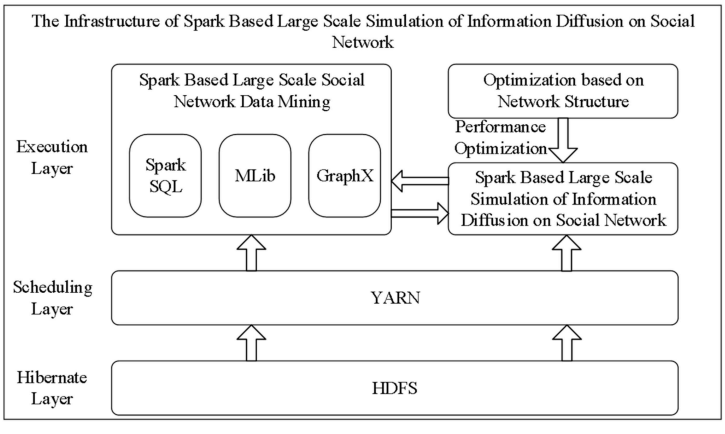


Figure 1. The infrastructure of a Spark-based large-scale simulation of information diffusion in a social network.

The large-scale social network simulation based on Spark is the core module that supports the execution of the agent-based simulation on distributed clusters. The agent model and the implementation of the simulation kernel on Spark are illustrated below.

3.2. The Agent Model

This paper uses ABMS to model information diffusion in large-scale social networks. Each agent represents a node in the social network. In a large-scale social network diffusion simulation, information is transmitted between nodes in the network according to a certain mechanism. In this paper, the SIR (Susceptible Infected Recovered) model is used for the large-scale social network diffusion simulation. First, the definition of the agent model is given as follows:

$$<\text{ID, ASN, S, I, R, M}>. \quad (1)$$

ID represents the unique identifier of the agent; ASN represents the set of out-degree neighbors, that is, the set of agents to whom the current agent will send messages. S, I, and R are the states of the agents. S means susceptible—it represents the agent who has not received the message or has received the message but the message is not enough to attract his attention; I means infected—it represents the agent who has received the message and would like to forward this message to his neighboring agents; R means recovered, which is always after S and never transforms to other states again—it represents the agent who does not care about the message and will never send this message to his neighboring agents. M is the message transmitted by the agent.

The behavior of the agent follows the mechanism of the SIR model. The SIR model is essentially the process by which an infected node infects its neighboring susceptible nodes and the self-immune. As for information diffusion in large-scale social networks, the switching of the model states is shown in Figure 2. The left side of the figure shows a simple case of a four-agent network. A1, A2, A3, and A4 stand for the agent ID. The letters in the circle represent the agent state. The states of A1 and A2 are I, which means that they have received the message and will forward the message. The A3 status is S, indicating that if he receives the message, he may turn to I. A4 is R, which means that he is immune to this message and does not care or will not forward it. Because A3 is a neighbor of A1, when A1 forwards the message, A3 will receive the message and transform to the I state according to a certain probability. Meanwhile, A1 will also automatically transform to the R state according to a specific probability. The state switching process is shown on the right side of Figure 2.

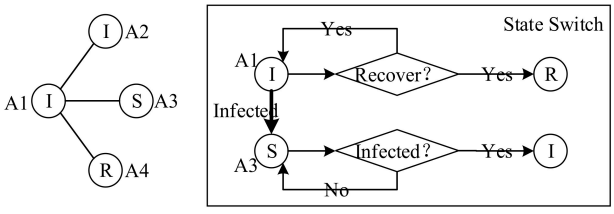


Figure 2. The state switch of the SIR model.

3.3. The Implementation of a Simulation Kernel in Spark

A simulation kernel is composed of the calculation of agents and the scheduling algorithm. As the simulation kernel is implemented in Spark, it is necessary to partition the social network reasonably and design the scheduling algorithm based on the Spark programming model. As a result, the Spark data structure, Resilient Distributed Data (RDD), is used to organize the large-scale social network, and the Spark programming model is also used to build the processing flow of the simulation kernel. Actually, the functions of the simulation kernel are implemented by a series of operations and transformations on the social network model stored in RDD in the cluster. Each computing node in the cluster performs local computing on a part of the network stored in the node and then aggregates to achieve parallel computation. Therefore, the design of a large-scale social network simulation kernel based on Spark can be divided into two parts: the design of the data structure and the design of the scheduling algorithm.

3.3.1. The Design of Data Structure

Data in Spark is stored, transformed, and calculated in the form of RDD. The broadcast variables and accumulators to process global data are also designed to optimize the simulation performance. Therefore, in the design of the data structure of the simulation kernel, it is necessary to consider the design of the RDD involved in the simulation process firstly, then broadcast variables and accumulators are considered too.

RDD is divided into two types: <Key> and <Key, Value>, which are used to organize different types of data. In a large-scale social network simulation, the data includes nodes and edges in the network, messages generated during the simulation, and temporary variables involved in the calculation process

of the model. The main data structures of the simulation kernel and their descriptions are shown in Table 1.

Table 1. The sesign of the data structure.

Name	Stored Information	Data Type	Description
AgentStateRDD	Agent state	<Key,Value>RDD	Record the state change of agent
AgentLinkListRDD	Neighboring Agents of Agent	<Key,Value>RDD	Record the neighboring agents list of agent
MsgRDD	Message	<Key,Value>RDD	Record the messages list received by agent
InfectedAgentBroadcast	Infected Agent	Broadcast	Record infected agents
InfectedAgentRDD	Infected Agent	<Key,Value>RDD	Record infected agents

The descriptions of the data structures are detailed below.

(1) AgentStateRDD

AgentStateRDD records the state information of Agents. The RDD is in the form of <Key, Value>. Key is the agent number, and value is the state of the agent. As mentioned in Section 3.2, the agent state includes three types—S, I, and R—in the SIR model.

(2) AgentLinkListRDD

AgentLinkListRDD records the list of neighboring agents. The key of AgentLinkListRDD is the agent number, and value is the list of neighboring agents. In the existing large-scale social network research, the links in the network are mostly stored in the form of edges. However, the storage of edges has certain problems. In large-scale social networks, agents have many connected edges, and each node corresponds to many neighboring agents. For example, young users of WeChat have an average of 128 friends. The edges corresponding to each user have an average of 128 pieces of data, and the start number will be recorded 128 times. In Sina Weibo, some verified influencers may have tens of millions of fans. Managing edge information in this form will inevitably bring data redundancy. In addition, in the simulation agents will send information to neighboring agents as needed, which requires frequent iterative access to the agent edge information.

Based on the above problems, the links between agents are stored in the form of the neighboring agents list of the agent, AgentLinkListRDD. All neighboring agents can be easily obtained by the start number in AgentLinkListRDD, which can greatly improve the efficiency of the simulation operation.

(3) MsgRDD

MsgRDD records all messages generated during the simulation. Each message represents an infection event between two agents. In the event, the target agent is the agent that receives the message, and the message needs to be passed to the target agent to update the state of the target agent. Therefore, set the key as the target agent number, and the content of the message and other information are recorded in the value.

(4) InfectedAgentBroadcast and InfectedAgentRDD

Both InfectedAgentBroadcast and InfectedAgentRDD are the data structures used to record infected agents during the simulation. In the SIR model, each infected agent sends messages to its neighboring agents to infect them. This operation needs to link the infected agent and AgentLinkListRDD. The linking is really time-consuming when the scale of information diffusion is large. At first, the InfectedAgentRDD is designed in the form of <Key, Value>, where key is the agent number. In Spark, the InfectedAgentRDD and the AgentLinkListRDD are connected through the same key to obtain MsgRDD. However, it was actually found that in the Spark programming model, the join operation needed to cache a large amount of data and the operation efficiency was low. Furthermore, it was found that in the initial stage of information diffusion on social networks, the number of infected agents was small. Performing a join operation, in this case, led to unnecessary computation, so the two data structures, InfectedAgentBroadcast and InfectedAgentRDD, were designed to manage the infected Agents together. When a large number of agents were infected, InfectedAgentRDD was used to connect with AgentLinkListRDD, and when the number of agents was not large, the broadcast

variable InfectedAgentBroadcast was set in each computing node. Thus, the communications were reduced greatly during the simulation.

3.3.2. Scheduling Mechanism

The design of the data structure basically includes the management of data and variables involved in the simulation process. However, for a large-scale social network simulation based on Spark, the scheduling mechanism needs to be converted into the operations of RDD. In the diffusion of information in large-scale social networks, messages are sent and received very frequently. The topology of large-scale social networks is also changing rapidly. Therefore, it is more efficient to use a time step-based method to schedule the simulation, and the changes of the network are observed within a certain time interval. The workflow of scheduling is shown in Figure 3. Because data is organized and managed in the form of RDD in Spark, an RDD operation task is allocated to computing nodes first, and the next operation is performed when the tasks on all computing nodes are completed. This mode implements the time synchronization of the agents in the simulation. On this basis, the large-scale social network simulation process can be divided according to time steps. In a simulation step, information is diffused based on the message passing and state updating of agents, which are actually the generations of MsgRDD and the update operations of AgentStateRDD.

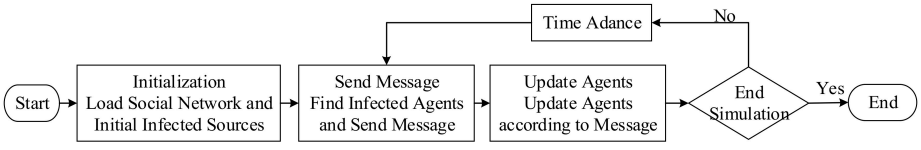


Figure 3. Simulation flow chart.

The simulation flow in each simulation step is represented as the RDD conversion flow frame shown in Figure 4.

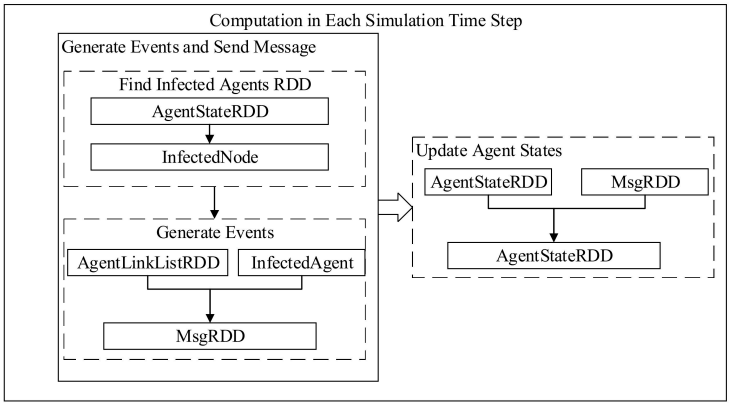


Figure 4. The computation in each simulation time step.

The corresponding simulation logic inside the dotted box in the figure is implemented in RDD operator, which is the core module in Spark programming. The reasonable design of the RDD operator can greatly reduce the execution time of the work and improve the performance of the simulation. The message-passing mechanism and state-updating mechanism in the RDD operator are detailed below.

The Design of the Message Passing Mechanism

Within each simulation step, each infected agent sends a message to its neighboring agent with a certain rate according to the parameter settings. This message-passing mechanism is implemented by the two RDDs: AgentStateRDD and AgentLinkListRDD. In addition, MsgRDD is generated by these two RDDs.

First, the filter operator is used to find the infected agents from AgentStateRDD according to the node state, then the infected agent is obtained. Then, the usage of the broadcast solution is determined by the number of nodes in the infected agent mentioned in Section 3.3.1. If necessary, broadcast the InfectedAgentBroadcast to each computing node. Next, each computing node obtains the local InfectedAgentBroadcast list and processes the AgentLinkListRDD through the filter operator according to the list and filters out all possible message source agents and their neighboring agents. If the broadcast is not required, connect InfectedAgentRDD and AgentLinkListRDD to obtain the possible message source agents and their neighboring agents. Then, according to the model parameters initially set, select the target agents and then generate MsgRDD through the FlatMap operation in Spark. Finally, the join operator is used to connect MsgRDD with AgentStateRDD, which means that the agent has received the messages. Then, merge the current state of all agents with the received messages; the messages are processed and the states of the agents are updated accordingly.

The Design of the State Update Mechanism

It can be seen from Figure 4 that the state switching in the agent model is to update the next state based on the current state with the received messages. The join operator in Spark is used to connect MsgRDD with AgentStateRDD, which means that the agents receive messages. Because the entire RDD is processed at the same time in Spark, the updating of the agent state is not operated during the message-receiving stage. Then, the ReduceByKey operator is used to aggregate all the messages received by a single agent and process messages according to the different agent state stored in AgentStateRDD. Agents with state I and R drop the messages. Agents with state I determine whether it is necessary to update the node state to R themselves. Agents with state S process the messages (infection events) and set the latest state according to the model parameters. Finally, all agents return the latest updated state to RDD. During the state update stage of the model, Spark's join operator is used to connect MsgRDD to AgentStateRDD for calculation. The join operator is a time-consuming operator because it involves matching the key values between different RDDs. The key values are matched one by one and then calculated. During the matching process, the partition of the RDDs also has a great impact on the calculation time.

When the connected agents are partitioned in the same computing node, the join operator is relatively simple, which means that the network communication and data passing are not necessary. When the connected agents are partitioned in different computing nodes, a large amount of communication is required. Data corresponding to the same key are should be transmitted together before calculation. According to the derived relationship of RDD, the parent RDD of MsgRDD is AgentLinkListRDD. Thus, the MsgRDD also follows the partitioning of AgentLinkListRDD—that is, the message data sent by the same computing node is processed in the same partition. MsgRDD is connected to AgentStateRDD in the message-passing mechanism. The partition of AgentStateRDD is determined by the agent number corresponding to the agent state. Obviously, MsgRDD and AgentStateRDD are not always partitioned together. Figure 5 uses two computing nodes as examples to illustrate the impact of whether MsgRDD and AgentStateRDD are partitioned together or not.

In the example shown in Figure 5, the cluster is composed of two computing nodes, which are represented by dashed boxes. The network involved in the calculation contains only two agents (A1, A2). The RDDs are allocated to the two computing nodes for calculation according to the different key values.

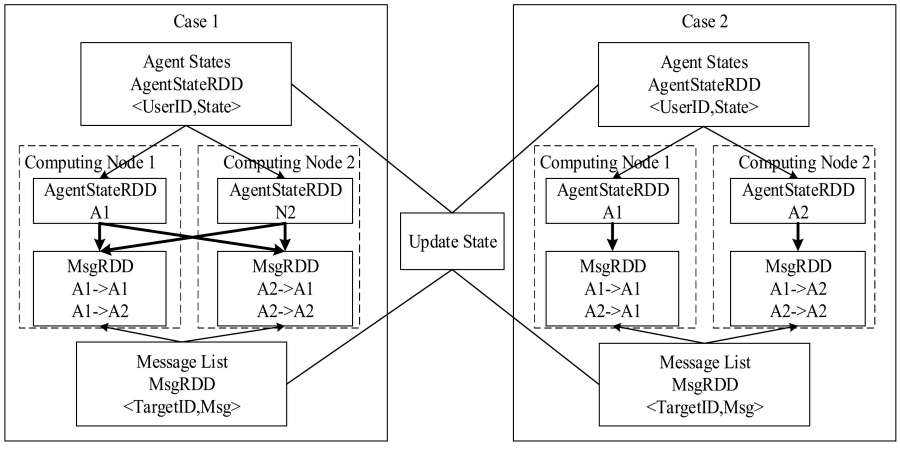


Figure 5. The sifferences in communication cost caused by the different partitions of MsgRDD and AgentStateRDD.

MsgRDD should be connected to AgentStateRDD in order to update AgentStateRDD. In these two cases, AgentStateRDD follows the partition of AgentLinkListRDD. Case 1 on the left indicates that MsgRDD follows the partition of AgentStateRDD according to the message sender. For example, MsgRDD (A1 > A2) is located in the same computing node with AgentStateRDD (A1). Case 2 on the right indicates that MsgRDD is re-partitioned according to the message receiver. For example, by being re-partitioned, MsgRDD(A1- > A2) is located in the same computing node as AgentStateRDD (A2). MsgRDD (A1- > A2) and AgentStateRDD (A2) can be combined together to update the AgentStateRDD of A2. In case 1, MsgRDD (A1- > A2) needs to combine with AgentStateRDD (A2) by communication, while this is not needed in case 2. Thus, when the states of the agents are updated, the data communications (indicated by bold arrows in the figure) for the two cases are quite different.

In case 1, since the message received by A1 is distributed to two computing nodes respectively, in addition to the local calculation, the message received by A1 on computing node 2 needs to be passed to computing node 1 to perform the calculation. Similarly, the message received by A2 on computing node 1 needs to be passed to computing node 2. When the number of agents and computing nodes are not large, the data communications are acceptable. However, when the size of the network reaches 1 billion and the number of computing nodes is hundreds, a large number of communications among different computing nodes in the cluster are needed and it becomes time-consuming. In case 2, because the state and the received messages of the same agent can be assigned to the same computing node, no other data passing is required at this step. Thus, compared to case 1, the performance of the simulation is improved. As a result, MsgRDD is re-partitioned before connecting MsgRDD and AgentStateRDD in the state update mechanism of agents in order to improve the performance of the simulation.

As mentioned before, the partition is a key factor in the performance of the large-scale social network simulation. The performance differs greatly in the different partitions. As a result, the two-tier optimization based on the network structure is described in detail in the next section.

4. Optimization

Compared with the large data analysis, the large number of interactions between agents will inevitably bring huge communication consumption between computing nodes in a large-scale social network simulation. Spark is a memory-based distributed computing framework, and the performance of Spark is closely related to the usage efficiency of the distributed clusters. The cost of network communication and the load balance of computing nodes are the two key points in the performance of

the simulation. Minimizing communication consumption while ensuring load balance can greatly improve the performance.

4.1. Figures, Tables, and Schemes

According to the partitioning strategy, the simulation kernel designed in this paper improves the performance of simulation by distributing the social network to different computing nodes in the cluster. As mentioned before, almost all the calculation of the simulation kernel is composed of message passing and agent state updating in each simulation step. During the message passing, the interconnected network agents produce message transmissions. The message passing becomes the communication among computing nodes if the source agent and target agent are partitioned in different computing nodes. The scale of communication changes greatly in different partition strategies. An example of nine agents is used in Figure 6 to illustrate how communications are determined by the partition strategy. The social network in Figure 6 contains nine network agents, A0 to A8. The directional arrows between the nodes represent the connection relationship between the agents (friend relationship in the social network). The agents are partitioned in two RDDs, and Hash partitioning is used based on the agent number.

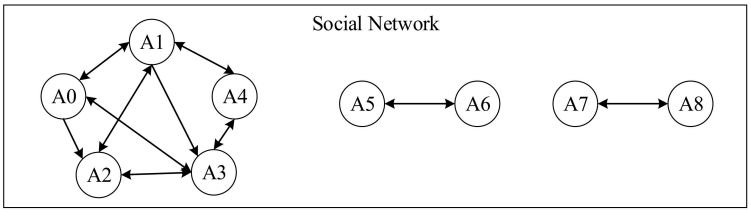


Figure 6. The structure of the social network.

Assuming that all agents in the network can send messages to their next Aaent, Figure 7 gives the data flow diagram in the message passing.

According to the partitioning strategy, Partition0 processes agent A0, A2, A4, A6, and A8, and Partition2 processes Agent A1, A3, A5, and A7. Then, this process is divided into the map stage and the reduce stage. The map stage traverses the AgentLinkList to generate the <Key, Value> of the target and source agents. Then, the output of the map is sorted by key after shuffle and sent to the computing nodes of reduce. The reduce side also partitions according to the same partitioning strategy, combines the data transmitted by shuffle, and generates a list of messages received by each target agent in MsgRDD. The arrows in the shuffle stage in the figure represent the data-passing route. It can be seen that, according to the current partitioning strategy, in this case it will generate a maximum of 16 cross-partition communications during a simulation step. When the network scale increases greatly, the number of edges in the network may reach tens of billions. The communication consumption caused by an unreasonable partitioning strategy may make the simulation impossible. This is because the characteristics of the social network were not taken into account in the partitioning, and the closely connected agents were divided into different partitions.

If the partition strategy is modified and the closely connected network nodes are allocated to the same partition, the data flow diagram shown in Figure 8 is obtained. Compared with the Hash partition in Figure 7, it can be seen that the message passing during the simulation is concentrated inside the partition in the changed partition. Thus, in the social network simulation, as many messages as possible should be passed inside the same computing node. In a cluster or cloud environment, different computing nodes may be deployed in different physical locations and network environments. Therefore, the cost of communication across partitions is much higher than the cost of communication within a partition. Using the partitioning strategy of Figure 8 to concentrate communication within the partition will reduce the communication consumption greatly.

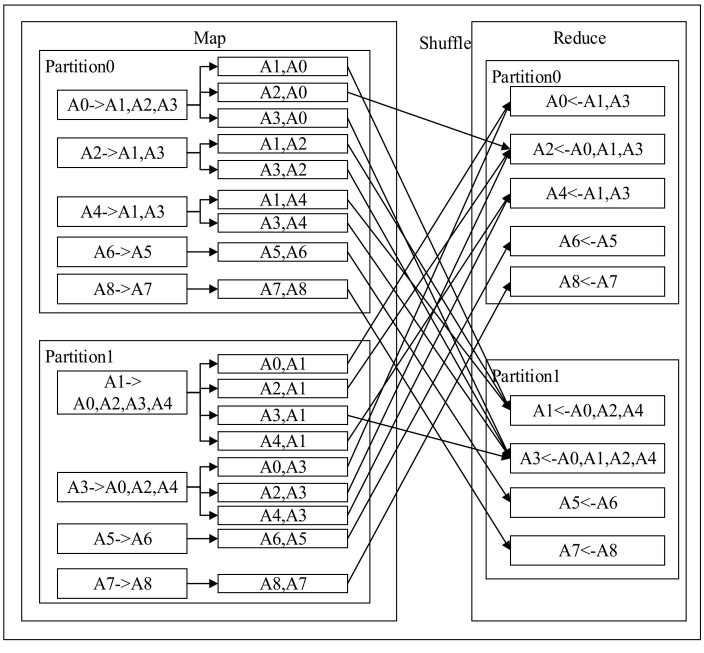


Figure 7. The data flow in Hash partition.

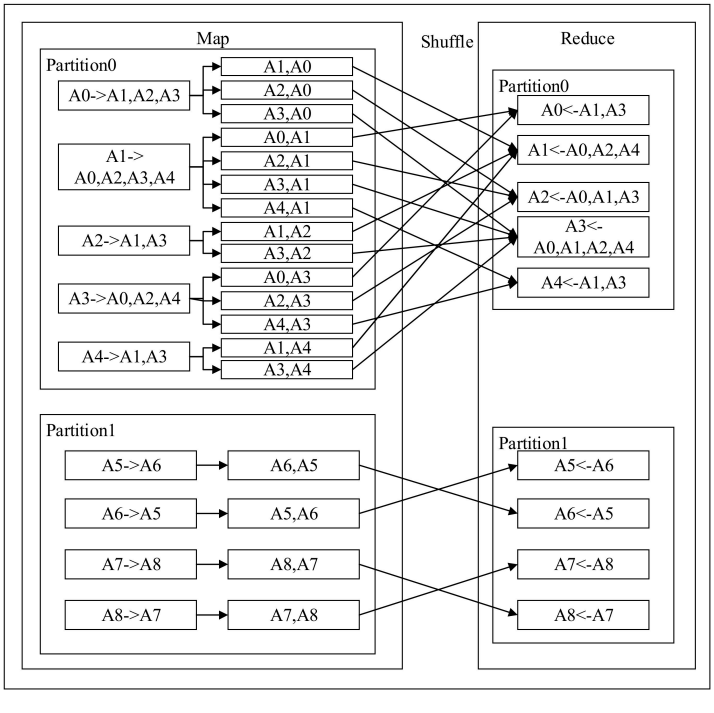


Figure 8. The data flow in community-based partition.

However, the actual social network structure is complex, and it is not easy to assign all the connected agents to the same partition. Therefore, the core goal of the partition algorithm is to allocate the closely related agents to the same partition, if possible. At the same time, load balance should also be considered for the reasonable calculation of partitions in the cluster. The partition based on network structure is detailed in the next section.

4.2. The Partition Based on Network Structure

4.2.1. The Principle of the Two-Tier Partition

An agent in a social network is the smallest calculating unit in a large-scale social network simulation. The calculation includes checking the agent state, sending messages to neighboring agents, and updating the agent state. Assume that the network is divided into K partitions, $\{P_1, P_2, \ldots P_K\}$. Each partition, P_i, needs to manage N_i agents $\left\{v_{i1}, v_{i2}, \ldots v_{ij}, \ldots, v_{iN_i}\right\}$. Each agent v_{ij} has w_{ij} neighbors. The neighbors of an agent can be divided into the same partition as v_{ij} and a different partition from v_{ij}, and their sizes are represented by $w_{ij}{}^{-}$ and $w_{ij}{}^{+}$, respectively.

For a large-scale social network simulation, the behavior of agents in the network is very simple. It can be considered that the computing load of each node when processing a message is the same, and the number of messages received by an agent is determined by the state and number of its neighboring agents. The state of the neighboring agents changes with simulation time, while the number of neighboring agents is always constant during the simulation. Statistically speaking, the number of messages received by a node is positively related to the number of its neighbors. Therefore, in each simulation step, the computational load L_{P_i} to be processed on a partition Pi is positively related to the sum of the neighboring agents of all agents in the partition. The communication between partitions is determined by the relationship and state of agents in different partitions. When an infected agent sends messages to its neighboring agents, a message communication between different partitions will be generated if the target agents are allocated in a different partition. Therefore, the total communication, C, between different partitions at a certain time is closely related to the current state of each agent in the partition and the state of its neighboring agents. C cannot be predicted in advance, but statistically, C is proportional to the number of connected edges across the different partitions. As a result, a partition algorithm for large-scale social networks is proposed in this paper to reduce the amount of communication, C, between partitions, based on ensuring the load, L_{P_i}, is balanced for all the partitions. In other words, under the constraint of load balance, an optimal network-cutting algorithm is found to minimize the network communication among computing nodes, and the network size of segmentation is N.

Consider the social network as a graph, the agents in the social network as the nodes in the graph, and the relationships between the nodes in the social network as the edges. The social network graph is divided into K sub-graphs of the same size; minimizing the number of the edges that are cut during the segmentation process is a typical application of the graph-cut algorithm [25]. However, the graph-cut problem is an NP(Non-deterministic Polynomial) problem. Although there have been many studies on graph cut, the current online social networks with a scale of over a billion is still a difficult problem. Partitioning directly based on the graph-cut algorithm is obviously time consuming, and cannot meet the requirements of performance improvement. At the same time, the goal of the graph-cut algorithm is to divide a given network into several parts of almost equal scale, while ignoring the structural characteristics of the graph and the similarity between nodes. It is not reasonable to directly use graph cut in the partition of the simulation of large-scale social networks.

Additionally, the community-detection algorithm is used to find nodes of similar structures in the network and then aggregate the nodes into communities. At the same time, in social networks the community shows the characteristics of close internal connection and sparse external connection [26,27]. Social networks can be divided into several independent sub-networks. The division of the network based on the community can not only preserve the characteristics of the network structure but also

reduce communication in the network. However, the result of the divisions in the community cannot guarantee load balance among the various communities.

Taking the advantages and disadvantages of the graph-cut algorithm and community-detection algorithm into consideration, a two-tier partition algorithm is proposed in this paper. The principle of the partition algorithm is shown in Figure 9. Through community detection, the large-scale social network is simplified into a network composed of several communities, which reduces the scale of the network. In addition, it ensures a reduction in communication while maintaining the characteristics of the network structure. Then, the graph-cut algorithm is used to partition based on the simplified network to ensure load balance among the partitions. Since the input of the graph cut is the simplified network, the efficiency of the graph cut will be greatly improved. Finally, a partition case can be obtained that ensures load balance while preserving the network structure characteristics and reducing network communications to a certain extent. The community detection and graph cut are detailed next, then the implementation of the two-tier partition algorithm is also illustrated in the following subsections.

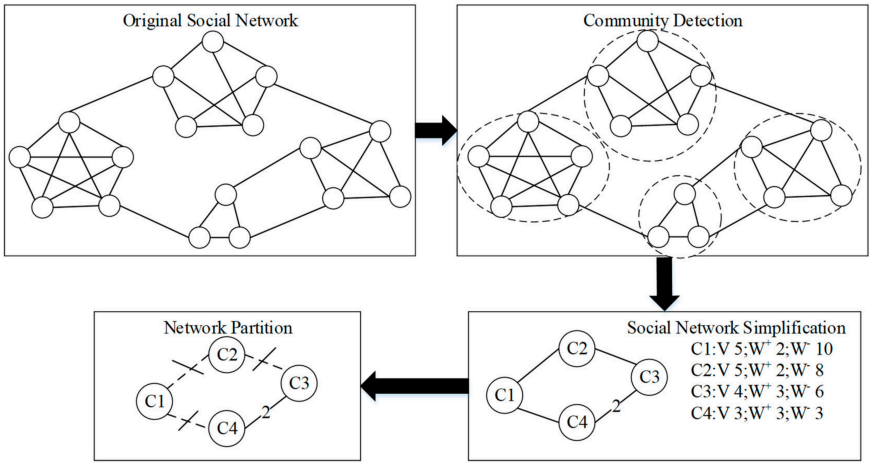


Figure 9. The principle of the two-tier partition algorithm.

4.2.2. Community Detection

Community is one of the most important characteristics of a complex network. Communities can be regarded as equal and independent sub-parts of the network [28], which represent user organizations with close internal connections and sparse external connections in a social network. In the process of network simplification, communities can be used to represent nodes in the simplified network. The community-detection algorithm is used to segment a large-scale social network into several communities, which ensures that the result of the segmentation is tightly connected internally and sparsely externally connected. Therefore, a relatively smaller number of edges are cut during the network simplification.

The community detection algorithms can be divided into non-overlapping community-discovery algorithms and overlapping community-discovery algorithms, which are used to detect whether there are overlapping nodes in the community in the network. The purpose of our work is to distribute the nodes into different partitions; each node only exists in one computing partition. Thus, non-overlapping community-discovery algorithms are selected in our work. The research on non-overlapping community discovery includes an optimization algorithm based on modularity, a community-discovery algorithm based on spectral analysis, a community-discovery algorithm based on label propagation, and so on. Girvan and Newman et al. proposed the GN(Girvan Newman)

algorithm [29]. In order to improve the GN algorithm, Newman proposed a fast algorithm which greatly reduced the time complexity of the GN algorithm [30]. Based on Newman's fast algorithm, Clauset et al. proposed the CNM (Clauset–Newman–Moore) algorithm [31], which used a heap data structure to store and calculate the modularity of the updated network. The above algorithms are community-discovery algorithms based on modularity which have a high computational complexity and cannot solve the problem of different community sizes in large-scale social networks. A community-discovery algorithm based on spectral analysis [32] uses the Laplace transformation to represent the graph and then uses the eigenvector of its Laplacian matrix on clusters to get communities. This algorithm also has the problem of high computational complexity when dealing with large-scale social networks. The label-propagation algorithm updates the state of nodes in the network through the propagation of several tags in the network and divides the communities according to the final state value of the nodes [33]. This algorithm is suitable for community detection in large-scale complex networks, and the computational complexity is low.

In the label-propagation algorithm, each node in the network is given a label value during initialization, and the label values are propagated continuously in the network according to the network structure. In each iteration, the label value of the node is updated by the largest number of label value owned by its neighboring nodes. In the iteration process, the nodes that are closely related will finally converge into groups with the same label value. According to the label value of the nodes, the network can be divided into corresponding communities. In the initialization stage, the label-propagation algorithm needs $O(n)$ time, where n is the number of nodes in the network. In the iterative process of label propagation, each step takes $O(m)$ time, where m is the number of nodes connected to this node. Compared with other community detection algorithms, the time complexity of the Fast-Newman algorithm is $O((m+n)n)$, while that of the CNM algorithm is $O(n\log^2 n)$. The performance of the label-propagation algorithm is excellent in the community detection of large-scale networks.

In this paper, the label-propagation algorithm is selected as the community-detection algorithm in the simplified network. M communities are obtained, $\{C_1, C_2, \ldots, C_i, \ldots, C_M\}$, and each community C_i contains N_i nodes. Then, a simplified large-scale social network is constructed based on the M communities. The simplified network consists of M nodes, $\{C_1, C_2, \ldots, C_i, \ldots, C_M\}$. The size of the node W_i is defined as the number of edges of all nodes within the community, and the edge between nodes e_{ij} is defined as the connection between the node C_i and node C_j.

$$W_i = \sum_{j=1}^{N_i} w_j \tag{2}$$

$$e_{ij} = \left\{(v_a, v_b) \middle| \forall (v_a, v_b), v_a \in C_i, v_b \in C_j\right\} \tag{3}$$

4.2.3. Graph Cut

The graph-cut algorithm is a common method to deal with load balance. The algorithm cuts a graph into sub-graphs of the same size with the specified number and minimizes the number of edges removed in cutting. For load balance, the multi-path cut algorithm [34] proposed by Kernighan and Lin is a kind of graph-cut algorithm. In the multi-path cut algorithm, a graph G with n nodes is defined; the size of each node in the graph is represented by w_i. For the given parameter K, the graph G is cut into K parts of the same size. The size of each part is the sum of the sizes of the nodes contained in the part. Based on the large-scale social network simplified by community detection, the goal of this paper is to cut the network into K partitions with the same load, ensuring that the communications among the partitions are as small as possible.

In the classical multi-path cut algorithm, a node with the size of p is regarded as a community composed of p nodes with size 1, and edges with high weight are added to this community to ensure

that the community will not be split in the cutting. However, it cannot guarantee that the communities are not cut in the case of communities with many kinds of sizes in this paper. At the same time, the classical multi-path cut algorithm needs to iterate many times to find the global optimal solution when $K \geq 3$ and the time complexity of the algorithm is far greater than $O(N^2 \log_2 N)$. Considering the above problems, Cheol H. Lee et al. proposed an efficient multi-path cut algorithm. In this algorithm, the problems are transformed into the problem of seeking the maximum K times cutting of the graph. Then, the generic algorithm is used to solve the problem. The complexity of the improved algorithm is $O(k|V|^2)$, where $|V|$ is the number of nodes in the subpart of the graph.

In the two-tier partition algorithm proposed in this paper, the multi-path cut algorithm is used to cut the simplified network graph composed of communities. The simplified network G contains M nodes, $\{C_1, C_2, \ldots, C_i, \ldots, C_M\}$. Each node represents a community, and the community contains n_i nodes, $\left\{v_{i1}, v_{i2}, \ldots, v_{ij}, \ldots, v_{in_i}\right\}$. The size of the nodes corresponds to the load of the nodes, L_{C_i}. The simplified network is divided into K partitions,$\{P_1, P_2, \ldots P_K\}$. Each partition P_i contains S_i communities, and the corresponding partition load, L_{P_i}, is the sum of all the community loads in the partition.

$$L_{C_i} \propto \sum_{i=1}^{n_i} w_{ij} \tag{4}$$

$$L_{P_i} = \sum_{j=1}^{s_i} L_{C_{ij}} \tag{5}$$

According to Equations (3)–(5) can be obtained—that is, the load on each partition is directly proportional to the number of neighboring nodes of all the nodes in the communities distributed in the partition. L_{C_i} is defined as the size of the node in the simplified network, and the goal of the multi-path cut algorithm of the simplified network is to ensure that L_{P_i} is the same for each partition.

The edge in the simplified network is defined as e_{ij}, and the simplified network is cut into K partitions with a similar load by cutting edges with the least cost. ${E_i}^-$ and ${E_i}^+$ are used to represent the inner edges of partition, P_i, and the edges out of the P_i. The corresponding communication consumption C after cutting is as shown in Equation (8):

$${E_i}^- = \{e_{ab} | C_a \in P_i, C_b \in P_i\} \tag{6}$$

$${E_i}^+ = \{e_{ab} | C_a \in P_i, C_b \notin P_i\} \tag{7}$$

$$C = \beta \cdot \sum_{i=1}^{K} {E_i}^+. \tag{8}$$

As mentioned before, the graph cut is transformed into the finding of an optimal network partition algorithm under the constraint of load balance (Equation (10)) so as to minimize the communications between computing nodes (Equation (9)). The problem of cutting an original N agents social network is transformed into the problem of cutting a simplified M nodes network, and $M << N$.

$$\min NC = \beta \cdot \sum_{i=1}^{K} {E_i}^+ \tag{9}$$

$$s.t. \forall i \in (1 \ldots K), L_{P_i} \leq B_{\max} := (1 + \varepsilon) \frac{\sum_{i=1}^{K} L_{P_i}}{K} \tag{10}$$

As a result, the two-tier partition algorithm of the large-scale social network is divided into two stages. The first stage is the community detection based on the label-propagation algorithm, and

the second stage is the graph cut of the simplified network. Finally, the partition results are obtained. The details of the implementation of the two-tier partition algorithm are illustrated in the next section.

4.2.4. The Flow of Two-Tier Partition

The flow of the algorithm consists of three steps: community detection, network simplification, and network cutting. The pseudocode is shown in Table 2.

Table 2. The pseudocode of the two-tier partition algorithm.

Algorithm the Two-tier Partition
Input: Adjacent set $N = \{n_1, n_2, \ldots, n_n\}$; n_i represents the neighbor list of vertex i. K, K represents the number of partitions.
Output: $P = \{p_1, p_2, \ldots, p_n\}$, p_i represents the index of located partition for vertex i.
1. $C = \{C_1, C_2, \ldots C_n\} \leftarrow$ find community of all vertexes with LPA(N)
2. for each community do
3. statistic compute load of all vertexes in this community as cl_i
4. end
5. $CL = \{cl_1, cl_2, \ldots cl_M\}$
6. for each edge in N do
7. construct community relation of source community and target community from CL
8. end
9. $CR \leftarrow$ statistic community relations
10. $SN(VC, EC) \leftarrow$ construct simplify network with CL and CR
11. $P \leftarrow$GraphPartition(SN,K)

First, communities are obtained by the label-propagation algorithm. Second, the load on each community and the connections between the communities are calculated, and a simplified network is obtained. Third, the multi-path cut algorithm is used to cut the simplified network, then the partitions of the social network are output.

The time complexity of the label propagation algorithm is $O(m + n)$, where m is the number of edges and n is the number of agents. The time complexity of the simplification is $O(mM)$. The time complexity of the cutting is $O(KM^2)$, where $M << n$. Compared with the time complexity of direct graph cut $O(Kn^2)$, the calculation of the two-tier partition algorithm is reduced greatly.

5. Experiments

In this section, two groups of experiments are designed to testify the improvement of the optimization of the simulation kernel based on the two-tier partition algorithm.

5.1. Experiments of Performance in Different Scales

In order to analyze the performance of the simulation kernel, a network following the power-law distribution is generated to test the proposed algorithm. The size of the network increases from 1000 to 1,000,000, and the minimum degree of the node is 100, while the minimum degree of the 100-node network is 99. The size of the nodes and the number of edges in the networks are listed in Table 3. The experiments are executed in a workstation with a CPU of Intel Xeon 24 core processor; the frequency of each core is 2.60 G and the size of the memory is 128 G.

Table 3. The scale of the synthesized social network.

Order of Magnitude of Network Size	Hundred	Thousand	Ten Thousand	Hundred Thousand	Million
The number of nodes	100	1000	10,000	100,000	1,000,000
The number of edges	1143	10,739	111,392	112,0837	11,183,970

Due to the limited network scale that traditional simulation tools and platforms can support, the large-scale social network information diffusion simulation framework based on Pregel designed

by Chuan Ai et al. [18] is selected for the comparison with the simulation kernel implemented in this paper. As described before, the SIR model is used as an example in the experiments. The initial infection number is set as 10, the infection probability is set as 0.05, and the recovery probability is set as 0.02. The experiments under two simulation frameworks are executed 50 times and the mean values are shown in Figure 10.

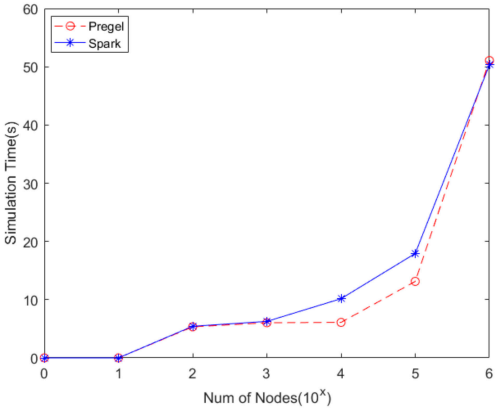


Figure 10. The comparison of results between the Spark-based simulation of information siffusion and the Pregel-based simulation.

As can be seen from Figure 10, there are almost no differences in performance between the two simulation frameworks. However, because of the tight encapsulation of Pregel, the simulation framework based on Pregel cannot perform statistical analysis of the temporary data generated by each simulation step in the simulation process. Actually, in order to analyze the mechanism of the diffusion process, the changes of all agents and edges should be recorded. Although the time consumption for the recording of temporary data is counted, the simulation kernel based on Spark designed in this paper still achieves the same level of computing performance as Pregel. In each simulation step, the infections of agents are recorded and output along with the simulation.

The simulation time recorded in the experiment includes the statistics of the time of the number of infected and recovered agents in each step. Figure 11 shows a sample of the statistical results of the number of infected and recovered agents in the experiments. In the case of similar performance, the simulation kernel proposed in this paper is flexible in programming and the real-time data statistics are supported, so time is saved in the results analysis.

In the large-scale social network simulation application, the real-time statistics of the data in the simulation process is of great significance. For example, when the information diffusion simulation in the large-scale social network is used to support decisions about rumor management, different rumor management strategies could be simulated in a short time. Through the real-time data statistics, the effect of rumor management policies can be observed clearly during the simulation. The most optimal policy can be selected for decision-making as soon as possible.

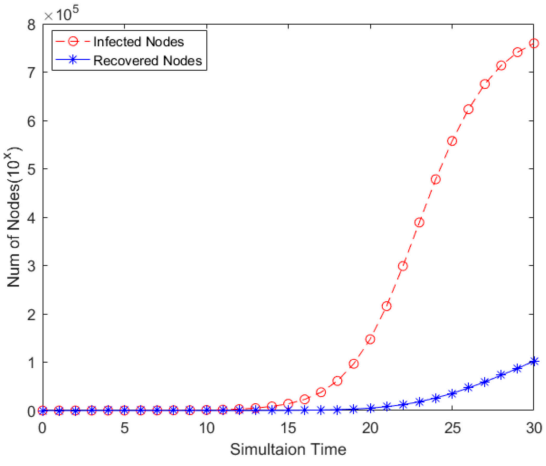


Figure 11. The count of infected and recovered agents.

5.2. Experiments of Two-Tier Partition

In order to verify the performance improvement of the two-tier partition algorithm, the empirical data from SNAP (Stanford Network Analysis Project) are used to design a series of experiments. The dataset from Twitter records a large-scale social network which has 81,306 users and 1,768,149 directed edges. Four different partition algorithms are designed to test the performance of the algorithms: (1) partition with Hash algorithm (Hash Partition); (2) partition with multi-path cut algorithm (Graph Partition); (3) simplify the network with the label propagation algorithm, and then partition the simplified network with the Hash algorithm (Community Hash); (4) simplify the network with the label propagation algorithm, and then partition the simplified network with a = the multi-path cut algorithm (Community Graph).

The parameters of the SIR model in agents are set in the experiments. The infection rate is set as 0.001, while the recovery probability is set as 0.01. A group of 10 agents is randomly selected as the initial infection sources to simulate the information diffusion in the social network from Twitter. Fifty simulation experiments are carried out for each partition algorithm, and the communications and loads of the partitions in the simulation are compared.

Figure 12 shows the comparison between the cross-partition communications and inter-partition communications of four different partition algorithms. In the figure, the solid line represents the cross-partition communication, the dotted line represents inter-partition communication, and the top left to bottom right corresponds the Hash Partition, Graph Partition, Community Hash, and Community Graph. It can be seen from the figure that in the partition mode of Hash Partition and Graph Partition, the cross-partition communications exceed the inter partition communications. However, when the community-detection algorithm is used to simplify the network, the inter-partition communications are obviously more numerous than cross-partition communications. This shows that community detection reduces communication across partitions greatly.

Figure 13 shows the comparisons of the communications of four different partition algorithms. The number of cross-partition communications is given at the top of the figure, while the number of communications between the partitions is given at the bottom. It can be seen that in the Hash Partition, the amount of cross-partition communications is the most, the amount of inter-partition communications is the least, and the communication consumptions are also the largest. Compared with the Hash Partition, the cross-partition communications of Graph Partition are reduced, but it still accounts for a higher proportion in the total communication. Although the Graph Partition cuts as few edges as possible, due to the need for the load balance of partitions, the consumptions are still large. It can be

seen that the simplification of the network by using the community-detection algorithm significantly reduces cross-partition communications. In the latter two partition algorithms, the inner-partition communications are far greater than the cross-partition communications.

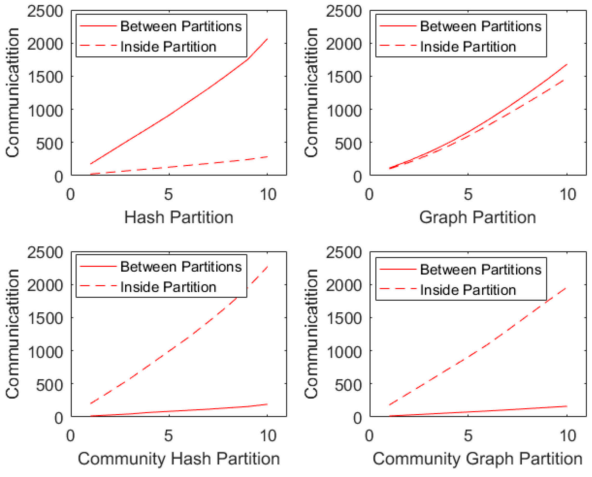


Figure 12. The communications in four partitions.

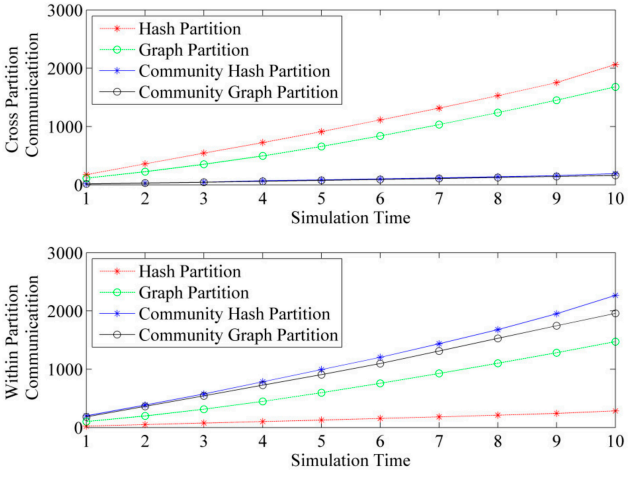


Figure 13. The comparisons of the communication of four partitions.

Due to the influences of random factors in the simulation, the amount of communication in various partition algorithms is different. For further comparison, the proportion of cross-partition communications in the total communication of various partition algorithms is listed in Table 4. It can be seen that the Graph Partition reduces the proportion of cross-partition communication by about 35% compared with the Hash Partition. However, the load balance should be considered in the Graph Partition. After using the community-detection algorithm, the community Hash and community both perform well in reducing cross-partition communications, which are less than 10% of the total communications. Because community detections are usually achieved before the simulation, the performance of latter partition algorithms is influenced.

Table 4. The communications in four partitions.

Partition Algorithm	Hash Partition	Graph Partition	Community Hash	Community Graph
The proportion of cross-partition communication	87.9%	53.3%	7.8%	7.6%
The proportion of inter-partition communication	12.1%	46.7%	92.2%	92.4%

The load of different partitions, LP, is recorded in the simulation. Figure 14 shows the comparisons of the partition loads of four different algorithms. It can be seen that the number of agents allocated to each partition after the Hash Partition is almost the same. Although the number of the neighboring agents of each agent is different, when the scale of the network is large, the Hash Partition can still acquire a good load balance. The Graph Partition can achieve better results by considering the load balance while reducing communications. When community detection is applied, the sizes of the communities located in different partitions are different. This is not good for the load balance of partitions. The Community Hash of the simplified network brings an obviously unbalanced load. The Community Graph improves this to a certain extent.

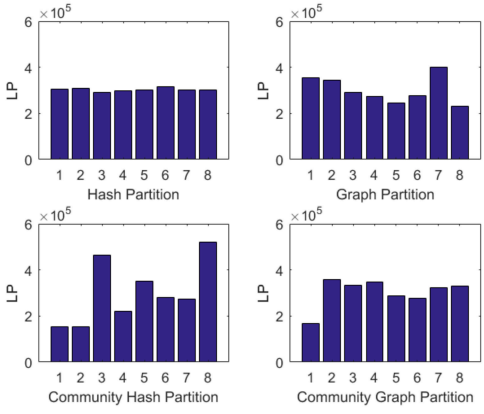


Figure 14. The comparisons of load in four partitions.

Figure 15 shows the load imbalance of the partitions of the four partition algorithms. It can be seen that the Hash Partition is the best in terms of load balance. The Community Graph gives a relatively better performance compared to Graph Partition and Community Hash.

Based on the analysis of the experimental results of the four partition algorithms, their features can be concluded below. Hash Partition is easy to use because the algorithm is embedded inside Spark. The load balance of this partition is good, but the network communications are huge in the simulation. Agents with frequent interactions in the social network are randomly assigned to different partitions. The cross-partition communications are much larger than the inter-partition communications. When different computing nodes in the cluster are located in different network locations, the communication consumption cannot be afforded.

Graph Partition reduces cross-partition communications to a certain extent, and the performance is not bad in load balance. However, when the scale of the social network is large, the graph-cut algorithm has a high time complexity, and the time computation of cutting cannot be accepted.

The application of community-detection algorithms reduces the cross-partition communications significantly. Almost 90% of the communication is within the partitions in the simulation, which reduces the communication consumption greatly. However, how the simplified network is allocated to different partitions has a great influence on the load balance. Due to the various sizes of the communities, Hash Partition brings a great imbalance, followed by large quantities of waiting time in the calculation.

The proposed two-tier partition algorithm of low time complexity not only effectively reduces the cross-partition communications, but also guarantees load balance between the partitions. In addition, according to the characteristics of the network, the calculation of edge weights in the graph-cut algorithm can be customized to achieve load balance in different social networks.

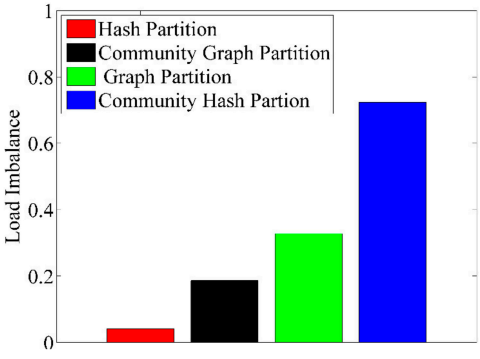


Figure 15. The imbalances of load in four partitions.

6. Conclusions and Future Work

The large-scale social network simulation is an important means to study the topology characteristics, evolution rules, and information diffusion characteristics of large-scale social networks. ABMS is a "bottom-up" modeling and simulation paradigm. Using ABMS to simulate large-scale social networks can model individuals to evolve the overall characteristics of social networks through the interaction between individuals. However, with the continuous development of the Internet, the scale of social networks has become larger and larger. Mainstream large-scale social networks even contain hundreds of millions of network nodes, and each node has hundreds of neighboring nodes. In the process of ABMS, a large number of agents need to be modeled and simulated, and the interactions between agents are very complex. There exist many difficulties in data access, data processing, and simulation. However, if the whole social network is replaced by the sampled network or synthesized network, the accuracy of simulation cannot be guaranteed. Therefore, this paper proposes a large-scale social network simulation framework based on Spark, which improves the performance of simulation using two aspects. First, through the Spark programming model, a large-scale social network simulation kernel based on Spark is implemented, which greatly expands the processing capacity and computing performance. Second, according to the characteristics of huge communications in large-scale social network simulation and based on the structure of social networks, a two-tier partition algorithm is designed. This optimizes the distribution of large-scale social networks in the computing cluster and reduces cross-partition communications during the simulation, thus reducing the network communication consumption. Finally, the performance of the two-tier partition algorithm is verified by the experiments of the empirical dataset from Twitter.

Author Contributions: Conceptualization, B.C., H.C., and D.N.; methodology, B.C. and H.C.; software, H.C., C.A., and D.N.; validation, B.C., M.Z., X.Q., and W.D.; formal analysis, B.C. and D.N.; investigation, D.N. and C.A.; writing—original draft preparation, B.C. and H.C.; writing—review and editing, B.C. and D.N.; visualization, D.N. and H.C.; supervision, B.C.; project administration, B.C.; funding acquisition, B.C. B.C. and H.C. contributed to the work equally and should be regarded as co-first authors. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by the National Key Research and Development (R&D) Plan [grant numbers 2018YFC0806900]; the National Natural Science Foundation of China [grant number 71673292, 21808181, 61673388, 71673294]; and the National Social Science Foundation of China [grant number 17CGL047].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wechat 2018. In *Wechat Data Report 2018*; Wechat: Singapore, 2019; Available online: https://support.weixin.qq.com/cgi-bin/mmsupport-bin/getopendays (accessed on 19 May 2020).
2. Fang, B.; Jia, Y.; Han, Y. Core scientific issues of social network analysis. research status and future prospects. *CAS Bull.* **2015**, *2*, 187–199.
3. Liu, W.; Sidhu, A.; Beacom, A.M.; Valente, T. Social Network Theory. In *The International Encyclopedia of Media Effects*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017.
4. Fang, B. *Online Social Network Analysis*; Electronic Industry Press: Beijing, China, 2014.
5. Zhang, J.; Li, X. Artificial society -sociological simulation based on Agent. *Syst. Eng.* **2005**, *23*, 13–20.
6. Sklar, E. Software Review: NetLogo, a Multi-Agent Simulation Environment. *Artif. Life* **2007**, *13*, 303–311. [CrossRef]
7. Minar, N.; Burkhart, R.; Langton, C.; Askenazy, M. *The Swarm Simulation System–A Toolkit for Building Multi-agent Simulations*; Santa Fe Institute: Santa Fe, NM, USA, 1996.
8. North, M.J.; Collier, N.; Vos, J.R. Experiences creating three implementations of the repast agent modeling toolkit. *ACM Trans. Model. Comput. Simul.* **2006**, *16*, 1–25. [CrossRef]
9. Railsback, S.F.; Lytinen, S.; Jackson, S.K. Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation* **2006**, *82*, 609–623. [CrossRef]
10. Zhen, L.; Gang, G.; Bin, C.; Xiaogang, Q. Accelerating Large Scale Artificial Society Simulation with CPU/GPU Based Heterogeneous Parallel Method. In Proceedings of the IEEE/ACM International Symposium on Distributed Simulation & Real Time Applications, London, UK, 21–23 September 2016.
11. Walker, S.J. Big Data: A Revolution That Will Transform How We Live, Work, and Think. *Int. J. Advert.* **2014**, *33*, 181–183. [CrossRef]
12. Google. Sorting 1 PB with MapReduce. Available online: http://googleblog.blogspot.com/2008/11/sorting-1pb-with-mapreduce.html,20081111 (accessed on 21 November 2019).
13. Matei, Z.; Mosharaf, C.; Franklin, J.; Shenker, M.; Ion, S. Spark: Cluster Computing with Working Sets. In Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, Berkeley, CA, USA, 22–25 June 2010.
14. Charlse, M.; Michael, J. agent-based modeling and simulation. In Proceedings of the 2009 Winter Simulation Conference, Austin, TX, USA, 13 December 2009; pp. 86–98.
15. Allan, R. *Survey of Agent Based Modelling and Simulation Tools*; Science and Technology Facilities Council: Washington, DC, USA, 2010.
16. Mittal, A.; Krejci, C.C.; Dorneich, M.C. An agent-based approach to designing residential renewable energy systems. *Renew. Sustain. Energy Rev.* **2019**, *112*, 1008–1020. [CrossRef]
17. Griffin, F.A.; Stanish, C. An Agent-Based Model of Prehistoric Settlement Patterns and Political Consolidation in the Lake Titicaca Basin of Peru and Bolivia. *Struct. Dyn.* **2007**, *2*, 2.
18. Testa, J.W.; Mock, K.J.; Taylor, C.; Koyuk, H.; Coyle, J.R.; Waggoner, R. Agent-based modeling of the dynamics of mammal-eating killer whales and their prey. *Mar. Ecol. Prog. Ser.* **2012**, *466*, 275–291. [CrossRef]
19. Lee, K.; Park, J. A Hadoop-Based Output Analyzer for Large-Scale Simulation Data. In Proceedings of the The Fourth IEEE International Conference on Big Data & Cloud Computing, Sydney, Australia, 3 December 2014.
20. Sethia, P.; Karlapalem, K. A multi-agent simulation framework on small Hadoop cluster. *Eng. Appl. Artif. Intell.* **2011**, *24*, 1120–1127. [CrossRef]
21. Kärger, J.; Ruthven, D.M.; Theodorou, D.N. Molecular Dynamics Simulations. In *Diffusion in Nanoporous Materials*; Wiley: Hoboken, NJ, USA, 2012; pp. 227–273.
22. Youbo, L.; Yang, L.; Junyong, L.; Yong, L.; Jianting, L.; Su, D. Hadoop Based Distributed Computing Framework for Large-scale Cascading Failure Simulation and Analysis of Power System. *Autom. Electr. Power Syst.* **2016**, *40*, 90–97.
23. He, L.; Fen, D.; Wang, R.; Su, P.-R.; Ying, L.-Y. Worm simulation of large-scale online social network based on MapReduce. *J. Softw.* **2013**, *7*, 1666–1682. [CrossRef]
24. Chuan, A.; Bin, C.; Liang, L.; Jian, J.; He, L.; Lai, K.; Qiu, X. Design and implementation of large-scale network propagation simulation method inspired by Pregel mechanism. *Sci. Sin. Inf.* **2018**, *48*, 932–946.
25. Kang, L.; Zhang, X.; Li, F.; Tian, Y. Graph partitioning method for social networks based on communication load balancing. *Comput. Eng. Appl.* **2018**, *54*, 66–71.

26. Clauset, A.; Newman, M.E.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111. [CrossRef]
27. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440–442. [CrossRef]
28. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2009**, *486*, 75–174. [CrossRef]
29. Girvan, M.; Newman, M.E.J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821–7826. [CrossRef]
30. Newman, M.E.J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **2004**, *69*, 066133. [CrossRef]
31. Clauset, A. Finding local community structure in network. *Phys. Rev. E* **2005**, *72*, 1–7. [CrossRef]
32. Fieldler, M. Algebraic connectivity of graphs. *Czechoslov. Math. J.* **1973**, *23*, 298–305.
33. Raghavan, U.N.; Albert, R.; Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **2007**, *76*, 036106. [CrossRef] [PubMed]
34. Kernighan, B.W.; Lin, S. An Efficient Heuristic Procedure for Partitioning Graphs. *Bell Syst. Tech. J.* **1970**, *49*, 291–307. [CrossRef]

symmetry

MDPI

Article

A New LSB Attack on Special-Structured RSA Primes

Amir Hamzah Abd Ghafar [1], Muhammad Rezal Kamel Ariffin [1,2,*] and Muhammad Asyraf Asbullah [1,3]

[1] Institute for Mathematical Research, Universiti Putra Malaysia, Serdang 43400, Selangor Darul Ehsan, Malaysia; amirghafar87@gmail.com (A.H.A.G.); ma_asyraf@upm.edu.my (M.A.A.)
[2] Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Serdang 43400, Selangor Darul Ehsan, Malaysia
[3] Centre of Foundation Studies for Agricultural Science, Universiti Putra Malaysia, Serdang 43400, Selangor Darul Ehsan, Malaysia
* Correspondence: rezal@upm.edu.my

Received: 13 February 2020; Accepted: 17 March 2020; Published: 20 May 2020

Abstract: Asymmetric key cryptosystem is a vital element in securing our communication in cyberspace. It encrypts our transmitting data and authenticates the originality and integrity of the data. The Rivest–Shamir–Adleman (RSA) cryptosystem is highly regarded as one of the most deployed public-key cryptosystem today. Previous attacks on the cryptosystem focus on the effort to weaken the hardness of integer factorization problem, embedded in the RSA modulus, $N = pq$. The adversary used several assumptions to enable the attacks. For examples, p and q which satisfy Pollard's weak primes structures and partial knowledge of least significant bits (LSBs) of p and q can cause N to be factored in polynomial time, thus breaking the security of RSA. In this paper, we heavily utilized both assumptions. First, we assume that p and q satisfy specific structures where $p = a^m + r_p$ and $q = b^m + r_q$ for a, b are positive integers and m is a positive even number. Second, we assume that the bits of r_p and r_q are the known LSBs of p and q respectively. In our analysis, we have successfully factored N in polynomial time using both assumptions. We also counted the number of primes that are affected by our attack. Based on the result, it may poses a great danger to the users of RSA if no countermeasure being developed to resist our attack.

Keywords: cryptography; RSA cryptosystem; RSA cryptanalysis; partial key exposure attack

1. Introduction

One of the earliest asymmetric key cryptosystems is the Rivest–Shamir–Adleman (RSA) cryptosystem, introduced by Rivest, Shamir and Adleman in 1978 [1]. Its simple and easy-to-understand mathematical design makes it compelling to be used in the early ages of digital cyberspace technology. Since then, it is considered as the most widely known asymmetric key cryptosystem. In its key generation algorithm, an RSA modulus, $N = pq$ is computed where p and q, called RSA primes are two distinct primes such that $p < q < 2p$. From the values of p and q, another parameter called RSA public exponent, e is obtained which satisfies $e < \phi(N)$ and $\gcd(e, \phi(N)) = 1$ where $\phi(N) = (p-1)(q-1)$. An RSA private exponent, d that satisfies $ed \equiv 1 \pmod{N}$ then is computed. One of the security strength of RSA is integer factorization problem and it is embedded in the RSA modulus since p and q are very large $n-$bit primes (typically, $n = 1024$). The problem is deemed infeasible to be solved by current computing machines and the best algorithm to solve the problem, called general number field sieve (GNFS) [2] is still running in sub-exponential time.

Past attacks on RSA by Pollard in 1974 [3] have shown that primes with particular structures are vulnerable to be factored in polynomial time, which is easily computed by any modern computers.

In his attacks, Pollard showed that if $p-1$ or $q-1$ are constituted of small primes, then there is a factoring algorithm to factor $N = pq$ in polynomial time. Another method in attacking RSA assumes that several bits of p and q are known by the adversary and this weakens the hardness of factoring N. Particularly, ref. [4] showed that $1/2$ least significant bits (LSBs) of the RSA primes are sufficient to factor N in polynomial time. Random reconstruction algorithm by Heninger and Shacham also showed that it can efficiently recover all of the RSA keys given 0.57 fraction of the random bits of each p and q [5]. Later, Maitra et al. [6] provided a combinatorial model of Heninger's work and was able to reconstruct the LSBs of RSA primes using modified brute-force by shortening the total search space.

The LSBs discussed in the prior attacks of RSA are commonly gathered by side-channel attack. It is one of the prominent methods to collect the physical outputs or side-effects of cryptographic devices during the computing processes [7]. The outputs or side-effects include but are not limited to the computational time and power of decryption [8,9], emission heat and electromagnetic radiation of the devices [10], cache behavior [11] and sound of processor during computations [12].

About This Paper

The results in this paper are the extensions from our papers in [13] and [14]. In this paper, we assume that certain LSBs of the RSA primes are known. We show that only a small amount of LSBs are required in our attack to factor N in polynomial time given that the RSA primes satisfy specified structures. We also show the abundance of primes that can satisfy the structures and no proper checking mechanism has been done in any standard RSA libraries to hinder the usage of such primes. This shows the risks inherent in the existing method to generate RSA keys may produces RSA modulus that falls under our attack.

2. Preliminaries

In this section, we provide some helpful lemmas which results are applied to make our attack successful.

Lemma 1. *Let $a, r \in \mathbb{Z}^+$ and $m \geq 2$ be an even number. If $\sqrt{a^m + r} = a^{m/2} + \epsilon$ then $\epsilon < \frac{r}{2a^{m/2}}$.*

Proof. Let $a^m + r$ be an integer where $a \in \mathbb{Z}^+$. Then

$$\sqrt{a^m + r} < \sqrt{a^m + \frac{r^2}{4}a^{-m} + r} = \sqrt{(a^{m/2} + \frac{r}{2}a^{-m/2})^2} = a^{m/2} + \frac{r}{2}a^{-m/2}$$

Since $\sqrt{a^m + r} = a^{m/2} + \epsilon$ then $\epsilon < \frac{r}{2a^{m/2}}$. This terminates the proof. □

Suppose $N = pq$ is a valid RSA modulus where $p = a^m + r_p$ and $q = b^m + r_q$. Let $a, b \in \mathbb{Z}^+$, we can see that ab is unknown if p and q are secret values. Using the result from Lemma 1, we find the lower and upper bounds of $N^{1/2} - (ab)^{m/2}$ in the following lemma.

Lemma 2. *Let $a, b \in \mathbb{Z}^+$ and $m \geq 2$ be an even number such that $a < b < (2a^m + 1)^{\frac{1}{m}}$. Suppose $N = (a^m + r_p)(b^m + r_q)$ where $r_p \leq r_q < N^\gamma$. If $r_p < 2a^{m/2}$ and $r_q < 2b^{m/2}$ then $(r_p r_q)^{1/2} < N^{1/2} - (ab)^{m/2} < \frac{r_q}{2} + 2^{\frac{m}{2}-1}r_p + 1$.*

Proof. To prove the lower bound, first we need to show that $a^m r_q + b^m r_p > 2(ab)^{m/2}(r_p r_q)^{1/2}$. Observe that

$$\left(a^{m/2}r_q^{1/2} - b^{m/2}r_p^{1/2}\right)^2 = a^m r_q + b^m r_p - 2(ab)^{m/2}(r_p r_q)^{1/2}.$$

Since $\left(a^{m/2}r_q^{1/2} - b^{m/2}r_p^{1/2}\right)^2$ will always be positive value, it implies that $a^m r_q + b^m r_p > 2(ab)^{m/2}(r_p r_q)^{1/2}$. Then

$$\begin{aligned}\sqrt{(a^m+r_p)(b^m+r_q)} &= \sqrt{(ab)^m+a^m r_q+b^m r_p+r_p r_q}\\ &> \sqrt{(ab)^m+2(ab)^{m/2}(r_p r_q)^{1/2}+r_p r_q}\\ &= \sqrt{\left(ab^{m/2}+(r_p r_q)^{1/2}\right)^2}\\ &= (ab)^{m/2}+(r_p r_q)^{1/2}\end{aligned}$$

Thus, $\sqrt{(a^m+r_p)(b^m+r_q)}-(ab)^{m/2}=N^{1/2}-(ab)^{m/2}>(r_p r_q)^{1/2}$. To prove the upper bound, since $\sqrt{a^m+r_p}=a^{m/2}+\epsilon_1$ and $\sqrt{b^m+r_q}=b^{m/2}+\epsilon_2$. Then, based on Lemma 1,

$$\begin{aligned}N^{1/2} &= \sqrt{(a^m+r_p)(b^m+r_q)}=\sqrt{(a^m+r_p)}\sqrt{(b^m+r_q)}\\ &= (a^{m/2}+\epsilon_1)(b^{m/2}+\epsilon_2)=(ab)^{m/2}+a^{m/2}\epsilon_2+b^{m/2}\epsilon_1+\epsilon_1\epsilon_2\\ &< (ab)^{m/2}+a^{m/2}\frac{r_q}{2b^{m/2}}+b^{m/2}\frac{r_p}{2a^{m/2}}+\frac{r_p}{2a^{m/2}}\frac{r_q}{2b^{m/2}} \qquad (1)\end{aligned}$$

If $r_p<2a^{m/2}$ and $r_q<2b^{m/2}$ then

$$\begin{aligned}\frac{r_p}{2a^{m/2}}\frac{r_q}{2b^{m/2}} &= \frac{r_p r_q}{4(ab)^{m/2}}<\frac{4(ab)^{m/2}}{4(ab)^{m/2}}\\ &= 1. \qquad (2)\end{aligned}$$

If $a<b<(2a^m+1)^{\frac{1}{m}}$, then Equation (1) becomes

$$\begin{aligned}N^{1/2}-(ab)^{m/2} &< a^{m/2}\frac{r_q}{2b^{m/2}}+b^{m/2}\frac{r_p}{2a^{m/2}}+1\\ &= \left(\frac{a}{b}\right)^{m/2}\frac{r_q}{2}+\left(\frac{b}{a}\right)^{m/2}\frac{r_p}{2}+1\\ &< (1)^{m/2}\frac{r_q}{2}+(2)^{m/2}\frac{r_p}{2}+1\\ &= \frac{r_q}{2}+2^{\frac{m}{2}-1}r_p+1.\end{aligned}$$

This terminates the proof. □

By obtaining the lower and upper bounds of $N^{1/2}-(ab)^{m/2}$ in Lemma 2, we have gathered a result that can be useful in our attack later. Throughout this paper, we focus on the RSA primes in the forms of $p=a^m+r_p$ and $q=b^m+r_q$. Therefore, we define LSBs in the next definition based on these forms.

Definition 1 (Least Significant Bits (LSBs) of Primes)**.** *Let $l_1,l_2,m\in\mathbb{Z}^+$. Suppose $p=a^m+r_p$ and $q=b^m+r_q$ are primes. Suppose there exist unknown a_0 and b_0 such that*

$$p=(2^{l_1}\cdot a_0)^m+r_p \qquad (3)$$

and

$$q=(2^{l_2}\cdot b_0)^m+r_q. \qquad (4)$$

Then we define r_p and r_q to be k-many LSBs of p and q respectively where $k\leq l_1 m, l_2 m$ satisfies

$$r_p\equiv p \pmod{2^{l_1 m}} \qquad (5)$$

and

$$r_q \equiv q \pmod{2^{l_2 m}}. \tag{6}$$

To identify primes that satisfy Equations (3) and (4), we observe the binary representations of a^m and b^m. Their LSBs must have k many consecutive 0's to satisfy $p = a^m + r_p$ and $q = b^m + r_q$. Particularly, let r_{p_i} be the binary representation of a and r_{q_i} be the binary representation of b where $i = 1, 2, \ldots, n$. Observe

$$a^m = \underbrace{r_{p_1} r_{p_2} \ldots r_{p_{(n-k)}}}_{n-k \text{ many bits of 1 and 0's}} \overbrace{r_{p_{(n-k+1)}} \ldots r_{p_n}}^{k \text{ many bits of 0's}} \tag{7}$$

$$b^m = \underbrace{r_{q_1} r_{q_2} \ldots r_{q_{(n-k)}}}_{n-k \text{ many bits of 1 and 0's}} \overbrace{r_{q_{(n-k+1)}} \ldots r_{q_n}}^{k \text{ many bits of 0's}} \tag{8}$$

The random reconstruction algorithm [5], which was improved by [6], is one of the efficient algorithms used to find the LSBs of RSA primes. Thus, it can be utilized to find the values of r_p and r_q that satisfy Equations (5) and (6).

3. Our Attack

Before we proceed to show how N can be factored in polynomial time using previous results, we define the term 'sufficiently small' that is used to justify our attack.

Definition 2. *We define* ***sufficiently small*** *value in this paper to be a value smaller than the largest feasible value of the lowest security level to be brute forced by current computing machine.*

Remark 1. *The latest recommendation for key management by NIST [15] stated that the lowest security level is 112-bit. This implies that the largest feasible value of this security level to be brute forced by current computing machine is* 2^{112}. *Based on Definition 2, a value lower than* 2^{112} *is considered sufficiently small. This value can be changed in the future, depends on the future advancements of computing technology.*

Now we are ready to show how RSA modulus can be factored in polynomial time by using this next theorem.

Theorem 1. *Let* $a, b \in \mathbb{Z}^+$ *and* $m \geq 2$ *be an even number such that* $a < b < (2a^m + 1)^{\frac{1}{m}}$. *Suppose* $N = pq = (a^m + r_p)(b^m + r_q)$ *is a valid RSA modulus. Let* $r_p \equiv p \pmod{2^m}$ *and* $r_q \equiv q \pmod{2^m}$ *where* $r_p < 2a^{m/2}$ *and* $r_q < 2b^{m/2}$ *such that* $\max\{r_p, r_q\} < 2^k$. *If* $2^{k-1}\left(2^{\frac{m}{2}} + 1\right)$ *is a sufficiently small value as defined in Definition 2 and k many LSBs of p and q are known then N can be factored in polynomial time.*

Proof. From Lemma 2 we can see that $(r_p r_q)^{1/2} < N^{1/2} - (ab)^{m/2} < \frac{r_q}{2} + 2^{\frac{m}{2}-1} r_p + 1$. Thus,

$$N^{1/2} - \left(\frac{r_q}{2} + 2^{\frac{m}{2}-1} r_p + 1\right) < (ab)^{m/2} < N^{1/2} - (r_p r_q)^{1/2}. \tag{9}$$

Suppose r_p and r_q are known LSBs of p and q respectively. The LSB values may be obtained from side-channel attacks described previously in Section 1. Since $\max\{r_p, r_q\} < 2^k$, then the difference between the upper and lower bounds of Equation (9) is

$$\begin{aligned} N^{1/2} - (r_p r_q)^{1/2} - N^{1/2} + \frac{r_q}{2} + 2^{\frac{m}{2}-1} r_p + 1 &< 2^k \left(2^{\frac{m}{2}-1} + \frac{1}{2} \right) - \left(\left(\min\{r_p, r_q\} \right)^2 \right)^{1/2} + 1 \\ &= 2^k \left(\frac{2^{\frac{m}{2}} + 1}{2} \right) - \min\{r_p, r_q\} + 1 \\ &= 2^{k-1} \left(2^{\frac{m}{2}} + 1 \right) - \min\{r_p, r_q\} + 1 \end{aligned} \tag{10}$$

which is the size for set of integers to find $(ab)^{m/2}$. If $2^{k-1}\left(2^{\frac{m}{2}}+1\right)$ is sufficiently small as defined in Definition 2, then we can find $(ab)^{m/2}$ in polynomial time. By computing $\left((ab)^{m/2}\right)^2$, we find $(ab)^m$. Then

$$\begin{aligned} N - r_p r_q &\equiv (a^m + r_p)(b^m + r_q) - r_p r_q \\ &\equiv (ab)^m + a^m r_q + b^m r_p \\ &\equiv a^m r_q + b^m r_p \pmod{(ab)^m}. \end{aligned}$$

Observe that from $r_p < 2a^{m/2}$ and $r_q < 2b^{m/2}$, then we can have $a^m r_q + b^m r_p < (ab)^m$. Thus, we obtain the full integer $a^m r_q + b^m r_p$ without modular reduction. Since the values of $r_p, r_q, (ab)^m$ and $a^m r_q + b^m r_p$ are known, we can find the roots of the following quadratic equation

$$X^2 - (a^m r_q + b^m r_p)X + ((ab)^m r_p r_q).$$

We find that $x_1 = a^m r_q$ and $x_2 = b^m r_p$. Since r_p and r_q are known, we can can obtain

$$a^m = \frac{x_1}{r_q} \quad \text{and} \quad b^m = \frac{x_2}{r_p}.$$

Thus we can factor N by calculating

$$\frac{N}{b^m + r_q} = a^m + r_p.$$

□

The next remark justifies our selection criteria on parameter m.

Remark 2. *Let $\mathbb{A}$ be the set of possible value of $(ab)^{m/2}$. From Equation (9), we know that $\mathbb{A}$ will yield a set of numbers between $N^{1/2} - \left(\frac{r_q}{2} + 2^{\frac{m}{2}-1} r_p + 1\right)$ and $N^{1/2} - (r_p r_q)^{1/2}$. If $m \geq 2$ is an even integer, then $(ab)^{m/2}$ will be an integer and causes $\mathbb{A}$ to be a finite set. However, if m is a positive odd integer, then $(ab)^{m/2}$ will be a real value and causes $\mathbb{A}$ to be an infinite set. The latter consequence will make our method to be infeasible since there are infinite possible values of $(ab)^{m/2}$ to be tried on. Therefore, m must be an even integer equals or greater than 2.*

The following is an example to illustrate the result from Theorem 1.

Example 1. *We use RSA-2048 modulus in this example. Specifically, we are given*

$$\begin{aligned} N = {} & 25443213484803330676546636060506767271319211956273880351374351825 \\ & 46256158013255117739836500456730264902937246910852858138318236603 \\ & 28796126064275138262348021411229982061934595317738337964801727892 \\ & 54233470084592231117946043667803816674367149523326731127008733355 \\ & 36182425074366173327195127004160399499185526019310064433935140944 \\ & 60366015740466980367515605709366458027738329608044170750026717443 \\ & 54815841155246667831512956948961180313537576080810878904128457697 \\ & 49463326499780838181084411701695971249384738323330037734781899087 \\ & 42844727615199026762546947725863259415895257407078268520959081886 \\ & 49384624121217162949627607660163 \end{aligned}$$

Suppose from side-channel attack described previously, we know the 12 LSBs of p and q. Particularly,

$$p = \underbrace{1\ldots0000000000}_{\textit{unknown } 1024 \textit{ bits}} + \underbrace{101111001001}_{\textit{known } 12\textit{-bits}}$$

and

$$q = \underbrace{1\ldots0000000000}_{\textit{unknown } 1024 \textit{ bits}} + \underbrace{100111101011}_{\textit{known } 12\textit{-bits}}$$

where

$$r_p = (101111001001)_2 = 3017 \tag{11}$$

and

$$r_q = (100111101011)_2 = 2539 \tag{12}$$

Then we set

$$\begin{aligned} i &= \left\lceil (r_p r_q)^{1/2} \right\rceil \\ &= 2768. \end{aligned}$$

Then we calculate

$$\sigma = \left(\left\lceil \sqrt{N} \right\rceil - i \right)^2 \quad \textit{and} \quad z \equiv N - (r_p r_q) \pmod{\sigma} \tag{13}$$

and solve the equation

$$x_{1,2} = X^2 - zX + \sigma r_p r_q = 0 \tag{14}$$

We find that neither $\frac{x_1}{r_q} + r_p$ *nor* $\frac{x_2}{r_p} + r_q$ *are integers. This means* x_1 *and* x_2 *are not our final solutions. It also means* $\sigma \neq (ab)^m$ *at this point. To find the correct* σ*, we have to iterate the computation of Equations (13) and (14) using iterations of increasing values of i. This search can be done in polynomial time as i should be less than* $\frac{r_q}{2} + 2^{\frac{m}{2}-1} r_p + 1 = 7304$ *as stated in Lemma 2. In this case, we find the correct* σ *when* $i = 2811$*. That is, we compute*

$$\begin{aligned}\sigma &= \left(\left\lfloor\sqrt{N}\right\rfloor - i\right)^2 \\ &= 254432134848033306765466360605067672713192119562738803513743518254625615801325511773983650045673026490293724691085285813831823660328796126064275138262348021411229982061934595317738337964801727892542334700845922311179460436678038166743671495233267311270087333553618242507436617332719512700416039949918552592962195579273096721757093357794065292733692579733017882760046777578179801403516768246292468519680986384686120264517134998212638327726468550707830214040511896758874144335396538824539148844087137816346245328888518360373902790724858882651191332644704993553711430100366047804022517832604599334389104100000000000000000$$

and

$$\begin{aligned}z &= N - (r_p r_q) \pmod{\sigma} \\ &= 896881086412041737270327265794640168763382302597634857526769155202986436934650994919725568989187148029362900930497247680492273743308164023833345436293443443589110393948271190234563044828085133601598675844458967154836894193689034014411135561508115826586218382730671222071693656405388924690682306752949627600000000.\end{aligned}$$

Using values of σ and z, we solve the equation

$$x_{1,2} = X^2 - zX + \sigma r_p r_q = 0. \tag{15}$$

The solutions of Equation (15) are used to compute

$$\begin{aligned}\frac{N}{\frac{x_1}{r_q} + r_p} &= p \\ &= 207632566695348090325106198564354306872362493463538154841386314580707220972445801440409737589803024013035554181699335224061662229162879643933792870833231736875142501533422110427899095351781206012327937258761409973123340262144886588093314114536052456895922041585909651666335476791456709509341751911472100003017\end{aligned}$$

and

$$\begin{aligned}\frac{N}{\frac{x_2}{r_p}+r_q} &= q\\ &= 1225396087413168498292617260986889571145024632726919066571061\\ &\quad 6588749446565648362779666067127897821347705191543359716126834\\ &\quad 5944097932917669169852614268434890176706523882967335716979529\\ &\quad 9071636233133238459212674004750005745005313778479423967599274\\ &\quad 3740090403457711105290569800062341129610183840357926739210000\\ &\quad 2539.\end{aligned}$$

Hence, N has been successfully factored in polynomial time.

Remark 3. *From Example 1, we show that as small as 12-bits of LSBs are required to successfully execute our attack. Hence, this put our method in advantage since it does not necessarily depend on side-channel attack [7] to gather the LSBs. Instead, by using our method, an adversary can use brute-force approach to find the correct LSBs since the required LSBs can be very small.*

4. Numbers of Primes with Vulnerable Specialized Structures Against Random Reconstruction Algorithm

From Equations (7) and (8) we can see that r_{p_1} until $r_{p_{(n-k)}}$ must be another binary representation of a squared number. The same case also applies on r_{q_1} until $r_{q_{(n-k)}}$ In the next Theorem, we count the number of squared numbers with $n-k$ bit.

Theorem 2. *If n is any large positive integer and k is a small positive integer then there are at least* $\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)\right\rfloor$ *squared numbers between* 2^{n-k-1} *and* $2^{n-k}-1$.

Proof. Let $X=\{x_i^2\}$ for $i=\{1,2,3,\ldots\}$ be the set of all squared numbers between 2^{n-k-1} and $2^{n-k}-1$. Particularly,

$$2^{n-k-1}<x_i^2<2^{n-k}-1.$$

Then

$$2^{\frac{1}{2}(n-k-1)}<x_i<\left(2^{n-k}-1\right)^{\frac{1}{2}}\Rightarrow 2^{\frac{1}{2}(n-k-1)}<x_i<\left(\left(2^{\frac{n-k}{2}}-1\right)\left(2^{\frac{n-k}{2}}+1\right)\right)^{\frac{1}{2}}. \tag{16}$$

To find the least number of i, the amount of squared numbers between 2^{n-k-1} and $2^{n-k}-1$, we compute the difference between the upper bound and the lower bound of Equation (16) in integer form. That is,

$$\begin{aligned}\left\lfloor\left(\left(2^{\frac{n-k}{2}}-1\right)\left(2^{\frac{n-k}{2}}+1\right)\right)^{\frac{1}{2}}-2^{\frac{1}{2}(n-k-1)}\right\rfloor &> \left\lfloor\left(\left(2^{\frac{n-k}{2}}-1\right)\left(2^{\frac{n-k}{2}}-1\right)\right)^{\frac{1}{2}}-2^{\frac{1}{2}(n-k-1)}\right\rfloor\\ &= \left\lfloor\left(\left(2^{\frac{n-k}{2}}-1\right)^2\right)^{\frac{1}{2}}-2^{\frac{1}{2}(n-k-1)}\right\rfloor\\ &= \left\lfloor 2^{\frac{n-k}{2}}-1-2^{\frac{1}{2}(n-k-1)}\right\rfloor.\\ &= \left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)-1\right\rfloor.\end{aligned}$$

If n is any large positive integer and k is a small positive integer then

$$\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)-1\right\rfloor\approx\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)\right\rfloor.$$

This terminates the proof. □

Theorem 3. *Let $a, b \in \mathbb{Z}^+$ and $m \geq 2$ be an even number such that $a < b < (2a^m + 1)^{\frac{1}{m}}$. Suppose $N = pq = (a^m + r_p)(b^m + r_q)$ be a valid RSA modulus. Let $r_p \equiv p \pmod{2^m}$ and $r_q \equiv q \pmod{2^m}$ where $r_p < 2a^{m/2}$ and $r_q < 2b^{m/2}$ such that $\max\{r_p, r_q\} < 2^k$. Let $x > 0$ be an integer where x^2 is the smallest squared number with n-bit size. If $2^{k-1}\left(2^{\frac{m}{2}} + 1\right)$ is a sufficiently small value as defined in Definition 2 and k many LSBs of p and q are known, then there are at most*

$$\frac{\left\lfloor 2^{\frac{n-k}{2}}\left(1 - 2^{-\frac{1}{2}}\right)\right\rfloor}{2}\left(\frac{2^k}{\log(x)^2} + \frac{2^k}{\log\left(x + \left\lfloor 2^{\frac{n-k}{2}}\left(1 - 2^{-\frac{1}{2}}\right)\right\rfloor\right)^2}\right)$$

candidates of p and q with size of n-bit such that $p = a^m + r_p$ and $q = b^m + r_q$ satisfy Theorem 1.

Proof. Let $x > 0$ be an integer where x^2 is the smallest squared number with $n - k$-bit. Let $f(x)$ be the prime-counting function between x^2 and $x^2 + \max\{r_p, r_q\}$. Then

$$\begin{aligned}\pi_1^*(x) = \frac{x^2 + \max\{r_p, r_q\}}{\log\left(x^2 + \max\{r_p, r_q\}\right)} - \frac{x^2}{\log x^2} &\approx \frac{x^2 + \max\{r_p, r_q\}}{\log x^2} - \frac{x^2}{\log x^2} \\ &= \frac{x^2 + \max\{r_p, r_q\} - x^2}{\log x^2} = \frac{\max\{r_p, r_q\}}{\log x^2} \\ &< \frac{2^k}{\log x^2}.\end{aligned}$$

From Theorem 2, we know there are approximately $\left\lfloor 2^{\frac{n-k}{2}}\left(1 - 2^{-\frac{1}{2}}\right)\right\rfloor$ squared numbers with $n - k$-bit size where $n - k$ is a large integer suitably used in RSA. Thus, $\pi_1^*(x)$ for the consecutive squared numbers are as follows:

$$\begin{aligned}\pi_1^*(x) &< \frac{2^k}{\log(x)^2} \\ \pi_1^*(x+1) &< \frac{2^k}{\log(x+1)^2} \\ \pi_1^*(x+2) &< \frac{2^k}{\log(x+2)^2} \\ &\vdots \\ &\vdots \\ \pi_1^*\left(x + \left\lfloor 2^{\frac{n-k}{2}}\left(1 - 2^{-\frac{1}{2}}\right)\right\rfloor\right) &< \frac{2^k}{\log\left(x + \left\lfloor 2^{\frac{n-k}{2}}\left(1 - 2^{-\frac{1}{2}}\right)\right\rfloor\right)^2}.\end{aligned} \tag{17}$$

The summation of Equation (17) can be represented in the sum of arithmetic progression formula where the number of i terms is multiplied by the sum of the first and last number in the progression and dividing by 2. That is,

$$\pi_2^* = \sum_{i=0}^{\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)-1\right\rfloor} \frac{2^k}{\log(x+i)^2} < \frac{\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)\right\rfloor}{2}\left(\pi_1^*(x)+\pi_1^*\left(x+\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)\right\rfloor\right)\right)$$
$$< \frac{\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)\right\rfloor}{2}\left(\frac{2^k}{\log(x)^2}+\frac{2^k}{\log\left(x+\left\lfloor 2^{\frac{n-k}{2}}\left(1-2^{-\frac{1}{2}}\right)\right\rfloor\right)^2}\right) \quad (18)$$

This terminates the proof. □

Result from Theorem 3 shows there is a significant amount of primes that satisfy Theorem 1.

5. Comparative Analysis

Here we compare our results with the existing attacks with known bits of primes. The authors of [16] introduced partial key exposure attacks with assumption that certain bits of primes can be known by the adversary. They showed that $2/3$ bits of p or q are sufficient to factor N using integer programming technique. Later, ref. [17] reduced this value to $1/2$ using LLL algorithm. The attack from Herrmann and May later on required the known bits to be arranged in random blocks [18].

Heninger and Shacham's attack is motivated by the so-called cold boot attack which targets the memory in electronic chips to reconstruct the bits of the private keys given that the bits are from random positions [5]. They successfully conducted the attack if 0.57 random bits of the primes are known. It should be noted here their fraction value is much lower if they consider the random bits of RSA private exponent, d (d_p and d_q in the case of CRT-RSA). Using a similar method, ref. [6] proved that if the total LSBs from both p and q known is at least 50% of the total length of N, then N can be factored using lattice-based method. Our method, unlike existing methods, utilize k-many LSBs of the primes where k is less than the value of $2^{k-1}\left(2^{\frac{m}{2}}+1\right)$ which is sufficiently small as defined in Definition 2, as shown in Theorem 1.

The summaries of all the attacks are compiled in Table 1.

From Table 1, we can see that our method required less LSBs for the attack to be successful when compared to [5,6]. That is, the attack required less computational time and space to be executed. It is easy to see that if $N \approx 2^{2048}$ and $k = 80$, then $r_p, r_q < N^{0.039}$. This is a substantial improvement from previous works.

We would like to point out the trade-off of our attack, namely the characteristics as mentioned in Theorem 1. Nevertheless, our analysis shows that if r_p and r_q are bounded to 2^k where k is stated as in Definition 2, the side-channel attack can be conducted in reasonable time in order to identify whether the primes in physical devices fall under the category as mentioned. This results in our research to be of importance for real-world implementation of the RSA cryptosystem. Moreover, we have shown in Section 4 that the number of primes satisfying our conditions are exponentially many. This shows the importance of our attack.

Table 1. Comparison of our method against existing attacks with known bits of primes.

Attacks	Position of Known Bits	Bits of Primes Need to Be Known	Comments/ Remarks	Advantages/ Disadvantages
Rivest and Shamir (1985)	LSBs or MSBs	2/3 of the bits of p or q	Solving integer programming problem	**Advantages:** Fast speed **Disadvantages:** Requires a lot of known bits
Coppersmith (1996)	LSBs or MSBs	1/2 of the bits of p or q	Using lattice-based method	
Herrmann and May (2008)	Any position (in blocks)	$\log_e(2) \approx 0.7$ of the bits of p or q	Number of blocks $\approx \log\log N$	
Heninger and Shacham (2009)	Any position	$r_p = N^{\delta_1}$ $r_q = N^{\delta_2}$ $\delta_1 + \delta_2 \geq 0.57$ of the bits of p or q	Using random reconstruction algorithm (RRA)	
Maitra et al. (2010)	LSBs	$r_p = N^{\delta_1}$ $r_q = N^{\delta_2}$ $\delta_1 + \delta_2 \geq 0.5$ of the bits of p or q	Using RRA together with lattice-based method	
Our method: Theorem 1	LSBs	$r_p, r_q < 2^k$ where 2^k is sufficiently small as in Definition 2. That is $r_p, r_q < N^{\frac{k}{\log_2 N}}$.	Side-channel attack of complexity $O(2^k)$ where 2^k is sufficiently small as in Definition 2.	**Advantages:** Fast speed, requires less known bits **Disadvantages:** Requires specific hardware to conduct side-channel attack

6. Countermeasure of the Attack

Although the attack seems to target a niche set of primes, there is no immediate noticeable detection that can be implemented to overcome the attack. This means the prevention from utilizing the weak primes must be applied in the RSA key generator with the full knowledge of the secret parameters, p and q. The countermeasure is depicted in Figure 1.

Given N, p and q, if

$$\left\lceil N^{1/2} - \left\lfloor p^{1/2} \right\rfloor \cdot \left\lfloor q^{1/2} \right\rfloor \right\rfloor$$

is a sufficiently small integer as defined in Definition 2, then RSA key generator must find new p or q.

Figure 1. Countermeasure of the Attack.

Since the computation is minimal, the prevention of the attack can be applied in the real-world RSA implementation.

Example 2. *For a toy example of this countermeasure method, we revisit the values in Example 1. Given N, p, q from Example 1, we compute*

$$\left\lceil N^{1/2} - \left\lfloor p^{1/2} \right\rfloor \cdot \left\lfloor q^{1/2} \right\rfloor \right\rfloor = 2811.$$

Since 2811 is definitely sufficiently small based on Definition 2, an RSA key generator must find new p and q. Let

$$
\begin{aligned}
p &= 10373821590420718162568315912935402272816716250952617784159371685\\
&\quad 44340371332193665789760371540571568043597631052985984619935841269\\
&\quad 00533099600902588040933556878478965238617603915696057625198338769\\
&\quad 03361223061009707594893117366305299494205202223327617461773922102\\
&\quad 75482121239772860175086815490154038705222031363 01\\
q &= 11233601978358194938103618628808793989586489373749842937474042065\\
&\quad 13933235347992919444792393988509367460666790358619415756939475813\\
&\quad 80412937835561807122090537966641130001194088391044588117638361372\\
&\quad 99643968716613613967481916652898906661611644105170965584735585835\\
&\quad 33313981952793800787986603919026942776013275383 53
\end{aligned}
$$

be the the new p and q. Then,

$$
\begin{aligned}
N &= 11653538274128513578568669090454309990749271193335847349122392459\\
&\quad 01318960034317752307651515404527551518430900334308748335133453988\\
&\quad 21286310578795557118148985154417613224899775560303891043729606906\\
&\quad 29637177530605885689603305847327219925303871989047949044982302417\\
&\quad 19652217537589201420247464831069631221516545858847199510976358555\\
&\quad 34569641991568190286013308968767353183943188900880965338613790529\\
&\quad 14898692740675146768914029502466472816780769463189924714976665682\\
&\quad 15047424802978071513075475252664886423135404769620269065551233781\\
&\quad 80576090100374515694019647558981694450446331689603531906067965349\\
&\quad 37648446600588401959096464052253
\end{aligned}
$$

be the new RSA modulus, N. We compute

$$
\begin{aligned}
\left\lceil N^{1/2} - \left\lfloor p^{1/2} \right\rfloor \cdot \left\lfloor q^{1/2} \right\rfloor \right\rceil &= 91788620433890001811698154984784049754386699417980052\\
&\quad 34196964320832189804911338215937374325313217127978801\\
&\quad 05034402880821593305374615932152728008 1664264988.
\end{aligned}
$$

which is larger than 2^{112}. *Hence N is safe from our attack.*

7. Conclusions

We have shown an attack on RSA modulus, $N = pq$ where $p = a^m + r_p$ and $b^m + r_q$ for r_p and r_q are k LSBs of p and q respectively. Our attack can be mounted successfully in polynomial time if the LSBs of the primes are known and satisfy the conditions. We also show that there is a significant number of primes with respect to their sizes that are vulnerable to our attack. This imposes a great threat to the RSA users who might not realize that their RSA primes may fall under these vulnerable primes. However, our suggestion on how to detect the vulnerable primes during the key generation process may help to overcome this problem so that the RSA cryptosystem can still be applied.

Author Contributions: Conceptualization, A.H.A.G., M.R.K.A. and M.A.A.; methodology, formal analysis, investigation, writing—original draft preparation, A.H.A.G.; writing—review and editing, A.H.A.G., M.R.K.A. and M.A.A.; supervision and funding acquisition, M.R.K.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by Ministry of Education of Malaysia with Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UPM/02/08).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LSB	Least significant bits
MSB	Most significant bits
RRA	random reconstruction algorithm
RSA	Rivest–Shamir–Adleman

References

1. Rivest, R.L.; Shamir, A.; Adleman, L. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **1978**, *21*, 120–126. [CrossRef]
2. Buhler, J.P.; Lenstra, H.W.; Pomerance, C. Factoring integers with the number field sieve. In *The Development of the Number Field Sieve*; Springer: Berlin/Heidelberg, Germany 1993; pp. 50–94.
3. Pollard, J.M. Theorems on factorization and primality testing. *Math. Proc. Camb. Philos. Soc.* **1974**, *76*, 521–528. [CrossRef]
4. Boneh, D.; Durfee, G.; Frankel, Y. An attack on RSA given a small fraction of the private key bits. In *International Conference on the Theory and Application of Cryptology and Information Security*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 25–34.
5. Heninger, N.; Shacham, H. Reconstructing RSA private keys from random key bits. In *Advances in Cryptology-CRYPTO 2009*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–17.
6. Maitra, S.; Sarkar, S.; Gupta, S.S. Factoring RSA modulus using prime reconstruction from random known bits. In *International Conference on Cryptology in Africa*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 82–99.
7. Kocher, P.; Jaffe, J.; Jun, B.; Rohatgi, P. Introduction to differential power analysis. *J. Cryptogr. Eng.* **2011**, *1*, 5–27. [CrossRef]
8. Kocher, P.C. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *Annual International Cryptology Conference*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 104–113.
9. Kocher, P.; Jaffe, J.; Jun, B. Differential power analysis. In *Annual International Cryptology Conference*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 388–397.
10. Martinasek, Z.; Zeman, V.; Trasy, K. Simple electromagnetic analysis in cryptography. *Int. J. Adv. Telecommun. Electrotech. Signals Syst.* **2012**, *1*, 13–19. [CrossRef]
11. Cho, J.; Kim, T.; Kim, S.; Im, M.; Kim, T.; Shin, Y. Real-Time Detection for Cache Side Channel Attack using Performance Counter Monitor. *Appl. Sci.* **2020**, *10*, 984. [CrossRef]
12. Genkin, D.; Shamir, A.; Tromer, E. RSA key extraction via low-bandwidth acoustic cryptanalysis. In *Annual Cryptology Conference*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 444–461.
13. Ghafar, A.H.A.; Ariffin, M.R.K.; Asbullah, M.A. Extending Pollard Class of Factorable RSA Modulus. In Proceedings of the 6th International Cryptology and Information Security Conference 2018 (CRYPTOLOGY2018), Port Dickson, Negeri Sembilan, Malaysia, 9–11 July 2018; p. 103.
14. Ghafar, A.; Ariffin, M.; Asbullah, M. A New Attack on Special-Structured RSA Primes. *Malays. J. Math. Sci.* **2019**, *13*, 111–125.
15. Barker, E.; Dang, Q. *Recommendation for Key Management, Part 1: General*; NIST Special Publication 800-57 Part 1, Revision 4; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2016.
16. Rivest, R.L.; Shamir, A. Efficient factoring based on partial information. In *Workshop on the Theory and Application of of Cryptographic Techniques*; Springer: Berlin/Heidelberg, Germany, 1985; pp. 31–34.
17. Coppersmith, D. Finding a small root of a bivariate integer equation; factoring with high bits known. In *International Conference on the Theory and Applications of Cryptographic Techniques*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 178–189.
18. Herrmann, M.; May, A. Solving linear equations modulo divisors: On factoring given any bits. In *International Conference on the Theory and Application of Cryptology and Information Security*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 406–424.

symmetry

MDPI

Article

Pancreatic Cancer Early Detection Using Twin Support Vector Machine Based on Kernel

Wismaji Sadewo [1], Zuherman Rustam [2,*], Hamidah Hamidah [2] and Alifah Roudhoh Chusmarsyah [2]

[1] Department of Neurosurgery, Faculty of Medicine, Universitas Indonesia, Jakarta Pusat, DKI Jakarta 10430, Indonesia; wisma30@hotmail.com

[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Jawa Barat 16424, Indonesia; hamidah61@sci.ui.ac.id (H.H.); alifah.roudhoh@ui.ac.id (A.R.C.)

* Correspondence: rustam@ui.ac.id

Received: 8 March 2020; Accepted: 11 April 2020; Published: 23 April 2020

Abstract: Early detection of pancreatic cancer is difficult, and thus many cases of pancreatic cancer are diagnosed late. When pancreatic cancer is detected, the cancer is usually well developed. Machine learning is an approach that is part of artificial intelligence and can detect pancreatic cancer early. This paper proposes a machine learning approach with the twin support vector machine (TWSVM) method as a new approach to detecting pancreatic cancer early. TWSVM aims to find two symmetry planes such that each plane has a distance close to one data class and as far as possible from another data class. TWSVM is fast in building a model and has good generalizations. However, TWSVM requires kernel functions to operate in the feature space. The kernel functions commonly used are the linear kernel, polynomial kernel, and radial basis function (RBF) kernel. This paper uses the TWSVM method with these kernels and compares the best kernel for use by TWSVM to detect pancreatic cancer early. In this paper, the TWSVM model with each kernel is evaluated using a 10-fold cross validation. The results obtained are that TWSVM based on the kernel is able to detect pancreatic cancer with good performance. However, the best kernel obtained is the RBF kernel, which produces an accuracy of 98%, a sensitivity of 97%, a specificity of 100%, and a running time of around 1.3408 s.

Keywords: pancreatic cancer; twin support vector machine; linear kernel; polynomial kernel; RBF kernel

1. Introduction

According to the World Health Organization (WHO), the second leading cause of death in the world is cancer [1]. In 2018, around 17 million new cases of cancer occurred in the world and 9.6 million of the world's population died from cancer [2]. Cancer has many types depending on location. If cancer occurs in the pancreas, then the cancer is called pancreatic cancer. Pancreatic cancer is the seventh leading cause of cancer deaths in the world and ranks as the 14th most common cancer [3]. Based on the Global Cancer Observatory in 2018, the estimated number of diagnoses of this cancer in the world is 458,918 and the estimated number of deaths is 432,242 [4]. This cancer is expected to be the second leading cause of cancer deaths in the world in 2030 [5].

Pancreatic cancer is a type of cancer that is difficult to detect physically. This is because the pancreas is an organ deep in the body; there are no external lumps or external skin changes such as in cases of breast lesions [6]. In addition, non-specific symptoms such as nausea, anorexia, jaundice, weight loss, and abdominal pain are also factors in the difficulty of detecting pancreatic cancer early [7]. The difficulty in detecting pancreatic cancer early results in many cases of pancreatic cancer being diagnosed late. When pancreatic cancer is detected, cancer is usually well developed, whereas in one study, only 7% of pancreatic cancer was diagnosed as localized disease [6]. If pancreatic cancer

is detected too late, then the cancer can develop properly and spread to other body parts so that the cancer is difficult to treat. Therefore, early detection of pancreatic cancer is an important problem.

The problem of detecting pancreatic cancer early is a classification problem in machine learning. Machine learning has been accepted for various classification problems in various fields, one of which is in the field of medicine. In the field of medicine, several machine learning methods are used to detect several types of cancer, namely breast cancer [8–11], cervical cancer [12,13], ovarian cancer [14,15], colon cancer [16], prostate cancer [17], and lung cancer [18].

Machine learning has also been implemented to detect pancreatic cancer. Qiu et al. [19] implemented several methods of machine learning, namely the decision tree, the *k*-nearest neighbor, and the support vector machine. From the results of the implementation, the support vector machine method is the method that produces the best performance, which is 70% sensitivity and 70% specificity.

The support vector machine (SVM) is one of the well-known machine learning methods; it uses the concept of "maximum margin", and this concept reduces generalization errors by maximizing margins between two half-separated planes [20]. Many researchers have proposed the development of SVM to obtain better performance. One such development is the twin support vector machine proposed by Jayadeva et al. in 2007.

The twin support vector machine (TWSVM) aims to find two symmetry planes such that each plane has a distance close to one data class and as far as possible from another data class [20]. On several benchmark data sets, TWSVM is not only fast, but shows good generalization [20]. At present, TWSVM has become one of the popular methods because of its excellent learning performance [21]. However, both SVM and TWSVM indirectly need a kernel method to classify data. The kernel method is a method that uses a kernel function that allows an algorithm to operate in a feature space that has a higher dimension by using product operations between images of all data pairs in the feature space [22]. The kernel functions commonly used for SVM methods are the linear kernel, polynomial kernel, and radial basis function (RBF) kernel.

This paper proposes the TWSVM method as a novel approach for early detection of pancreatic cancer. The kernel functions used are the linear kernel, the polynomial kernel, and the RBF kernel. This paper compares the performance of TWSVM with each kernel to get the best kernel for early detection of pancreatic cancer.

2. Materials and Methods

2.1. Kernel Method

The kernel method uses a kernel function that allows an algorithm to operate in a feature space that has a higher dimension by using product operations between images of all data pairs in the feature space [22].

Let X^n be an input space; F is a feature space, and $\varphi : X^n \rightarrow F$. Kernel function is defined by

$$K\left(x_i, x_j\right) = \varphi(x_i).\varphi\left(x_j\right) \tag{1}$$

where $x_i, x_j \in X^n$.

Kernel functions that are often used are the linear kernel, the polynomial kernel, and the radial basis function (RBF) kernel. The formulas for the kernel functions are listed below.

2.1.1. Linear Kernel

Linear kernels are the simplest kernel functions, represented by the product $\langle x, y \rangle$ [23]. This kernel is the basic kernel that is most often used by SVM because with this kernel, SVM divides data linearly. The linear kernel formula is presented in Equation (2) [24].

$$K\left(x_i, x_j\right) = x_i^T x_j + c \tag{2}$$

2.1.2. Polynomial Kernel

The polynomial kernel is a kernel that is suitable for problems where training data is normalized [24]. From Equation (3), σ is the parameter that must be settled. Variable c is the constant that is set and variable d is the degree of polynomial that is set.

$$K\left(x_i, x_j\right) = \left(\sigma x_i^T x_j + c\right)^d;\ \sigma > 0 \tag{3}$$

2.1.3. Radial Basis Function (RBF) Kernel

The RBF kernel is a kernel family where distance measurements are smoothed by radials function (exponential function) [23]. The RBF kernel is denoted as in Equation (4),

$$K\left(x_i, x_j\right) = \exp\left\{-\sigma\left\|x_i - x_j\right\|^2 + c\right\};\ \sigma > 0 \tag{4}$$

where σ is the adjustable parameter.

2.2. Twin Support Vector Machine (TWSVM)

The twin support vector machine (TWSVM) is one of the developments of the support vector machine proposed by Jayadeva et al. in 2007. TWSVM aims to find two symmetry planes such that each plane has a distance close to one data class and as far as possible from another data class. [20]. Figure 1 shows an illustration of TWSVM.

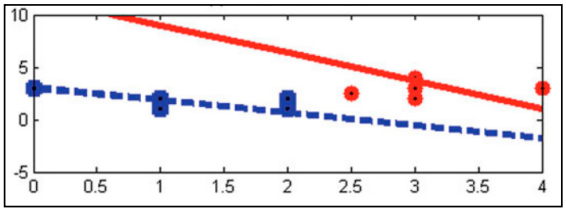


Figure 1. Illustration of the twin support vector machine (TWSVM) [20].

Let $D = \left\{\left(\mathbf{x_i},\ y_i\right) | \ x_i \in X^n,\ y_i \in \{-1, +1\},\ i = 1,\ 2,\ \ldots,\ N\right\}$ is the training data. Then let there be d_1 data points in class +1 and d_2 data points in class −1 such that $d_1 + d_2 = d$. We form $(d_1 \times n)$ matrix A, which contains the data points in class +1, and $(d_2 \times n)$ matrix B, which contains the data points in class −1. The two non-parallel hyperplanes are [20]:

$$\mathbf{x}^T\mathbf{w}_1 + b_1 = 0 \tag{5}$$

$$\mathbf{x}^T\mathbf{w}_2 + b_2 = 0 \tag{6}$$

where $\mathbf{x}$ is data vector, $\mathbf{w}_1$ is weight parameter for first hyperplane, b_1 is bias parameter for first hyperplane, $\mathbf{w}_2$ is weight parameter for second hyperplane, and b_2 is bias parameter for second hyperplane.

The TWSVM method is obtained by solving the following pair of quadratic programming problems [20]:

TWSVM 1:

$$\min_{\mathbf{w}_1, b_1, \xi_2} \frac{1}{2}\|A\mathbf{w}_1 + e_1 b_1\|^2 + c_1 e_2^T \xi_2 \tag{7}$$

subject to

$$-\left(B\mathbf{w}_1 + e_2 b_1\right) \geq e_2 - \xi_2 \tag{8}$$

$$\xi_2 \geq 0 \tag{9}$$

and

TWSVM 2:

$$\min_{\mathbf{w}_2, b_2, \xi_1} \frac{1}{2}\|B\mathbf{w}_2 + e_2 b_2\|^2 + c_2 e_1^T \xi_1 \tag{10}$$

subject to

$$-(A\mathbf{w}_2 + e_2 b_2) \geq e_1 - \xi_1 \tag{11}$$

$$\xi_1 \geq 0 \tag{12}$$

where $c_1 > 0$ and $c_2 > 0$ are penalty parameters, ξ_1 and ξ_2 are slack variables, and e_1 and e_2 are vectors of 'ones', i.e., each component is 'one' only [20].

The two hyperplanes of TWSVM with kernel [20]:

$$K\left(x^T, C^T\right)\mathbf{u}_1 + b_1 = 0 \tag{13}$$

$$K\left(x^T, C^T\right)\mathbf{u}_2 + b_2 = 0 \tag{14}$$

where $C^T = [A, B]^T$, $\mathbf{u}_1, \mathbf{u}_2 \in R^d$, and K is the kernel matrix corresponding to an appropriately chosen kernel function [11].

The kernel TWSVM can be obtained by solving the optimization problems [20]:

KTWSVM 1:

$$\min_{\mathbf{w}_1, b_1, \xi_2} \frac{1}{2}\left\|K\left(A, C^T\right)\mathbf{u}_1 + e_1 b_1\right\|^2 + c_1 e_2^T \xi_2 \tag{15}$$

subject to

$$-\left(K\left(B, C^T\right)\mathbf{u}_1 + e_2 b_1\right) \geq e_2 - \xi_2 \tag{16}$$

$$\xi_2 \geq 0 \tag{17}$$

and

KTWSVM 2:

$$\min_{\mathbf{w}_2, b_2, \xi_1} \frac{1}{2}\left\|K\left(B, C^T\right)\mathbf{u}_2 + e_2 b_2\right\|^2 + c_2 e_1^T \xi_1 \tag{18}$$

subject to

$$-\left(K\left(A, C^T\right)\mathbf{u}_2 + e_2 b_2\right) \geq e_1 - \xi_1 \tag{19}$$

$$\xi_1 \geq 0 \tag{20}$$

where $c_1 > 0$ and $c_2 > 0$ are penalty parameters, ξ_1 and ξ_2 are slack variables, e_1 and e_2 are vectors of 'ones', i.e., each component is 'one' only, $C^T = [A, B]^T$, $\mathbf{u}_1, \mathbf{u}_2 \in R^d$, and K is the kernel matrix corresponding to an appropriately chosen kernel function [20].

2.3. k-Fold Cross Validation

In this paper, to obtain a model and evaluate the model obtained, the dataset was divided into training data and testing data. Training data are data used by machines to recognize and study pancreatic cancer data patterns, while testing data are data used to evaluate models obtained after a machine learns data patterns. The dataset was divided into training data and testing data using the *k*-fold cross validation method. *k*-fold cross validation is a method for selecting training data samples. The *k*-fold cross validation method divided the dataset into *k* sections of equal size [25]. Each subsample was taken as validation data to test the model and repeat the process *k* times [25]. The advantage of this method is the repetition of random samples as training data and validation [25].

2.4. Proposed Method

In this paper, the method proposed for early detection of pancreatic cancer consisted of four stages. In the first stage, the data were divided into training data and testing data using *k*-fold cross

validation. The k value chosen was 10. This means that the dataset was divided into 10 samples of the same size. In this dataset, 9 samples were used as training data, and 1 sample was used as testing data. In the second stage, the training data were used by the TWSVM method based on linear kernels, polynomial kernels, and RBF kernels to study data patterns and build classification models. In the third stage, the classification model obtained was evaluated based on the parameters of accuracy, sensitivity, specificity, and required running time. This evaluation used testing data. After that, evaluation parameters generated by each kernel were compared to find out the best kernel to detect pancreatic cancer early using the TWSVM method. The stages carried out in this paper used the Python 3 programming language.

3. Results and Discussions

3.1. Data

In this paper, the data used were pancreatic cancer data obtained from Al Islam Hospital Bandung, Indonesia. The data consisted of 203 samples and six features. The six features were the diagnosis of patient, namely pancreatic cancer (Y) and not pancreatic cancer (N), and blood tests which consisted of cancer antigens, hemoglobin, leukocytes, hematocrit, and platelets. The diagnosis feature became a target feature in detecting pancreatic cancer. Table 1 shows part of the data.

Table 1. Part of the pancreatic data.

No	CA (U/mL)	Hemoglobin (g/dL)	Leukocytes (sel/uL)	Hematocrit (%)	Platelets (sel/uL)	Diagnosis
1	5.73	12.1	10,200	36.7	143,000	N
2	8.05	11.8	11,300	35.9	222,000	N
3	86.21	10.1	12,800	34.1	346,000	Y
4	87.13	12	11,700	36.7	612,000	Y

In the data shown in Table 1, the diagnosis feature is a categorical feature. This feature must be changed into a numeric feature for the proposed method to work. Therefore, based on the TWSVM method, the Y category in the diagnostic feature was transformed into +1 and the N category in the diagnostic feature was transformed into −1.

3.2. Confusion Matrix

In this paper, a confusion matrix was used to assist in calculating the evaluation parameters of the classification model. Table 2 shows the confusion matrix used to evaluate the TWSVM classification model based on the kernel for early detection of pancreatic cancer.

Table 2. Confusion matrix.

		Predict	
		Pancreatic Cancer (Y)	**Not Pancreatic Cancer (N)**
Actual	Pancreatic Cancer (Y)	True Positive (TP)	False Negative (FN)
	Not Pancreatic Cancer (N)	False Positive (FP)	True Negative (TN)

Explanation:

TP = Many cases of pancreatic cancer are predicted to be correct
TN = Many cases of not pancreatic cancer are predicted to be correct
FP = Many cases of not pancreatic cancer are predicted to be wrong (predicted as pancreatic cancer)
FN = Many pancreatic cancer cases are predicted to be wrong (predicted as not pancreatic cancer)

3.3. Evaluation Parameters

The parameters to evaluate the performance of the TWSVM classification model were accuracy, sensitivity, specificity, and required running time. Table 3 shows the formula for accuracy, sensitivity, and specificity.

Table 3. Parameter formulae to evaluate the classification model.

Parameter	Formula	Explanation
Accuracy	$\frac{(TN+TP)}{(FN+TP+FP+TN)} \times 100\%$	Comparison between the number of cases of pancreatic cancer and not pancreatic cancer that identified correctly with the total number of all cases
Sensitivity	$\frac{TP}{(FN+TP)} \times 100\%$	Proportion of pancreatic cancer cases identified correctly
Specificity	$\frac{TN}{(FP+TN)} \times 100\%$	Proportion of not pancreatic cancer cases identified correctly

3.4. Results

In this section, we discuss the performance evaluation of the TWSVM classification model with a linear kernel, polynomial kernel, and RBF kernel. The TWSVM classification model based on kernel that is proposed in this paper refers to research conducted by [20] which detects hepatitis using TWSVM with a linear kernel and a RBF kernel. In research conducted by the authors of [20], the accuracy produced by the RBF kernel is superior to that of a linear kernel. This indicates that the RBF kernel is the appropriate kernel in detecting hepatitis using TWSVM.

In this paper, we have built the TWSVM classification model with a linear kernel, a polynomial kernel, and a RBF kernel in detecting pancreatic cancer. Table 4 presents a comparison of TWSVM performance with linear kernels, polynomial kernels with $d = 4$, and RBF kernels with $\sigma = 0.05$. The performance evaluation parameters compared are accuracy, sensitivity, specificity, and running time.

Table 4. Results of TWSVM classification model based on kernel.

Classification Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Running Time (seconds)
TWSVM with Linear Kernel	92%	86%	95%	1.2811
TWSVM with Polynomial Kernel with $d = 4$	80%	75%	83%	1.2040
TWSVM with RBF Kernel with $\sigma = 0.05$	98%	97%	100%	1.3408

From Table 4, it can be seen that for accuracy, TWSVM models with RBF kernel have the most superior accuracy compared to the linear kernel and the polynomial kernel, reaching 98%. That is, the TWSVM model with the RBF kernel correctly detected 98% of the total cases. The lowest accuracy is the TWSVM model with a polynomial kernel, with a resulting percentage of 80%. In addition, for consideration of sensitivity and specificity, TWSVM models with RBF kernel have percentages that are also superior than the linear kernel and polynomial kernel, which are 97% sensitivity and 100% specificity. This means that the TWSVM model with the RBF kernel is able to detect 98% of cases correctly, with a truth of 97% of all cases of pancreatic cancer, and 100% of all cases of non-pancreatic cancer.

Based on consideration of accuracy, sensitivity, and specificity produced by the TWVM model with a linear kernel, a polynomial kernel, and a RBF kernel, overall the TWSVM model with RBF kernel is the most superior kernel. This means that the pancreatic cancer dataset can be separated almost precisely by RBF function. However, for consideration of running time, the TWSVM model with RBF kernel has the longest running time compared to linear and polynomial kernels, which is around 1.3408 s. The TWSVM model with the polynomial kernel actually produces the fastest running time,

which is around 1.2040 s. Even so, the running time produced by the RBF kernel is quite good and acceptable for detecting pancreatic cancer early. Thus, the RBF kernel is the best kernel for TWSVM in detecting pancreatic cancer early.

4. Conclusions

Early detection of pancreatic cancer is very important so that the handling of pancreatic cancer does not occur too late, before the cancer spreads to other organs in the body. However, early detection of pancreatic cancer is difficult because this cancer has non-specific symptoms. The twin support vector machine method based on the kernel can help detect pancreatic cancer early, based on blood tests. The most appropriate kernel for the TWSVM method in detecting pancreatic cancer is the RBF kernel which produces an accuracy of 98%, sensitivity of 97%, and 100% specificity, and the required running time is 1.3408 s.

Author Contributions: Conceptualization, W.S. and Z.R.; methodology, W.S. and H.H.; software, H.H. and Z.R.; validation, W.S., H.H., Z.R., and A.R.C.; formal analysis, W.S., H.H., and Z.R.; investigation, W.S. and Z.R.; resources, W.S.; data curation, W.S. and Z.R.; writing—original draft preparation, H.H.; writing—review and editing, A.R.C.; visualization, W.S.; supervision, Z.R.; project administration, Z.R.; funding acquisition, W.S. and Z.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universitas Indonesia with a DRPM PUTI Q2 2020 research grant scheme and The APC received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cancer. Available online: https://www.who.int/cancer/en/ (accessed on 19 January 2020).
2. Worldwide Cancer Statistics. Available online: https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer (accessed on 25 January 2020).
3. McGuigan, A.; Kelly, P.; Turkington, R.C.; Jones, C.; Coleman, H.G.; MacCain, R.S. Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World J. Gastroenterol.* **2018**, *24*, 4846–4861. [CrossRef] [PubMed]
4. Global Cancer Observatory 2018. Available online: http://gco.iarc.fr/ (accessed on 23 December 2019).
5. Rahib, L.; Smith, B.D.; Aizenberg, R.; Rosenzweig, A.B.; Fleshman, J.M.; Matrisian, L.M. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* **2014**, *74*, 2913–2921. [CrossRef] [PubMed]
6. Kim, V.M.; Ahuja, N. Early Detection of Pancreatic Cancer. *Chin. J. Cancer Res.* **2015**, *27*, 321–331. [PubMed]
7. Badger, S.A.; Brant, J.L.; Jones, C.; McClements, J.; Loughrey, M.B.; Taylor, M.A.; Diamond, T.; McKie, L.D. The role of surgery for pancreatic cancer: A 12-year review of patient outcome. *Ulster Med. J.* **2010**, *79*, 70–75. [PubMed]
8. Octaviani, T.L.; Rustam, Z. Random Forest for Breast Cancer Prediction. In Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018), Depok, Indonesia, 30–31 October 2018.
9. Rustam, Z.; Putri, R.A. Comparison between stochastic support vector machine (stochastic SVM) and Fuzzy Kernel Robust C-Means (FKRCM) in breast cancer classification. In Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018), Depok, Indonesia, 30–31 October 2018.
10. Rustam, Z.; Hartini, S. Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel. In Proceedings of the 9th Annual Basic Science International Conference 2019 (BaSIC 2019), Malang, Indonesia, 20–21 March 2019.
11. Fijri, A.L.; Rustam, Z. Comparison between Fuzzy Kernel C-Means and Sparse Learning Fuzzy C-Means for Breast Cancer Clustering. In Proceedings of the ICAITI 2018—1st International Conference on Applied Information Technology and Innovation: Toward A New Paradigm for the Design of Assistive Technology in Smart Home Care, Padang, Indonesia, 4–5 September 2018.

12. Rustam, Z.; Hapsari, V.A.W.; Solihin, M.R. Optimal cervical cancer classification using Gauss-Newton representation based algorithm. In Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018), Depok, Indonesia, 30–31 October 2018.
13. Zahras, D.; Rustam, Z. Cervical Cancer Risk Classification Based on Deep Convolutional Neural Network. In Proceedings of the ICAITI 2018—1st International Conference on Applied Information Technology and Innovation: Toward A New Paradigm for the Design of Assistive Technology in Smart Home Care, Padang, Indonesia, 4–5 September 2018.
14. Arfiani, A.; Rustam, Z. Ovarian cancer data classification using bagging and random forest. In Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2018), Depok, Indonesia, 30–31 October 2018.
15. Octaviani, T.L.; Rustam, Z.; Siswantining, T. Ovarian Cancer Classification using Bayesian Logistic Regression. In Proceedings of the 9th Annual Basic Science International Conference 2019 (BaSIC 2019), Malang, Indonesia, 20–21 March 2019.
16. Salmi, N.; Rustam, Z. Naïve Bayes Classifier Models for Predicting the Colon Cancer. In Proceedings of the 9th Annual Basic Science International Conference 2019 (BaSIC 2019), Malang, Indonesia, 20–21 March 2019.
17. Huljanah, M.; Rustam, Z.; Utama, S.; Siswantining, T. Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. In Proceedings of the 9th Annual Basic Science International Conference 2019 (BaSIC 2019), Malang, Indonesia, 20–21 March 2019.
18. Rustam, Z.; Kharis, S.A.A. Comparison of Support Vector Machine Recursive Feature Elimination and Kernel Function as feature selection using Support Vector Machine for lung cancer classification. In Proceedings of the Basic and Applied Sciences Interdisciplinary Conference, Depok, Indonesia, 18–19 August 2017.
19. Qiu, Y.; Jiang, H.; Shimada, K.; Hiraoka, N.; Maeshiro, K.; Ching, W.K.; Aoki-Kinoshita, K.; Furuta, K. Towards Prediction of Pancreatic Cancer Using SVM Study Model. *J. Clin. Oncol. Res.* **2014**, *2*, 1031.
20. Jayadeva; Khemchandani, R.; Chandra, S. Twin Support Vector Machines for Pattern Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 905–910. [CrossRef] [PubMed]
21. Huang, H.; Wei, X.; Zhou, Y. Twin support vector machines: A survey. *Neurocomputing* **2018**, *300*, 34–43. [CrossRef]
22. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
23. Suo, H.; Li, M.; Lu, P.; Yan, Y. Using SVM as Back-End Classifier for Language Identification. *EURASIP J. Audio Speech Music Process.* **2008**, *1*, 674859. [CrossRef]
24. Chidambaram, S.; Srinivasagan, K.G. Performance evaluation of support vector machine classification approaches in data mining. *Clust. Comput.* **2018**, *22*, S189–S196.
25. Raju, K.S.; Murty, M.R.; Rao, M.V.; Satapathy, S.C. Support Vector Machine with K-fold Cross Validation Model for Software Fault Prediction. *Int. J. Pure Appl. Math.* **2018**, *118*, 321–334.

symmetry

MDPI

Article

Symmetry Analysis in Analyzing Cognitive and Emotional Attitudes for Tourism Consumers by Applying Artificial Intelligence Python Technology

Guoxia Sun [1,2]

[1] School of Tourism Culture, Tourism College of Changchun University, Changchun 130607, China; sgx@tccu.edu.cn

[2] Northeast Asia Research Center on Leisure Economics, Tourism College of Changchun University, Changchun 130607, China

Received: 13 March 2020; Accepted: 2 April 2020; Published: 11 April 2020

Abstract: Symmetries play very important roles in the analysis of cognitive and emotional attitudes. The analysis with Python technology, including optimized artificial intelligence technology, is designed on the basis of symmetry principles. Destination image perception as a branch of destination image research is of great significance to tourists' decision-making and destination image building. Ice-snow tourism is a hot topic nowadays, and research on perceptions of images of ice-snow tourism has become a focus. In this paper, python programming was used to crawl online travel journals and reviews about Jilin province's ice-snow tourism on the Internet to analyze the frequency of frequently used words, their classification, word cloud and co-occurrence network, and other aspects of image perception, and proceed to the emotional perception of and emotional attitude to the emotional images and an overall image analysis. The study found that: (1) Perceptions of images of ice-snow tourism can be divided into five categories: tourism attractions, tourism activities, tourism facilities, tourism features and the tourism service environment. The frequency of tourism attractions is the highest, followed by tourism facilities and the tourism service environment. "Changbai Mountain" and "rime" are the core words, that is, tourists are most impressed by the scenic spot and landscape of "Changbai Mountain and rime." (2) Positive emotional expressions accounted for 67.23% of perceptions of images of ice-snow tourism. Tourists gave a positive evaluation for Changbai Mountain, the snow landscape of Tianchi and skiing facilities. Meanwhile, passive emotional expressions accounted for 21.07% and tourists gave passive evaluations for travel, transportation, accommodation and catering. (3) Tourists spoke highly of overall images of ice-snow tourism in Jilin Province but few were willing to revisit. In the conclusion, strategies are put forward to improve image perceptions of ice-snow tourism and promote the sustainable development of ice and snow tourism.

Keywords: artificial intelligence; ice-snow tourism; sustainable development; Python; text mining

1. Introduction

In 1992, the China National Tourism Administration (CNTA) listed "ice-snow scenery tours" as one of the 14 special tourism products in China for the first time. Since then, ice-snow tourism has become highly popular in China. In recent years, relevant ministries and commissions have jointly issued a series of industrial policies to support the development of ice-snow tourism, in response to General Secretary Xi's important instructions that "snow and ice are also invaluable assets" and to "encourage 300 million people to participate in the ice-snow sport". Ice-snow tourism is becoming a new pillar of national strategic convergence and local economic and social development. From the perspective of policy, China's ice-snow tourism has ushered in a golden period of leapfrog development. According to a report on the development of China's ice-snow tourism for 2020 released by the China Tourism

Academy, participation in ice-snow tourism in China during the 2018–2019 ice and snow season totaled 224 million, achieving a revenue of about 386 billion yuan. Ice-snow tourism continues to grow rapidly. From the perspective of the market, China has become an important country in the field of ice-snow tourism economy. Jilin province, by virtue of its abundant snow and ice resources and unique geographical advantages, occupies an important position in China's ice-snow tourism market and has become the main battlefield in China's ice-snow tourism market. Studying the perception of images of Jilin province's ice-snow tourism will greatly help to improve the overall image of China's ice-snow tourism. As the 2022 winter Olympic Games is around the corner, tourists from all over the world will come to experience China and Jilin province's ice and snow tourism. Jilin province will take this opportunity to build itself into a province with booming ice and snow tourism and a world-class tourist destination. However, the outbreak of NCP (novel coronavirus pneumonia) at the end of 2019 has seriously affected snow tourism in the province during the snow season. Jilin province should seize the opportunity, face up to the crisis, avoid the risk, study the demands of ice and snow tourists, and utilize modern artificial intelligence Python technology to obtain big data from online texts. The experience and perception of snow and ice tourists should be analyzed based on the data, so as to provide products and services that meet the needs of tourists, and to conduct precise marketing.

The term "destination image" was first used in a doctoral dissertation titled "Image: A Factor in Tourism" by J.D. Hunt of Colorado State University (CSU) in 1971. Hunt believed that destination image is people's impression of a place they do not live in, and discussed the significance of development of destination image [1]. Since then, destination image has been widely studied and applied by scholars at home and abroad. Destination image perception research is a branch of destination image research, which is the overall impression on tourists of relevant resources and features of tourist destinations. Destination image perception is by nature a psychological behavior, as it is tourists' experience, perception and emotional evaluation of various elements of a tourist destination, as well as a kind of social perception of tourists' understanding of a tourist destination. Destination image perception is of great guiding significance for tourists to make tourism decisions, and for image construction, dissemination and promotion of tourism destinations.

1.1. Study of Tourism Image Perception

The study of tourism image perception at home and abroad is reflected in research content, research cases and research methods.

1.1.1. Research contents

Domestic and foreign research focuses on the constituent elements of and factors influencing tourism image perception.

Elements of Tourism Image Perception

Foreign scholars believe that the constituent elements of tourism image perception are from the perspective of tourists, and that tourism destination image perception is composed of pre- and post-tour images. Gunn Clare proposed that the formation of tourist destination image includes two levels, i.e., original image and induced image. The original image refers to the information accumulated by the tourist before the trip, and the induced image is formed after the field trip [2]. Kolter and Barich (1991) divided the perception of destination image into emissive image and receptive image. Emissive image refers to the fact that tourism destinations actively integrate their own resources to selectively deliver to tourists. Received image refers to the comprehensive impression of the tourist destination obtained by tourists on the spot or by external information [3]. On this basis, Fakeye and Crompton summarized the tourist perception image formed by tourists (including potential tourists) as the original image, induced image and mixed image (the comprehensive impression of tourist destinations after travel) [4]. Gartner continued to refine this on the basis of his predecessors. According to the formation process of tourism perception image, tourism perception image has been divided into 8 types:

obvious induction, hidden induction, and autonomous native [5].Martin Selby and others proposed the concepts of native image and re-evaluation image from the perspective of tourists' cognitive time. Additionally, they distinguish the constituent elements of the tourist image from the perspective of the information source of the tourists. They believe that the original image of the tourist is formed by the mechanism image of the media, education and other channels that are not directly related to tourism and the induced image of tourism business channels such as advertising [6]. In addition, Seyhmus Baloglu and others agreed to divide the perceived image of tourists into cognitive image and emotional image and proposed a "cognitive-emotion" model of tourist destination image, and they considered that the two images together constitute a comprehensive image of the destination. This view has also been recognized by many scholars [7]. Tapachai and Waryszak (2000) used consumption value theory to introduce research, and introduced the tourist destination image's perceived goals covering the five major modules: attraction function, social characteristics (safety, residents friendly), conditions (comfort, hygiene), emotion and awareness [8]. Beerli & Martin (2004) summarize relevant literature and propose that tourists' perceptions of tourist destination images should be analyzed in four aspects: tourist attractions (natural and cultural attractions); reception facilities and services; emotions; and social environment and atmosphere [9]. Wang and Hsu (2010) explore the relationship between tourist destination image, satisfaction, and behavioral intent, showing that the overall state is reflected in both cognitive and emotional aspects [10]. Perceived image is a conceptual model reflection from the perspective of tourism image psychology.

Domestic scholars have also actively explored the constituent elements of tourism image perception, but due to different research objects, the constituent elements of tourism perception image are different. Xie Chaowu and Huang Yuanshui believe that five parts: tourism resources, tourist destination facilities, tourist destination services, industry management, and community participation together form the perception of tourist destination image [11]. Qi Huangxiong and others believe that social environment, folk culture, landscapes, urban planning, and economic construction jointly produce a tourism perception image system in Linhai City [12]. Li Xi, Ye Sheng, and Wang Dong analyzed the perception characteristics of business tourists visiting Australia, and put forward the perception of tourist destination image including tourism experiences such as room and board, transportation, shopping and overall perceptions of politics, economy, culture and history [13]. Wu Jinfeng (2014) believes that tourism image perception can be divided into five dimensions: tourism attraction; infrastructure; leisure and entertainment; environment; and local atmosphere. Each dimension contains different attributes [14]. Bai Dan, Ma Yaofeng, and Liu Junsheng believe that travel expectations, travel parades, tourist attraction experience, supporting facilities, service experience, and post-tour evaluation jointly form tourists' perception of the Terracotta Warriors and Horses of Qin Shihuang [15]. Li Ping [16], Feng Qing [17], Lu Lijun [18], and Tu Wenhui [19] discussed the perception of urban tourism community image and the perception of tourist destination image in Shaanxi considering three aspects: cognitive image, emotional image and overall image and Nanyue Hengshan tourist destination image perception and Beijing Fangshan District tourist image perception.

Factors influencing tourism image perception

Foreign scholars have carried out research on cognitive image, emotional image and overall image, which basically comprise three factors that affect the tourist destination, tourists and tourism media. In terms of cognitive image, Mayo believes that scenery, climate, traffic and other factors will affect the perceived image of tourism [20]. Stabler believes that demand factors such as tourists' motivation, cognition, and personal attributes, together with supply factors such as the tourism market and intermediary media, act on the perceived image of tourism [21]. Seyhmus and David point out that the perceptual influence factors include stimulus conditions and self-factors [22]. Hanlan and Kelly reported that tourism experience is an important factor affecting the perceived image of tourism, and tourism marketing organizations should use multiple methods to convey brand information [23]. In terms of emotional image, Beerli and Martín verified that the factors affecting tourists' perception

evaluation include tourism, information sources, stimuli, and various demographic characteristics [9]. In the overall dimension, Olivia believes that the image of the tourist destination is the result of the tourists' internal and external conditions [24]. Dimitrios and Amir (2017) studied the perceived image of residents and tourists in the resort city of Eilat based on three levels of cognition, emotion and overall image, and verified the applicability of the behavioral intention model. Moreover, they found that the emotional component influences the overall tourist destination image to a greater extent than the cognitive evaluation [25].

Domestic scholars have carried out research on objective factors, subjective factors, destinations, time and space, transportation and information. Studies highlighting the combined effect of all influencing factors include those by Cheng Jinlong and Wang Fa, who built models of influencing factors for tourism image perception. These include tourist factors (individual and group factors), tourist destination factors (tourist destinations themselves, between tourist destinations, and tourist sources and destinations), information factors (personal, interpersonal, and business factors) and external forces factors [26]. Other scholars have emphasized one or two of these factors. Cheng Wei and Sui Lina believe that stimulus factors such as first-hand and second-hand sources, and individual factors such as tourism motivation and socio-demographic characteristics, are the main factors affecting the perception of tourism in Korea in the Yangtze River Delta [27]. Gan Lu and Lu Tianling concluded that tourists with different motivations had different evaluations of their tourist destination image [28]. Zheng Peng empirically pointed out that the urban environment has the greatest effect on the recognition of tourism image [29], Zhang Hongmei et al. [30] and Tang Yufeng et al. [31] pointed out that distance is an important factor affecting tourism image perception.

1.1.2. Research Case

As far as foreign research is concerned, many studies have selected a certain country as the research area. Jeong studied the image of Korean tourism perceived by Russian tourists [32], Steven studied the image perception of Brazilian, Argentina, and Chile in the hearts of Australian tourists [33], and Eran studied the restoration and reconstruction of perception of tourist images after the earthquake in Nepal [34].

Domestic tourism image perception cases are generally concentrated in tourist destinations and tourist attractions.

Most of the literature on tourism destination image perception research focuses on strategies for improving tourism destination perception image. Relevant studies include empirical research by Shan Linyao and others on image perception of Qingdao's tourism image [35], Yang Jie and others' study of Chongqing citizens' perception of Shanghai's tourism image [36], Zhu Cuilan and others' study of Xiamen's tourist destination image [37], Shi Kunbo et al.'s study of mainland university students' travel motivations and tourism image in Taiwan [38], and Yang Min et al.'s study of Xi'an tourism image perception [39].

Concerning tourist attractions, there are tourism image perception studies of Huashan Scenic Area by Zhang Gaojun et al. [40], studies on tourism cognition, emotion, and overall image of Shaolin Temple by Pi Rui et al. [41], tourism image perception studies of Yongding Tulou, Fujian Province by Zhang Wenting, et al., [42] and Xu Yayuan et al.'s study on tourism image perception of the Huangshan Scenic Area [43].

1.1.3. Research Method

The foreign research methods on tourism perception image mainly include factor analysis and structural equation models, network text mining, and content analysis methods. Hao Zhang and Taeyoung Cho and others used factor analysis and structural equation models to analyze the risk perception of tourism destinations, which negatively affected the tourism image, cultural image and stability image of the destinations [44]. Stella Kladou and Eleni Mavragani evaluated the image of tourist destinations on TripAdvisor [45]. Hunter and William Cannon used network information

and traditional print media to study the image of Seoul's tourist destinations [46]. Jinah Park and Alastair M. Morrison and others used content analysis to analyze the different perception characteristics of Chinese golf courses and their relation to Korean society [47].

Domestic research methods mainly include regression analysis, comparison of questionnaire surveys and web text analysis, and a combination of web text analysis and IPA models. Peng Huijun and others used the regression analysis method to analyze tourists' perception of the multiple tourism image positioning of mountain-type scenic spots, using Hengshan in Nanyue as an example [48]. Li Yan and others used the network text analysis method to compare the tourism image perception of desert-type scenic spots with Shapotou and Shahu scenic spots in Ningxia as examples [49]. Zhang Zhenzhen and others used questionnaire surveys and web text methods to analyze tourism image perception using Xi'an as an example [50]. Zhang Rui et al. analyzed the perception of tourism image of Shanghai Chenshan Botanical Garden by using web text and IPA models [51].

1.2. Research on Ice and Snow Tourism at Home and Abroad

Foreign ice and snow tourism research has basically focused on spatial differences in the application of ice and snow tourism resources, the demand for ice and snow tourism markets, and the suitability of ice and snow tourism. Arvid explores the value of sustainable tourism development in winter [52]; Yu Zhang predicts Harbin ice and snow tourism market demand with a time series univariate linear regression model [53]; Jun Yang and Ruimeng Yang used the Delphi analytic hierarchy process and the spatial analysis method to study spatial differences in the suitability of snow and ice tourism in China [54].

Domestic snow and ice tourism research is basically focused on research in the Northeast region. The research content mainly focuses on the analysis and forecast of ice and snow tourism status, ice and snow tourism marketing, snow and ice tourism resource development, and sustainable development. Dong Xia mainly uses the logistic model to predict the passenger flow of ice and snow tourism in Jilin Province, and makes suggestions for the development of ice and snow tourism in Jilin Province [55]. Song Hongjuan analyzed the potential source market of Yabuli Ski Resort and proposed a strategy for tapping the potential source market; [56] Xu Yijun and others analyzed the development status of the Harbin ice and snow tourism market and divided the ice and snow tourism source market into four levels, and proposed corresponding development strategies for each level [57].

In the study of ice and snow tourism image, there is Wang Hairong's image of ice and snow tourism in Heilongjiang Province [58]. Han Zhenkun and others have promoted Harbin's ice and snow tourism image in two aspects: ice and snow landscape and ice and snow culture [59]. There is also Liang Shuang's research on the influence factors of Harbin's ice and snow tourism season tourism image [60].

1.3. Deficiencies of Existing Researches

From the research literature summarized above, it is found that the research content of domestic and foreign scholars in tourism image perception mainly includes tourism perception components and influencing factors. The research area focuses on countries, destinations, and tourist attractions; in the research methods, factor analysis and structural equation models, network text mining, content analysis, regression analysis, and network text analysis and IPA models are combined. Domestic and foreign scholars' research on ice and snow tourism has basically focused on the development of ice and snow tourism resources, the application of ice and snow tourism resources, the analysis and prediction of the status of ice and snow tourism, the marketing of ice and snow tourism, and spatial differences in the suitability of ice and snow tourism.

From the perspective of tourism image perception, domestic and foreign scholars' research on tourism image perception of a certain destination or a certain scenic spot has gradually matured, but little attention has been paid to image perception of certain types of tourism products. In the future, the scope of application of tourism image perception research should be expanded, and it can be

applied to coastal tourism image, rural tourism image and so on. This study selects snow and ice tourism as the research object. It will theoretically open up new research fields and broaden the scope of tourism image perception, which will help scholars to think about differentiating and establishing different types of tourism image perception theories and models.

From the perspective of ice and snow tourism research, existing research focuses on ice and snow tourism resources, ice and snow tourism markets, and the suitability of ice and snow tourism. Based on image perception, this study opens up a new research direction for ice and snow tourism, which is helpful for scholars to study ice and snow tourism from the perspective of the experience economy.

Concerning research method, there are existing studies using traditional methods and tools. This research uses the most modern Python technology for big data text mining and analysis. This can make up for the shortcomings of other text data mining methods in fine cleaning and deep mining, and can extract some hidden information, thereby improving the accuracy and scientificity of tourism image perception research.

From the perspective of interdisciplinary studies, there have been many studies focusing on tourism geography or management. This study combines management science, economics, and psychology perfectly, and applies it to the new ice and snow tourism industry. This is conducive to the intersection and fusion of disciplines to create new research directions and fields.

In summary, based on existing research, in the era of big data where travel and reviews are king, text mining methods and the Python language will be used to study emerging tourism formats around the three aspects of cognitive image, emotional image and overall image of snow and ice tourism image perception. In addition, this paper accurately grasps and judges the factors that influence the image of ice and snow tourism, and proposes strategies for improving the image perception of ice and snow tourism, with a view to providing a reference for image perception research on ice and snow tourism and other emerging tourism formats in Jilin Province.

2. Research Design

2.1. Research Object

2.1.1. Current Status of Ice and Snow Tourism Resources in Jilin Province

Jilin Province is located in the core area of Northeast Asia in the world's golden ice and snow tourism belt. Ice and snow tourism resources have the characteristics of resources and climate, of globalization (all types), high taste, good combination, little wind, and relatively warm winters. The characteristics of their spatial distribution are outstanding. Jilin Province's ice and snow tourism resources are led by Changbai Mountain, Songhua River, and Chagan Lake, showing the spatial distribution characteristics of "ice in the west, town in the middle, and snow in the east". The excellent snow and ice tourism resources include Changbai Mountain, Tianchi, Jilin rime, Changbai Mountain Hot Spring, Moon Lake, Chagan Lake (winter catch), Songhua Lake, Lianhua Mountain in Changchun, etc. There are four types of ice and snow leisure and vacation resources: ice and snow hot spring health resources, ice and snow tourism cultural resources, and ice and snow folk experience resources. The most popular ice and snow leisure and vacation resources are skiing vacations, such as Wanda Changbai Mountain International Resort and Vanke Songhua Lake Resort, Town, Tianmu Hot Spring and other representatives; ice and snow sightseeing experience resources are represented by Changbai Mountain Scenic Area, Shidao Daogou Scenic Area, Devil World, Old Rick Lake, Liuding Mountain, etc. Among the snow and ice folk cultural resources, the historical and cultural categories include Goguryeo cultural sites, Jingyue Snow World, Changying Century City, and the Puppet Manchurian Palace. Abundant ice and snow resources have laid a very advantageous resource foundation for the development of global ice and snow tourism, key ice and snow tourism industry clusters, characteristic ice and snow tourism towns and famous villages, and the creation of characteristic boutique ice and snow tourism routes.

2.1.2. Current Development of Ice and Snow Tourism Market in Jilin Province

The development status of the ice and snow tourism market in Jilin Province adopts a questionnaire survey method. It is calculated that the longest period of time is 180 days (from 1 November to 1 April of the following year). A total of 4020 questionnaires were distributed this time, with 4000 questionnaires being retrieved and 4000 complete questionnaires. Foreign tourists who come to Jilin Province and local citizens in the province (travelers who travel more than 10 km and travel more than 6 h) are divided into two groups: overnight tourists and day tourists. The survey results are analyzed as follows: Jilin Province's ice and snow tourism source market is dominated by domestic tourists, with the largest number of tourists in the province, followed by Liaoning, Beijing, Heilongjiang, Shandong, Guangdong, Zhejiang, and Shanghai. The entry market is dominated by tourists from South Korea, Russia, Hong Kong, Macao and Taiwan, followed by Japan, Germany, Singapore, Australia, the United States, the United Kingdom, and France. In terms of gender and age, female tourists are fewer than men. The majority are people in their 80 s and 90 s, followed by those in their 40 s and 60 s, and finally students. In terms of occupational composition, private enterprises, foreign-funded enterprises, state-owned enterprises/institutions, and trade staff are the main forces; in terms of travel and stay time, the travel time is mainly concentrated on winter holidays, New Year's Day long holidays and the Golden Week of the Spring Festival. Around 1–2-day surrounding tours are more popular, and the needs of tourists from other provinces for 2–3-day medium and long-term itineraries are increasing year by year; the average stay time of tourists is 2 days, and the stay time in Changchun, Jilin and Yanbian is relatively long. In terms of consumption composition, transportation costs, accommodation costs and food and beverage expenses are concentrated; forms of travel mainly include agency teams and self-help tours, only a small number of tourists are organized by companies or units.

In view of the fact that the ice and snow tourism market in Jilin Province is mainly dominated by Chinese tourists, this paper studies the image perception of ice and snow tourism mainly by Chinese tourists who enter the Jilin Province.

2.2. Research Methodology and Tools

With the popularization of mobile Internet technology, more and more tourists choose to express their travel feelings by posting online travel journals and reviews. Therefore, it is of great significance to research destination image perception to mine these online texts [61]. Text mining is a process of extracting hidden, previously unknown and potentially meaningful patterns from text data in order to discover knowledge [62]. Commonly used text mining tools include the Rost Content Mining software developed by Professor Shenyang of Wuhan University and the Language Technology Platform (LTP) developed by the Social Computing and Information Retrieval Research Center of Harbin Institute of Technology. However, given the fact that there are many punctuation marks and irrelevant words in the content of travel journals and reviews, neither tool is capable of fine cleaning and analysis of text data. Therefore, this paper chooses the Python language, which is a flexible and easy-to-use programming language and has been widely used in data mining, machine learning and other fields in recent years. A large number of high-quality third-party Python modules are available for facilitating the analysis of this study.

2.3. Research Technology Route

The first step is to use Python programming to crawl travel journals and reviews related to the ice-snow tourism in Jilin province from the major travel websites. Secondly, regular expressions were used to extract Chinese content from the original travel journals' text data, and then the jieba module was used for Chinese word segmentation and part-of-speech tagging. After the deactivated words were removed, the statistical results for word-part-of-speech and word frequency were obtained, so as to obtain high-frequency words and their corresponding parts of speech. Then the word cloud module

was used to visualize the word frequency. Programs were written to segment the original travel journal data, to count all pairs of co-occurring words in each sentence and the word frequency matrix of co-occurring words was obtained. Then, the most frequent co-occurring words were extracted from them, and the netwulf module was used to conduct visual analysis of the text co-occurrence network and complete the cognitive image analysis. Thirdly, adjectives frequently used in online travel journals and reviews were used to describe emotional perception. According to the distribution of emotion score, the snownlp module was used to obtain the negative, neutral and positive emotional attitude comment analysis by using the natural break point method, and to do emotional image analysis. The last step was to analyze the perception of the image of ice-snow tourism in Jilin province and put forward conclusions and suggestions on improving the image of ice-snow tourism in Jilin province.

2.4. Data Source

In this paper, the keywords "Jilin" and "snow and ice" were used to search travel journals on the internet and 526 travel journals were selected, with 476 from Ctrip and 50 from mafengwo.cn. However, there are few text data of reviews of ice-snow tourism in Jilin Province. A search on tuniu.com based on the keywords "Jilin" and "snow and ice" found only 842 reviews about the two routes. Eventually, in view of the limited research period and actual review data, 203 travel journals and all 842 review texts from November 2017 to April 2018 and November 2018 to April 2019 were selected. At last, 219,426 words of travel journals and 28,441 words of reviews were obtained after deleting highly viewed repeated reviews of the same account on the same topic, text descriptions containing news, advertisements, business promotions etc., and invalid reviews consisting of photos only without text, and selecting travel journals emphasizing personal feelings, cleaning and removing stop words.

3. Result Analysis

In this paper, the Cognition-Emotion-Overall Image Model proposed by Baloflul is adopted to divide destination image perception into cognitive image, emotional image and overall image. Cognitive image refers to tourists' understanding of the attributes of tourist destinations, while emotional image refers to tourists' feelings and attitudes towards tourist destinations, and the overall image is a combination of the two.

3.1. Analysis of Destination Image Perception

3.1.1. Analysis of Frequency of Frequently Used Words

Python regular expressions were used to extract Chinese content from the online travel journals obtained through network crawling, and the jieba module was used for Chinese word segmentation and part-of-speech tagging. After the stop words were removed, the statistical results of word, part-of-speech and word use frequency were obtained, so as to obtain frequently used words and their corresponding part-of-speech. The jieba module was used for Chinese word segmentation and part-of-speech tagging. Single words were deleted, and 25,648 words were finally outputted. Due to the length of the paper, the top 120 high-frequency words (including only nouns, verbs and adjectives) were extracted. These words reflect tourists' cognition of various elements of the image of ice-snow tourism in Jilin province, as shown in Table 1.

Table 1. Top 120 Most Commonly Used Terms in Network Travel Journals.

No.	Feature Words	Frequency	Part-of-Speech	No.	Feature Words	Frequency	Part-of-Speech	No.	Feature Words	Frequency	Part-of-Speech
1	Rime	2096	Noun	41	Baihe	262	Noun	81	Freedom	172	Adjective
2	Changbai Mountain	1623	Noun	42	Located in	257	Verb	82	Charter	170	Verb
3	Ice and snow	1141	Noun	43	Park	254	Noun	83	Life	170	Verb
4	Harbin	1106	Noun	44	Arrive	253	Verb	84	Yangcao	169	Noun
5	Hometown of snow	1097	Noun	45	Waterfall	251	Noun	85	Scenic spot	169	Noun
6	Northeast China	1021	Noun	46	Ticket	251	Noun	86	Beautiful	168	Noun
7	Skiing	938	Noun	47	Erdao	250	Noun	87	Snow sculpture	167	Noun
8	Hotel	872	Noun	48	Culture	248	Noun	88	Kids	166	Noun
9	Jilin	762	Noun	49	Photography	240	Noun	89	Mobile phone	166	Noun
10	Time	711	Noun	50	Activities	239	Verb	90	Check in	165	Verb
11	China	698	Noun	51	Suggestions	238	Noun	91	Songhua Lake	162	Noun
12	World	673	Noun	52	Manchu	231	Noun	92	North	161	Noun
13	Place	651	Noun	53	Architecture	227	Noun	93	Breakfast	159	Noun
14	Tianchi	627	Noun	54	Special	225	Noun	94	Keep warm	159	Verb
15	Tourism	564	Verb	55	Project	221	Noun	95	Art	159	Noun
16	Changchun	503	Noun	56	Hiking	215	Verb	96	Fun	158	Noun
17	Scenic spot	500	Noun	57	Mountaintop	212	Noun	97	Airplane	158	Noun
18	Snow valley	490	Noun	58	Take photo	209	Verb	98	Ice sculpture	158	Noun
19	Hour	472	Noun	59	Morning	204	Noun	99	More River	158	Noun
20	Hot spring	465	Noun	60	Good	200	Adjective	100	Camera	157	Noun
21	Songhua Lake	456	Noun	61	Subzero	198	Noun	101	Arrangement	156	Verb
22	Ski resort	416	Noun	62	Scenery	197	Noun	102	Photos	154	Noun
23	Depart	413	Verb	63	Price	197	Noun	103	Cold	152	Adjective
24	Feeling	408	Noun	64	Weather	197	Noun	104	Jilin province	151	Noun
25	Jilin City	381	Noun	65	Accumulated snow	196	Noun	105	Driver	150	Noun
26	Cross	379	Verb	66	Ula	195	Noun	106	Xipo	144	Noun
27	Season	367	Noun	67	International	195	Noun	107	Discover	144	Verb
28	Tourists	321	Noun	68	Feeling	194	Verb	108	Beijing	144	Noun
29	Schedule	321	Noun	69	City	192	Noun	109	Ride	144	Verb
30	Experience	311	Noun	70	Wanda	191	Noun	110	Temperature	143	Noun
31	Scenic spot	300	Noun	71	Beipo	187	Noun	111	Rest	142	Verb
32	Airport	299	Noun	72	Street	184	Noun	112	Jingyuetan	141	Noun
33	Accommodation	297	Noun	73	South	183	Noun	113	Train	141	Noun

Table 1. *Cont.*

No.	Feature Words	Frequency	Part-of-Speech	No.	Feature Words	Frequency	Part-of-Speech	No.	Feature Words	Frequency	Part-of-Speech
34	Museum	277	Noun	74	Vocation	183	Verb	114	Delicacies	140	Noun
35	Forest	276	Noun	75	Scenery	182	Noun	115	Snow	140	Noun
36	Friends	275	Noun	76	Recommendation	179	Verb	116	Days	140	Noun
37	Travel	273	Verb	77	Partners	174	Noun	117	Gloves	139	Noun
38	Inn	273	Noun	78	Beautiful scenery	174	Noun	118	Changbai	138	Noun
39	Devildom	263	Noun	79	History	173	Noun	119	Return	137	Verb
40	Resort area	262	Noun	80	Appreciation	172	Verb	120	Luggage	136	Noun

As can be seen from Table 1, "rime", "Changbai Mountain", "snow and ice", "Harbin", "hometown of snow" and "Northeast China" have the highest frequency, occurring between 2092 and 1021 times, and are things attractive to tourists. Terms including scenic spots, tourism resources, place names and locations are the most intense part of tourists' image perception, leaving a deep impression on tourists. These attractions have become the first choice for tourists to visit Jilin province for snow and ice. The frequent use of such terms as "Harbin" and "hometown of snow" indicates that tourists often compare the ice-snow tourism in Jilin province with that in Heilongjiang province. Meanwhile, the frequency of terms on travel accommodation and travel transportation such as "hotel", "time" and "hours" is between 872 and 472, which reflects that tourists pay close attention to the accommodation and transportation elements related to scenic spots. Frequency of terms on tourist activities such as "hot springs", "ski resorts", "museums", "forests", "resorts" and "waterfalls" is average, between 465 and 251. These tourist activities are very attractive to tourists and are an important part of perception of the image of ice-snow tourism. In addition, frequency of terms on travel services and features such as "tickets", "culture", "Manchu", "architecture", "features", "prices", "ula", "international", "history" and "art" is between 251 and 159, which is the most important part of tourist perceptions. Frequency of nouns related to tourism climate is between 152 and 138, which is the most direct perception of the image of the ice-snow tourism in Jilin province.

3.1.2. Analysis of Classification of Frequently Used Terms

Based on analysis of the frequency of the above frequently used terms, the perception of image of ice-snow tourism in Jilin province was classified into five main categories and 12 sub-categories based on the 120 extracted frequently used terms. The five main categories were tourism attractions, tourism activities, tourism facilities, tourism features and tourism service environment. The 12 sub-dimensions were: tourist sites, scenic spots, tourist resources, ice-snow themed activities, auxiliary themed activities, tourist accommodation, tourism transportation, tourism catering, ethnic, international, art and culture, tourism services, tourism climate conditions, etc. According to the classification system, frequently used terms unrelated to the primary and secondary categories were excluded, and the selected feature words were classified into the corresponding primary and secondary categories, forming a statistical table classifying frequently used terms (see Table 2).

As can be seen from Table 2, of the main categories, tourism attractions have the highest frequency, followed by tourism facilities and tourism services. Among the sub-categories, tourists pay the closest attention to scenic spots, followed by tourist destinations, tourism transportation, tourism resources, ice and snow themed activities, tourism accommodation, auxiliary themed activities, internationalization, tourism services, art and culture, and tourism catering. Seen in this light, scenic spots are the most important part of tourists' perception of the destination image, and tourists pay close attention to tourist destinations, tourism transportation, tourism resources, and ice-snow themed activities.

Table 2. Classification of Frequently Used Terms in the Perception of Image of Ice–Snow Tourism in Jilin Province.

Main Categories (Frequency/Percentage)	Sub-Categories (Frequency/Percentage)	Frequently Used Terms
Tourism attractions (13027/53.19%)	Tourist sites (4136/31.75%)	Northeast China, Jilin, Changchun, Jilin City, Baihe, Erdao, South, Jilin Province, North, Beijing, Changbai
	Tourist attractions (6673/51.22%)	Rime, Changbai Mountain, Tianchi, Songhua River, museum, devildom, park, waterfall, Beipo, Wanda, Songhua Lake, Xipo, Jingyuetan
	Tourist resources (2218/17.03%)	Snow, hot spring, forest, accumulated snow, snow
Tourism activities (3117/12.73%)	Snow and ice themed activities (1679/53.87%)	Skiing, ski resorts, snow sculptures, ice sculptures
	Auxiliary themed activities (1438/46.13%)	Traversing, photography, hiking, mountaintop, photo taking, vacation
Tourism facilities (4160/16.99%)	Tourist accommodation (1607/38.63%)	Hotel, lodging, inn, check in
	Tourism transportation (2254/54.18%)	Time, hour, airport, chartered bus, plane, driver, ride, train
	Tourism catering (299/7.19%)	Breakfast, delicious food
Tourism features (2153/8.79%)	Ethnic (231/10.73%)	Manchu
	International (1063/49.37%)	World, ula, international
	Art and culture (859/39.80%)	Culture, characteristics, art, architecture
Tourism service environment. (2033/8.30%)	Tourism services (1045/51.40%)	Schedule, tickets, prices, days, luggage
	Tourism climate conditions (988/48.60%)	Subzero, weather, cold, temperature, warmth, gloves

3.1.3. Word Cloud of Frequent Elements (Word Cloud)

The wordcloud module was used to visualize the word frequency, and the Chinese word segmentation results of ice-snow tourism in Jilin province were visualized in the form of a word cloud. The higher the word frequency, the larger the font size, and vice versa (see Figure 1). The word cloud clearly shows the frequent terms in tourists' perception of the destination image of ice-snow tourism in Jilin province. Tourists are most impressed by the scenic spot and landscape of "Changbai mountain and rime", which were compared with the hometown of snow. Tourists are most impressed by the tourist destination of Jilin and compare it with Harbin, followed by tourist resources and activities such as snow, ice and skiing. They also pay close attention to "hotel", "time", "Tianchi", "Changchun" and other tourist accommodation, transportation and other tourist sites and scenic spots.

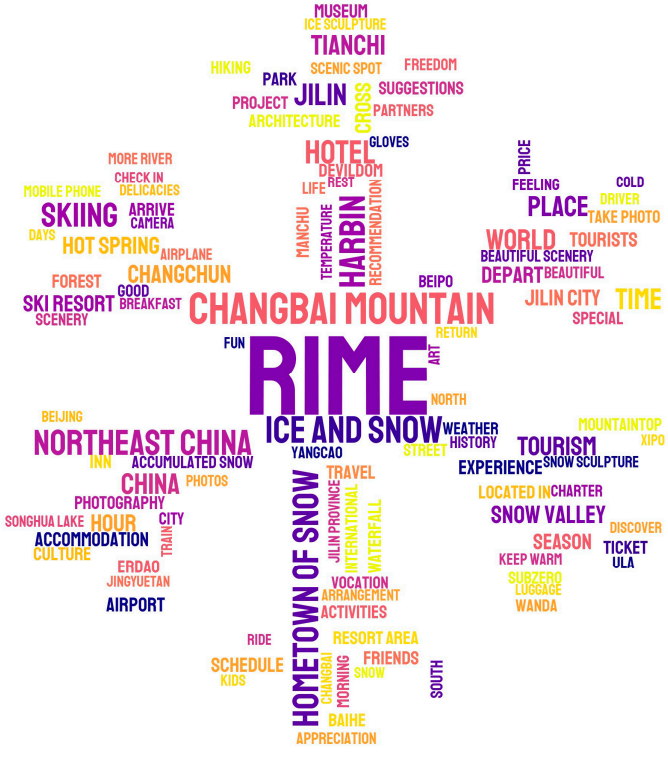


Figure 1. Wordcloud of Ice-Snow Tourism in Jilin Province.

3.1.4. Analysis of the Co-occurrence Network of Frequent Elements

A program was written to segment the original travel journal data and to count all pairs of co-occurring words in each sentence, and the word frequency matrix of co-occurring words was obtained. Then, the most frequent co-occurrence words were extracted from them, and the netwulf module was used to visualize the text co-occurrence network. The direction of the line linking two frequently used terms indicates the semantic connection between the two words, and the length of the lines indicates the degree of semantic connection between the two words, so as to analyze the underlying meaning of the frequently used terms.

It can be seen that the results in Figure 2 are the same as those of the classification of frequently used words and word cloud analysis. "Changbai Mountain" and "rime" are the core words of the network of co-occurring frequently used words, which are the most profound impressions and memories in

tourists' perceptions, and the most important factor of winter ice-snow tourism in Jilin. These core words form a network of internal and external connections. Co-occurring words such as "Changbai Mountain—skiing", "Changbai Mountain—Tianchi", "Changbai Mountain—hotels", "Changbai Mountain—hot spring", "rime—enjoy" and "rime—Jilin city" have a high degree of connection. This is a further cognition of the evaluation subject, which is in the sub-core position, indicating that tourists are most impressed by the core resource elements of the scenic spot and the city where the scenic spot is located. These are the most prominent representatives of Jilin ice-snow tourism, with brand characteristics. Peripheral terms include "ski resort," "Wanda," "Changchun," "snow and ice tourism", and all other high-frequency words involving the specific tourism sites and range of reputation, and they provide further expansion and enrichment of the core words and sub-core words.

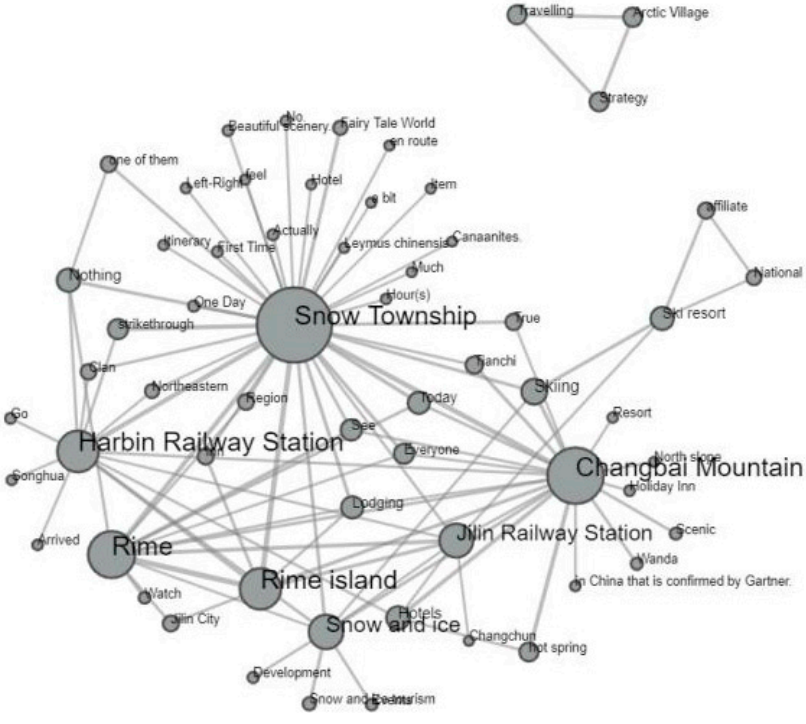


Figure 2. Diagram of Network of Frequently Co-occurring Words in Ice-Snow Tourism in Jilin Province.

3.2. Analysis of Emotional Image

Tourists' perception of emotional image is an important part of Jilin's ice and snow tourism image perception. Emotion image perception can be characterized from two dimensions: emotion perception and emotion attitude.

3.2.1. Emotional Perception

Emotional perception is described by the top 30 high-frequency adjectives screened in online travel notes and reviews, as shown in Table 3.

Table 3. Top 30 High-Frequency Adjectives in Samples of Web Travel and Reviews.

Sequence	Word	Frequency	No.	Word	Frequency	Sequence	Word	Frequency
1	Good	200	11	Beautiful	81	21	Mysterious	57
2	Free	172	12	Full	73	22	Exactly	56
3	Cold	152	13	Regret	73	23	Magical	53
4	Most Beautiful	136	14	Warm	72	24	Happy	53
5	Magnificent	120	15	Perfect	69	25	Crystal	51
6	Cheap	113	16	Simple	65	26	Fun	49
7	Unique	99	17	Freezing	65	27	Exquisite	47
8	Famous	98	18	Look-looking	62	28	Lively	44
9	Big	87	19	Warm	59	29	Comfortable	41
10	Gorgeous	86	20	Luck	58	30	Cozy	39

From Table 3 and crawled web texts, we can see that "nice, comfortable, and cozy" is the emotional perception of hotel accommodation and the overall itinerary; "free" is the emotional experience of the emerging travel style of "free journey" in Changbai Mountain; "cold and freezing" is an emotional perception of the tourist climate; "most beautiful", "magnificent", "unique", "famous", "big" and "gorgeous" are tourists' emotional experience of Changbai Mountain, Tianchi, and misty scenery, emotional cognition in "Smart Choice" accommodation, package price and transportation. "Beautiful" is a compliment to moments of playing in snow, happy holidays and overall travel memories; "full" is an emotional interpretation of the oriental charm of Changbai Mountain, the mystery of snow, the childlike snow fun, and the joy of playing in snow; snow means the beautiful Tianchi, and the emotional perception evaluation caused by the lateness of the tourist traffic plane. "Perfect" is a fake "perfect" emotional experience that tourists expect from a richer breakfast and a guided tour throughout the Tianchi service; "simple" is a "simplified" emotional sigh for the convenience of free online ticket purchase, skiing, and experience. "Warm" is an emotional experience of hot springs, hotel service environment, and the rustic and hospitable folk customs of Northeast China; "pretty" is the praise of snow, scenery, blue sky, Baiyun and Changbai Mountain Tianchi, and the small town, and the emotional experience of seeing the Changbai Mountain Tianchi, wild animals and the misty mist of Jilin. "Mysterious" is the tourist's emotional experience of Changbai Mountain, Tianchi, Wuluo Island, Beauty Songyuan Park, and Songhua Lake winter fishing and hunting culture and ancient techniques; "indeed" expresses the emotional experience of tourists choosing the right time to see the beauty. It is an emotional expression that is worth visiting after experiencing Changbai Mountain; "crystal" is the description of snow, fog, and ice sculpture; "fun" is the emotional expression of skiing, skating, winter fishing and adventure crossing.

3.2.2. Emotional Attitude

The snownlp module was used to analyze the emotion in the reviews, and according to the distribution of emotion score, the natural break point method was used to obtain the negative, neutral and positive review data, as shown in Table 4.

Table 4. Statistical Table of Emotional Attitude.

Emotional Types	Frequency (Time/Percentage)	Segmented Statistical	Frequency (Time)	Percentage
Positive emotion	278/67.31%	Low (0.797–0.893)	29	7%
		Middle (0.893–0.97)	45	10.86%
		High (>0.97)	204	49.3%
Neutral emotion	48/11.62%		48	11.83%
Passive emotion	87/21.07%	High (0–0.058)	57	13.8%
		Middle (0.058–0.165)	12	2.9%
		Low (>0.165)	18	4.35%

As can be seen from the reviews in Tables 4 and 5, positive emotions account for a large proportion (67.13%), and high-degree positive emotional expressions accounted for 204 reviews (49.3%). These positive reviews use the adjectives "so beautiful", "very beautiful", "first-class," and "best", used by tourists when describing the scenic spot Changbai Mountain and Tianchi snow scenery, hotel accommodation and skiing facilities. Negative reviews only accounted for 21.07%, and the frequency of high-degree negative emotional expressions was 57 (13.8%). Negative evaluation refers to insufficient breakfast, hotel accommodation without sound insulation, geographical location, inconvenient transportation and high cost. However, it should be pointed out that the emotional attitude is affected by tourists' degree of satisfaction in travel consumption needs, knowledge and experience, personality characteristics and group, and so on, and there are subjective differences in the experience evaluation of the same thing by different tourists. Therefore, in practice, it is necessary to specifically analyze the influencing factors of the formation of tourists' emotional attitude and make targeted changes.

3.3. Overall Image

The overall image is tourists' evaluation of the overall image perception after they separately evaluate the "eating, living, traveling, shopping and entertainment" of the ice-snow tourism in Jilin province. The words selected from frequently used words such as "spectacle," "most beautiful," "spectacular," "magical," "shocking," "scenic spot," and "not bad" are tourists' overall evaluation of the ice-snow tourism in Jilin province. This shows that ice-snow tourism in Jilin province has formed its own characteristics and generated brand effect. Tourists use positive emotion words such as "recommend merit," "worthwhile trip," "satisfied," and valuable" to describe the overall image of ice-snow tourism in Jilin province. Only a small number of negative words were used by tourists to describe the image of ice-snow tourism in Jilin province such as "pity" and "bad" in relation to accommodation, catering, transportation and service management. Among the frequently used terms, "next time" indicates tourists' willingness to visit again, but its frequency is relatively low, indicating that the structure of ice-snow tourism products in Jilin province is single and lacks a continuous attraction.

Table 5. Representative Positive and Negative Evaluation Texts.

Positive Evaluation	Negative Evaluation
Tianchi, skiing is worth visiting, beautiful and breathtaking.	The hotel is a little better than the express hotel, the sound insulation is poor, you can clearly hear the speech in the corridor when lying on the bed. There are few types of breakfast. The least favorite is the airline of Juneyao Airlines. It also has to stop in Tianjin, wasting time and energy! Of course, if you just go for skiing, the price/performance ratio of the Smart Choice Hotel is still okay. The hotel is only 5 minutes' walk from the ski resort, and the snow card process is now very convenient.
The Changbai Mountain is so beautiful. It's a great time to go to the cold snow while enjoying the snow while enjoying the cold snow. Unfortunately, we did not see Tianchi in heavy snow, so we will go there next time. The ski slopes are good, but you must be able to get hit or you won't master it.	During the Chinese New Year, take children to travel, the overall itinerary is mainly skiing, the hotel with a good snow slope is slightly bad for breakfast.
Very nice, the snow is beautiful and skiing is even cooler. Convenient transportation, prices in the scenic area are naturally higher than normal, but almost the same as in Nanjing. Great! The only drawback is: I set off from Nanjing at noon to Changbai Mountain Hotel (Nanjing has no direct flight to Baishan)	The plane arrives at 9 pm and the return trip is 11 noon. This arrangement is almost a waste of the whole day, and the schedule is not very satisfactory. Hyatt's restaurant is good, the resort has KFC McDonald's Pizza Hut, so you can also eat anything if you are tired. The traffic is not very good. The location of the resort should be relatively biased; it is not convenient to go anywhere. We reported a group to Changbai Mountain locally; the overall feeling is OK; the feeling of skiing is great.
The flight time is good. The Park Hyatt Hotel is first-rate in the Wanda Resort. The rooms are very large, the facilities are very advanced, and they are very comfortable and worth a stay. The ski resort facilities are also the best in the country. The ski tracks and hardware facilities are all excellent. Although it is a beginner, please hire a coach, but you can get started soon.	Changbai Mountain is free of charge in winter, but you need to take a scenic bus + ascent to the off-road vehicle, and you have to spend a lot. It is found that the trip to the Northeast is a big traffic, and accommodation and meals are not expensive.

4. Conclusions and Discussions

4.1. Conclusions

In this paper, python technology was used to crawl, analyze and mine travel journals and reviews on famous travel portals like the OTA (Online Travel Agent) and UGC (User Generated Content) platforms related to ice-snow tourism in Jilin province to explore the perception of the image of ice-snow tourism in Jilin province. The following conclusions are drawn:

4.1.1. Cognitive Image

The perception of the image of ice-snow tourism in Jilin province can be classified into five main categories and 12 sub-categories according to the 120 extracted frequently used terms. The five main categories were tourism attractions, tourism activities, tourism facilities, tourism features and tourism service environment. The 12 sub-dimensions were: tourist sites, scenic spots, tourist resources, snow and ice themed activities, auxiliary themed activities, tourist accommodation, tourism transportation, tourism catering, ethnic, international, art and culture, tourism services, tourism climate conditions, etc. Among the five main categories, tourism attractions have the highest frequency, followed by tourism facilities and tourism service environment. According to the word

cloud and text co-occurrence visualization network, "Changbai Mountain" and "rime" are core words, indicating that tourists are most impressed by "Changbai Mountain" and "rime".

4.1.2. Emotional Image

As for the emotional image of ice-snow tourism in Jilin province, positive emotional expressions account for 67.31%, neutral emotional expressions 11.62%, and negative emotional expressions 21.07%. These positive evaluations include the positive adjectives "so beautiful", "very beautiful", "first-class", and "best" used by tourists when describing the scenic spot Changbai Mountain and Tianchi snow scenery and skiing facilities. Negative evaluations refer to inconvenient transportation and unsatisfying accommodation and catering.

4.1.3. General Image

Tourists tend to speak highly of the image of ice-snow tourism in Jilin province but few intend to revisit. Tourists used words such as "spectacle," "most beautiful," "spectacular," and "not bad" in evaluating ice-snow tourism in Jilin province. Tourists use positive emotion words such as "recommend" and "worthwhile trip" to describe the overall image of ice-snow tourism in Jilin province. Tourists used negative terms such as "pity" and "bad" in relation to accommodation, catering, transportation and service management. In addition, the low frequency of the term "next time" indicates tourists' willingness to revisit is low.

4.2. Suggestions

The following suggestions are proposed based on the above conclusions to improve the image of ice-snow tourism in Jilin province.

4.2.1. Develop Special Ice-Snow Tourism Products and Embark on a Road of "Transformation and Upgrading"

In view of the classification of perception of the image of ice-snow tourism in Jilin province, the four product systems of ice-snow leisure vacation, ice-snow hot spring health care, ice-snow sightseeing culture and ice-snow folk experience should be further improved, with attention paid to tourist sites, scenic spots, tourist resources, snow and ice themed activities, auxiliary themed activities, tourist accommodation, tourism transportation, tourism catering, ethnic, international, art and culture, tourism services and tourism climate conditions. The core Jilin ice-snow tourism product portfolio features long-term and short-term "in-depth ice playing," "thick snow entertainment," "warm hot springs," "heat" and "folk customs" to meet diverse demands, extend the industrial chain and improve supporting services. These measures aim to create a richer, more comfortable and more concentrated ice-snow tourism experience through transformation and upgrading, meet the increasingly diversified travel needs of tourists, and improve the willingness of tourists to visit again.

4.2.2. Improve the Brand Effect of Ice-Snow Tourism and Embark on a Road of "Multi-Elemental Creative Marketing"

The marketing strategy of the ice-snow brand should be implemented, with a focus on cultivating four ice-snow tourism brands: ice-snow Changbai Mountain, fishing and hunting Chagan Lake and fairy tale rime island. It is suggested to make use of the advantageous resources of ice-snow tourism in Jilin province, to carefully plan the products of the ice-snow tourism festival, improve and enhance the Jingyue Wasa International Cross-country Skiing Festival, expand its influence, and make it an international ice-snow tourism event. In addition, the "national ice-snow season" should be continued to cultivate the ice-snow market. The destination image of Jilin shall be established, and the brand influence of "Jilin ice-snow tourism" developed in collaboration with major media and network resources for extensive publicity and influential mass media. Propaganda utilizing CCTV's golden advertising time, airports, railways and other must-bypass traffic scenes should be

continued. It is necessary to guide and encourage tourism enterprises to utilize the mainstream network platforms to establish the Internet platform marketing system of ice-snow tourism marketing.

4.2.3. Build a Brand-New Ice-Snow Tourism Environment and Optimize Services

For an optimized and attentive service, it is necessary to design a whole set of travel experiences that take into consideration tourists' diversified needs for food, accommodation, travel, shopping and entertainment. An ice-snow theme hotel should be built, so that tourists in the ice and snow can enjoy a variety of unexpected and novel accommodation. There should be enrichment of food types, characteristic catering, accommodation and travel products for different periods. The business district, accommodation area and entertainment area should be properly planned around the snow field to integrate the commercial street, food street and entertainment area facilities. A special ice and snow food bar or food stall should be set up to sell special food or snacks of northeast China and enjoyable gourmet dinners. Tourist souvenirs with northeast characteristics should be developed to integrate local customs, historical culture and folk customs into tourist souvenirs. The indoor temperature of the ski resort should be raised, sufficient rest facilities should be provided, panoramic pictures of the ski resort should be improved, ski guide signs should be perfected, and the parking lots, tourist centers, hot springs and facilities for swimming should be updated. Infrastructure such as transportation and entertainment should be constantly improved. Transportation and entertainment infrastructures should be constantly improved and regulatory platforms should be established to achieve full-coverage supervision.

4.2.4. Improve the Safety Mechanism for Ice-Snow Tourism to Ensure Safety

"Exciting and thrilling" refers to the emotional experience of snow and ice tourists in terms of tourism safety. Safe ice-snow tourism is the focus of continuing attention and the construction of ice-snow tourism in Jilin province. Therefore, tourists' safety awareness should be enhanced, and business operators should attach great importance to the safety warning of tourists, so as to strengthen tourists' attention to the safety of themselves and others. It is necessary to improve the guarantee level of ice and snow sports and ice and snow tourism facilities and improve the preventive mechanisms of facility maintenance, management and investigation. Measures should be taken to actively guide related industries such as medical care, insurance, education, science and technology, and equipment to participate in promoting the snow and ice industry and establish a snow and ice safety system. We should improve the safety and development of the market, continue to maintain the order of the ice and snow tourism market, crack down on violations in accordance with the law, and promote the establishment of a comprehensive regulatory system for the modern ice and snow tourism market. The level of market security and development should be improved to continuously maintain the order of the ice and snow tourism market, and more efforts should be made to crack down on illegal behaviors in accordance with the law, so as to promote the establishment of a comprehensive regulatory system for the modern ice and snow tourism market.

4.3. Discussion

The perception of tourism image is of great significance to the image formation of tourism destinations and the sustainable development of tourism. Research on the perception of ice and snow tourism image, especially the specific promotion strategy of tourism image perception, is of great significance for the formation of ice and snow tourism image and the high-quality development of ice and snow tourism in Jilin Province, and even further for improving the overall tourism ice and snow image and high-quality development of tourism in China.

This research has theoretically opened up a new field of tourism image perception, which has played a role in attracting scholars to study different types of tourism image perception issues. It opens up a new perspective on ice and snow tourism, and introduces image perception into the field of ice and snow tourism instead of traditional ice and snow resources and snow sports. It realizes the

perfect combination of new liberal arts and new engineering, uses the artificial intelligence big data mining technology of engineering, studies cognitive image and emotional image in psychology from the perspective of consumer behavior, and uses product management and marketing of management science. Additionally, management theories such as brand philosophy and safety management have put forward specific solutions. Ultimately, it has promoted the development of the industrial economy in the region of ice and snow tourism and ice and snow economy in Jilin Province, realized the cross fusion of disciplines such as engineering, psychology, management, and economics, and created new research directions and fields.

However, this study has certain limitations. First, there are deviations in the network text data analysis. According to the report of iUserTracker, among the users who book travel products online, users aged 19–35 account for up to 79%, and users with college degrees or above account for up to 84%. However, OTA online travel review users are not representative of the overall travel audience. Some social groups make less use of online travel services. The second issue is in the segmentation and comparison of image perception multidimensional data. This paper does not subdivide the ice and snow tourism image perception from the perspective of users. In addition, this study considers the impact of external environmental factors on the perception of snow and the snow tourism image in the future. Also, the image-aware text analysis method is not detailed enough. With the combination of the Python language and text mining method, although it is possible to accurately find high-frequency words and visualize word clouds and co-occurrence networks, there is still a problem that classification studies are not performed by topic and type, and the corpus data is not accurate enough in sentiment analysis.

In the future, research will continue to be based on Python technology. With the multi-dimensionalization of network text data, it will further subdivide research users. This paper conducts targeted image perception research around users of different genders, ages, nationalities, and regions, so as to more accurately propose strategies for improving the image perception of snow and ice tourism in different market segments, and to realize the comprehensive and high-quality development of ice and snow tourism. At the same time, with the approach of the 2022 Winter Olympics, tourists from all over the world will come to China to experience China's ice and snow. At that time, it will be possible to study the perception of ice and snow tourism from the perspective of inbound tourists. The status of tourism power has important practical significance. In addition, whether external environmental factors such as global warming, haze and the new coronavirus, and the full passenger flow on holidays, will have a significant negative impact on the perception of ice and snow tourism image is also a question for further discussion. Finally, regarding the research method, in future research, methods like LDA, BTM and other topic models can be used to extract different topics in text data, and draw word cloud diagrams to study the frequency of words more specifically. In terms of sentiment analysis, corpus data that better fits the travel scene can be used in the future. In addition, more advanced deep learning methods such as Bi-LSTM will be used to improve the model, to obtain real-time text data from the Internet, to improve research on the image perception of snow and ice tourism in real time, and to promote the sustainable development of snow and ice tourism.

Funding: This research received the Social Science Fund Project of Jilin Province (A Study into the Deep Development of Ice and Snow Tourism Products in Jilin Province, No. 2018B75).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Hunt, J.D. Image as a Factor in Tourism Development. *J. Travel Res.* **1975**, *13*. [CrossRef]
2. Gunn, C. *Vacation Scape: Designing Tourist Regions*; AustinzBureau of Business Research; University of Texas: Austin, TX, USA, 1972; pp. 445–447.
3. Kotler, P.; Barich, H. A Framework for Marketing Image Management. *Sloan Manag. Rev.* **1991**, *3232*, 94–104.
4. Fakeye, P.; Crompton, J. Image Differences between Prospective, First-Time, and Repeat Visitors to the Lower Rio Grande Valley. *J. Travel Res.* **1991**, *2929*, 10–16. [CrossRef]

5. Gartner, W.C. Image Formation Process. *J. Travel Tour. Mark.* **1993**, *2*, 199–212. [CrossRef]
6. Selby, M.; Morgan, N.J. Reconstruing place image: A case study of its role in destination market research. *Tour. Manag.* **1996**, *1717*, 287–294. [CrossRef]
7. Baloglu, S.; McCleary, K.W. A Model of Destination Image Formation. *Ann. Tour. Res.* **1999**, *2626*, 868–897. [CrossRef]
8. Tapachai, N.; Waryszak, R. An Examination of the Role of Beneficial Image in Tourism Destination Choice. *J. Travel Res.* **2000**, *39*, 37–44. [CrossRef]
9. Beerli, A.; Martin, J.D. Factors influencing Destination Image. *Ann. Tour. Res.* **2004**, *31*, 657–681. [CrossRef]
10. Wang, C.Y.; Hsu, M.K. The relationships of destination image, satisfaction, and behavioral intentions: An integrated model. *J. Travel Tour. Mark.* **2010**, *27*, 829–843. [CrossRef]
11. Xie, C.W.; Huang, Y.S. On the Participatory Organization Model of Tourism Destination Image Planning. *Tour. J.* **2002**, *1717*, 63–67.
12. Qi, H.X.; Cai, Y.L.; Wei, X. Regional Tourism Image Construction and Landscape Planning: A Case Study of Linhai City. *Chin. J. Ecol.* **2003**, *2222*, 84–88.
13. Li, X. Research on the Application of Unstructured Measurement of Perceived Image of Tourist Destinations—Taking Image Perception of Business Tourists to Australia as an Example. *Tour. Trib.* **2011**, *26*, 57–63.
14. Wu, J.F. Tourist Destination Image "Puzzle" and Evaluation Method. *J. Shaanxi Norm. Univ. (Nat. Sci. Ed.)* **2014**, *4242*, 85–93.
15. Bai, D.; Ma, Y.F.; Liu, J.S. Research on Tourists' Perception Evaluation of World Heritage Tourist Sites Based on Grounded Theory—Taking Terracotta Warriors and Horses Scenic Spot of Qin Shihuang as an Example. *J. Arid Land Resour. Environ.* **2016**, *3030*, 198–203.
16. Li, P.; Chen, T.; Wang, F.Y.; Wang, X.G. Research on Image Perception of Urban Tourism Communities Based on Text Mining—Taking Beijing as an Example. *Geogr. Res.* **2017**, *3636*, 1106–1122.
17. Feng, Q.; Tian, Y.J.; Sun, G.N. Study on Image Perception of Tourism Destinations in Shaanxi Based on Online Travel Travels—Taking the Eight Major 5A Tourist Attractions in Shaanxi Province as Examples. *Resour. Dev. Mark.* **2018**, *3434*, 1623–1628.
18. Lu, L.J.; Liao, X.P. Study on Image Perception of Hengshan Tourist Destination in Nanyue Based on UGC Data. *Econ. Geogr.* **2019**, *39*, 221–229.
19. Tu, W.H.; Yue, J.; Dai, X.Y. Research on Tourism Image Perception in Fangshan District of Beijing—Based on the Perspective of Web Text Analysis. *Manag. Manag.* **2020**, *1*, 132–138.
20. Mayo, E.J. Regional images and regional travel behavior. In Proceedings of the fourth annual Conference Travel Research Association Research for Changing Travel Patterns: Interpretation and Utilization, Sun Valley, 12–15 August 1973; pp. 211–218.
21. Stable, M.J. The image of destination regions: Theoretical and empirical aspects. In *Marketing in the Tourism Industry: The Promotion Destination*; Regions, B., Goodall, G., Ashworth, Eds.; CroomHelm: London, UK, 1988; pp. 133–161.
22. Baloglu, S.; Brinbg, D. Affective Images of Tourism Destination. *Travel Res.* **1997**, *35*, 11–15. [CrossRef]
23. Hanlan, J.; Kelly, S. Image formation, information sources and an iconic Australian tourist destination. *J. Vacat. Mark.* **2005**, *11*, 163–177. [CrossRef]
24. Jenkins, O.H. Understanding and Measuring Tourist Destination Images. *Tour. Res.* **1999**, *1*, 1–15. [CrossRef]
25. Stylidis, D.; Shani, A. Testing an integrated destination image model across residents and tourists. *Tour. Manag.* **2017**, *58*, 184–195. [CrossRef]
26. Cheng, J.L.; Wang, F.Z. Influential Factors and Shaping Strategies of Tourism Image. *Econ. Geogr.* **2009**, *2929*, 1753–1758.
27. Cheng, W.; Sui, L.N. Research on Tourism Image Perception Model and Its Application—Taking Residents of the Yangtze River Delta to Perceive Korean Tourism Image as an Example. *Tour. Sci.* **2007**, *26*, 7–12.
28. Gan, L.; Lu, T.L.; Wang, X.H. An Empirical Study of Domestic Tibetan Tourists' Perception of Tibet Tourism Image. *Tour. Sci.* **2013**, *32*, 73–82.
29. Zheng, P. A Study on the Influential Factors and Differences of the Overall Image Identity of Tourist Destinations—Taking the Domestic Market of Zhengzhou as an Example. *J. Arid Land Resour. Environ.* **2014**, *27*, 200–204.

30. Zhang, H.M.; Lu, L.; Zhang, J.H. An Analysis of the Impact of Perceived Distance on the Image of Tourist Destinations—Taking Tourists from Five Major Tourist Sources to Perceive Suzhou Zhouzhuang's Tourist Image as an Example. *Hum. Geogr.* **2006**, *20*, 25–30.
31. Tang, Y.F.; Zhang, H.M. A Comparative Study of Destination Space Imagery Before and After Tourism—Taking Chinese Tourists in Korea as an Example. *Areal Res. Dev.* **2018**, *1*, 103–109.
32. Jeong, G.C.; Tamara, T.; Shomir, S. On the destination image of Korea by Russian tourists. *Tour. Manag.* **2009**, *3232*, 193–194.
33. Steven, P. Destination image: Identifying baseline perceptions of Brazil, Argentina and Chile in the nascent Australian long haul travel market. *J. Destin. Mark. Manag.* **2015**, *55*, 164–170.
34. Eran, K. Destination image restoration on facebook: The case study of Nepal's Gurkha Earthquake. *J. Hosp. Tour. Manag.* **2016**, *28*, 66–72.
35. Shan, L.Y.; Wu, J. Image of Qingdao tourist destination based on online travel notes. *J. Qufu Norm. Univ.* **2019**, *45*, 88–93.
36. Yang, J.; Hu, P.; Yuan, B.H. Research on the Impact of Familiarity on Tourism Image Perception Behavior: A Case Study of Chongqing Citizens' Perception of Shanghai Tourism Image. *Tour. Trib.* **2009**, *24*, 56–60.
37. Zhu, C.L.; Hou, Z.Q. Tourism Destination Image Perception Based on Internet Word of Mouth: A Case Study of Xiamen City. *Trop. Geogr.* **2013**, *33*, 489–495.
38. Shi, K.B.; Yang, Y.C. Evaluation of Tourism Motivation and Tourism Image of Mainland College Students in Taiwan. *Resour. Sci.* **2015**, *37*, 593–597.
39. Yang, M.; Li, X.Y. Research on the perception of Xi'an tourism image based on Weibo data analysis. *J. Qufu Norm. Univ.* **2017**, *11*, 88–92.
40. Zhang, G.J.; Li, J.Y.; Zhang, L. Research on Tourism Image Perception of Huashan Scenic Area: A Text Analysis Based on Tourists' Web Logs. *Tour. Sci.* **2011**, *25*, 87–94.
41. Pi, R.; Zheng, P. "Online Review of Shaolin": Study on Tourism Cognition, Emotion and Overall Image of Shaolin Temple. *J. Arid Land Resour. Environ.* **2017**, *31*, 201–207.
42. Zhang, W.T.; Luo, P.C. A Comparative Study on Tourists' Perception and Official Communication of Destination Tourism Image Based on Web Texts: Taking the Building of Yongding, Fujian. *J. Fujian Norm. Univ. (Nat. Sci. Ed.)* **2017**, *33*, 90–98.
43. Xu, Y.Y.; Yao, G.Y. Study on Tourism Image Perception of Huangshan Scenic Area Based on Online Reviews. *Res. World Geogr.* **2016**, *25*, 158–168.
44. Zhang, H.; Cho, T.; Wang, H. The Impact of a Terminal High Altitude Area Defense Incident on Tourism Risk Perception and Attitude Change of Chinese Tourists Traveling to South Korea. *Sustainability* **2019**, *12*, 7. [CrossRef]
45. Kladou, S.; Mavragani, E. Assessing destination image: An online marketing approach and the case of Trip Advisor. *J. Destin. Mark. Manag.* **2015**, *44*, 187–193.
46. Hunter, W.C. The social construction of tourism online destination image: A comparative semiotic analysis of the visual representation of Seoul. *Tour. Manag.* **2016**, *54*, 221–229. [CrossRef]
47. Park, J.; Morrison, A.M.; Wu, B.; Kong, Y. Korean Golf Tourism in China: Place, Perception and Narratives. *Sustainability* **2018**, *10*, 1055. [CrossRef]
48. Peng, H.J.; Huang, C.Q.; Zhou, L.X. Research on Tourists' Perception of Multiple Tourism Image Positioning in Mountain-type Scenic Areas: Taking Nanyue Hengshan as an Example. *Tour. Forum* **2016**, *9*, 21–27.
49. Li, Y.F.; Li, L.T. A Comparative Study on Perception of Tourism Image in Desert-type Scenic Spots: Taking Shapotou and Shahu Scenic Spots in Ningxia as Examples. *Soc. Sci. Ningxia* **2016**, *4*, 128–133.
50. Zhang, Z.Z.; Li, J.Y. Comparison of questionnaire survey and web text data in tourism image research: A case study of tourism image perception in Xi'an. *Tour. Sci.* **2014**, *28*, 73–81.
51. Zhang, R.; Zhang, J.G. Research on tourism image perception of Shanghai chenshan botanical garden based on network text and IPA model analysis. *Chin. Landsc. Archit.* 2019; 35, 83–87.
52. Flagestad, A.; Hope, C.A. Strategic Success in Winter Sports Destinations: A Sustainable Value Creation Perspective. *Tour. Manag.* **2001**, *2222*, 445–461. [CrossRef]
53. Zhang, Y. Analysis and Prediction of the Total Number of Harbin Ice-Snow Tourism Based on Times Series. *Adv. Intell. Soft Comput.* **2012**, *115*, 495–501.
54. Yang, J.; Yang, R.; Sun, J.; Huang, T.; Ge, Q. The Spatial Differentiation of the Suitability of Ice-Snow Tourist Destinations Based on a Comprehensive Evaluation Model in China. *Sustainability* **2017**, *9*, 774. [CrossRef]

55. Dong, X. *Research on the Operation of Jilin Ice and Snow Tourism Market*; Huaqiao University: Quanzhou, China, 2008.
56. Song, H.J.; Yu, H.X. Survey on the potential tourist market of Yabuli Ski Resort. *China For. Econ.* **2007**, *3*, 42–45.
57. Xu, Y.J.; Wei, Y.P. Analysis on the Status and Position of Harbin Ice and Snow Tourist Source Market. *China High-tech Enterp.* **2007**, *3*, 13.
58. Wang, H.R. Research on Tourism Image Design of Heilongjiang Province to Russia. *J. Suihua Univ.* **2016**, *9*, 35–38.
59. Han, Z.K.; Tang, Z.Z. Analysis of Harbin Ice and Snow Tourism Image Promotion Mechanism. *Design* **2013**, *12*, 159–160.
60. Liang, S. *Study on Tourism Image of Harbin Ice and Snow Tourism Season*; Harbin Normal University: Harbin, China, 2016.
61. Shen, C.-W.; Min, C.; Wang, C.-C. Analyzing the trend of O2O commerce by bilingual text mining on social media. *Comput. Human Behav.* **2019**, *101*, 474–483. [CrossRef]
62. Li, S.H.; Hao, Q. A Comparative Study of Content Analysis and Text Mining in the Application of Information Analysis. *Res. Libr. Sci.* **2015**, *23*, 37–42.

symmetry

MDPI

Article

Analysis of Structural Changes in Financial Datasets Using the Breakpoint Test and the Markov Switching Model

Seuk Wai Phoong [1,*], Seuk Yen Phoong [2] and Kok Hau Phoong [3]

[1] Department of Operation and Management Information System, Faculty of Business and Accountancy, University of Malaya, Kuala Lumpur 50603, Malaysia
[2] Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim 35900, Malaysia; phoong@fsmt.upsi.edu.my
[3] Faculty of Management and Information Technology, Sultan Azlan Shah University, Kuala Kangsar 33000, Malaysia; kokhauphoong@gmail.com
* Correspondence: phoongsw@um.edu.my; Tel.: +60-379-673-831

Received: 14 February 2020; Accepted: 2 March 2020; Published: 4 March 2020

Abstract: The price movements of commodities are determined by changes in the expectations about future economic variables. Crude oil price is non-stationary, highly volatile, and unstructured in nature, which makes it very difficult to predict over short-to-medium time horizons. Some analysts have indicated that the difficulty in forecasting the crude oil price is due to the fact that economic models cannot consistently show evidence of a strong connection between commodities and economic fundamentals, and, as a result, regarded the idea that economic fundamentals help predict price values as random luck. This study aimed to overcome the limitations of the economic models through the detection of structural changes as well as breaks in the data, using a breakpoint test. The Markov switching model is used to address the price patterns that led to a different market state. The results show that there are several changes as well as breaks in the estimated model. Moreover, there is an asymmetric correlation between the crude oil price and the GDP.

Keywords: markov switching; breakpoint test; crude oil price; structural change

1. Introduction

Since the year 1980, the non-stationarity of financial data series has increasingly been taken into consideration. The non-stationarity of time series might be affected by time series components, such as the stochastic trend, the cyclical variation, or the seasonal variation. The non-stationary properties of financial data cannot simply be calculated through the application of filters in order to continue within the framework of stationary models [1,2]. A Markov switching model has been developed and thoroughly discussed [3,4]. However, it is analytically intractable. As a consequence, the conditions for covariance stationarity were not established. A multivariate Markov switching model that is analytically tractable was subsequently proposed [5]. The model also allows us to derive the stationarity conditions and further dynamic properties of the financial data. In the Markov switching model, all parameters allow us to shift between a high-volatility and a low-volatility regime. Moreover, different parameters for different regimes are allowed and defined as a regime-dependent volatility. A study revealed that the Markov model provides more reliability and availability in dealing with the problem, especially when the parameters can either fail or be repaired instantly [6].

The generalization of two Markov regression switching models is as follows:

$$\begin{aligned} Y_t &= X_1 B_1 + U_1 \text{ for regime 1} \\ Y_t &= X_2 B_2 + U_2 \text{ for regime 2} \end{aligned} \quad (1)$$

where U_1 and U_2 are the random errors with normally distributed $\left(0, \sigma_1^2\right)$ and $\left(0, \sigma_2^2\right)$, and B_1 and B_2 are the coefficients of vector regression with the assumption that $\left(_1, \sigma_1^2\right) \neq \left(_2, \sigma_2^2\right)$. The probabilities λ and $1 - \lambda$ are a random or natural selection for regimes 1 and 2, where the probability λ is an independent state of the system.

There are numerous studies that examine the effect of oil price on the economy through different channels and methodology frameworks. The effects of oil price shocks and of the exchange rate volatility on inflation in Malaysia were studied, using data from January 2005 to November 2011, by adopting the Granger Causality Model [5]. The findings suggest that inflation does not granger cause the exchange rate, but granger causes the oil price. Likewise, the oil price granger causes inflation but does not granger cause the exchange rate. This shows that the oil price has a significant impact on inflation. In addition, a study was conducted to investigate the impact of oil price shocks on the output, inflation, and real exchange rate by drawing evidence from selected Association of Southeast Asian Nations (ASEAN) countries [7]. The results showed that the oil price fluctuations only had short-run effects on the countries. It was concluded that oil price shock does not induce much of the fluctuations in the ASEAN-5 economies. Another study [8] examined the price of ethanol and commodities in Brazil using the Bai–Perron test, the Johansen test, and the vector error correction model. The results showed that there was a short-term relationship between ethanol and commodities. However, to date, the nature of the relationship remains inconclusive, based on the past studies [9–12]. The volatility or the structural change is the common characteristic of the crude oil price. Crude oil prices are noisy, non stationary, non-linear, and unstructured in nature, which makes them very difficult to examine [13]. In order to evaluate the structural change of the time series, the Quandt–Andrews breakpoint test and the Bai–Perron test were used in this study to detect one or more unknown structural changes in the data. The Quandt–Andrews breakpoint test is derived from the Chow breakpoint test, to allow for a single Chow breakpoint test to be performed at every observation between two observations, and then summarized into one-test statistics.

This study had three aims: Firstly, it aimed to identify the structural changes in the crude oil price and the GDP using breakpoint tests. Secondly, the variables were tested using a co-integration test to investigate the short-run or long-run relationships. Thirdly, this study utilized the Markov switching regression to examine the effects of changes in the crude oil price on the GDP.

The remaining part of this article is presented in several sections: Section 2 discusses the theoretical aspect of the Markov switching model, Section 3 presents the data and the method, Section 4 shows the most relevant results and the discussion, and, finally, Section 5 summarizes the findings and provides the conclusions of our study.

2. Markov Switching Model

Linear regression is a popular statistical tool for econometric and statistical analyses. However, the nonlinearity of the time series might sometime cause less-precise and low-accuracy findings. A simple regression analysis is employed to explain the impact of changes in independent variables on a dependent variable.

A regression switching idea that uses a square root for each equation to represent the subsample was first introduced in 1958 [14]. The idea was then extended [15], by proposing a λ probability to avoid wasting information. The probabilities λ and $1 - \lambda$ were used in a two-state function $g\left(Y_j \middle| X_j\right)$ [15], and the function is shown below:

$$\begin{aligned} h\left(Y_j \middle| X_j\right) &= \lambda\left[g_1\left(Y_j \middle| X_j\right)\right] + (1-\lambda)\left[g_2\left(Y_j \middle| X_j\right)\right] \\ &= \frac{\lambda}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{\left(Y_j - X_j\beta_1\right)^2}{2{\sigma_1}^2}\right] + \frac{1-\lambda}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{\left(Y_j - X_j\beta_2\right)^2}{2{\sigma_2}^2}\right] \end{aligned} \quad (2)$$

where $j = 1, 2 \ldots, M$. The natural logarithm is used to maximize Equation (2) to obtain the values of the parameters in the model $\left(\beta_1,\ \beta_2,\ \sigma_1^2, \sigma_1^2 \text{ and } \lambda\right)$ by using the equation ($j = 1, \ldots, M$) below:

$$L = \sum_{j=1}^{M} \text{In } h\left(Y_j \middle| X_j\right) \tag{3}$$

The Quandt model was then reviewed and enhanced [16], and it incorporated the Markov chain properties. A study revealed that the state changes in the variable series cannot be observed directly [16]. Each state in the estimated model is independent from each other, and the probability of a state is constant. The likelihood function in the Markov switching regression equation was corrected by suggesting a recursive algorithm to replace the likelihood function [17]. An algorithm was used to develop a filtering algorithm to calculate the conditional densities and the probability of the unobserved state value, S_t [3,17].

$$Y_{t-1}, P(s_t = j, s_{t-1} = i | Y_{t-1}) \text{ when } i, j = 1, \ldots, M \tag{4}$$

This model has been further investigated and extended in two studies [4,18]. One study introduced a smoothing algorithm for the unobserved state variable [18], and the other extended it to the multivariate Markov switching model [4]. Another study examined the performance of the linear and the Markov switching model on the analysis of the stock price and the commodity price [19]. The Markov switching model performed better than the linear model, because it was able to detect the asymptotic behavior, and identified the expected duration for each state of the estimated model. It has been found that the Markov switching model outperforms when forecasting value at risk and expected shortfall of assets' return [20]. It has also been reported that the Markov switching generalized autoregressive conditional heteroskedasticity (GARGCH) model is able to provide a more appropriate result in forecasting the volatility index [21].

3. Methodology

3.1. Data and Variables

The sample data were obtained from the United States (U.S.) Department of Energy, from Quarter 1 (Q1) 2010 to Quarter 4 (Q4) 2018. The benchmark crude oil price serves as a pricing reference to the sellers and buyers. The West Texas Intermediate (WTI), the Brent blend, and Dubai crude oil are the three primary benchmarks for crude oil. The oil that is traded in the United States of America (USA) is priced using the WTI as a benchmark, while most of the oil traded outside of the USA and the Far East is priced using the Brent blend as a benchmark. Dubai, however, is the main benchmark for the oil exported from the Middle East to Asia. The WTI crude oil, which is extracted from the wells in the southern part of the USA, specifically Oklahoma and Texas, has a very high quality of crude oil, with an American Petroleum Institute gravity of 39.6° and 0.24% of sulfur. Given these qualities, the WTI is the benchmark for light or sweet crude oil. The Brent blend crude oil is combined from different fields located in the North Sea, and it has an API gravity of 38.3° and 0.37% of sulfur, making it a light or sweet crude oil as well, although slightly less so compared with the WTI crude oil. The quality characteristics of both the WTI and Brent blend crude oil are quite similar, with the only difference being that the WTI crude oil results in rather more gasoline and rather less heating oil than the Brent blend crude oil. Consequently, the WTI crude oil has a slight price advantage compared with the Brent blend. On the other hand, Dubai's crude oil has an API gravity of 32°, and it has a high sulfur content of roughly 2%; hence, it is considered to be the benchmark for heavy or sour crude oil. The light or sweet crude oil is usually traded at a premium to heavy or sour crude oil (Dubai). We use the Brent blend crude oil price as a proxy for the oil price, quoted in U.S. dollars in this study.

To measure the financial development, we used the country's domestic variable, the GDP. Among the major economies of Asia, Malaysia is the second largest liquefied natural gas exporter in the world. According to the BP Statistical Review of World Energy, in 2015, 693,000 barrels of oil were produced

in Malaysia and the average output over the past five years was 654,000 a day. The shipments of crude petroleum, petroleum products, and liquefied natural gas accounted for approximately 14% of Malaysia's total exportations in the first half of the year 2016. Energy is a key sector of Malaysia's GDP. As Malaysia is an oil net exporter, a high crude oil price may benefit the country in the short term. However, it can also be a disadvantage, since the rising prices of oil will have an impact on the world's growth, which could affect the world's consumption and income. Therefore, the focus of this study was to investigate the structural changes in and correlations between the price of crude oil and the GDP in Malaysia, using the Quandt–Andrews and the Bai–Perron breakpoint tests. The Markov switching regression model was used to analyze the correlations between oil price and GDP.

The Brent blend crude oil price has been widely used in previous studies [7,22,23], and the GDP used in this study was GDP per capita (in current U.S. dollars). These proxies have been widely used in previous studies [24–26]. The descriptive values for skewness and kurtosis in Table 1 indicated that all of the parameters were not normal. These results are consistent with the findings of the Jarqua–Bera test, and the statistics are non-negative and far from zero. Thus, there might be a structural change as well as breaks in the datasets.

Table 1. Descriptive statistics for the parameters.

	Mean	Standard Deviation	Skewness	Kurtosis	Jarqua–Bera
LOGOP	4.32	0.37	−0.34	1.81	2.85
MYGDP	71,410.50	5296.73	−0.39	2.29	1.66

3.2. Breakpoint Test

3.2.1. Quandt–Andrews Breakpoint Test

To identify the possible structural breaks, the Quandt–Andrews breakpoint test was used to specify the null hypothesis of no break is occurring within 15% of the trimmed data. Trimming the sampling data is mainly used to ensure that the subsample is not close to the endpoint of the sample [2].

The test statistic for examining the null hypothesis of no break at time period T, $T_0 \le T \le T_1$ is as follows:

$$QLR = \max\left[c(T_0), c(T_{0+1}), \ldots, c(T_{1\text{-}1}), c(T_1)\right] \tag{5}$$

where T_0 and T_1 are usually the inner 70% of the sample that is excluded from the first 15% and the last 15% of the sample.

3.2.2. Bai–Perron Test

The Bai–Perron test calculates the sup F statistics on no structural change ($p = 0$) on the null hypothesis and $p = r$ changes. Let $\mathbf{M}$ be a conventional matrix, such that $(\mathbf{M}\lambda)' = \left(\lambda_1' - \lambda_2', \ldots, \lambda_r' - \lambda_{r+1}'\right)$. Then

$$F_T(\beta_1, \ldots, \beta_r; q) = \frac{1}{T}\left(\frac{T-(r+1)q-p}{rq}\right)\hat{\lambda}'\mathrm{M}'\left(\mathrm{M}'\hat{\mathrm{v}}\left(\hat{\lambda}\right)\mathrm{M}'\right)^{-1}\mathrm{M}\hat{\lambda} \tag{6}$$

where r is a break, and $\hat{\mathbf{v}}\left(\hat{\boldsymbol{\lambda}}\right)$ estimates the covariance matrix of $\hat{\lambda}$ robust to serial correlation and heteroskedasticity. It is a generalization of the sup F test following Andrews (1993) and others,

$$\sup F_T(r, q) = F_T\left(\hat{\beta}_1, \ldots, \hat{\beta}_r, q\right) \tag{7}$$

where $\left(\hat{\beta}_1, \ldots, \hat{\beta}_r\right)$ is the global sum of squared residuals under the chosen trimming. This is equivalent to maximizing the F test since the estimated break dates are consistent, even in the presence of a serial correlation.

The breakpoint *F*-test is:

$$F = \frac{[\tilde{u}\tilde{u} - (u_1u_1 + u_2u_2)]/k}{(u_1u_1 + u_2u_2)/(T - 2k)} \quad (8)$$

where $\tilde{u}'\tilde{u}$ is the residual of the restricted sum of squares, $u_j'u_j$ is the sum of squared residuals from subsample j, k is the number of parameters, and T denotes the total number of observations.

3.3. Co-Integration Test

The Johansen maximum likelihood estimator is a co-integration test that is powerful for analyzing the existence of a co-integration in the series. This estimator is able to provide asymptotically efficient estimates of the co-integrating vectors and adjustment parameters. Therefore, the Johansen test was applied in this study to examine the existence of co-integrating vectors among the variables [2].

3.4. Markov Switching Regression Model

The time series for all of the variables in this study vary with dynamics that are state-dependent [27]. The coefficients of the parameters may be different for each state, since the state can be in low or high volatility, or recession or expansion. Moreover, the time of transition and the duration in a particular state are both random. Therefore, the Markov switching regression was used to estimate the state-dependent parameters. The framework for the Markov model is to be memoryless in each individual state [27]. Thus, the switching properties can be calculated by using the following equation:

$$MYGDP_t = \mu_i + \alpha logOP_t + \varepsilon_t$$

where $\mu_i = \mu_1$ if $i = 1$ (state 1), and $\mu_i = \mu_2$ if $i = 2$ (state 2). The transition probabilities for a two-state model are:

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

where $p_{11} + p_{12} = 1$, $p_{21} + p_{22} = 1$. The expected duration for each state is important to identifying the asymmetric properties for the business cycle. The expected duration can be estimated by using the following formula: $E(D) = \frac{1}{1-p_{ii}}$, where i is the state/regime.

4. Results and Discussions

The results of the analysis are shown and discussed below.

Figure 1 shows the changes in the data series. All of the variables that are plotted in Figure 1 are irregular, suggesting that the variables series are not stationary. Additionally, there are several structural changes present in the series, including the 2011 quarter 2, the 2012 quarters 1 and 2, the 2015 quarters 1, 3, and 4, the 2017 quarter 2, and the 2018 quarter 3. All these changes can be related to an economic crisis such as the Global Economic Crisis of 2012, witnessing the European debt crisis, with a dramatic depression in 2011, the Chinese stock market crash, the Russian financial crisis of 2014–2017, etc. Malaysia is an emerging market that is undergoing rapid growth, and is keen to be affected by other countries, especially trading partners, as mentioned in [28]. Therefore, we can conclude that these series have regime shifts due to the uncertainty in the parameters' series.

The *p*-value for the Quandt–Andrews breakpoint test in Table 2 is at a less than 0.05 level of significance. Thus, we can conclude that there are breaks in the sample data. We next identified the break date by using the Bai–Perron test.

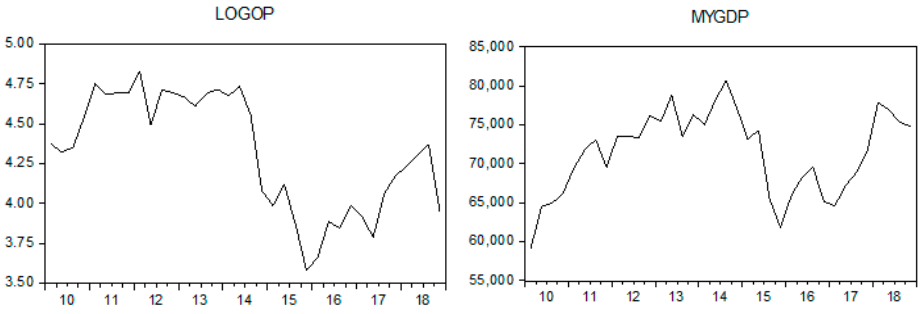


Figure 1. The variables' plot.

Table 2. Findings of the Quandt–Andrews breakpoint test.

Quandt–Andrews Unknown Breakpoint Test **Null Hypothesis: No Breakpoints within 15% Trimmed Data** **Varying Regressors: All Equation Variables**		
Statistic	Value	Prob.
Maximum LR F-statistic (2012Q2)	21.39673	0.0000
Maximum Wald F-statistic (2012Q2)	42.79346	0.0000
Exp LR F-statistic	8.147680	0.0000
Exp Wald F-statistic	18.32268	0.0000
Ave LR F-statistic	9.538625	0.0001
Ave Wald F-statistic	19.07725	0.0001

Note: probabilities calculated using Hansen's (1997) method. Prob.—Probability.

The estimated break dates suggested by the Bai–Perron tests in the Table 3 are Quarter 2, 2012 and Quarter 3, 2015. This might be related to the great recession in the United States, and the 2015 Chinese stock market crash. China was the largest trading partner of Malaysia between 2009 and 2017 (16.4 per cent or 67,229 million U.S. dollars (USD); and the United States was Malaysia's third largest trading partner in 2017, which contributed 8.9 percent or 36,623 million USD. Unstable economic conditions in these two countries can affect the GDP in Malaysia.

Table 3. The Bai–Perron test outputs.

Multiple Breakpoint Tests **Bai–Perron Tests of L+1 vs. L Sequentially Determined Breaks** **Break Test Options: Trimming 0.15, Max. Breaks 5, Sig. Level 0.05**			
Sequential F-statistic determined breaks:			2
Break Test	F-statistic	Scaled F-statistic	Critical Value **
0 vs. 1 *	21.39673	42.79346	11.47
1 vs. 2 *	7.279554	14.55911	12.95
2 vs. 3	4.667017	9.334033	14.03
Break dates:			
	Sequential	Repartition	
1	2012Q2	2012Q2	
2	2015Q3	2015Q3	

* Significant at the 0.05 level, ** Bai–Perron (Econometric Journal, 2003) critical values. Max.—Maximum; Sig.—Significance.

To understand the co-integration impact of the time series variables, the Johansen co-integration test was used to evaluate the variables series. The findings are presented as follows.

Table 4 shows that all the *p*-values are greater than 0.05. There is no co-integration between Malaysia's GDP and the oil price. There is a short term relationship between the changes in oil price and the GDP. This finding is consistent with previous results [7], indicating that the oil price fluctuations only have short-run effects on the inflation, real exchange rate, and GDP in Malaysia.

Table 4. The co-integration test outputs.

Unrestricted Co-Integration Rank Test (Trace)				
Hypothesized		Trace	0.05	
No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None	0.226176	12.01652	20.26184	0.4473
At most 1	0.092458	3.298539	9.164546	0.5264
Hypothesized		Max. Eigenvalue	0.05	
No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None	0.226176	8.717976	15.89210	0.4647
At most 1	0.092458	3.298539	9.164546	0.5264

Trace test indicates no co-integration at the 0.05 level, * denotes rejection of the hypothesis at the 0.05 level, ** MacKinnon–Haug–Michelis (1999) *p*-values, Unrestricted Co-integration Rank Test (Maximum Eigenvalue), Max. eigenvalue test indicates no co-integration at the 0.05 level.

Table 5 reported the outputs for Markov switching regression. Two regimes of the Markov switching models were selected in this study as suggested by previous studies [29,30] indicating that two-regime models can represent the recession and growth states in the business cycle. A Marquardt step is used in the Markov switching regression model to estimate the parameters of an unobserved state. The findings show that there is a positive correlation between the oil price and the GDP. These results are consistent with previous findings [31], but contradict others [32,33]. It has been argued that the negative effect of oil price shocks on GDP growth is greater than the time of oil price increases [32]. In addition, a study has revealed that the relationship between the GDP and the oil price is relatively turbulent [32]. The changes in the relationship rely on the economic conditions, such as the stability, recession, and growth conditions. Previous studies [32,33] agree that the correlations between the oil price and the GDP are unstable and vary in different phases over time. The present study is expected to provide better insights into the relationship between the oil price and the GDP, since we have divided the datasets into two regimes (regime 1: growth, and regime 2: recession) while measuring the relationship between the oil price and the GDP.

Table 5. The Markov switching regression outputs.

Method: Markov Switching Regression (BFGS/Marquardt Steps)				
Number of states: 2 Initial probabilities obtained from ergodic solution Ordinary standard errors and covariance using a numeric Hessian Random search: 25 starting values with 10 iterations using 1 standard deviation (rng = kn, seed = 1,560,858,386) Convergence achieved after 28 iterations				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
		Regime 1		
LOGOP	10,254.71	1877.429	5.462102	0.0000
C	29,104.07	7933.635	3.668440	0.0002
		Regime 2		
LOGOP	21,525.61	6420.504	3.352635	0.0008
C	−30,461.34	29431.43	−1.034993	0.3007

Std.—Standard.

The transition probabilities in the Table 6 reported from regime 1 to regime 2 are higher than the transition probabilities from regime 2 to regime 1. This indicates that the recovery variable needs a longer time than the stagnant variable. p_{11} and p_{22} have a high value; thus, we rejected the null hypothesis of no shifts in the regime.

Table 6. The transition probabilities and expected durations.

Constant Markov Transition Probabilities			
Sample: 2010Q1 2018Q4 Included observations: 36 P(i, k) = P(s(t) = k | s(t − 1) = i) (row = i/column = j)			
		1	2
All periods	1	0.975638	0.024362
	2	0.043480	0.956520
Expected duration: Constant Markov transition probabilities Sample: 2010Q1 2018Q4 Included observations: 36 Constant expected durations:			
		1	2
All periods		41.04773	22.9920

Based on the findings of expected durations, we can conclude that there is an asymmetric business cycle, since the expected durations in regimes 1 and 2 are approximately 41.048 and 22.999 quarters, respectively. The first regime is more prevalent than the second regime. The results are consistent with previous findings [34–36], indicating that there is an asymmetric oil shock.

5. Conclusions

The crude oil price is non-stationary, highly volatile, and unstructured in nature, which makes it difficult to forecast over short-to-medium time horizons. Past studies [29–36] have also shown that the inconsistency may lie in the intrinsic limitations of the theoretical framework, or in ignoring the time series components. The present study aimed to overcome the limitations of the economic models through the detection of the structural changes in the data series, using the breakpoint test and the Markov switching regression model to address the price patterns that led to different market states. The results of the Quandt–Andrews test show the existence of breaks in the data. Moreover, the estimated break dates are Quarter 2, 2012 and Quarter 3, 2015. Based on the Markov switching regression outputs, the oil price has a positive relationship with the GDP, where the increase of the oil price impacts the increase of the GDP. However, according to the transition probabilities and the expected duration results, there is an asymmetric relationship between the oil price and Malaysia's GDP. Even though Malaysia is an oil exporter, it is not a member of Organization of the petroleum exporting countries (OPEC). This indicates that Malaysia has no influence on the determination of the oil price internationally.

This study has some limitations that future research can address. One such limitation is that we used quarterly data, while daily or monthly data may provide a greater understanding of the breaks in or changes of the time series. Future research may also include daily or monthly data for a better understanding of the changes in the oil price. The results of this study can also be expanded to other financial or economic variables, such as the stock price and exchange rate, to expand the multivariate framework. Further studies with this method can be extended to a three-regime Markov switching model to measure three states: depression, high appreciation, and low appreciation, as suggested previously [37].

Author Contributions: Conceptualization, S.W.P. and S.Y.P.; methodology, S.W.P.; software, S.W.P.; validation, S.W.P., S.Y.P., and K.H.P.; formal analysis, S.W.P. and S.Y.P.; investigation, S.W.P. and S.Y.P.; resources, K.H.P.; data curation, S.Y.P. and K.H.P.; writing—original draft preparation, S.W.P.; writing—review and editing, S.W.P., S.Y.P, and K.H.P.; project administration, S.W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Grant Scheme provided by the ministry of education of Malaysia, grant number FRGS/1/2019/STG06/UM/02/9, and the APC was funded by the University of Malaya and Fundamental Research Grant Scheme (grant number: FRGS/1/2019/STG06/UM/02/9) provided by the ministry of education of Malaysia.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fan, J.; Yao, Q. *Nonlinear Time Series: Nonparametric and Parametric Methods*; Springer: Berlin, Germany, 2005.
2. Phoong, S.W.; Phoong, S.Y.; Moghavvemi, S.; Phoong, K.H. Multiple Breakpoint Test on Crude Oil Price. *Found. Manag.* **2019**, *11*, 187–196. [CrossRef]
3. Hamilton, J.D. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **1989**, *57*, 357–384. [CrossRef]
4. Krolzig, H.M. *Markov-Switching Vector Autoregression*; Springer: Berlin, Germany, 1997.
5. Shaari, M.S.; Hussain, N.E.; Abdullah, H. The effects of oil price shocks and exchange rate volatility on inflation: Evidence from Malaysia. *Int. Bus. Res.* **2012**, *5*, 106–112. [CrossRef]
6. Siddiqui, M.A.; Butt, S.I.; Gilani, O.; Jamil, M.; Maqsood, A.; Zhang, F. Optimizing Availability of a Framework in Series Configuration Utilizing Markov Model and Monte Carlo Simulation Techniques. *Symmetry* **2017**, *9*, 96. [CrossRef]
7. Basnet, H.C.; Upadhyaya, K.P. Impact of oil price shocks on output, inflation and the real exchange rate: Evidence from selected ASEAN countries. *Appl. Econ.* **2015**, *47*, 3078–3091. [CrossRef]
8. David, S.A.; Inácio, C.M.C., Jr.; Tenreiro Machado, J.A. Ethanol Prices and Agricultural Commodities: An Investigation of Their Relationship. *Mathematics* **2019**, *7*, 774. [CrossRef]
9. Alghalith, M. The interaction between food prices and oil prices. *Energy Econ.* **2010**, *32*, 1520–1522. [CrossRef]
10. Nazlioglu, S.; Soytas, U. World oil prices and agricultural commodity prices: Evidence from an emerging market. *Energy Econ.* **2011**, *33*, 488–496. [CrossRef]
11. Nazlioglu, S. World oil and agricultural commodity prices: Evidence from nonlinear causality. *Energy Policy* **2011**, *39*, 2935–2943. [CrossRef]
12. Rafiq, S.; Salim, R.; Bloch, H. Impact of crude oil price volatility on economic activities: An empirical investigation in the Thai economy. *Resour. Policy* **2009**, *34*, 121–132. [CrossRef]
13. Zhou, Z.; Jin, Q.; Peng, J.; Xiao, H.; Wu, S. Further Study of the DEA-Based Framework for Performance Evaluation of Competing Crude Oil Prices' Volatility Forecasting Models. *Mathematics* **2019**, *7*, 827. [CrossRef]
14. Quandt, R.E. Estimation of the parameters of a linear regression system obeying two separate regime. *J. Am. Stat. Assoc.* **1958**, *53*, 873–880. [CrossRef]
15. Quandt, R.E. A new approach to estimating switching regression. *J. Am. Stat. Assoc.* **1972**, *67*, 306–317. [CrossRef]
16. Goldfeld, S.M.; Quandt, R.E. A Markov model for switching regressions. *J. Econom.* **1973**, *1*, 3–16. [CrossRef]
17. Cosslett, S.R.; Lee, L.F. Serial correlation in the latent discrete variable models. *J. Econom.* **1985**, *27*, 79–97. [CrossRef]
18. Kim, C.J. Dynamic linear models with Markov-switching. *J. Econom.* **1994**, *60*, 1–22. [CrossRef]
19. Phoong, S.W.; Ismail, M.T.; Sek, S.K. Linear Vector Error Correction Model versus Markov Switching Vector Error Correction Model to Investigate Stock Market Behaviour. *Asian Acad. Manag. J. Account. Financ.* **2014**, *10*, 133–149.
20. Ardia, D.; Bluteau, K.; Boudt, K.; Leopoldo, C. Forecasting risk with Markov-switching GARCH models:A large-scale performance study. *Int. J. Forecast.* **2018**, *34*, 733–747. [CrossRef]
21. Caporale, G.; Zekokh, T. Modelling volatility of cryptocurrencies using Markov-Switching GARCH models. *Res. Int. Bus. Financ.* **2019**, *48*, 143–155. [CrossRef]
22. Berument, M.; Ceylan, N.; Dogan, N. The Impact of Oil Price Shocks on the Economic Growth of Selected MENA Countries. *Energy J.* **2010**, *31*, 149–176.

23. Arezki, R.; Jakab, Z.; Laxton, D.; Matsumoto, A.; Nurbekyan, A.; Wang, H.; Yao, J. Oil Prices and the Global Economy. *Int. Monet. Fund* **2017**, *17*, 1–30. [CrossRef]
24. Shahbaz, M.; Lean, H.H. Does financial development increase energy consumption? The role of industrialization and urbanization in Tunisia. *Energy Policy* **2012**, *40*, 473–479. [CrossRef]
25. Gómez, M.; Ciarreta, A.; Zarraga, A. Linear and nonlinear causality between energy consumption and economic growth: The case of Mexico 1965–2014. *Energies* **2018**, *11*, 784. [CrossRef]
26. Gómez, M.; Rodríguez, J.C. Energy Consumption and Financial Development in NAFTACountries, 1971–2015. *Appl. Sci.* **2019**, *9*, 302. [CrossRef]
27. De Martino, I. Decaying Dark Energy in Light of the Latest Cpsmological Dataset. *Symmetry* **2018**, *10*, 372. [CrossRef]
28. Farah, P.D. Five Years of China WTO Membership: EU and US Perceptives about China's Compliance with Transparency Commitments and the Transitional Review Mechanism. *Leg. Issuses Econ. Integr.* **2006**, *33*, 263–304.
29. Neftci, S.N. Are economic time series asymmetric over the business cycle? *J. Political Econ.* **1984**, *92*, 306–328. [CrossRef]
30. Brunner, A.D. Conditional symmetries in real GNP: A semi nonparametric approach. *J. Bus. Econ. Stat.* **1992**, *10*, 65–72.
31. Laredic, S.; Mignon, V. The impact of oil prices on GDP in European countries: An empirical investigation based on asymmetric cointegration. *Energy Policy* **2006**, *34*, 3910–3915. [CrossRef]
32. Gadea, M.D.; Gómez-Loscos, A.; Montañés, A. Oil Price and Economic Growth: A Long Story? *Econometrics* **2016**, *4*, 41. [CrossRef]
33. Benhmad, F. Dynamic cyclical comovements betweem oil prices and US GDP: A wavelet perspective. *Energy Policy* **2013**, *57*, 141–151. [CrossRef]
34. Rafiq, S.; Bloch, H. Explaining commodity prices through asymmetric oil shocks: Evidence from nonlinear models. *Resour. Policy* **2016**, *50*, 34–48. [CrossRef]
35. Miao, D.W.C.; Wu, C.C.; Su, Y.K. Regime-switching in volatility and correlation structure using range-based models with Markov-switching. *Econ. Model.* **2013**, *31*, 87–93. [CrossRef]
36. Balcilar, M.; Eyden, R.; Uwilingiye, J.; Gupta, R. The Impact of Oil Price on South African GDP Growth: A Bayesian Markov Switching VAR Analysis. *Afr. Dev. Rev.* **2017**, *29*, 319–336. [CrossRef]
37. Ayodeji, I.O. A Three-State Markov-Modulated Switching Model for Exchange Rates. *J. Appl. Math.* **2016**, *2016*, 5061749. [CrossRef]

symmetry

MDPI

Article

Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results

Huseyin Polat * and Saadin Oyucu

Department of Computer Engineering, Faculty of Technology, Gazi University, 06560 Ankara, Turkey; saadinoyucu@gazi.edu.tr
* Correspondence: polath@gazi.edu.tr

Received: 5 February 2020; Accepted: 13 February 2020; Published: 17 February 2020

Abstract: To build automatic speech recognition (ASR) systems with a low word error rate (WER), a large speech and text corpus is needed. Corpus preparation is the first step required for developing an ASR system for a language with few argument speech documents available. Turkish is a language with limited resources for ASR. Therefore, development of a symmetric Turkish transcribed speech corpus according to the high resources languages corpora is crucial for improving and promoting Turkish speech recognition activities. In this study, we constructed a viable alternative to classical transcribed corpus preparation techniques for collecting Turkish speech data. In the presented approach, three different methods were used. In the first step, subtitles, which are mainly supplied for people with hearing difficulties, were used as transcriptions for the speech utterances obtained from movies. In the second step, data were collected via a mobile application. In the third step, a transfer learning approach to the Grand National Assembly of Turkey session records (videotext) was used. We also provide the initial speech recognition results of artificial neural network and Gaussian mixture-model-based acoustic models for Turkish. For training models, the newly collected corpus and other existing corpora published by the Linguistic Data Consortium were used. In light of the test results of the other existing corpora, the current study showed the relative contribution of corpus variability in a symmetric speech recognition task. The decrease in WER after including the new corpus was more evident with increased verified data size, compensating for the status of Turkish as a low resource language. For further studies, the importance of the corpus and language model in the success of the Turkish ASR system is shown.

Keywords: automatic speech recognition; speech corpus; text corpus; data acquisition; multi-layer neural network; natural language processing

1. Introduction

The primary function of an automatic speech recognition (ASR) system is to automatically convert human speech into transcribed text. Command and control systems, dictation software, broadcast news transcription (for indexing news content), and telephone speech transcriptions are at the forefront of the applications of ASR systems [1]. In addition, ASR can also be used to analyze social media data. The social data contains not only textual data, but also a large number of audio, video, and image data. ASR can be used to transcribe various audio recordings and videos on social media [2].

Typical ASR works with models trained by machine learning algorithms based on statistical pattern classification [3]. Machine learning algorithms use two main approaches: supervised and unsupervised learning. Supervised learning is usually used for classification and uses labelled data as a training set. Unsupervised learning is usually used for clustering with unlabelled data. ASR uses supervised learning to training speech classifiers, such as hidden Markov models (HMMs) [4]. In training,

the statistical pattern classification technique Gaussian mixture model (GMM) can be used with HMM. In the early development era of ASR, GMM with HMM were used successfully. Despite the success of GMM and HMM approaches, the accuracy of ASR systems still lags behind human-level performance. This implies that more research is required.

The first use of neural networks in speech applications in the 1990s never produced improved performance compared to the traditional GMM and HMM technology [5]. Some key problems appeared with the use of artificial neural networks (ANNs), such as the vanishing gradient, lack of sufficient amounts of training data, lack of significant computation power, and weak temporal correlation structure in the neural predictive models. To overcome these problems, studies have been conducted in different fields [6]. As a result of these studies, ANN-based approaches have shown accurate performance in many domains of research, such as image processing, speech recognition, language modelling, parsing, information retrieval, speech synthesis, and speech translation [7,8]. When an ANN is applied to these domains, the results obtained exceed those of other state-of-the-art approaches. The reason why an ANN-based approach produces this high performance is the capability to find and learn compound structures in a large amount of data [9]. These deep neural nets have been used in speech recognition to develop acoustic models (AMs). However, ANN-based approaches require large amounts of data to produce accurate results. Therefore, to develop a speaker-independent and high-quality performance (i.e., low word error rate) ASR, an extensive collection of speech samples must be gathered from various speakers along with the transcriptions of those speeches [10].

Developing a useful and practical process to create a Turkish transcribed speech corpus, consisting of audio and text data, poses a substantial challenge. The major problem with the existing commonly used methodologies to collect transcribed speech data is that the process is usually costly, both in terms of time and budget, as well as cumbersome. In the classical approach, native speakers of a particular language transcribe pre-recorded speech documents. Williams et al. showed that six hours of work is needed on average for transcribing one hour of speech [11].

Unfortunately, most of the spoken languages, including all Turkish-based Asian languages, are classified as low-resource for ASR applications due to the limited amount of available transcribed speech, which is required for training AMs and language models (LMs). Turkish language ASR systems also suffer from a low accuracy rate; hence, it is very crucial to create a symmetric Turkish transcribed speech corpus according to the corpus of the high resources languages.

The accuracy rate of ASR is not the same for different languages. Even for the same language, this rate may widely vary for different environments with different acoustic characteristics. However, even under the same conditions, ASR performance for English is known to be superior to that of Turkish due to the two main constraints in Turkish ASR. The first and more restrictive is the above-mentioned shortage of transcribed speech. The second reason is the agglutinative nature of the language with free word order and sequence, which leads to the generation of a relatively high number of new words through the use of suffixes. These characteristics complicate accurate prediction under conditional probabilities.

In this article, the two fundamental problems described above are addressed by developing a Turkish corpus with a large vocabulary. The Turkish speech corpus was obtained with three different approaches. The first approach used a feature film, the second one, a mobile application, and the third one, the transfer learning method. Also, the obtained Turkish speech corpus was presented for the approval of real users, so that more accurate data were obtained. The tests conducted by comparing the speech recognition performance of a system trained with the newly collected corpus to that of a system trained on an existing corpus showed that the results are comparable, which validates the effectiveness of the newly collected corpus. We completed studies to solve the difficulties in the processing of Turkish. We illustrate the procedures and the steps involved in ASR and present how different choices in the design can influence Turkish ASR performance. Finally, we show that ANN approaches could be more profitable.

The remainder of this paper is organized as follows. In the next section, a brief literature review of speech recognition research, with an emphasis on the Turkish-language-related efforts, is provided. Section 3 provides the details of the corpus collection process and the characteristics of the Turkish language. Section 4 describes the experimental setup. Section 5 outlines the joint results on the progress and evaluation of test sets, followed by our conclusions in Section 6.

2. Related Works

Numerous studies have aimed to recognize speech using a computer. The most important advances in this domain, however, were obtained after the use of HMMs, especially between the 1960s and 1970s. The theoretical bases of HMMs used in contemporary speech recognition systems were described in [12]. The historical developments of speech recognition systems were detailed in [13]. A tutorial on HMM and selected applications in speech recognition was published [4].

The first LM approach for improving speech recognition results was reported [14]. The authors introduced a component called the store of linguistic knowledge, which can be considered a precursory LM. Speech analysis results obtained from an acoustical analyzer were combined with the information provided by this LM using a computer, and a text output or transcript was produced.

During the early years of speech recognition activities, the comparison of speech recognition systems was tricky as there were no standards for database (corpus) creation or for the properties of the corpus. To solve these standardization issues, researchers from the Massachusetts Institute of Technology (MIT) and Texas Instruments (TI) joined efforts to create a recorded corpus to be used for training speech recognition systems. The corpus created as a result of this collaboration was named TIMIT and has since become a standard for benchmarking speech recognition results in English [15]. The Defense Advanced Research Projects Agency was also interested in collecting a corpus for speech recognition and recorded a corpus based on the Wall Street Journal (WSJ) [16]. The goal was to recognize read speech from the WSJ with a vocabulary size as large as 60,000 words.

Traditionally, two approaches are employed for collecting a transcribed speech corpus to be used for training speech recognition models. In the first approach, selected texts extracted from newspapers or books are read aloud by different speakers. In the second and more pervasive approach, pre-recorded television and radio programs are collected and manually transcribed [17–19]. The former approach is faster as it does not require a transcription effort; however, it has two main drawbacks. The first disadvantage is that spoken texts may not include enough samples for all the phonemes in the language. The second disadvantage is that only a limited number of speakers are used, and the environment is usually clean without any noise, which does not represent the test conditions well. The lack of speech samples from different types of speakers and different environments leads to poor recognition results for large vocabulary tasks. The second approach, transcription, is usually a costly and a rather slow process that is conducted by professional transcribers. Such experts demand approximately six hours of work for one hour of speech and the hourly cost of transcription is between USD $90 and $150 in the USA [20,21]. The slow pace and high expense of acquiring transcriptions are significant deterrents to further improvements in ASR for low-resourced languages. Additionally, procuring as many expert transcribers as needed for transcription projects of high-volume is difficult, mainly when the volume of work tends to fluctuate. One the other hand, more natural and spontaneous speech fragments from a large number of speakers can be collected in the corpus using the transcription approach.

Since the beginning of the 2010s, the speech recognition activity for Turkish has increased [22–24]. Attempts have been made to develop text and speech corpora for Turkish. The first transcribed Turkish speech corpus published by the Linguistic Data Consortium (LDC) contains eight hours of text, speech, and alignments [23]. This corpus contains 120 speakers (60 men and 60 women, all native Turkish speakers whose ages ranged between 19 and 50 years, with an average of 24 years) delivering 40 sentences each (approximately 300 words per speaker), which required approximately 500 min of clean and noise-free speech in total. The 40 sentences were selected randomly for each speaker from a triphone balanced set of 2462 Turkish sentences. In a more recent work, Oflazoglu and Yildirim

concentrated only on detecting emotions by creating a corpus from movies and labelling them with emotions, which is unsuitable for large vocabulary speech recognition tasks [25].

Recently, crowdsourcing has risen as a new method for the large-scale economic transcription of spoken documents [20–25]. In one study, large spoken documents were divided into smaller utterances, which were then distributed to a pool of staff through a coordinating web service, such as Amazon's Mechanical Turk [26]. The staff worked concurrently, accelerating the culmination of the original transcription task. The economics of such work incurred costs ranging from USD $2.25 to $22.50 per hour of transcription. Several studies used closed captions for the collection of speech databases from broadcast news [27–29]. Several other studies used movie subtitles for building parallel text datasets for aligning texts between different language pairs [30].

When previous studies were examined, we found that many field-specific studies were conducted. Some of these studies focused on a particular dialect. The study on the Goalparia dialect is one of them [31]. Records of spontaneous talks were taken from 27 speakers (19 men, 8 women). In total, six hours and eight minutes of speech data were collected for the Assam language. For the Bengali language, in which the dialect of Goalparia was spoken, records were collected from 30 speakers (20 men, and 10 women). In a study on South Korea, about 40 h of speech were obtained from 40 speakers [32].

Another study aimed to label a 20-h conversation by creating a model with an 11-h speech corpus, previously tagged in Slovakia [33]. In a study on Urdu, a speech corpus was created for travel. An interactive response system was used when constructing a corpus consisting of a total of 250 words, such as city names, days, and times [34]. In another study, a corpus was prepared for the Bahasa Indonesian language [35], records being captured in a studio environment. A corpus was created for the Japanese language to understand what older adults were saying. Recordings were collected from old people with the help of a microphone [36]. A comprehensive study was carried out on Hindi [37]. In this study, recordings were captured in a studio environment. However, the texts required for registration were subdivided and presented to the speaker. Thus, a balanced distribution was provided for different topics in the dataset. In a different study, the need for tag social media was first explained, and then a dataset containing ASR transcripts and social media data information was presented [2]. The data is predominantly English conversation records, but there are also small numbers of Czech, Dutch, French, Italian, German, and Spanish present.

When the literature was examined, we observed that speech data for different purposes, different language origins, and different dialects have been collected. When collecting speech data, different sources were used: movie subtitles, a studio environment, social media data, an interactive response system, and portable microphones, and audio files in different areas were transferred and labelled using existing speech recognition systems. When the Turkish language was considered, we found that the existing datasets were insufficient. For this reason, we conducted a comprehensive speech corpus creation work on the Turkish language. Film subtitles, mobile application, and transfer learning methods were used together when creating the corpus. In the selection of the film subtitles, the content of the genre and speech were considered. Studies were conducted to ensure that sex distribution was equal. All data were presented to real users through a web interface for approval. Developed using a combination of different methods, this study presents an innovative approach that may be useful for other languages.

3. Description of the Corpus and Collecting Process

In this study, the film subtitles, mobile application, and transfer learning methods were used together to create a speech corpus. The use of these methods is explained in detail under separate headings.

3.1. Use of Film Movies and Time-Bound Subtitle Documents

In this study, we used the audio collected from Turkish movies and the text extracted from the corresponding time-aligned subtitle documents as sources for spoken data and transcriptions, respectively. Using our proposed process, it is possible to collect massive amounts of transcribed speech for any given language, provided that a large collection of movies and their subtitles in that language are accessible.

Figure 1 shows a block diagram of the main modules for the movie speech corpus (MSC) collection process presented in this work. Our process has four main modules that were implemented as MSC collection tools.

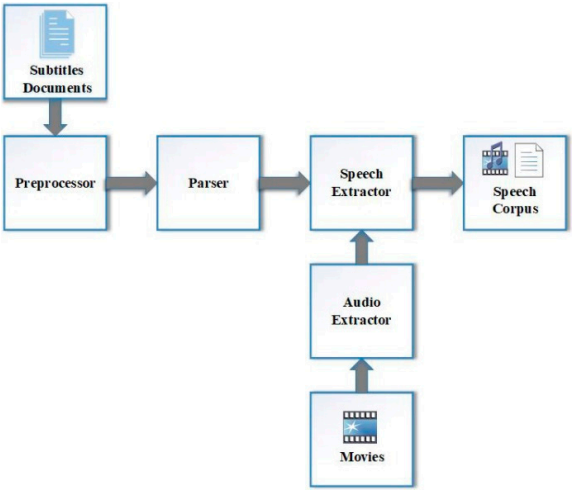


Figure 1. Block diagram of the movie speech corpus (MSC) collection process.

The preprocessor module is responsible for optimizing subtitle documents for automatic alignment. It performs pre-processing steps in the subtitle documents for this task. Each subtitle document is converted into a format that can be used in the training of the ASR system. The transcription of each speech section in the subtitle document is just below the time interval tag. The punctuation marks in the transcription prepared for each section were removed. Also, the pronunciation of the numbers and symbols were presented in written form. The transcriptions in the subtitle documents are ready for the parser module after these processes.

The parser module extracts the text that corresponds to the speech in subtitles. In this module, each speech section is transcribed in the subtitle document. Then, it is saved as a different text file. The ID provided during this process is saved to be provided to the speech file. The speech files corresponding to the text units obtained from the subtitle documents are created using the movie audio extractor.

The output of the parser and the movie audio extractor are used as inputs for the speech-extractor module, which cull speech segments from the given movie's audio. Before this process, audio information is extracted from film videos in the audio extractor module. In the speech extractor module, the speech information corresponding to the text information obtained by the parser module is extracted. The time interval of the speech information is provided at the beginning of each speech section in the subtitle document. The speech extractor process is performed according to this time interval information. The ID provided to the text file saved in the parser output is also given to the audio file in the speech extractor output. Thus, the necessary speech–text matching for the corpus was achieved.

To build a transcribed speech corpus, a total of 150 Turkish movies were collected. The film collection procedure consisted of choosing movie data by checking the summaries and the genre of movies. However, not all of the collected movies were used in the corpus collection procedure because the speech data extracted from some of these movies were found to be unsuitable due to a low speech-to-noise ratio or numerous speakers. After the screening process, 120 of these movies were selected for building the speech corpus.

Generally, the length of a speech corpus in hours is regarded as an indicator of how useful it may be to train an AM. The system developed here extracted 90 h of Turkish speech from subtitled Turkish movies, and the transcription of this extracted speech was collected from the prealigned subtitles associated with each movie. Table 1 summarizes some of the statistics of the speech data extracted from the selected movies.

Table 1. Statistics about movies used in the study.

Movie No.	No. of Words	No. of Unique Words	Length (min)
M1	4056	1308	55.0
M2	2306	900	62.0
...	...	...	...
M120	5440	1605	44.0
Avg.	4842	1229	45.0

As shown in Table 1, an average of 45.0 min of speech and transcription per processed movie were extracted using our MSC tools in as quickly as a few seconds. Table 1 also presents the number of words and the number of unique words (vocabulary size) to demonstrate the word diversity per movie. The number of words and vocabulary size together with the length of speech parts were used during the pre-selection process for eliminating movies with little speech content.

3.1.1. Pre-Processing of Subtitles

The goal of the preprocessor module is to convert each subtitle into a format that can be used by the speech recognition training tools for which the corpus is being prepared. Figure 2 presents an example of these subtitles for a Turkish movie.

```
1
00:01:31,540 --> 00:01:32,256
Evet burası çok güzel !!

2
00:01:33,860 --> 00:01:38,058
Peki siz ne zaman geliyorsunuz???

3
00:01:49,060 --> 00:01:54,373
[MUSIC]

4
00:01:55,420 --> 00:02:02,019
Aslında ben arabayı alıp gitmek istiyordum.

5
00:02:02,380 --> 00:02:02,892
<i>Buyurun gidelim.</i>
```

Figure 2. Sample subtitle document for Turkish movies.

As shown in Figure 2, each segment (time interval) of transcribed speech was numbered, and a new line was placed between successive segments. The transcription of each speech segment was immediately below its corresponding time interval tags.

The Turkish language also makes frequent use of umlaut letters, which are particularly prone to spelling, encoding, and formatting errors. For example, different forms of time interval specification,

unnecessary new lines, misspelled words, text encoding errors, and a varying number formats are the most frequently observed inconsistencies in the subtitles. In addition, subtitle documents are prepared in different text editors with varying Unicode standards. These differences cause some Turkish characters to be displayed incorrectly. For example, the Turkish characters Ü, Ç, Ö, İ, Ş, and Ğ have different encoding standards in different text editors. Therefore, for consistency and accuracy in the present study, all subtitle documents were re-encoded in a single Unicode standard: UTF-8. Also, in the pre-processing step, non-alphanumeric characters that were not spoken were removed from the given text during the pre-processing phase. All numbers in a subtitle document were converted to their official word forms because AM training requires a phonetic representation of numbers. For example, the number 86 was converted to "seksen altı", which is the corresponding written form in Turkish.

3.1.2. Misspelled Words in the Subtitle Document

Misspelled words distort the statistical distributions that are used for training AM and LM. Hence, we had to detect and remedy misspellings during pre-processing in our study. Three different categories of frequent spelling errors occurred in the subtitle documents:

- Insertion errors: One or more extra letters in the word. For example, in the misspelled word "merrhabaa," there are two letter insertion errors and the correct form should be "merhaba" (hello).
- Deletion errors: One or more letters are missing from the correct form of the word. For example, in "arbala," there are two letter deletions, and the correct form should be "arabalar" (cars).
- Substitution errors: One or more letters of the original word are replaced with other letters. For example, in "çanda," there is one substitution error, and the correct form is "çanta" (bag).

A proportion of these errors were fixed automatically using a Turkish spell checker framework. To correct the misspelled words, we used the Zemberek tool [38]. Devised for Turkic languages, mainly Turkish, Zemberek is a set of open-source natural language processing libraries and tools. Table 2 shows the statistics for the transcription corpus extracted from subtitles.

Table 2. Misspelled word statistics.

Corpus Name	Total Number of Words	Vocabulary Size	Misspelled (%)
MSC	670,546	100,128	5.2

As shown in Table 2, 5.2% of the total non-unique words were misspelled; therefore, fixing misspelled words enhanced the quality of transcription corpus by improving the accuracy of the transcription.

3.1.3. Speech Segmentation for Movies

Once the pre-processing steps were completed, time-aligned subtitle texts were ready to be used for the segmentation of speech data. First, the audio extracted from each movie was saved as a separate sound file in WAV format. The sampling rate and sampling size were set to 16 kHz and 16 bits, respectively, avoiding the need for re-adjusting the sampling rate since 16 kHz is the most common sampling rate used for speech recognition.

All the extracted audio files along with their subtitles needed to be checked to identify and correct misalignments between speech segments and their matched transcriptions in subtitle documents. Misalignment occurred when the start and end times of a speech segment in an audio file did not match the start and end times of the corresponding annotated transcription in the subtitle document. For example, a sentence started at the 120th second of the audio file and ended at the 150th second, but, in the subtitles document, this sentence was annotated between the 125th and 155th seconds. Misalignments were detected by randomly picking sample subtitles and verifying if start times were correctly marked by applying a speech recognition test that used our baseline AM and checked if

the error rate was above a certain threshold. Movies that failed this test were examined carefully by manually conducted a listening test to determine the time shift causing this misalignment. If the time shift was consistent throughout the movie, the correction of the subtitle time annotation was necessary. If not, the movie was marked as unreliable and excluded from the corpus.

3.2. Collection of Speech Data with a Mobile Application

Turkish is an agglutinating language, and the vocabulary is very large. A corpus containing only speech from films would not be sufficient for general-purpose ASR. For this reason, a mobile application was developed to increase the size of the corpus and to obtain speech and text data in different areas (magazine, sports, technology, agenda, etc.). Figure 3 shows a block diagram of the operation of the main modules for mobile data collection applications.

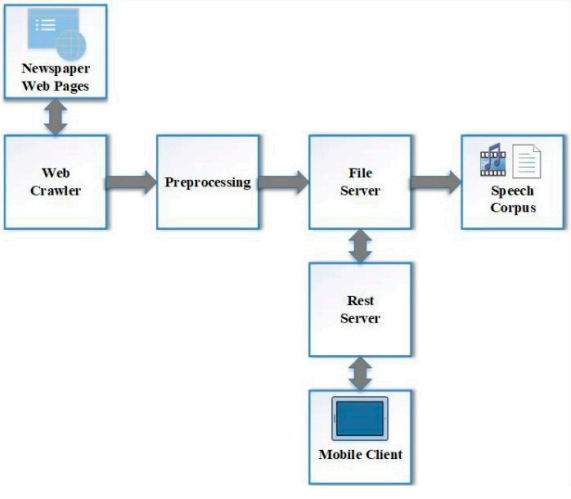


Figure 3. Collection of speech data with a mobile application.

Newly published newspaper articles were obtained with the help of a web crawler. The web crawler module had access to the published Turkish-language newspaper web pages at specific times. This procedure was limited to once a day. The identified newspaper articles were sports, magazines, and general articles. The pre-processing module analyzed the obtained articles with the help of a web crawler as a parser. Attention had to be paid to the absence of word loss at this stage. Texts separated info sentences in the preprocessor module were identified and uploaded to the file server. A representational state transfer (Rest) application programming interface (API) running on the Rest server sent previously prepared texts to a mobile client [39]. The mobile client showed the text to the user and asked the user to read the text. The created audio file created was identified and transferred to the file server. Thus, the text and voice files the users vocalized were matched and stored. A screenshot of the mobile application developed in this study and used by the real users is provided in Figure 4.

With the mobile application shown in Figure 4, the necessary text information was provided to the users. However, the user was asked to enter some information before starting the vocalization process. This information was sex, age, and name. The information received was necessary for sex distribution to be proportional when preparing the corpus. Additionally, in this study, a web interface was developed to measure the contribution of mobile application users to the corpus. Mobile users were selected from volunteer students, faculty, and staff at Gazi University (Ankara, Turkey). The native language of all mobile users was Turkish. The web interface that provides mobile user statistics is depicted in Figure 5.

Statistical information was gathered via the information from the user provided by the mobile application, such as who logs in and how many sessions they log in, the last time they sent data, how much data they processed, and how long data they vocalized. This information was followed through the prepared web interface. We found that the users created approximately 55 min of voice recordings. Also, we observed that mobile users were aged 19–52 years and the average age of mobile users was 26. The collection of the corpus via the mobile application was completed in 86 days. The developed mobile application was used by a total of 130 real people (50 men and 80 women). A total of 120 h of speech data were obtained by this method.

Figure 4. Mobile application interface developed for corpus collection.

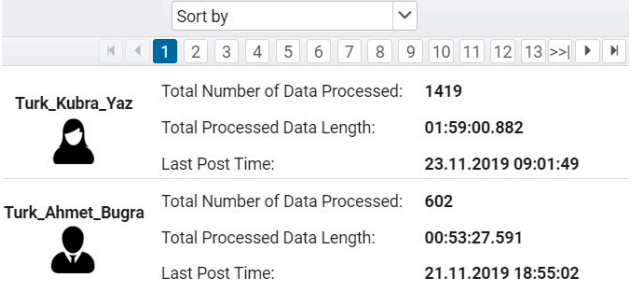


Figure 5. Web interface for mobile usage statistics.

Some difficulties were encountered when collecting the corpus with the mobile application. Different ambient noises, the speaking style of the speaker, and the motivation of the mobile application user to use the application were among the major problems. Due to the different ambient noises in the environments in which the users were located, various acoustic information was also added to the system. However, the recordings, including inaudible speeches and high ambient noise, were not added to the corpus. Some issues occurring during corpus collection were eliminated due to the selection of voluntary users from university and training provided before the use of mobile applications.

3.3. Collection of Speech Data with Transfer Learning

Transfer learning is a standard method used in machine learning to transfer information from one model to another [40]. The purpose of transfer learning is to adapt the existing resource model

to a new task with limited target data. Transfer learning on the Grand National Assembly of Turkey (GNAT) session recording (videotext) was performed.

For the transfer learning, HMM-based AMs were trained using a seven-layer time-delay neural network (TDNN). The number of cell units in the layers was selected as 600. Rectified Linear Unit (ReLU) was used as the activation function. Acoustic model training was performed with the open-source Kaldi speech recognition toolkit. A Titan X Pascal GPU (Nvidia, Santa Clara, CA 95051, USA) was used for training. Resource model training was completed in four epochs using cross-entropy. The training of the source model was completed in 94 h. The source corpus was obtained from the film subtitles and mobile applications. The corpus consisted of a total of 210 h of speech data. For the resource LM training, a text collection with shared statistics was used, as shown in Table 2. Using this collection, a 3-g-based language model was trained and this LM was used as the source. The obtained source model was applied to the speech data of the Parliament as target data. Figure 6 shows a block diagram of the main modules used for data acquisition with the transfer learning presented in the study.

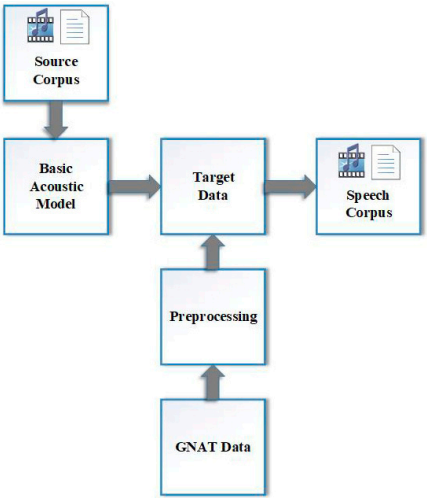


Figure 6. Collection of speech data with transfer learning.

The GNAT data module contained two different files. These files were GNAT session video recordings and session reports. The GNAT manually matched session reports and related videos, facilitating the process. In the pre-processing module, audio files were first extracted from video images. The separated audio files were 15–25 s. These audio files were used as input to the previously trained ASR system. The obtained texts from the ASR system output were investigated in the GNAT reports and the related texts (the original transcriptions) were found. A total of 250 h of speech data were obtained by the transfer learning method.

3.4. Corpus Verification Procedures

The all corpus (MSC, mobile-acquired speech and GNAT speech) obtained in this study was verified manually by real users. A web interface was created for this process. Figure 7 shows a block diagram of the main modules of corpus verification.

Audio-to-text mappings on the file server were received one by one and were transferred to a web interface with the help of Rest web service. Through the web interface, real users could control the received audio-to-text mappings. In this control process, users listened to the incoming audio data. They also controlled text information that was received with the audio file. The web interface created for the verification process is depicted in Figure 8.

As seen in Figure 8, the users could check and verify the audio–text mapping. The users could correct the text provided to them and send it back to the file server. In addition, if noise, overlap, and uncertainty existed within the audio data, they could label the data according to the specified situation. As a result of the verification process, we obtained a corpus that was controlled by real users.

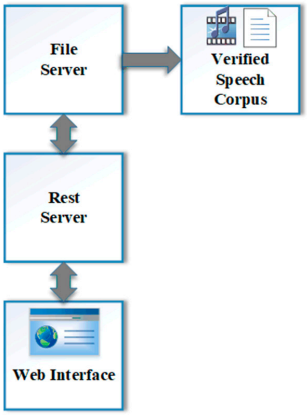


Figure 7. Corpus verification processes.

Figure 8. Web interface for the corpus verification process.

The human-based verification process requires extra time. However, it is necessary to perform the verification task to construct a totally error-free corpus. A web interface was created to obtain statistical information regarding the verification process, similar to the tracking of mobile application users. According to the statistical information obtained, an average of 2 h and 46 min of study was required to confirm one hour of speech. This indicates that the time spent on data verification was less than with the conventional transcription process. We also considered the possibility that users made mistakes. Some samples were selected from corpus data verified by each user and these samples were sent to real users again. The performance of the user performing the verification process was evaluated through this process.

4. Experimental Setup

This section explains the development of our Turkish ASR system in detail. We conducted experiments on an existing corpus and the newly collected corpus. We aimed to overcome the difficulties in the processing of Turkish. We illustrate the procedures and the steps involved for ASR and present how different choices in the design influence the Turkish ASR system performance.

4.1. Speech Corpus for Experiments

The first Turkish speech corpus was prepared by the Middle East Technical University (METU) in 2006 [23]. The METU corpus is formed from the speech of 120 different people, including 60 women and 60 men. Each person spoke 40 sentences on average and approximately 300 words. The average age of the speakers was 24 years. Another Turkish Corpus was prepared by Bogazici University in 2012 [17]. Both corpora were set to 16 kHz sampling frequencies with a 16-bit sample size. Thus, the acoustic information similarity was provided to produce more accurate results. The obtained corpus information and corpus data in this study are detailed in Table 3. The corpus developed within the scope of this study is called the HS corpus. When creating the HS corpus, three different methods were used. The real users using the mobile application included 50 men and 80 women. Since the target in the transfer learning process was corpus GNAT data, we had 490 male and 96 female speakers. Since the age of the deputies working at the GNAT was confidential, that information was unavailable. Sex distribution is a difficult task for the MSC because the same actor may act in different films. Therefore, identifying the number of different speakers was difficult. For this task, the number of actors in the films was obtained. We found that eight different people spoke in each movie on average. Classification studies on the HS corpus in the future will have the necessary statistical information.

Table 3 provides information about the Turkish corpora considered in the study. The verified HS corpus refers to a corpus verified by real users. The unverified HS corpus refers to a corpus that had not passed through the filter of real users. The speech signals were sampled at 16 kHz and parsed up to a maximum of 15 s. A text file with the transcription corresponding to each speech file was also provided.

Table 3. Information on Turkish corpora.

Corpus Name	Duration (h)	Total Number of Utterances
METU	8.33	8002
Bogazici	94.44	82,033
Unverified HS	460.12	780,014
Verified HS	350.27	565,073

The Mel frequency scale was used in feature extraction operations. The Mel frequency cepstral coefficient (MFCC) is a feature extraction technique commonly used in speech recognition systems [41]. The frequency bands are logarithmically located in the MFCC. The MFCC calculation is based on short-term analysis. In this study, for MFCC feature extraction, speech signals were divided into overlapping segments and each segment is windowed. Each segment's length is 25 ms and the overlapping ratio was taken as 40%. We calculated 13 MFCCs per segment using a Mel-scaled filterbank with 23 triangular filters distributed between 20 and 16,000 Hz.

4.2. Development of the Turkish ASR System for Experiments

The Kaldi toolkit was used for the development of the Turkish ASR system. Kaldi is an open-source toolkit for speech recognition applications written in C++ and licensed under Apache License v2.0 [42]. The Kaldi toolkit is basically connected to two external libraries. The first one is OpenFst and the other is the digital algebra library. The digital algebra library is divided into: BLAS and LAPACK. The Kaldi library modules are positioned so that each module is connected to only one of the external libraries. Access to library functions is provided by code snippets written in C++. The code snippets and libraries prepared in Kaldi are called by the scripting language to create and run the ASR system.

A Turkish ASR system was developed using version 5.0 of the Kaldi toolkit. The AM in the classical GMM-HMM-based ASR system was prepared using Gauss blend models. Based on phoneme units, all basic GMM models were created and their initial values were assigned. Then, following the coding of the texts in the training corpus, these models were properly integrated. The classical GMM model to be created from a training sample, from which phoneme models are combined, is determined

by the word-based analysis of the text on the training sample and the phoneme analysis corresponding to the words in a pronunciation lexicon. GMM model parameters for all phonemes must be trained on a large number of observations. The datasets prepared for training AM in this study were large enough to estimate model parameters.

Acoustic models can be generated in many ways using HMM, GMM, and deep neural networks (DNNs). In this study, the subspace Gaussian mixture model (SGMM)-based acoustic models were used. SGMM is an efficient method for GMM acoustic modelling by means of strong a speaker adaptation. The SGMM uses mixtures of Gaussians as the underlying state distribution. In addition, it contains the state-dependent parameters, like mean, covariance, and mixture weights. The state-dependent parameters are estimated for low dimensional subspace. In this study, the mean of the Gaussian components was calculated to create a low dimensional subspace. A maximum likelihood linear transform (MLLT) and linear discriminant analysis (LDA) for speaker adaptation training were used. To create an LDA-MLLT model, MLLT was applied on top of LDA features. Then, speaker adaptation training was performed on top of the LDA–MLLT models. Our GMM models have a total of 3500 context-dependent (CD) triphone states. In addition, they have a mean of 14 Gaussian components per state. Every triphone was modelled with three CD states.

In this study, the ASR system was developed primarily based on the classical GMM-HMM and that improved with a DNN-based system. The ASR system using neural networks is different from the classical GMM-based ASR system. In the ASR system using classical GMM-HMM, the probability of observation of each HMM state was calculated using GMM. However, in the DNN-based AM, the probability of observation of each HMM triphone state was calculated using a DNN. In the DNN-based AM, a phoneme tag was assigned to these features by taking the sound attributes. Technically, when training the GMM-HMM system, the monophone model was trained from the expression-level transcripts. To train the DNN, we obtained phoneme-sound alignments produced by the traditional GMM-HMM system. Therefore, DNN acoustic modelling depends on the properties used to train a GMM-HMM as well as the decision tree produced by the GMM-HMM. In this study, acoustic feature frames in DNNs were placed on the first layer, and the neural network assigned a phoneme label to each acoustic frame.

The DNN-based ASR was developed by taking MFCC as input into the neural network, phoneme state posteriors per frame (HMM states) were estimated, and then stochastic gradient descent (SGD) was applied to compare the estimated phoneme state posteriors with the HMM state obtained from GMM system. Then, the weights of the neural network were adjusted. To speed up the training of our neural network, we computed the gradient on a small portion of training data, known as a mini-batch, then we updated the weights soon after every mini-batch. The DNN network used five hidden layers containing 500 neurons each. In our experiment, we used either a hyperbolic tangent activation function or a P-norm generalized maxout function. Also, a 3-g LM was applied for ASR. We trained our DNN system using a Nvidia Titan X Pascal GPU.

5. Experimental Results

Performance is expressed in terms of word error rate (WER) for different experiments. Several measurement methods are available to evaluate ASR performance. The most accurate measurement method is to evaluate the differences between the hypothesis and the reference words [43]. For this reason, we did WER as a performance metric in this study. WER is calculated as

$$WER = \frac{D + S + I}{N} \times 100 \quad (1)$$

where N is the total number of symbols in the reference word, D represents the number of deleted symbols in the hypothesis with respect to the reference word, S is the number of changed symbols, and I represents the number of additional symbols.

In the study, two approaches were compared: GMM and DNN. However, we also compared the multiplicity and variety of data. Two different ASR systems were developed, and similar vocabularies and language models were used, but their acoustic models were different. In all the tests, the DNN-based system performed better than the GMM-based system. This result shows that DNN is a better choice than GMM for acoustic modelling in Turkish ASR systems. DNN performed better because the DNN was able to capture current context information by adding adjacent frames. Table 4 shows the effect of HS corpus and other corpora regarding the GMM-based Turkish ASR system.

Table 4. Results of Gaussian mixture model (GMM) based Turkish automatic speech recognition (ASR) systems.

Corpus Name	Word Error Rate (WER%)
METU	70.71
Bogazici	27.70
Unverified HS	55.27
Verified HS	24.70

For all experiments, one-third of the corpus was reserved for test purposes. As shown in Table 4, the METU corpus had the worst WER in the tests performed. The verified HS corpus produced the best WER. When the details of the corpus were examined, we observed many trivial examples within the METU corpus. These pointless examples and limited information adversely affected the performance of the ASR system. When we analyzed the case of the Bogazici corpus, we concluded that the acoustic diversity was insufficient even though the content was smooth.

The performance of the ASR system depends not only on the phonetic information obtained from the AM but also on the representation of the lexicon and syntax obtained by the LM. Therefore, in addition to the preparation of the text corpus, each developed model was tested using an LM with different n-gram values. Tests were carried out with 2-, 3-, 4- and 5-g. The most successful results were recorded at 3- and 4-g values. Table 5 lists the effect of different n-gram values on ASR systems developed on different datasets. Experiments were conducted without changing the GMM-based AM.

Table 5. Results of different n-gram values on Turkish ASR systems.

Corpus Name	N-Gram	Word Error Rate (WER%)
METU	2-Gram	78.06
METU	3-Gram	70.71
METU	4-Gram	72.46
METU	5-Gram	74.92
Bogazici	2-Gram	31.79
Bogazici	3-Gram	27.70
Bogazici	4-Gram	28.23
Bogazici	5-Gram	30.02
Unverified HS	2-Gram	63.14
Unverified HS	3-Gram	55.27
Unverified HS	4-Gram	58.93
Unverified HS	5-Gram	61.02
Verified HS	2-Gram	29.41
Verified HS	3-Gram	24.70
Verified HS	4-Gram	26.19
Verified HS	5-Gram	28.01

The results demonstrated that the phonetic content of the HS corpus performed well for balanced and different language models. The nature of Turkish explained the decline in the word-error rate for larger n-gram language models. The use of a larger set of texts for language model training led to better results.

In the tests of the DNN-based ASR system, the size of the corpus is crucial. We observed that different parameters affect the system. The depth of a DNN has a significant effect on performance. However, both the training and the decoding processing of large models are slow. Therefore, we recommend selecting the appropriate amount of hidden layers rather than choosing a large model size. Theoretically, deeper DNNs must be capable of modelling more complex functions than simple neural networks. However, optimizing the site of models in real applications is considered a problem.

In addition to the number of hidden layers and the hidden layer size, the learning rate, which is an important parameter, also affects the performance of the system. However, the size of the corpus is important in the selection of the learning rate. A high learning rate for a large corpus will require a longer training period. Another important parameter for DNN is the size of the minibatch. The size of the minibatch should be selected in intervals such as 128, 256, or 512. Using a large minibatch is useful because it considers interaction with the optimizations used in matrix multiplication, especially when a GPU is used. In the case of CPU usage, a large minibatch size can cause instability. For these reasons, the size of the minibatch was set to 128 for multi-threaded CPU-based training, and 512 for GPU-based training. The performance of DNN-based ASR systems, trained and tested on different datasets, is outlined in Table 6. The ASR system was trained on 158 h when using the verified HS corpus.

Table 6. Results of deep neural networks (DNN)-based Turkish ASR systems.

Corpus Name	Word Error Rate (WER%)
METU	64.55
Bogazici	22.63
Unverified HS	49.20
Verified HS	18.70

The results presented in Table 6 show that results on the LDC database are comparable with those obtained on the HS corpus. Experiments on the METU corpus led to a much higher WER rate. This situation can be explained by two reasons: the size and the quality of the corpus. Turkish is a language with an extensive vocabulary. Therefore, the size of the corpus and the number of unique words should be high. The speech files are very short in the METU corpus. Therefore, sufficient acoustic information in the METU corpus has not been completely captured; on the other hand, the Bogazici corpus contains news broadcasts. Therefore, the vocabulary of the corpus is small and most of the voice recordings are presenter speeches, limiting the variety of the acoustic environments.

The HS corpus includes the data obtained from different environments. Therefore, more information about the acoustic environment is provided in the HS corpus than in the other two corpora. The HS corpus not only contains political or news speeches but also those from other fields. For this reason, when an ASR system developed with HS corpus is applied to different areas, its recognition performance will be higher than that of other available corpora. A new set of experiments indicated the generalization ability of HS corpus is much better. In these experiments, ASR models developed with the HS corpus were tested on METU and Bogazici. In the experiments, one-third of both corpora was reserved for test purposes. The results obtained are provided in Table 7.

Table 7. Results of DNN-based Turkish ASR system (trained on the verified HS corpus) on different test corpora.

Training Corpus	Test Corpus	Word Error Rate (WER%)
Verified HS	METU	1.8
Verified HS	Bogazici	2.3

When the results in Table 7 are analyzed, we found that the ASR system developed with the verified HS corpus produced better results in different corpus recognition tasks. METU and Bogazici corpora

include transcriptions of broadcast news and newspaper news. Therefore, the acoustic information to be obtained from these two corpora is nearly symmetric. Most of these transcriptions have political content. The HS corpus, including particularly GNAT data, more successfully recognized political news speeches.

In this paper, a quick method was presented for creating a detailed speech recognition corpus. This method is a suitable alternative to the costly corpus preparation techniques used to construct a corpus. When tests were performed with the corpus, deep-learning-based multi-layer DNN approaches were found to be more efficient than classical GMM-based approaches. In addition, the importance of corpus size in deep learning-based multi-layer DNN approaches was demonstrated.

6. Conclusions and Future Work

In this study, we created a viable alternative procedure for collecting Turkish speech data to classical transcribed corpus preparation techniques. In the presented approach, three different steps were used. In the first step, instead of using selected pre-recorded utterances, we used movies as the source of speech utterances. Subtitle documents, which are mainly supplied for people with hearing difficulties, were then used as transcriptions for the speech utterances obtained from the movies. In the second step, speech data were collected via a mobile application from real users. Text data presented to users were taken from current newspaper texts. In the third step, a transfer learning approach was used. Thus, the most common corpus preparation and verification approaches were presented for Turkish speech recognition systems.

The main challenges during the study were the selection of appropriate movies, the construction of an appropriate LM, and creating a baseline speech recognition system. We selected appropriate movies based on their genre and their speech content inferred from the subtitle documents. We also manually verified the speech content of movies by randomly listening to parts of the movies. The results obtained from the speech recognition experiments showed that the proposed method provided an acceptable and relatively low-cost alternative for building a transcribed speech corpus. However, the cost of data collection through mobile applications increased because finding enough human resources was a difficult task. The environments in which the mobile application was used were noisy, and the inclusion of unwanted sounds complicated the process.

In transfer learning, we observed that pre-processing time on video data was a challenge. A web interface was developed to verify the obtained corpus. Actual users were asked to verify this corpus. This process was a challenging task due to the increase in human resource requirements. However, this step was crucial for training the basic model with a corpus.

The process was designed so that many other movies in Turkish could be added to increase the corpus size for further corpus improvement. For mobile applications, many articles from various fields could be collected from different web pages. Thus, diversity could be increased. The basic model could be successfully created, and the transfer learning method could be used in various areas. Thus, a corpus could be constructed for different areas and languages.

Looking at the Turkish ASR results, we observed that DNN-based approaches were better than classical GMM based approaches. In DNN-based approaches, the corpus size is quite important. To prepare the corpus, a unique approach was presented. Thus, a Turkish ASR was developed and tested with the largest available corpus. The two main outputs of this study were: the collection of the largest speech corpus in Turkish and highlighting the importance of the corpus and LM in the success of an ASR system. As a result, the Turkish corpus deficiency reported in many studies was resolved in this study. The superiority of DNN-based speech recognition approaches over classical approaches was evident in the study.

The process outlined here is also applicable for creating a domain-specific corpus for use in the training of specialized speech recognition systems. These specific systems could include the medical domain, law domain, call centres, or noisy environments. We think that this work on Turkish ASR will provide opportunities for more research on Turkish AM and LM. The designing of algorithms

for speaker and environment adaptation in DNN and designing an LM using deep neural networks are future paths to explore. Different models could be developed using the HS corpus, and with its increased use, further studies could be conducted to increase the performance of Turkish speech recognition systems.

Author Contributions: Conceptualization, H.P. and S.O.; methodology, S.O. and H.P.; software, S.O.; validation, S.O. and H.P.; formal analysis, S.O. and H.P.; investigation, S.O. and H.P.; resources, S.O. and H.P.; data curation, S.O. and H.P.; writing—original draft preparation, S.O. and H.P.; writing—review and editing, S.O. and H.P.; visualization, S.O. and H.P.; supervision, H.P.; project administration, H.P. All authors have read and agreed to the published version of the manuscript.

Funding: No additional funding was received.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Oyucu, S.; Sever, H.; Polat, H. Overview of automatic speech recognition, approaches and challenges: Way the future to Turkish speech recognition. *Gazi Univ. Sci. J. Part C Des. Technol.* **2019**, *7*, 834–854.
2. Schmiedeke, S.; Xu, P.; Ferrané, I.; Eskevich, M.; Kofler, C.; Larson, M.; Estève, Y.; Lamel, L.; Jones, G.; Sikora, T. Blip10000: A social video dataset containing SPUG content for tagging and retrieval. In Proceedings of the 4th ACM Multimedia Systems Conference, Oslo, Norway, 27 February–1 March 2013; pp. 96–101.
3. Braun, D.; Neil, D.; Liu, S. A curriculum learning method for improved noise robustness in automatic speech recognition. In Proceedings of the European Signal Processing Conference, Kos, Greece, 28 August–2 September 2017; pp. 548–552.
4. Paramonov, P.; Sutula, N. Simplified scoring methods for HMM-based speech recognition. *Soft Comput.* **2016**, *20*, 3455–3460. [CrossRef]
5. Bourlard, H.; Morgan, N. *Connectionist Speech Recognition a Hybrid Approach*, 1st ed.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1994.
6. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Pearson Education: London, UK, 2010.
7. Siniscalchi, S.M.; Svendsen, T.; Lee, C.H. An artificial neural network approach to automatic speech processing. *Neurocomputing* **2014**, *140*, 326–338. [CrossRef]
8. Jena, M.; Mishra, S. Review of Neural Network Techniques in the Verge of Image Processing. In *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications. Advances in Intelligent Systems and Computing*; Springer: Singapore, 2018.
9. Mahanty, R.; Mahanti, P.K. Unleashing artificial intelligence onto big data: A review. In *Research on Computational Intelligence Applications in Bioinformatics*, 1st ed.; Dash, S., Subudhi, B., Eds.; IGI Global: Hershey, PA, USA, 2016; pp. 1–16.
10. Aggarwal, R.; Dave, M. Acoustic modeling problem for automatic speech recognition system: Advances and refinements (Part II). *Int. J. Speech Technol.* **2011**, *14*, 1572–8110. [CrossRef]
11. Williams, J.D.; Melamed, I.D.; Alonso, B.; Hollister, T.; Wilpon, J. Crowd-sourcing for difficult transcription of speech. In Proceedings of the IEEE Workshop on Automatic Speech Recognition Understanding, Big Island, HI, USA, 11–15 December 2011; pp. 535–540.
12. Rabiner, L.R. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **1989**, *2*, 257–286. [CrossRef]
13. Janet, M.B.; Li, D.; Sanjeev, K.; Chin-Hui, L.; James, G.; Nelson, M. Historical development and future directions in speech recognition and understanding. In *Report of the Speech Understanding Working Group*; NIST: Gaithersburg, MD, USA, 2007; pp. 1–21.
14. Fry, D.B. Theoretical aspects of mechanical speech recognition. *J. Br. Inst. Radio Eng.* **1959**, *19*, 211–218. [CrossRef]
15. John, S.; William, M.; Jonathan, G. *Darpa Timit*; National Institute of Standards and Technology Computer Systems Laboratory: Gaithersburg, MD, USA, 1993; pp. 1–99.
16. Paul, D.; Douglas, B.; Baker, M. The design for the Wall Street Journal-based CSR corpus. *Assoc. Comput. Linguist.* **1992**, *6*, 357–362.

17. Arisoy, E.; Can, D.; Parlak, S.; Saraçlar, M.; Sak, H. Turkish broadcast news transcription and retrieval. *IEEE Trans. Audio Speech-Lang. Process.* **2009**, *17*, 874–883. [CrossRef]
18. Graff, D. An overview of broadcast news corpora. *Speech Commun.* **2002**, *37*, 15–26. [CrossRef]
19. Matsoukas, S. Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. *IEEE Trans. Audio Speech-Lang. Process.* **2006**, *14*, 1541–1554. [CrossRef]
20. Novotney, S.; Callison-Burch, C. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. *Hum. Lang. Technol.* **2010**, *6*, 207–215.
21. Christopher, C.; David, M.; Kevin, W. The Fisher corpus: A resource for the next generations of speech-to-text. In Proceedings of the International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004; pp. 69–71.
22. Salor, O.; Pellom, B.; Ciloglu, T.; Hacioglu, K.; Demirekler, M. Developing new text and audio corpora and speech recognition tools for the Turkish language. In Proceedings of the International Conference Spoken Language Processing, Denver, CO, USA, 16–20 September 2002; pp. 349–352.
23. Salor, O.; Pellom, B.; Ciloglu, T.; Demirekler, M. Turkish speech corpora and recognition tools developed by porting SONIC: Towards multilingual speech recognition. *Comput. Speech Lang.* **2007**, *21*, 580–593. [CrossRef]
24. Sak, H.; Saraçlar, M.; Güngör, T. Morpholexical and discriminative language models for Turkish automatic speech recognition. *IEEE Trans. Audio Speech-Lang. Process.* **2012**, *20*, 2341–2351. [CrossRef]
25. Oflazoglu, C.; Yildirim, S. Recognizing emotion from Turkish speech using acoustic features. *EURASIP J. Audio Speech Music Process.* **2013**, *1*, 1–11. [CrossRef]
26. Fort, K.; Adda, G.; Cohen, K.B. Amazon Mechanical Turk: Gold Mine or Coal Mine. *Comput. Linguist.* **2011**, *37*, 413–420. [CrossRef]
27. Schultz, T. Globalphone: A multilingual speech and text database developed at Karlsruhe University. In Proceedings of the Annual Conference of the International Speech Communication Association, Singapore, 16–20 September 2002; pp. 345–348.
28. Chan, H.Y.; Woodland, P. Improving broadcast news transcription by lightly supervised discriminative training. In Proceedings of the IEEE International Conference Acoustic Speech, Signal Processing, Montreal, QC, Canada, 17–21 May 2004; pp. 3–6.
29. Zhang, S.; Ling, W.; Dyer, C. Dual subtitles as parallel corpora. *Eur. Lang. Resour. Assoc.* **2014**, *5*, 1869–1874.
30. Lavecchia, C.; Smaïli, K.; Langlois, D. Building parallel corpora from movies. In Proceedings of the International Workshop on Natural Language Processing and Cognitive Science, Funchal, Madeira, Portugal, 12–16 June 2007; pp. 201–210.
31. Ismail, T.; Joyprakash, L. Development of speech corpora for Goalparia dialect and similar languages. In Proceedings of the IEEE International Conference on Signal and Image Processing Applications, Kuching, Malaysia, 12–14 September 2017; pp. 12–14.
32. Weonhee, Y.; Kyuchul, Y.; Sunwoo, P.; Juhee, L.; Sungmoon, C.; Ducksoo, K.; Koonhyuk, B.; Hyeseung, H.; Jungsun, K. The Korean corpus of spontaneous speech. *J. Korean Soc. Speech Sci.* **2015**, *7*, 103–109.
33. Koctúr, T.; Ondáš, S.; Juhár, J. Speech corpus generation based on N-gram confidence measure classification. In Proceedings of the International Symposium ELMAR, Zadar, Croatia, 18–20 September 2017; pp. 149–152.
34. Qasim, M.; Rauf, S.; Hussain, S.; Habib, T. Urdu speech corpus for travel domain. In Proceedings of the Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques, Bali, Indonesia, 26–28 October 2016; pp. 237–241.
35. Cahyaningtyas, E.; Arifianto, D. Development of under-resourced Bahasa Indonesia speech corpus. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 1097–1101.
36. Iribe, Y.; Kitaoka, N.; Segawa, S. Development of new speech corpus for elderly Japanese speech recognition. In Proceedings of the International Conference on Asian Spoken Language Research and Evaluation, Shanghai, China, 28–30 October 2015; pp. 27–31.
37. Magdum, D.; Shukla, M.; Patil, T.; Shah, R.; Belhe, S.; Kulkarni, M. Methodology for designing and creating Hindi speech corpus. In Proceedings of the International Conference on Signal Processing and Communication Engineering Systems, Guntur, India, 2–3 January 2015; pp. 336–339.
38. Akin, A.A.; Akin, M.D. Zemberek, an open source NLP framework for Turkic languages. *Structure* **2007**, *10*, 1–5. Available online: http://zemberek.googlecode.com/ (accessed on 5 January 2020).

39. Paik, H.; Lemos, A.L.; Barukh, M.C.; Benatallah, B.; Natarajan, A. Web Services—REST or Restful Services. In *Web Service Implementation and Composition Techniques*; Springer: Cham, Switzerland, 2017.
40. Lee, J.Y.; Dernoncourt, F.; Szolovits, P. Transfer learning for named-entity recognition with neural networks. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 7–12 May 2018; pp. 4470–4473.
41. Dave, N. Feature extraction methods LPC, PLP and MFCC. *Int. J. Adv. Res. Eng. Technol.* **2013**, *1*, 1–5.
42. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Vesely, K. The Kaldi speech recognition toolkit. In Proceedings of the Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, USA, December 2001; pp. 1–4.
43. Aksoylar, C.; Mutluergil, S.O.; Erdogan, H. The anatomy of a Turkish speech recognition system. In Proceedings of the IEEE Signal Processing and Communications Applications Conference, Antalya, Turkey, 9–11 April 2009; pp. 512–515.

symmetry

MDPI

Article

A Feasible Community Detection Algorithm for Multilayer Networks

Dongming Chen, Panpan Du, Qianrong Jiang, Xinyu Huang and Dongqi Wang *

Software College, Northeastern University, Shenyang 110169, Liaoning, China; chendm@mail.neu.edu.cn (D.C.); 17864179689@163.com (P.D.); 17854257001@163.com (Q.J.); neuhxy@163.com (X.H.)

* Correspondence: wangdq@swc.neu.edu.cn

Received: 10 January 2020; Accepted: 21 January 2020; Published: 2 February 2020

Abstract: As a more complicated network model, multilayer networks provide a better perspective for describing the multiple interactions among social networks in real life. Different from conventional community detection algorithms, the algorithms for multilayer networks can identify the underlying structures that contain various intralayer and interlayer relationships, which is of significance and remains a challenge. In this paper, aiming at the instability of the label propagation algorithm (LPA), an improved label propagation algorithm based on the SH-index (SH-LPA) is proposed. By analyzing the characteristics and deficiencies of the H-index, the SH-index is presented as an index to evaluate the importance of nodes, and the stability of the SH-LPA algorithm is verified by a series of experiments. Afterward, considering the deficiency of the existing multilayer network aggregation model, we propose an improved multilayer network aggregation model that merges two networks into a weighted single-layer network. Finally, considering the influence of the SH-index and the weight of the edge of the weighted network, a community detection algorithm (MSH-LPA) suitable for multilayer networks is exhibited in terms of the SH-LPA algorithm, and the superiority of the mentioned algorithm is verified by experimental analysis.

Keywords: multilayer network; community detection; label propagation algorithm; H-index

1. Introduction

Popular research in the field of network science is to mine hidden information under the network structure. Community detection is an important aspect of complex network research, and we can see the presence of the community in various fields, such as detecting the intensive group organization in a social network [1], the different muscle tissue composed by various genes found in the gene protein networks [2], and so on. However, effectively and accurately detecting the community structure for large-scale networks tends to be urgently addressed.

The community detection algorithms can be divided into non-overlapping community detection algorithms and overlapping community detection algorithms according to whether they contain overlapping communities or not. Non-overlapping community detection algorithm can be divided into the following categories. (1) The hierarchical clustering method defines the similarity or distance between network nodes by the topology of the given network, groups network nodes into a tree hierarchy by single-connection or full-connection hierarchical clustering, and cross-cuts the tree diagram according to actual needs to obtain the community structure. The most famous algorithm is the GN algorithm [3], which continuously deletes the edge in the network that has the maximum edge-betweenness with respect to all source nodes, and then the edge-betweenness number of the remaining edges relative to all source nodes in the network is recalculated, and the process is repeated until the network, all edges are deleted. (2) In the spectral clustering method, the objective is to find a method of dividing the nodes into disjoint sets by cutting the least-cut edges, such as the

algorithm in [4,5]. (3) In the modularity optimization method, a modularity optimization function Q is employed to describe the quality of the detected community. A larger Q value indicates a better community structure, such as the FN algorithm [6] that each node in the initialization network is a single community, and then select the most value-added community q module to merge, finally the network is merged into a community, the algorithm stopped. The overlapping community detection method allows one node to fall into one or more communities simultaneously, it can be mainly divided into the following categories: (1) the clique percolation method, such as the CPM algorithm, which is a kind of project plan management method based on mathematical calculation and belongs to positive network diagram [7], (2) an improved label propagation algorithm, such as the COPRA algorithm [8] that is an improvement of LPA, which makes the nodes with multiple tags overlap, and then discover the overlap community, and (3) methods based on local community optimization and extension, such as the LFM algorithm [9] that is to expand into a number of local associations, the completion of the community division.

The community detection algorithm has made great progress, but the time complexity of the existed algorithms is relatively high. In 2007, Raghavan et al. [10] first applied the label propagation algorithm (LPA) to community detection. Compared with the above-mentioned community detection algorithms, LPA relies only on the propagation characteristics of the network and has linear time complexity, which is suitable for the community detection and analysis for large-scale networks. However, LPA also has some disadvantages: (1) randomness of node updating order and (2) randomness of label selection. In response to the above problems, in 2014, Yan Xing et al. [11] proposed the NIBLPA algorithm, which uses k-shell decomposition to calculate the influence of each node; then, it updates and selects labels according to nodes influence. In 2015, Sun et al. [12] proposed the Cen_LP algorithm, which defines the central value and the bias value of the node, and the values are used to update and select the label. In 2017, Tamron et al. proposed the NILPA algorithm [13], where the node importance is judged according to the degree of the node, and the node similarity matrix is formed according to the random walk theory; then, these two points are combined to form new measure criteria to update the label. These algorithms improved the stability and accuracy, but at the cost of increasing the time complexity.

The research on the complex network community structure has achieved good results. However, the above-mentioned community detection algorithms mainly aim at the traditional single-layer network, but there are still no mature research achievements on multilayer networks. There are currently two methods for multilayer network community detection: merge analysis and multilayer combination analysis. Merge analysis. There exist two cases of merging analysis. (1) The first involves merging the multilayer network into a single-layer network, and then carrying out community detection using the existed community detection algorithm [14–16], but this method may ignore the topological information in each layer of a multilayer network [17]. (2) The second case involves detecting the community in each layer, and then merging the communities in different layers [18]. This method does not consider that the meaning of the nodes in each layer may be different [19]. Multilayer combination analysis directly detects the community in a multilayer network [20]. The cross-layer edge clustering coefficient (CLECC) used for multilayer network community detection is proposed based on the edge cluster coefficient, such as tensor decomposition [21,22], the method [23–25] based on modularity Qm. However, the number of communities must be an a priori condition for the tensor decomposition method, and the method based on modularity holds inherent higher time complexity.

In this paper, there are some contributions to the existing knowledge. First, by analyzing the instability of the label propagation algorithm (LPA), this paper concludes that the centrality of the node can be used to change the randomness of LPA update nodes and node labels, thereby improving the stability of the LPA algorithm. The shortcomings of the H index directly applied to the LPA algorithm are explained in detail, and the SH index is proposed. Based on this, the SH-LPA algorithm is proposed. The stability of the algorithm is verified by an example. The time complexity of the algorithm is $(NlogN)$, which is close to the linear time complexity. Secondly, in order to solve the problems such as the loss of a lot of network information when the previous multilayer network is merged into a

single-layer network, a new network fusion method is proposed in this paper. The edge weights of the fused network are determined by calculating the similarity of the nodes, and the multilayer network is fused into a weighted single-layer network. Considering the weight of the network as one of the methods to evaluate the centrality of the nodes, the MSH-LPA algorithm is proposed.

2. SH-Index-Based LPA Algorithm

2.1. The Idea of the Algorithm

The label propagation algorithm (LPA) is favored by researchers for its linear time complexity. However, the instability is a significant deficiency of the algorithm, which comes from the randomness of the order of node updating as well as the randomness of node label updating. To reduce the randomness of the LPA and simultaneously ensure that the algorithm retains linear time complexity, the influence of each node is calculated in this paper, which determines the order of node updating and node labels updating for the LPA algorithm. The basic idea of the LPA algorithm is to use the tag information of marked nodes to predict the tag information of unmarked nodes. The relationship between samples is used to build a complete relationship graph model. In a complete graph, nodes include labeled and unlabeled data, the edges represent the similarity of the two nodes, and the labels of the nodes are passed to other nodes according to the similarity. Label data are similar to a source, which can be labeled as unlabeled data. The more similar the nodes are, the easier it is for the label to spread.

By incorporating the node itself, the SH-index is proposed based on the H-index to calculate the influence of the node, which improves the robustness of the algorithm and ensures that the algorithm keeps the same efficiency with the LPA algorithm.

2.2. Related Issues and Definitions

To illustrate the process of the SH-LPA algorithm more clearly, the variables and functions employed in the algorithm are defined as follows.

2.2.1. LPA Algorithm

The idea of the LPA algorithm is that a unique label is first assigned to each node in the network, and each label just represents a community; then, the labels are updated by

$$l(i) = \underset{l}{argmax} \sum_{j \in N(i)} l_{(j)}, \tag{1}$$

where $N(i)$ represents the set of neighboring nodes of node i.

If there are multiple labels, randomly select a label until the maximum number of iterations or each label of the nodes is no longer changed; that is, the algorithm process is completed.

2.2.2. H-Index

A typical and representative indicator for describing a node's importance is degree, but this is often poorly performed when measuring the nodes that are taken as a bridge between communities; betweenness and coreness are shortest-path based indicators and are capable of evaluating the node's influence in most cases. However, this kind of computing requires the global topological information of the network, which is not applicable to large-scale networks. To find a compromised method to evaluate the influence of the node, in 2016, Zhou Tao et al. [26] expanded the H-index.

The H-index is an indicator for quantitatively evaluating the academic achievements of researchers, which was originally proposed by physicist Jorge E. Hirsh of the University of California, San Diego in 2005 [27]. The most primitive definition of a researcher's H-index is as follows: among N published papers, there are H papers that have been cited at least H times, and the remaining N-H papers were

all cited less than H times. The higher the H-index is, the stronger the influence of his paper will be. The H-index of a node means that a node has at least H neighboring nodes, and the degree of these neighboring nodes is not less than H.

Supposing a relational expression is represented as $y = F(x_1, x_2, \ldots, x_n)$, where F returns an integer number greater than 0, and the function is to find a maximum value y satisfying the condition that there exist at least y elements whose values are not less than y. Hence, the H-index of any node i is defined as

$$H(i) = F\left(k_{j1}, k_{j2}, \ldots, k_{jki}\right), \tag{2}$$

where $k_{j1}, k_{j2}, \ldots, k_{jki}$ represent the set of degrees of neighboring nodes of node i. The pseudo-code of calculating a node's H-index is presented in Algorithm 1.

Take the toy network in Figure 1 as an example; the calculated H-indexes of nodes are shown in Table 1.

Algorithm 1: H-Index

Input: network G, node n
Output: node's H-index h

```
nd = {};
h = 0;
for v in G.neighbors(n) do
    nd[v] = G.neighbors(v).length();
    snd = sorted(nd.values(), descending);
    for (i = 0; i < snd.length(); i++) do
        h = i;
        if snd[i] < i then
            break;
return h;
```

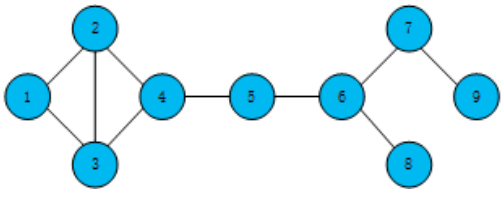


Figure 1. A simple network.

Table 1. The H-index of each node in Figure 1.

Node	1	2	3	4	5	6	7	8	9
H-index	2	2	2	2	2	2	1	1	1

2.2.3. SH-Index

Although the H-index can be applied to quickly calculate the influence of a node, the distinction of the node influence is very low, because the H-index only considers the neighboring nodes of a node but does not regard the node itself. In this paper, considering the node itself as well as its neighboring

nodes, the SH-index of node i (marked as $SH(i)$) is proposed, which is relevant to node's H-index and its neighboring nodes, and it is defined as

$$SH(i) = \frac{H(i) * \left(\prod_{j \in N(i)} H(j)\right)}{\left|N(i)\right|}, \tag{3}$$

where $N(i)$ is the set of node i's neighboring nodes, and $\left|N(i)\right|$ represents the degree of node i. The pseudo-code of calculating a node's SH-index is shown as Algorithm 2.

Algorithm 2: SH-Index

Input: network G, node n
Output: node's SH-index sh

```
1.  sh = n.H-index();
2.  N = G.neighbors(n).length();
3.  for v in G.neighbors(n) do
4.      sh *= v.H-index();
5.  sh /= G.neighbors(n).length();
6.  return sh;
```

Likewise, take the toy network in Figure 1 for instance; the H-index of node 1 is 2, the list of its neighboring nodes is [2,3], and the H-index list of neighboring nodes is [2,2]. According to Equation (3), node 1 has an SH-index of 4. Similarly, the SH-index of all nodes is shown in the following Table 2.

Table 2. The SH-index of each node in Figure 1.

Node	1	2	3	4	5	6	7	8	9
SH-index	4	5.3	5.3	5.3	4	1.3	1	2	1

In the toy network in Figure 1, we can calculate the degree, H-index, and SH-index of each node, as shown in Figure 2.

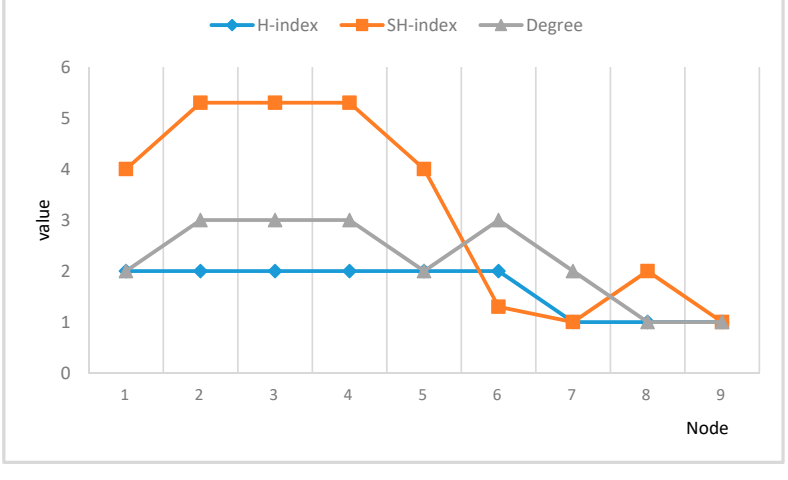


Figure 2. Comparison of degree, H-index, and SH-index.

Figure 2 shows that the SH-index can effectively solve the problem that the discrimination of nodes' H-index is not obvious for nodes with similar degrees.

By employing the SH-index for calculation, the influence of the nodes can be apparently distinguished. Therefore, according to the value of the SH-index, the order of node updating in the LPA algorithm can be improved, and ultimately the stability of the LPA algorithm can be enhanced.

2.2.4. Update Rules of the SH-LPA Algorithm

The randomness of the LPA algorithm updating comes from the randomness of the order of node updating and the randomness of node labels updating, so in order to reduce its randomness, the SH-LPA algorithm changes its updating rules from the following two aspects:

First, the order of node updating. By calculating the SH-index of each node in a graph G, sort them in ascending order, and then update the node labels following the sorted order. Updating the labels in ascending order can make the algorithm converge as soon as possible, because a node with a small SH-index is first updated to a node label with a large SH-index in the neighbor, so that when a node with a large SH-index is updated, the label of the neighboring node is exactly its label and resulted without being updated; therefore, the algorithm can converge more quickly.

Second, the order of node labels updating. The node label is first updated according to Equation (1). When there are multiple choices, we update the current node's label by selecting the node label with the maximal SH-index among the neighboring nodes of the current node rather than just randomly select one, as indicated by

$$l(i) = \underset{l}{argmax} \sum_{j \in N(i)} SH(j). \tag{4}$$

If there is still more than one result, then any one of them is randomly selected as the node label for updating.

2.3. Procedures of SH-LPA Algorithm

Given a network $G = (V, E)$, the process of the SH-LPA algorithm is as follows:

First step: calculate the SH-index of each node in G

(1) Traverse each node in G, calculate the H-index of each node in terms of Equation (2), then store each node and its H-index value as a dictionary node_h_index;

(2) Traverse each node in G again, calculate the SH-index of each node according to Equation (3) and the node H-index of node_index, and store each node and the corresponding SH-index into a dictionary node_sh_index;

(3) Sort node_sh_index in ascending order.

Second step: updating the process of the SH-LPA algorithm

(1) Initialize each node in G as a unique label;

(2) Obtain the SH-index list visit sequence of each node;

(3) Traverse each node in the visit sequence in turn and update the label of the node in terms of the update rules in Section 2.2.4;

(4) Repeat Step (3) until the label of each node reaches the maximum value of the neighboring node label or the algorithm iterates to the maximum number of times, and the algorithm terminates.

Third step: re-traverse each node in graph G, and then store them in the dictionary communities with the node label as the key and the node as the value, so that the nodes with the same label share the same key; that is, the community division is completed.

The pseudocode of the SH-LPA algorithm and method of calculating the SH-index are as shown in Algorithm 3.

Algorithm 3: SH-LPA

Input: a network G
Output: community C

```
initialize node's label in G and calculate node's SH-index;//according to Equation (3);
visitSequence;//sorting node's SH-index by ascending order;
i = 0;
while i < K or node's label != neighbor's maximum label do
    i += 1;
    for v in visitSequence do
        label = G.node(v).label;
        m = getMaxNeighborLabel(v);//according to Equation (1);
        if m.length() > 1 then
            L = getMaxNeighborSHIndex(v);//according to Equation (4);
            if L.length() > 1 then
                label = random.choice(L);
            label = L;
for n in G.nodes do
    l = n.label;
    C[l].append(n);
return C;
```

2.4. Complexity Analysis

Given a network G, the number of nodes is N, and the average number of neighboring nodes of each node is K.

2.4.1. Space Complexity

For this network G, the space required to store each node in the network is $O(N)$; during the execution of the algorithm, initializing a unique label for each node requires space $O(N)$; the space required to store the result of calculating H-index is $O(N)$. According to the H-index of the node, the space required to store the SH-index is $O(N)$; when sorting the SH-index result sequence, the required space complexity is $O(logN)$ by the fast sorting algorithm. Therefore, the total space complexity of the algorithm is $O(4N + logN)$, which is simplified as $O(N + logN)$.

2.4.2. Time Complexity

First, initialize a unique label for the node and traverse each node in the graph; the time complexity is $O(N)$. Then, calculate the H-index of each node and find the neighboring nodes of each node; the time complexity is $O(k)$, so the time complexity for finding the neighboring nodes of all nodes is $O(kN)$. The result of the calculated H-index is also stored as the data structure of the dictionary, and the SH-index of the node is calculated according to the H-index of the node. The time complexity of the H-index of the neighbor node of each node is $O(k)$, the time complexity of finding the H-index is $O(1)$, the total time complexity is $O(kN)$, and the data structure of the dictionary is stored. The SH-index sequence of the node is sorted in ascending order, and the time complexity is $O(NlogN)$. Then, the time complexity of the SH-LPA algorithm used in this part is $O(N + 2kN + NlogN)$, which is approximate to $O(kN + NlogN)$.

Then, according to the ascending sequence of the SH-index, the process of the LPA algorithm is executed, and the time complexity is $O(N)$. Assuming that the algorithm converges after m iterations, the time complexity is $O(mN)$. Then, the total time complexity of the SH-LPA algorithm is

$O(kN + mN + NlogN)$, which is simplified to $O(kN + NlogN)$. That is, the SH-LPA algorithm is still close to linear time complexity.

3. Community Detection Algorithm for Multilayer Networks (MSH-LPA)

3.1. Constructing the Model for Multilayer Networks

A multilayer network can be regarded as a combination of multiple single-layer networks, but with the same number of nodes in each layer, various edges between nodes in the different layers, and the possibility of isolated nodes. The nodes between any two layers are a one-to-one correspondence. Therefore, a multilayer network consisting of L layers can be represented as $G = \left\langle G^{(1)}, G^{(2)}, \ldots, G^{(l)} \right\rangle$, where $l \in L$ and $G^{(l)} = (V, E)$. At present, the main merging methods are as follows: Reference [28] defines a merged adjacency matrix based on a multilayer network. If in a layer or layers of a multilayer network, two nodes are connected by at least one edge, an edge exists between these two nodes in the matrix. This method is easy to understand but ignores the fact that the edges between the same nodes in different layers of a multilayer network represent different meanings. In addition, if community detection is performed using the merged adjacency matrix, the result may be inaccurate, because it does not well reflect the tightness between the multilayer network nodes. The authors in [29] proposed a method called Network Integration to integrate information by calculating the average interaction of nodes in a multilayer network. This method considers the fact that the interaction between the different layers of the network is different, but it treats each layer of the network as equivalent, which makes the network different from the actual situation. Strehl et al. [30] proposed Partition Integration, which first performs community detection at each layer and then constructs a structural similarity matrix for each layer. Within a multilayer network, if two nodes in each layer belong to the same community, then the similarity of these two nodes is 1; otherwise, it is 0. However, only 0 and 1 are insufficient to describe the similarity of each single-layer network because the similarity of the two nodes is different in each layer, but here, they are all set to 1. Some researchers consider the number of edges between two nodes in the process of merging, so that the number of edges is accumulated, and it is regarded as the weight of the edge after merging.

As we have known, in each layer, the meaning of the connected edges between two corresponding nodes in a multilayer network is different, such as the edge between two nodes in a layer representing a relative relationship, but in another level, the connection between the two corresponding nodes may represent a friend relationship, or it may also represent a business relationship, and so on. According to common sense, we know that the edges with a relationship of relatives and friends are more important than that of business, so the weight of the edges should be distinguished, and it is obviously not appropriate to simply accumulate the weights or the number of edges. The following describes the multilayer network merging method proposed in this paper.

In a complex network, the greater the similarity between two nodes, the more similar the two nodes tend to be, and naturally the closer the relationship of the two nodes will be. Therefore, the weight of the edge is obtained by calculating the similarity between two nodes of an edge. The larger the value of the similarity, the larger the weight of the edge will be. In this paper, the similarity is calculated using Jaccard similarity, which is formulated as

$$S_{a,b} = \frac{|A \cap B|}{|A \cup B|}, \tag{5}$$

where A represents the set of neighboring nodes of node a, and B represents the set of neighboring nodes of node b.

In the process of calculating similarity, two nodes in a multilayer network have no connected edges at each layer, so the similarity is not calculated even if the similarity is high, because in the process of merging the network, if there is no edge in each layer, then there must be no connected

edges after merging. Considering an edge that exists in one layer between two nodes but no edge in another layer between the two corresponding nodes, we define two different types of edges:

same_layer_edge: the edge that exists between the nodes in layer l of the multilayer network;

latent_edge: the edge that exists in layer l but does not exist in the other one or more layers.

Depending on the type of the edge, we define the weights of the edges of the merged network as follows:

$$w(a,b) = S_s(a,b) + S_l(a,b), \tag{6}$$

where $S_s(a,b)$ denotes the result by employing same_layer_edge, and $S_l(a,b)$ is the result by using latent_edge.

According to Equation (6), by looping through each layer of the multilayer network, the weights of all edges of the merged network can be calculated until a weighted network is ultimately obtained.

3.2. MSH-LPA Algorithm

After building the multilayer network model, we obtained a weighted network. The larger the sum of the weights of all the edges of a node, the greater the influence of the node will be. Therefore, based on the SH-LPA algorithm, the MSH-LPA algorithm considers the weight of the edge of the node. The influence of the node is calculated by the sum of the SH-index of the node and the weight of the node (indicated as the MSH-index), and the updating order of the nodes and labels of nodes in the network are determined in terms of the size of the MSH-index of the node.

3.2.1. SH-Index Processing

From the calculation of the weight of the merged network, the similarity between two nodes' ranges can be concluded $[0,\ 1]$. Assuming that each layer of the L-layer network is kept the same, and the maximal similarity of the two corresponding nodes is employed, the weight of the merged network is in the range of $\alpha \times [0,\ L]$, $\alpha \in [-1, 1]$, and therefore the weight ranges $[-L,\ L]$.

In this paper, the log function is employed to reduce the SH-index by a certain proportion, and a new SH-index (denoted as ($\acute{SH}$)) is obtained, which is formulated as

$$\acute{SH}(i) = \log(\mathrm{SH}(i)). \tag{7}$$

3.2.2. MSH-Index

After the normalization of the SH-index, the numerical ranges of the SH-index and the weight are approximately the same, so the weight and the $\acute{SH}$ index can be jointly used to evaluate the influence of the node, which is denoted as follows:

$$MSH(i) = \acute{SH}(i) + \sum_{j \in N(i)} \frac{w(i,j)}{\left|N(i)\right|}, \tag{8}$$

where $N(i)$ is the set of neighboring nodes of i, and $\left|N(i)\right|$ represents the number of neighbors. The metric for evaluating i is better because it considers the influence that comes from the neighbors of different layers more. $\acute{SH}(i)$ depicts the basic influence of node i in a conventional graph model, which dominates the updating order in the improved label propagation algorithm (i.e., MSH-LPA). The influence of neighboring nodes from different layers are represented by $w(i,\ j)$ in the transformed weighted network, and it is divided by the degree of node i, so the influence is described as $\sum_{j \in N(i)} \frac{w(i,j)}{\left|N(i)\right|}$, which is mainly used to distinguish the nodes with the same SH-index. The experiments conducting on SH-LPA have proved that the algorithm is more stable than LPA, and we have fully utilized the layers information and made the nodes easier to distinguish, so the metric is better than the previous one, as the comparison in experiment illustrated.

3.2.3. Updating Rules of MSH-LPA

The MSH-index is proposed based on the SH-LPA algorithm, so the MSH-index determines the order of node updating and node label updating in the MSH-LPA algorithm.

First, update the order of nodes. Here, we follow the same process as the order of node updating for the SH-index in Section 2.2.4, except that we replace the SH-index with the MSH-index.

Second, update the order of labels. Here, we still follow the same process as the order of node labels updating for the SH-index in Section 2.2.4, except that we replace the SH-index with the MSH-index, which is formulated as

$$l(i) = \underset{l}{argmax} \sum_{j \in N(i)} MSH(j), \tag{9}$$

where $N(i)$ is the set of neighboring nodes of node i.

If there is still more than one maximal neighboring labels at this time, then one of them is randomly selected as the node label for updating.

The detailed implementation process is essentially in agreement with the SH-LPA algorithm, except that SH is replaced by MSH.

3.3. Complexity Analysis

For a merged network MG, the number of nodes is defined as N, the average degree of nodes is k, and the number of edges is E.

3.3.1. Space Complexity

For this merged network MG, the space required to save each node in the network is $O(N)$); the space required to store the weight of the edge is $O(E)$.

Algorithm initialization phase: Initialize a unique label for each node, in which the required space is $O(N)$. after calculating the node's H-index, the result needs to be stored, and the required space is $O(N)$. According to the node's H-index, the space required to store the result of the SH-index is $O(N)$; the space complexity required to calculate the $S\acute{H}$ index is $O(1)$; and the space complexity required to store the MSH-index is $O(N)$. When sorting the MSH-index result sequence, the required space complexity is $O(logN)$ by the fast sorting algorithm. Therefore, the subtotal space complexity of the algorithm is $(E + 5N + logN)$, and it is approximated as $O(E + logN)$.

3.3.2. Time Complexity

Initializing the label of the node in the graph MG requires traversing each node in the graph with a time complexity of $O(N)$.

Calculating the MSH-index of each node: (1) For the H-index of each node, the time complexity required to traverse the neighboring nodes of the node is $O(k)$, and the H-index calculation result of the node is stored as the data structure of the dictionary. So, the time complexity of N nodes is $O(kN)$. (2) Then, we calculate the SH-index of the node according to the H-index of the node, and we also need to find the H-index of the neighboring node; here, the time complexity is $O(1)$, the time complexity of traversing the neighboring nodes is $O(k)$, and the time complexity for storing the SH-index as a dictionary data structure is $O(kN)$. (3) The data of the node's SH-index is normalized to obtain the $S\acute{H}$-index, and the time complexity is $O(N)$. (4) When calculating the MSH-index of a node, it is necessary to know the weights of all the edges of the node, and still traverse the neighboring nodes of the node; here, the time complexity is $O(k)$, and the time complexity is $O(kN)$ for N nodes. (5) The time complexity of sorting the MSH-index sequence in ascending order is $O(NlogN)$. Then, the partial time complexity of the MSH-LPA algorithm is $O(N + 3kN + NlogN)$, and it is approximated as $O(kN + NlogN)$.

The process of the LPA algorithm: Execute the LPA algorithm following the SH-index in ascending order, in which the time complexity is $O(N)$. Assuming that the algorithm converges after the algorithm iterates for m times, the time complexity is $O(mN)$.

After analyzing the time complexity in the three main stages of the MSH-LPA algorithm, the total time complexity of the algorithm is $O(N + kN + NlogN + mN)$, which can be approximated as $O(N + NlogN)$.

4. Experimental Results and Analysis

In this chapter, the SH-LPA algorithm and the MSH-LPA algorithm are compared and analyzed with the LPA algorithm and CDMN algorithm that divides communities by calculating the influence of nodes [31], respectively. We set up the following experimental environment: processor Intel (R) Core (TM) i7-2600CPU@3.40GHz, Memory 8GB, Hard disk 930G, Operating System Windows10, Programming Language Python 3.7.

4.1. SH-LPA Algorithm

The following five network datasets were employed for this experiment. The evaluation index is modularity, and the higher the modularity, the better the experimental results.

4.1.1. Dolphin Network

The dolphin network is a network of dolphins that Lusseau et al. used for seven years to observe the exchanges between 62 dolphins in the Doubtful Sound Channel; the network comprised 62 nodes and 159 edges, in which the average node degree was 5.1290.

Experimenting on the dolphin network, the modularity changing trends of LPA, SH-LPA, GN, and SCAN that nodes are clustered according to the way they share neighbors are shown in Figure 3.

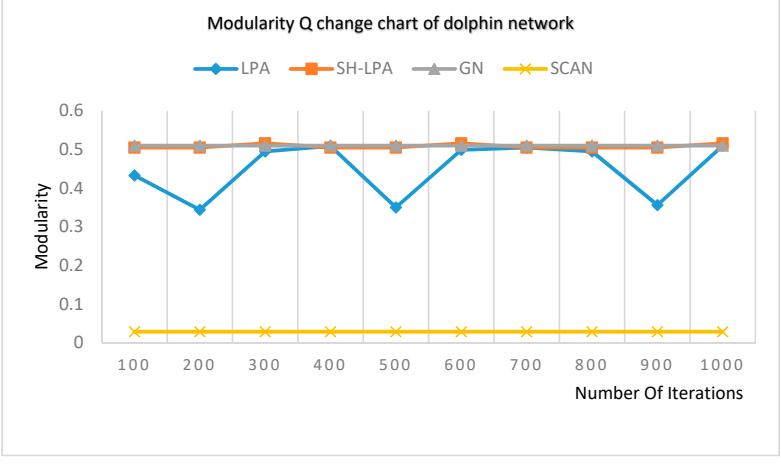


Figure 3. Modularity Q change chart of the dolphin network.

It can be seen from Figure 3 that the modularity of the LPA algorithm fluctuates when the number of iterations follows between 100 and 1000 because of the randomness of the LPA algorithm. The modularity of the improved SH_LPA is marked as an orange line and is relatively stable and even higher than LPA.

4.1.2. Email Network

The Enron email communication network (http://snap.stanford.edu/data/email-Enron.html) covers all the email communication within a dataset of around half a million emails. This data were originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. This dataset is the largest connected subgraph, comprising 291 nodes and 3099 edges, in which the average node degree equals 21.2990.

Experimenting on the email network, the modularity changing trends of LPA, SH-LPA, GN, and SCAN are shown in Figure 4.

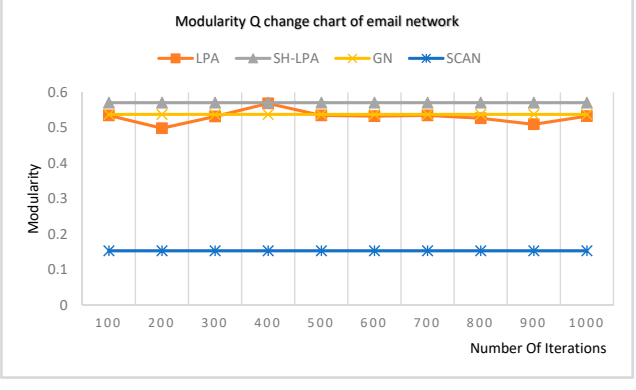


Figure 4. Modularity *Q* change chart of the email network.

It can be concluded from Figure 4 that the modularity of the LPA algorithm fluctuates from 100 to 400 on account of the randomness of the LPA algorithm. The modularity of the improved SH_LPA, which is marked as a gray line, is comparatively stable and even higher than that of the LPA.

4.1.3. Chengdu Bus Route Network

The network of the Chengdu bus route (https://www.neusncp.com/api/view_dataset?dataset_id=163) comprises 1895 nodes and 3051 edges, in which the average node degree is 3.2760. The dataset of the transportation system in Chengdu, China was collected by our team members manually.

Experimenting on the Chengdu bus route network, the modularity changing trends of LPA, SH-LPA, GN, and SCAN are shown in Figure 5.

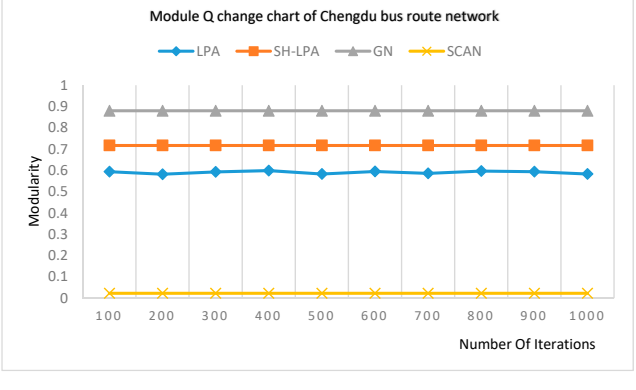


Figure 5. Modularity *Q* change chart of the Chengdu bus route network.

It can be seen from Figure 5 that the modularity of SH_LPA is reasonably stable and even higher than that of the LPA algorithm.

4.1.4. DBLP Collaboration Network

The network of the DBLP (Digital Bibliography & Library Project) collaboration (http://snap.stanford.edu/data/com-DBLP.html) comprises 3911 nodes and 6244 edges, in which the average node degree is 3.1930. Since the GN algorithm does not run out of results in the same time, we use the largest connection subgraph of the author's network.

Experimenting on the authors' network, the modularity changing trends of LPA, SH-LPA, GN, and SCAN are shown in Figure 6.

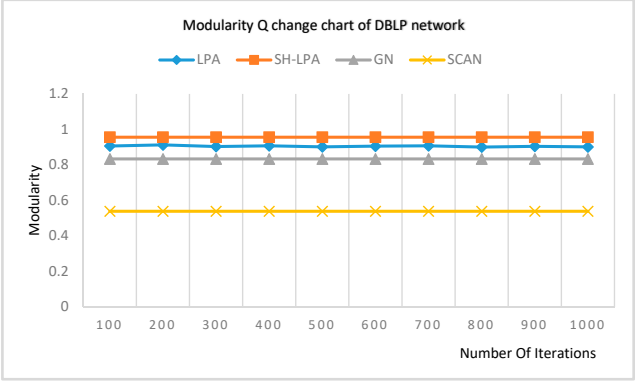


Figure 6. Modularity Q change chart of the network of the DBLP collaboration.

It can be seen from Figure 6 that the modularity of the SH_LPA is relatively stable and even higher than that of the LPA algorithm.

4.1.5. Network of Scientists Cooperation

The original dataset (http://www.umich.edu/~{}mejn/centrality) contains 1589 nodes and 2742 edges. This dataset is the largest connected subgraph, which contains 379 nodes and 914 edges, and the average node degree is 4.8232, mainly representing co-authorships between 379 scientists whose research centers on the properties of networks of one kind or another.

Experimenting on the scientists' cooperation network, the modularity changing trends of LPA, SH-LPA, GN, and SCAN are shown in Figure 7.

It can be seen from Figure 7 that the modularity of the LPA algorithm fluctuates from 100 to 300. This is because of the randomness of the LPA algorithm. The improved SH_LPA is relatively more stable than that of the LPA and simultaneously holds a higher modularity.

It can be seen from the above five figures that line charts of the SH-LPA algorithm close to a straight line and the line charts of the LPA algorithm are more complicated in the dolphin network, email network, Chengdu bus route network, authors' network of DBLP, and the scientists' cooperation network. In short, the variation range of modularity Q in the SH-LPA algorithm is smaller than that in the LPA algorithm, and the SH-LPA algorithm is smoother than the LPA algorithm. Therefore, the experimental results and analysis from the above five experimental datasets can sufficiently prove that the SH-LPA algorithm proposed in this paper improves the stability of the LPA algorithm.

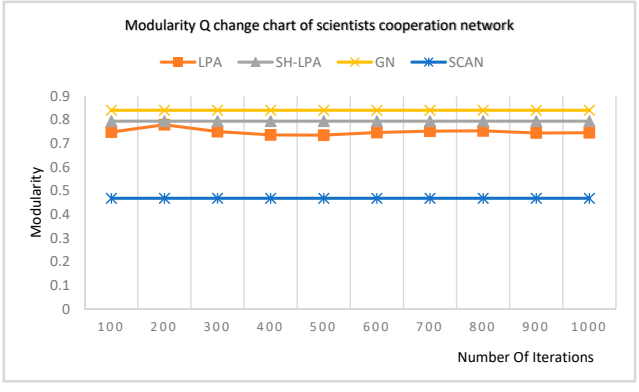


Figure 7. Modularity *Q* change chart of the scientists' cooperation network.

According to the modularity results of the above five experimental SH-LPA, LPA, GN, and SCAN algorithms, the average modularity is shown in Figure 8.

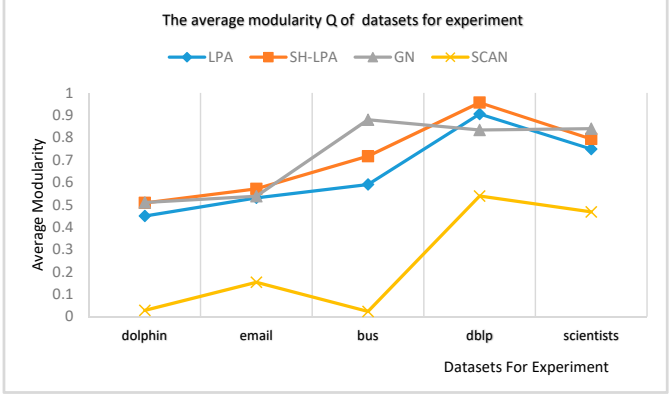


Figure 8. The average modularity *Q* of the dolphin, email, bus, DBLP, and scientists' networks.

It can be seen from Figure 8 that the average modularity of the SH-LPA algorithm in this paper is comparatively higher than the LPA and the SCAN algorithm, and it is even slightly higher than the GN algorithm accounting to the average modularity. It can be concluded that the SH-LPA algorithm outperforms the LPA algorithm in modularity comparison. It proves that the proposed SH-LPA algorithm improves the stability as well as the accuracy.

4.2. MSH-LPA Algorithm

The experimental results and analysis are based on modularity. The following four datasets are employed as the experimental multilayer networks.

4.2.1. Students' Cooperation Social Network (SCSN)

The dataset [31] is a social network built on the homework of 185 students in two different majors at Ben-Gurion University to complete the compulsory course of computer network security. The network has a total of 360 edges with three types—'time', 'computer' and 'partner'; here, 'time' denotes that two students link with each other if they submit assignments within the same period, 'computer'

means students submit the assignment on the same computer, and 'partner' indicates that students complete the assignment together.

4.2.2. Enron's Mail Network

The network [32] consists of 151 nodes and 266 edges, and there are two types of edges: mail exchanges between supervisors and subordinates and mail exchanges between colleagues.

4.2.3. Indonesian Terrorist Network

The Noordin top terrorism network [33] was drawn primarily from the "Terrorism in Indonesia: Noordin's Network", which is a publication of the International Crisis Group (2006) and includes relational data on the 79 individuals listed in Appendix C of that publication. The data were initially coded by Naval Postgraduate School students as part of the course "Tracking and Disrupting Dark Networks" under the direction of Professor Sean Everton, Co-Director of the Core Lab, and Professor Nancy Roberts. CORE Lab Research Assistant Daniel Cunningham reviewed and cleaned all the coding made by students. This paper extracts four types of edges from the relationship of the 79 people in the network dataset, namely, 'classmates' (175 edges of classmate relationship), 'training' (186 edges of training relationship), 'communication' (200 edges of communication), and 'business' (30 edges of business dealings).

4.2.4. 9/11 Terrorist Dataset

The 9/11 terrorist dataset [34] contains 62 nodes and 153 edges. In the real world, most terrorists of the dataset started as friends, colleagues, or relatives; they were drawn closer by bonds of friendship, loyalty, solidarity, and trust, and rewarded by a powerful sense of belonging and collective identity. The data are supplied in an edge-list file, in which two numbers signify the strength of tie (5 = strong tie, 1 = weak tie) and the level to which the tie has been verified (1 = confirmed close contact, 2 = various recorded interactions, 3 = potential or planned or unconfirmed interactions).

The modularity obtained by the MSH-LPA algorithm and CDMN algorithm on the above four network datasets is shown in Figure 9.

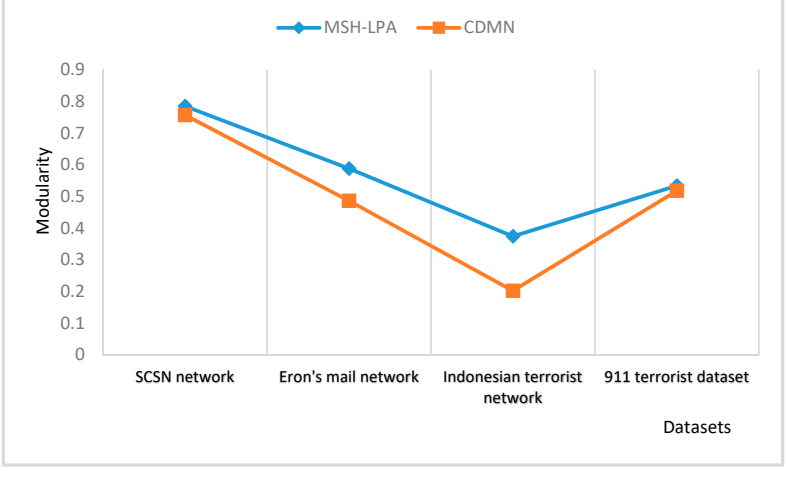


Figure 9. Comparison of the community detection algorithm suitable for multilayer networks (MSH-LPA) algorithm and CDMN algorithm.

As shown in Figure 9, the MSH-LPA algorithm obtains higher modularity conducting on the four real-world datasets than the CDMN algorithm.

5. Conclusions

By analyzing the instability of the label propagation algorithm (LPA), it is concluded that the randomness of node and node labels updating in the LPA algorithm can be changed by calculating the centrality of the node, and then improving the stability of the LPA algorithm. The deficiency of the H-index directly applied to the LPA algorithm is described in detail, and the SH-index is proposed. Based on the SH-index, the SH-LPA algorithm is presented. The stability of the algorithm is verified by experiments, as is the time complexity of the algorithm is $O(kN + NlogN)$, which is close to linear time complexity.

In order to solve the problem that much network information may be lost when merging a multilayer network into a single-layer network, the similarity of the nodes is employed to determine the weight of the edge of the merged network, and the multilayer network is merged into a weighted single-layer network, in which the SH-index and the weight of the node jointly determine the order of node and node labels updating. Here, we propose a more accurate MSH-LPA algorithm.

In order to verify the superiority of the SH-LPA algorithm and the MSH-LPA algorithm, the experimental results on five datasets show that the SH-LPA algorithm improves the stability of the LPA algorithm. Compared with the CDMN algorithm on the four multilayer network datasets, it is proved that the MSH-LPA algorithm proposed in this paper achieves larger modularity than the CDMN algorithm, which indicates its higher accuracy.

Author Contributions: Conceptualization, D.C. and Q.J.; methodology, Q.J. and D.C.; software, P.D. and X.H.; validation, X.H., D.C. and D.W.; formal analysis, D.P.; investigation, Q.J.; resources, P.D.; data curation, Q.J.; writing—original draft preparation, P.D. and Q.J.; writing—review and editing, D.C. and D.W.; visualization, X.H.; supervision, D.C.; project administration, D.C. and D.W.; funding acquisition, D.C. and D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Liaoning Natural Science Foundation under Grant No. 20170540320, the Doctoral Scientific Research Foundation of Liaoning Province under Grant No. 20170520358, and the Fundamental Research Funds for the Central Universities under Grant Nos. N161702001 and N172410005-2.

Acknowledgments: We would like to thank the anonymous reviewers for their careful reading and useful comments that helped us to improve the final version of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rozario, V.S.; Chowdhury, A.; Morshed, M.S.J. Community Detection in Social Network using Temporal Data. *arXiv* **2019**, arXiv:1904.05291.
2. Jeong, H.M.; Mason, S.P.; Barabási, A.L.; Oltvai, Z.N. Lethality and Centrality in Protein Networks. *Nature* **2001**, *411*, 41–42. [CrossRef] [PubMed]
3. Newman, M.E.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2004**, *69*, 026113. [CrossRef] [PubMed]
4. Shiga, M.; Takigawa, I.; Mamitsuka, H. A spectral clustering approach to optimally combining numericalvectors with a modular network. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 12–15 August 2007; pp. 647–656.
5. Jiang, J.Q.; Dress, A.W.; Yang, G. A spectral clustering-based framework for detecting community structures in complex networks. *Appl. Math. Lett.* **2009**, *22*, 1479–1482. [CrossRef]
6. Newman, M.E. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **2004**, *69*, 066133. [CrossRef] [PubMed]
7. Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814. [CrossRef]
8. Gregory, S. Finding overlapping communities in networks by label propagation. *New J. Phys.* **2010**, *12*, 103018. [CrossRef]

9. Lancichinetti, A.; Fortunato, S.; Kertesz, J. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* **2009**, *11*, 033015. [CrossRef]
10. Raghavan, U.N.; Albert, R.; Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **2007**, *76*, 036106. [CrossRef]
11. Xing, Y.; Meng, F.; Zhou, Y.; Zhu, M.; Shi, M.; Sun, G. A node influence based label propagation algorithm for community detection in networks. *Sci. World J.* **2014**, *2014*(5), 627581. [CrossRef] [PubMed]
12. Sun, H.; Liu, J.; Huang, J.B.; Wang, G.T.; Yang, Z.; Song, Q.B.; Jia, X.L. CenLP: A centrality-based label propagation algorithm for community detection in networks. *Phys. A Stat. Mech. Appl.* **2015**, *436*, 767–780. [CrossRef]
13. Ma, T.; Xia, Z. An improved label propagation algorithm based on node importance and random walk for community detection. *Mod. Phys. Lett. B* **2017**, *31*, 1750162. [CrossRef]
14. Berlingerio, M.; Coscia, M.; Giannotti, F. Finding and characterizing communities in multidimensional networks. In Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, Kaohsiung, Taiwan, 25–27 July 2011; pp. 490–494.
15. Kazienko, P.; Musial, K.; Kukla, E.; Kajdanowicz, T.; Bródka, P. Multidimensional social network: Model and analysis. In Proceedings of the International Conference on Computational Collective Intelligence, Gdynia, Poland, 21–23 September 2011; pp. 378–387.
16. Rossetti, G.; Berlingerio, M.; Giannotti, F. Scalable link prediction on multidimensional networks. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 979–986.
17. Tang, L.; Wang, X.; Liu, H. Uncoverning groups via heterogeneous interaction analysis. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, Miami, FL, USA, 6–9 December 2009; pp. 503–512.
18. Berlingerio, M.; Pinelli, F.; Calabrese, F. Abacus: Frequent pattern mining-based community discovery in multidimensional networks. *Data Min. Knowl. Discov.* **2013**, *27*, 294–320. [CrossRef]
19. De Domenico, M.; Lancichinetti, A.; Arenas, A.; Rosvall, M. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X* **2015**, *5*, 011027. [CrossRef]
20. Bródka, P.; Filipowski, T.; Kazienko, P. An introduction to community detection in multi -layered social network. *Ccis* **2012**, *278*, 185–190.
21. Kolda, T.; Dunlavy, D.; Kegelmeyer, W. Multilinear algebra for analyzing data with multiple linkages. In Proceedings of the Submitted to Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006.
22. Leginus, M.; Dolog, P.; Žemaitis, V. Improving tensor based recommenders with clustering. In Proceedings of the International Conference on User Modeling, Adaptation, and Personalization, Montreal, QC, Canada, 16–20 July 2012; pp. 151–163.
23. Radicchi, F.; Arenas, A. Abrupt transition in the structural formation of interconnected networks. *Nat. Phys.* **2013**, *9*, 717. [CrossRef]
24. Nicosia, V.; Bianconi, G.; Latora, V.; Barthelemy, M. Growing multiplex networks. *Phys. Rev. Lett.* **2013**, *111*, 058701. [CrossRef] [PubMed]
25. De Domenico, M.; Solé-Ribalta, A.; Cozzo, E.; Kivelä, M.; Moreno, Y.; Porter, M.A.; Gómez, S.; Arenas, A. Mathematical Formulation of Multi-Layer Networks. *Phys. Rev. X* **2013**, *3*, 4192–4195.
26. Lü, L.; Zhou, T.; Zhang, Q.-M.; Stanley, H.E. The H-index of a network node and its relation to degree and coreness. *Nat. Commun.* **2016**, *7*, 10168. [CrossRef] [PubMed]
27. Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569–16572. [CrossRef] [PubMed]
28. Magnani, M.; Micenkova, B.; Rossi, L. Combinatorial analysis of multiple networks. *arXiv* **2013**, arXiv:1303.4986.
29. Lusseau, D.; Schneider, K.; Boisseau, O.J.; Haase, P.; Slooten, E.; Dawson, S.M. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Soc.* **2003**, *54*, 396–405. [CrossRef]
30. Zhu, G.; Li, K. A Unified Model for Community Detection of Multiplex Networks. In Proceedings of the Web Information Systems Engineering—WISE 2014, Thessaloniki, Greece, 12–14 October 2014.

31. Huang, X.; Chen, D.; Ren, T. Community Discovery Algorithm for Multi-relationship Networks. *J. North. Univ. Nat. Sci.* **2018**, *39*, 1375–1379. [CrossRef]
32. Mucha, P.J.; Richardson, T.; Macon, K.; Porter, M.A.; Onnela, J.-P. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* **2009**, *328*, 876. [CrossRef]
33. Kitsak, M.; Gallos, L.K.; Havlin, S.; Liljeros, F.; Muchnik, L.; Stanley, H.E.; Makse, H.A. Identification of influential spreaders in complex networks. *Nat. Phys.* **2010**, *6*, 888–893. [CrossRef]
34. Lü, L.-Y.; Zhou, T.; Zhang, Q.M.; Stanley, H.E. The H -index of a network node and its relation to degree and coreness. *Nat. Commu.* **2016**, *7*, 10168.

symmetry

MDPI

Article

A Feasible Temporal Links Prediction Framework Combining with Improved Gravity Model

Xinyu Huang, Dongming Chen * and Tao Ren

Software College, Northeastern University, Shenyang 110169, China; neuhxy@163.com (X.H.); rent@swc.neu.edu.cn (T.R.)

* Correspondence: chendm@mail.neu.edu.cn

Received: 19 December 2019; Accepted: 2 January 2020; Published: 5 January 2020

Abstract: Social network analysis is a multidisciplinary study covering informatics, mathematics, sociology, management, psychology, etc. Link prediction, as one of the fundamental studies with a variety of applications, has attracted increasing focus from scientific society. Traditional research based on graph theory has made numerous achievements, whereas suffering from incapability of dealing with dynamic behaviors and low predicting accuracy. Aiming at addressing the problem, this paper employs a diagonally symmetrical supra-adjacency matrix to represent the dynamic social networks, and proposes a temporal links prediction framework combining with an improved gravity model. Extensive experiments on several real-world datasets verified the superiority on competitors, which benefits recommending friends in social networks. It is of remarkable significance in revealing the evolutions in temporal networks and promoting considerable commercial interest for social applications.

Keywords: social network; temporal links prediction; gravity model; multilayer network

1. Introduction

The analysis of social networks has drawn increasing attention in the field of sociology. It analyzes and explores the potential relations between social objects [1]. The rapid development of social media has brought us plentiful data sources, along with enormous challenges such as data incompletion and dynamic changes [2]. On the one hand, researchers are facing data incompletion problem since only part of social information can be collected from social platforms. On the other hand, the dynamic changes may lead the nodes and links to appear and disappear in the future, which makes the underlying graph longitudinal [3].

Link prediction [4], as fundamental research in social network analysis, is proposed to detect unobserved links from existing parts of the network [5] or forecast future links from current network structures [6]. The former research, also known as missing links prediction, has been fruitful during the last decade, whereas the prediction of future links is more challenging to estimate the upcoming connections with limited social information. This study is of great importance, not only in revealing the evolution of social networks, but also benefiting network management, such as promoting useful links or prohibiting harmful interactions. For instance, a recommendation system [7], as a typical application of temporal links prediction, is designed for individuals to make friends or purchase goods via efficient predicted results, which brings considerable benefits for corporations.

Numerous attempts have been made to address the problem of temporal links prediction, but it is a really difficult task. Firstly, the observed social information is quite limited, which leads to the smoothness assumption [8] being frequently adopted in studies, thereby the methods may be incapable when the network changes seriously. Secondly, longitudinal bias is inevitable, as the dynamic changes must be shifted towards future [3]. Finally, the different observation of extensive network changes

over time may also lead to various social structures, which may yield extremely different predicting results. To solve the problem of temporal links prediction, this paper proposes a dynamic similarity framework with an improved gravity model to estimate the future links in temporal networks.

The rest of this paper is organized as follows. Section 2 introduces the related works on link prediction. Section 3 presents the mathematical model, the improved gravity model, and the framework for predicting temporal links. Section 4 presents the experiments and analysis, including comparison experiments on real-world dataset separated by different levels, which verified the feasibility and veracity of the method. Section 5 summarizes the whole paper and provides concluding remarks.

2. Related works

Link prediction was first proposed on SIGKDD in 2005 [4]. Afterwards, Liben-Nowell and Kleinber reviewed the link prediction in social networks, which attracted more and more scholars devoting themselves to this field [9]. In 2011, Lü summarized the existing methods and classified them into three categories: structure similarity index, maximum likelihood approximation, and probabilistic model [10]. In 2017, Pech et al. [11] introduced robust principal component analysis (robust PCA) method into link prediction and estimated the missing links of the adjacency matrix. Experimental results show that, when the target network is connected and sufficiently dense, the proposed method achieves much higher accuracy compared of some the state-of-the-art algorithms. A brief summary of the classic predicators is shown in Table 1.

The prediction of future links [12,13], i.e., temporal links prediction, aims to predict the links in a network that would appear in the next state of period. As for temporal networks, a series of mathematic models are proposed, such as temporal graphs [14], evolving graphs [15], time-varying graphs, dynamic networks [16], etc. Three representative models are available to depict dynamic behaviors: Snapshots, Contact sequences and Interval graphs.

Snapshots, as a favorable model to exhibit dynamic behaviors, have been widely used in various application scenarios. By employing such representation, unsupervised learning methods are feasible to estimate the links at time t with the observed network structures at time $[1, t-1]$ [17–19]. Besides, statistical methods, such as Exponential Smoothing (EPS) [20] and Autoregressive Integrated Moving Average(ARIMA) [21], are also employed to predict temporal links with snapshots representation. However, Snapshots suffer from coarse-grained depiction of continuous changes, which probability result in poor predictive performance and misleading results [22]. Distinguished by the duration of interaction being negligible or not, Contact sequences and Interval graphs are proposed to illustrate the network dynamics. A temporal network can be represented by a series of triplet (i, j, t) where i and j are entities and t is the time of interaction. If the duration of interaction is not negligible, the framework can be represented by $(i, j) = (t_1, t_1^{'}), \ldots, (t_n, t_n^{'})$, namely Interval graphs. Although they encode more interactions information of dynamic behaviors, the above-mentioned two models are scarcely used in studying temporal link prediction for their complicated expressions.

By combining with temporal correlations and evolutions of link occurrences, Özcan and Öğüdücü [23] proposed Multivariate Time Series Link Prediction method. Experiments on real-world bibliographic datasets showed that the proposed method, which can incorporate covariance structures, achieved better results for temporal links prediction than classic competitors. Considering both temporal dynamics and multi-relational properties in bibliographic networks, Sett et al. [3] proposed a robust and efficient set of features named time-aware multi-relational link prediction (TMLP) features to predict future links using supervised learning framework in dynamic multi-relational network. They analyzed unsupervised performance of individual features, and then applied a supervised learning method that combines multiple features towards link prediction. To overcome the inherent problem of longitudinal bias, random forests supervised learning framework is utilized in the experiments. Experiments on bibliographic datasets showed the effectiveness. However, the above-mentioned methods rely on the plentiful data features and the performance on general

temporal network is uncertain. Overall, the study on temporal links prediction is blossoming and still requires great endeavors to achieve better performance.

Table 1. A brief summary of the existing link prediction methods.

Indicator	Topology	Definition [1]	Complexity [2]
CN [24]	Local	$S_{xy} = \|\Gamma(x) \cap \Gamma(y)\|$	$O(n^2)$
Salton [25]	Local	$S_{xy} = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{\sqrt{k_x k_y}}$	$O(n^2)$
Jaccard [26]	Local	$S_{xy} = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\cup\Gamma(y)\|}$	$O(2n^2)$
Sorenson [27]	Local	$S_{xy} = \frac{2\times\|\Gamma(x)\cap\Gamma(y)\|}{k_x+k_y}$	$O(n^2)$
HPI [28]	Local	$S_{xy} = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{\min(k_x,k_y)}$	$O(n^2)$
HDI [29]	Local	$S_{xy} = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{\max(k_x,k_y)}$	$O(n^2)$
LHN-I [29]	Local	$S_{xy} = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{k_x k_y}$	$O(n^2)$
AA [30]	Local	$S_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log k_z}$	$O(2n^2)$
RA [31]	Local	$S_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k_z}$	$O(2n^2)$
PA [32]	Local	$S_{xy} = k_x k_y$	$O(2n)$
LP [33]	Semi-Local	$S = A^2 + \alpha \cdot A^3$	$O(n^3)$
Katz [34]	Global	$S = (I - \alpha \cdot A)^{-1} - I$	$O(n^3)$
LHN-II [35]	Global	$S = 2m\lambda_1 D^{-1}(I - \frac{\phi}{\lambda_1})^{-1}D^{-1}$	$O(n^3)$
LRW [36]	Semi-Local	$S_{xy}^{LRW}(t) = q_x \cdot \pi_{xy}(t) + q_y \cdot \pi_{yx}(t)$	$O(nk^t)$
SRW [36]	Semi-Local	$S_{xy}^{SRW}(t) = \sum_{l=1}^{t} S_{xy}^{LRW}(l) = q_x \sum_{l=1}^{t} \pi_{xy}(l) + q_y \sum_{l=1}^{t} \pi_{yx}(l)$	$O(nk^t)$
RWR [37]	Semi-Local	$S_{xy}^{LRW}(t) = q_{xy} + q_{yx}$	$O(n^3)$
ACT [38]	Semi-Local	$S_{xy}^{ACT} = \frac{1}{l_{xx}^{+}\cdot l_{yy}^{+} - 2l_{xy}^{+}}$	$O(n^3)$
SimR [39]	Global	$S_{xy}^{SimR} = C\frac{\sum_{v_z\in\Gamma(x)}\sum_{v_{z'}\in\Gamma(y)} S_{zz'}^{SimR}}{k_x k_y}$	$O(n^3)$
Cos+ [40]	Semi-Local	$S_{xy}^{cos+} = \cos(x,y)^{+} = \frac{l_{xy}^{+}}{\sqrt{l_{xx}^{+}\cdot l_{yy}^{+}}}$	$O(n^3)$
TS [41]	Global	$S^{Tr} = (I - \varepsilon S)^{-1} S$	$O(n^3)$
LowRank [11]	Global	$S = \min_{X^*,E} \|\|X^*\|\|_* + \lambda\|\|E\|\|_1$	$O(nk^3)$
MFI [42]	Global	$S = (I + \alpha \cdot L)^{-1}, \alpha > 0$	$O(n^3)$

[1] $\Gamma(x)$ and $\Gamma(y)$ denote the neighbors of node x and y, respectively; k_x and k_y are the degrees of node x and y, respectively; m is the number of edges; l_{xy}^{+} denotes the element at row x, column y of the pseudo-inverse matrix L^{+}; q_{xy} represents the probability of random walk from node x to node y; ε represents an adjustable parameter; $\pi_{xy}(l)$ is the random walk probability from node x to node y at time l; [2] n denotes the number of nodes; k is the average degree of nodes; and t is the step of random walk steps.

3. Modeling and Methods

3.1. Model

The problem of link prediction is described as estimating the likelihood of all the possible links with a given network model $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ is the node set and $E = \{(v_i, v_j)\}$, $(v_i, v_j \in V)$ is the edge set. Suppose the likelihood (or similarity) of the link between every two nodes in G can be calculated by a certain algorithm, and then sorted by ascending order; the future links are obtained by top-k links, where k is the target links amount.

In this paper, we employ a generative multilayer network model [43], which can transform the temporal network into a considerable collection of snapshots. An illustration of the multilayer network model with data derived from a large European research institution [44] is shown in Figure 1 and the corresponding supra-adjacency matrix is shown in Figure 2.

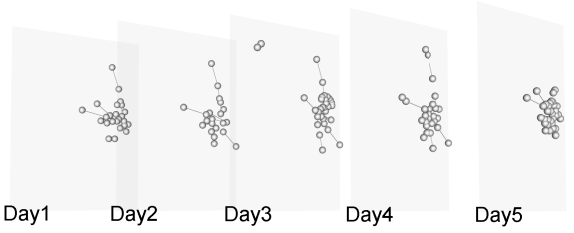


Figure 1. Snapshot of temporal network of the first five days.

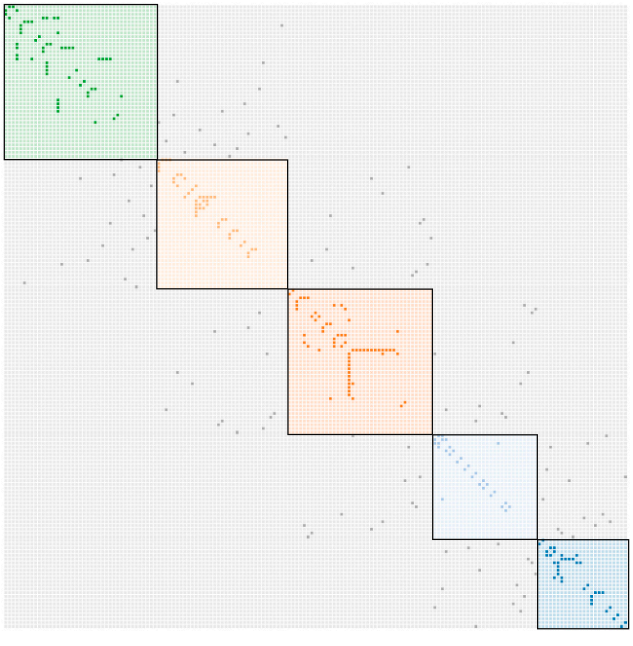


Figure 2. Supra-adjacency matrix representation. The diagonal blocks with different colors represent the network structures at different snapshots.

Suppose given a temporal network G^T with T separated slices, where each slice is modeled as a mono-layer network (i.e., single-layer network), T is the total number of layers, and $t = 1, 2, ..., T$; the model is denoted by

$$G^T = \{\sum V^t, \sum E^t\}, \tag{1}$$

where $\sum V^t$ and $\sum E^t$ are the union of the nodes and edges at each slice. To simplify the research problem, we utilize undirected graph to present the network structures in each snapshot. Thus, the temporal network can be represented by a diagonally symmetrical supra-adjacency matrix (denoted by $\widetilde{M}$), described as

$$\widetilde{M} = \begin{bmatrix} A_1 & I_{1,2} & & \\ I_{2,1} & A_2 & \ddots & \\ & \ddots & \ddots & I_{T-1,T} \\ & & I_{T,T-1} & A_T \end{bmatrix} \in \mathscr{R}^{N \times N}, \tag{2}$$

where A_1, A_2, …, A_T are the adjacency matrix of time 1, 2, …, T, respectively, representing the links (i.e., intra-layer edges). N is the total numbers of the nodes, which can be calculated by

$N = \sum_{1 \le l \le T} |V^l|$. I denotes the relationship of the node located in the continuous snapshots (i.e., the inter-layer edges). By utilizing the former temporal information from time 1 to $T-1$, our goal is to predict the links at the last time T. The problem is described as

$$P_{(u,v)} \otimes T = \sum_{1 \le t \le T-1} \tau(p_{(u,v)} \otimes t) \cdot E_{(u,v)} \otimes t, \tag{3}$$

where $P_{(u,v)} \otimes T$ is the predicted likelihood of link connected node u and v at time T, $p_{(u,v)} \otimes t$ is the likelihood indicator of node u and v at time t, τ depicts the varying function of determining the influence at different time t and $E_{(u,v)} \otimes t$ is the existence function of nodes u and v at time t, described as

$$E_{(u,v)} \otimes t = \begin{cases} \sum I_u I_v, & \text{if node } u \text{ links } v \text{ at time } t \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where I_u and I_v represent the interlayer edges of node u and v, respectively. The prediction of future links (at time T) relies on the former structures (i.e., from time 1 to time $T-1$). The links with the maximum score are the prediction results of future links. The prediction result is evaluated by precision, recall, $F_{\text{1-value}}$, accuracy, etc. Considering the significance of prediction, AUC [45] is employed for evaluation in this paper.

3.2. Definitions

Gravity is the force between objects, which relates to the qualities of the objects and the distance between the two objects. The gravity we focus on in this paper is utilized to describe the strength of the interactions between the two nodes. The gravity in networks [46] is defined as

$$G_{i,j} = \frac{k_i \cdot k_j}{d_{ij}^2}, \tag{5}$$

where $G_{i,j}$ represents the gravity between node i and j, d_{ij} depicts the shortest path length between node i and node j, and k_i is the degree of node i. Inspired by this model, we simplify the time-consuming process, namely the calculation of shortest path, by merely considering the neighbors within two steps (i.e., neighbors and second-order neighbors), denoted by GR. Thus, it can be reduced to the accumulation of gravity between common neighbors, and the likelihood of existing links between node i and j is given by

$$S_{i,j}^{GR} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{G_{i,z} \cdot G_{z,j}}{\delta^2}, \tag{6}$$

where $\Gamma(i)$ and $\Gamma(j)$ are the neighbors of node i and node j, respectively. δ is the steps between node i and j if there are common neighbors between i and j, and the number of nodes otherwise. Thus, the predicted likelihood of node i and j at time t (marked as $S_{i,j}^{GR} \otimes T$) is given by

$$S_{i,j}^{GR} \otimes T = \sum_{t=0}^{t-1} G_{i,j}^{GR|t} \times e^{-\alpha(t-p)}, \tag{7}$$

where $G_{i,j}^{GR|t}$ is the gravity of node i and j at time t. $e^{\alpha(t-p)}$ is the above-mentioned τ function to enforce temporal effect on the similarity evaluation [47], namely dynamic similarity process (for short DS). p is the existence duration of i and j. α is the attenuation constant in the range of [0, 1]. The values change with x under different α, as shown in Figure 3.

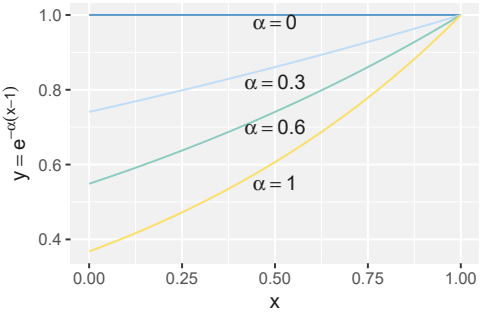


Figure 3. Illustration of τ function with different parameters. α varies in [0, 1], which depicts the attenuation level. The larger is α, the less effect there is on the current x.

Suppose k is the number of links to be predicted in the future. The algorithm of predicting future links is described as the following steps.

Step 1: Obtain all the node pairs at $T-1$ layer as the target-predicting links.

Step 2: Collect all the existing links from the start time to time $T-1$, marked as training set.

Step 3: Calculate the likelihood of node pairs in the training set in terms of Equation (7).

Step 4: Sort the possible links in descending order and obtain the top-k results as the predicted result.

The pseudo-code of the above process is shown as Algorithm 1.

Algorithm 1: Temporal links prediction framework.

Input: temporal network G^T, target links amount k
Output: future links L

T= len(G^T.layers());
g^{T-1} = G^T.layers($T-1$); // Obtain the network structures at time $T-1$.
targets = array();
candidates = {};
for $i = 0; i < len(g^{T-1}.nodes()); i++$ **do**
  for $j = i+1; j < len(g^{T-1}.nodes()); j++$ **do**
    if $(i,j) \in g^{T-1}.edges()$ **then**
      $targets += (i,j)$; // Step 1: Obtain target-predicting links.
for $t = 0; t \leq T-1; t++$ **do**
  g^t = G^T.layers(t);
  for $k = 0; k < len(targets); k++$ **do**
    (u,v) = $targets[k]$; // Step 2: Obtain testing set.
    if $(u,v) \in g^t.edges()$ **then**
      $S_{u,v} += S_{u,v}^{GR|t} * e^{-\alpha(t-p)}$; // Step 3: Calculate the likelihood of node pairs.
      $candidates[(u,v)] = S_{u,v}$;
L = sort($candidates$.values(), descending, k); // Step 4: Obtain the predictive results.
return L;

3.3. Complexity Analysis

Suppose m and n are the number of edges and nodes, respectively, the average degree of nodes is d, and the total layers of temporal networks is T. The complexity of calculating common neighbors is $O(d^2)$. Thus, the complexity of the proposed indicator $S_{i,j}^{GR}$ is lower than $O(d^2+d)$, which can

be simplified as $O(d^2)$. Traversing every two nodes at time $T-1$ needs complexity of $O(n^2)$, thus the total complexity is $O(n^2d^2)$. Actually, d^2 is much smaller than n, thus the proposed method can be simplified as $O(n^2)$. It is very close to competitive indicators, e.g. CN, AA, and RA are all with complexity of $O(n^2)$ [48]. The process of temporal links prediction our method $S_{i,j}^{GR} \otimes T$ is $O(n^2T)$, which is the same as the representative Linear Regression methods, EPS [20], and so on.

4. Experiments and Discussion

The experimental environment was a Intel(R) Core (TM) i5-7200U CPU @ 2.50 GHz (4 CPUs), 2.7 GHz, the memory was 8 GB DDR3, the operating system was Windows10 64 bit, the programming language was Python 3.7.1, and the relevant libs were NetworkX 2.2 and Multinetx. The goal of the experiments was to validate the performance of the proposed method and compare with competitive indicators.

4.1. Experimental Datasets

To verify the proposed indicator, seven real-world datasets were employed in the experiments, as shown in Table 2. The real-world email dataset from a large European research institution [44] was employed to check the proposed dynamic similarity framework and the data statistics are shown in Table 3.

Table 2. Statistics of seven real-world datasets.

Dataset Name	$\|V\|$	$\|E\|$	<k>	$\|C\|$	<c>	$\|D\|$	r
Zachary karate Club [49]	34	78	4.59	0.57	2.41	2.22	−0.48
Dolphins social network [50]	62	159	5.13	0.26	3.06	3.36	−0.04
Terriers of 9/11 [51]	69	159	4.61	0.47	1.76	3.22	−0.04
NEUSNCP dataset [1]	89	365	4.10	0.54	3.15	1.92	−0.40
Books about US politics [52]	105	411	8.40	0.49	5.26	3.08	−0.13
American college football network [53]	115	613	10.66	0.40	10.23	2.51	0.16
Scientist collaboration network [52]	1589	2742	4.60	0.64	0.08	5.99	−0.09

Note: $|V|$ denotes the number of nodes; $|E|$ denotes the number of edges; <k> is the average degree; $|C|$ is the average clustering index; <c> is the average connectivity; $|D|$ is the average shortest path; and r represents assortativity coefficient. [1] We developed an experimental social platform and invited hundreds of users to register. Data availability: https://www.neusncp.com/api/about.

Table 3. Data statistics of Email-Eu-core temporal network.

Dataset Name	$\|V\|$	$\|E\|$	Days
Email-Eu-core temporal network	986	332,334	803
Email-Eu-core-temporal-Dept1	309	61,046	803
Email-Eu-core-temporal-Dept2	162	46,772	803
Email-Eu-core-temporal-Dept3	89	12,216	803
Email-Eu-core-temporal-Dept4	142	48,141	803

The email dataset is convictive without providing any feature information. The e-mails only represent communication between institution members (the core), and the dataset does not contain incoming messages from or outgoing messages to the rest of the world. To illustrate the connections of each period, the dataset was separated by different intervals (daily, weekly, and monthly), as shown in Figure 4.

As shown in Figure 4, the more layers there are in the multilayer network model, the fewer connections there are in each slice. The fitting results of connections are marked with red lines in the above three panels. Here, we can see the number of edges is varying over time among all three separations. When the dataset is daily separated, the fitting result is not obvious and indicates poor predictive performance. When the dataset is weekly separated, the improved fitting result provides

a better predictive training set. When the dataset is monthly separated, an obvious fitting seems to reveal a better predictive result.

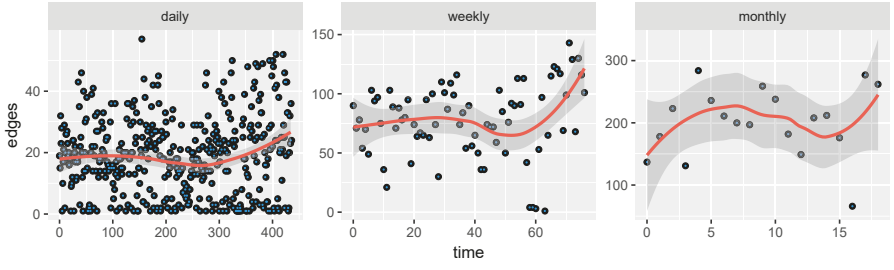


Figure 4. Email-Eu-core-temporal-Dept3 network for link prediction in the dynamic network (daily, weekly, and monthly separated).

4.2. Performance Comparison

First, we compared the proposed GR indicator with several classic methods. The results were evaluated by AUC via the average result of 100 independent experiments, as shown in Table 4.

Table 4. AUC of the GR and other indicators comparison.

Dataset Name	GR	AA	RA	JC	PA	CN
Zachary karate Club	**0.8790**	0.8784	0.8784	0.6281	0.8773	0.8433
Dolphins social network	**0.7442**	0.7428	0.7425	0.7431	0.6621	0.7379
Terriers of 9/11	**0.9374**	0.9339	0.9371	0.9151	0.7144	0.9103
NEUSNCP dataset	**0.9110**	0.9105	0.9096	0.8855	0.6725	0.9012
Books about US politics	0.8310	0.8299	0.8299	**0.8397**	0.2573	0.8304
American college football network	**0.8775**	0.8750	0.8769	0.7494	0.8381	0.8657
Scientist collaboration network	**0.9431**	**0.9431**	**0.9431**	0.9430	0.6725	0.9429

Table 4 shows that the AUC obtained by the proposed GR indicator generally outperforms the competitive methods, as marked by boldface. Although the performance of GR on Books about US politics dataset is inferior to that of JC indicator, it is still competitive. Generally, the experiments verified the performance of the proposed indicator in predicting unknown links.

Secondly, we compared the proposed method with the existing linear regression method, ARIMA, EPS methods, and the AUC results of six indicators are shown in Figure 5.

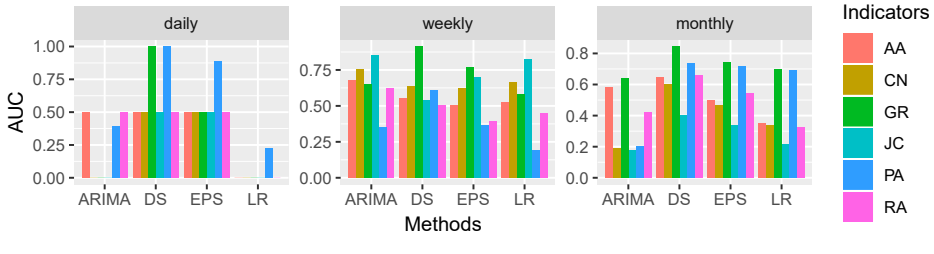


Figure 5. Comparison of temporal links prediction on Email-Eu-core-Dept3 network.

Obviously, the result of temporal links prediction obtained from our proposed method (DS combined with GR indicator) outperforms the competitive methods. When the datasets were daily separated, the proposed method obtained maximum AUC, achieving 0.913 and 0.8444 when weekly and monthly separated, respectively, which are greater than the result of the other methods.

4.3. Parameters Analysis

In the proposed framework, parameters are crucial to the prediction results. Thus, in this subsection, the parameter α is analyzed and the results are plotted in Figure 6.

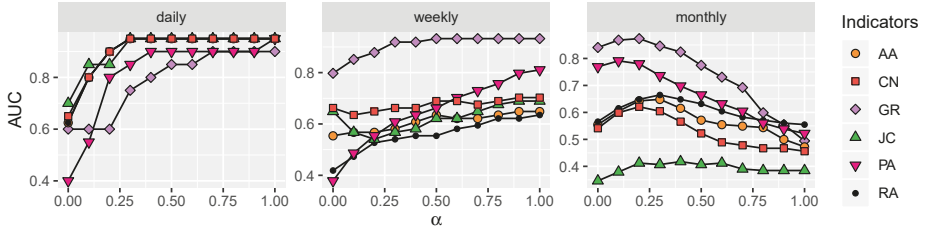


Figure 6. Analyzing parameter α in the temporal network with different separations. In general, the obtained AUC is larger with α increasing when the networks are daily and weekly separated.

As shown in Figure 6, in a temporal network with a large number of slices, a larger α is preferred to be selected. On the contrary, when the temporal network is separated with fewer slices, a smaller α contributes to obtaining better performance. Finally, the robustness of the proposed method was verified by conducting experiments on different scales of temporal networks, as shown in Figure 7.

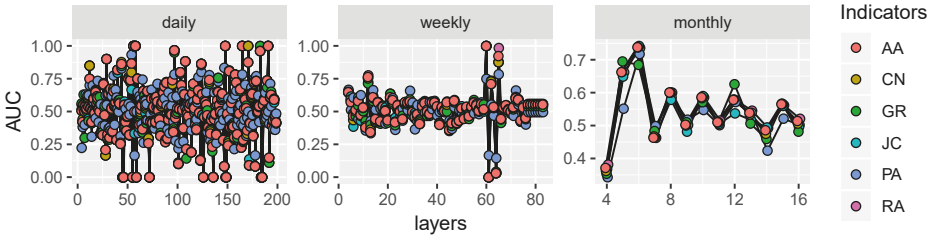


Figure 7. Robustness verification by varying layers (or time slices).

As shown in Figure 7, with the increasing of the number of slices (i.e., layers), the AUC result is changing periodically, and the varying range is stable generally. The results of the six indicators are in the same tendency, which verifies the robustness of the proposed method. To analyze the slicing effect on network features and performance, we conducted experiments with DS method and GR indicator. The results are shown in Figures 8 and 9.

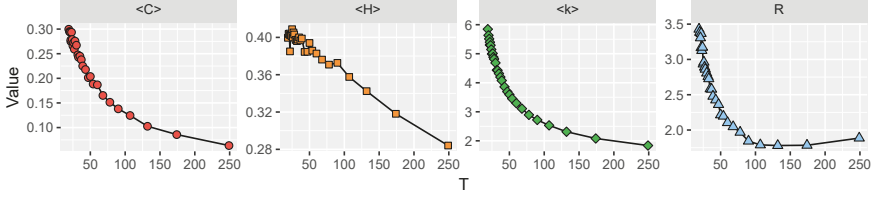


Figure 8. Illustration of network features with varying numbers of slicing. T is the number of slices ranging from 17 to 249. <C>, <H>, and <k> are the average clustering coefficient [54], average heterogeneity [55], and average degree of the nodes in all slices, respectively. R is the ratio of intralayer edges comparing interlayer edges. In general, R declines when the temporal network is separated into more slices.

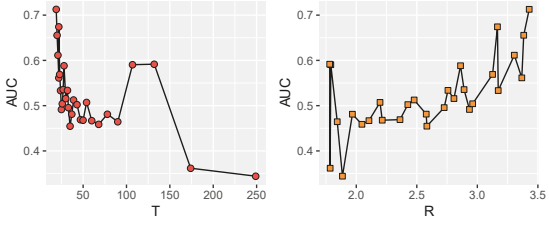


Figure 9. Relationship of performance with T (i.e., number of slices) and R (i.e., intralayer edges comparing interlayer edges). AUC declines with T increasing (or R declining) in general.

As shown in Figure 9 (left), the AUC result is declining as the number of slices increases. This results from the large numbers of intralayer edges of each time slice, which providing more edges to calculate the existence likelihood. Thus, we can utilize the network information more comprehensively. The performance is better with the ratio (i.e., intralayer edges comparing interlayer edges) increasing, as shown in Figure 9 (right).

5. Conclusions

Aiming at solving the problem of temporal links prediction in social networks, this paper proposes a novel dynamic similarity framework combining with an improved gravity model. Experiments on real-world datasets with different separations were conducted, and the experimental results show that the proposed method outperforms competitors. Afterwards, the determination of parameter α was analyzed by conducting a series of experiments, and we give the recommended selection for different temporal structures. Finally, the robustness of the proposed framework was also verified by comparing the obtained AUC results with varying time slices. Overall, the proposed framework is capable of predicting temporal links with reasonable results.

The contribution of this work is likely to benefit many real-world social applications, such as recommending new friends, protecting teenagers from harmful interactions, etc. Inspired by the multiple interactions among social actors, we have established a social platform, namely NEUSNCP (https://www.neusncp.com), for college students to make friends and share knowledge in various manners. By applying the proposed framework into recommending friends for newcomers, we have observed an obvious increase in user interactions. As part of future works, link prediction on more complicated models, i.e., multi-relational networks and bipartite networks on social platforms can be further studied. Notably, the research of recommendation for dynamic "user-blog" networks is in development, via a combination of collaborative filtering algorithm with temporal changes computation. In brief, the application of temporal links prediction is just unfolding.

Author Contributions: X.H. designed the framework and wrote the original draft; D.C. revised the manuscript; and T.R. checked the manuscript and made some modifications. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by Liaoning Natural Science Foundation under Grant No. 20170540320, the Doctoral Scientific Research Foundation of Liaoning Province under Grant No. 20170520358, the National Natural Science Foundation of China under Grant No. 61473073, and the Fundamental Research Funds for the Central Universities under Grant Nos. N161702001 and N172410005-2.

Acknowledgments: We would like to thank the anonymous reviewers for their careful reading and useful comments that helped us to improve the final version of this paper.

Conflicts of Interest: The author declares that there are no conflicts of interest regarding the publication of this paper.

Abbreviations

The following abbreviations are used in this manuscript:

AA	Adamic-Adar
ACT	Average commute time
ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under the receiver operating characteristic Curve
CN	Common Neighbors
DS	Dynamic Similarity
EPS	Exponential Smoothing
GR	Gravity
JC	Jaccard
LR	Linear Regression
LRW	Local Random Walk
MFI	Matrix-forest index
PCA	Principal Component Analysis
RA	Resource Allocation
RWR	Random Walk with Restart
SimR	SimRank
SRW	Superposed Random Walk
TMLP	Time-aware Multi-relational Link Prediction
TS	Transferring Similarity

References

1. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994; Volumn 8.
2. Antonacci, G.; Fronzetti Colladon, A.; Stefanini, A.; Gloor, P. It is rotating leaders who build the swarm: Social network determinants of growth for healthcare virtual communities of practice. *J. Knowl. Manag.* **2017**, *21*, 1218–1239. [CrossRef]
3. Sett, N.; Basu, S.; Nandi, S.; Singh, S.R. Temporal link prediction in multi-relational network. *World Wide Web* **2018**, *21*, 395–419. [CrossRef]
4. Getoor, L.; Diehl, C.P. Link mining: A survey. *ACM SIGKDD Explor. Newslett.* **2005**, *7*, 3–12. [CrossRef]
5. Srinivas, V.; Mitra, P. Link Prediction Using Thresholding Nodes Based on Their Degree. In *Link Prediction in Social Networks*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 15–25.
6. Oyama, S.; Hayashi, K.; Kashima, H. Cross-temporal link prediction. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining, Vancouver, BC, Canada, 11–14 December 2011; pp. 1188–1193.
7. Slokom, M.; Ayachi, R. A New Social Recommender System Based on Link Prediction Across Heterogeneous Networks. In Proceedings of the International Conference on Intelligent Decision Technologies, Sorrento, Italy, 17–19 June 2017; pp. 330–340.
8. Kim, W.; Kwon, K.; Kwon, S.; Lee, S. The identification power of smoothness assumptions in models with counterfactual outcomes. *Quantit. Econ.* **2018**, *9*, 617–642. [CrossRef]
9. Liben-Nowell, D.; Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 1019–1031. [CrossRef]
10. Lü, L.; Zhou, T. Link prediction in complex networks: A survey. *Phys. A Stat. Mech. Appl.* **2011**, *390*, 1150–1170. [CrossRef]
11. Pech, R.; Hao, D.; Pan, L.; Cheng, H.; Zhou, T. Link prediction via matrix completion. *EPL (Europhys. Lett.)* **2017**, *117*, 38002. [CrossRef]
12. Munasinghe, L.; Ichise, R. Time aware index for link prediction in social networks. In Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, Toulouse, France, 29 August–2 September 2011; pp. 342–353.
13. Yasami, Y.; Safaei, F. A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks. *Phys. A Stat. Mech. Appl.* **2018**, *492*, 2166–2197. [CrossRef]
14. Kostakos, V. Temporal graphs. *Phys. A Stat. Mech. Appl.* **2009**, *388*, 1007–1023. [CrossRef]

15. Alhajj, R.; Rokne, J. *Encyclopedia of Social Network Analysis and Mining*; Springer: Berlin/Heidelberg, Germany, 2014.
16. Casteigts, A.; Flocchini, P.; Quattrociocchi, W.; Santoro, N. Time-varying graphs and dynamic networks. *Int. J. Parallel Emerg. Distrib. Syst.* **2012**, *27*, 387–408. [CrossRef]
17. Hua, T.D.; Nguyen-Thi, A.T.; Nguyen, T.A.H. Link prediction in weighted network based on reliable routes by machine learning approach. In Proceedings of the 2017 4th NAFOSTED Conference on Information and Computer Science, Hanoi, Vietnam, 24–25 November 2017; pp. 236–241.
18. Zhou, J.; Huang, D.; Wang, H. A dynamic logistic regression for network link prediction. *Sci. China Math.* **2017**, *60*, 165–176. [CrossRef]
19. Tabourier, L.; Bernardes, D.F.; Libert, A.S.; Lambiotte, R. RankMerging: A supervised learning-to-rank framework to predict links in large social networks. *Mach. Learn.* **2019**, *108*, 1729–1756 . [CrossRef]
20. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts: Melbourne, Australia, 2018.
21. Hyndman, R.; Koehler, A.B.; Ord, J.K.; Snyder, R.D. *Forecasting with Exponential Smoothing: The State Space Approach*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
22. Divakaran, A.; Mohan, A. Temporal Link Prediction: A Survey. *New Gener. Comput.* **2019**. doi:10.1007/s00354-019-00065-z. [CrossRef]
23. Özcan, A.; Öğüdücü, Ş.G. Multivariate temporal link prediction in evolving social networks. In Proceedings of the 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), Las Vegas, NV, USA, 28 June–1 July 2015; pp. 185–190.
24. Lorrain, F.; White, H.C. Structural equivalence of individuals in social networks. *J. Math. Soc.* **1971**, *1*, 49–80. [CrossRef]
25. Worth, D. Introduction to modern information retrieval. *Aust. Acad. Res. Libr.* **2010**, *41*, 305–306. [CrossRef]
26. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547–579.
27. Sorensen, T.A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.* **1948**, *5*, 1–34.
28. Ravasz, E.; Somera, A.L.; Mongru, D.A.; Oltvai, Z.N.; Barabási, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **2002**, *297*, 1551–1555. [CrossRef]
29. Molloy, M.; Reed, B. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms* **1995**, *6*, 161–180. [CrossRef]
30. Adamic, L.A.; Adar, E. Friends and neighbors on the web. *Soc. Netw.* **2003**, *25*, 211–230. [CrossRef]
31. Zhou, T.; Lü, L.; Zhang, Y.C. Predicting missing links via local information. *Eur. Phys. J. B* **2009**, *71*, 623–630. [CrossRef]
32. Barabási, A.L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512. [CrossRef] [PubMed]
33. Lü, L.; Jin, C.H.; Zhou, T. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E* **2009**, *80*, 046122. [CrossRef] [PubMed]
34. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **1953**, *18*, 39–43. [CrossRef]
35. Leicht, E.A.; Holme, P.; Newman, M.E. Vertex similarity in networks. *Phys. Rev. E* **2006**, *73*, 026120. [CrossRef]
36. Liu, W.; Lü, L. Link prediction based on local random walk. *EPL (Europhys. Lett.)* **2010**, *89*, 58007. [CrossRef]
37. Vragović, I.; Louis, E. Network community structure and loop coefficient method. *Phys. Rev. E* **2006**, *74*, 016105. [CrossRef]
38. Klein, D.J.; Randić, M. Resistance distance. *J. Math. Chem.* **1993**, *12*, 81–95. [CrossRef]
39. Jeh, G.; Widom, J. SimRank: A measure of structural-context similarity. In Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 538–543.
40. Fouss, F.; Pirotte, A.; Renders, J.M.; Saerens, M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. Knowl. Data Eng.* **2007**, *19*, 355–369. [CrossRef]

41. Sun, D.; Zhou, T.; Liu, J.G.; Liu, R.R.; Jia, C.X.; Wang, B.H. Information filtering based on transferring similarity. *Phys. Rev. E* **2009**, *80*, 017101. [CrossRef]
42. Chebotarev, P.Y.; Shamis, E. A matrix-forest theorem and measuring relations in small social group. *Avtomatika i Telemekhanika* **1997**, *58*, 125–137.
43. Boccaletti, S.; Bianconi, G.; Criado, R.; Del Genio, C.I.; Gómez-Gardenes, J.; Romance, M.; Sendina-Nadal, I.; Wang, Z.; Zanin, M. The structure and dynamics of multilayer networks. *Phys. Rep.* **2014**, *544*, 1–122. [CrossRef]
44. Paranjape, A.; Benson, A.R.; Leskovec, J. Motifs in temporal networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 601–610.
45. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
46. Li, Z.; Ren, T.; Ma, X.; Liu, S.; Zhang, Y.; Zhou, T. Identifying influential spreaders by gravity model. *Sci. Rep.* **2019**, *9*, 8387. [CrossRef] [PubMed]
47. Chen, D.; Kong, L.; Wang, D.; Huang, X.; Fang, B. TNLCD: A Feasible Algorithm for Local Community Discovery in Temporal Networks. In *FSDM*; IOS Press: Amsterdam, The Netherlands, 2018; pp. 459–464.
48. Wang, P.; Xu, B.; Wu, Y.; Zhou, X. Link prediction in social networks: The state-of-the-art. *Sci. China Inf. Sci.* **2015**, *58*, 1–38. [CrossRef]
49. Zachary, W.W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **1977**, *33*, 452–473. [CrossRef]
50. Lusseau, D.; Schneider, K.; Boisseau, O.J.; Haase, P.; Slooten, E.; Dawson, S.M. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **2003**, *54*, 396–405. [CrossRef]
51. Tsvetovat, M.; Kouznetsov, A. *Social Network Analysis for Startups: Finding Connections on the Social Web*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2011.
52. Newman, M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef]
53. Girvan, M.; Newman, M.E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821–7826. [CrossRef]
54. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440. [CrossRef]
55. Hu, H.B.; Wang, X.F. Unified index to quantifying heterogeneity of complex networks. *Phys. A Stat. Mech. Appl.* **2008**, *387*, 3769–3780. [CrossRef]

symmetry

MDPI

Article

Learning Large Margin Multiple Granularity Features with an Improved Siamese Network for Person Re-Identification

Da-Xiang Li [1,2], Guo-Yuan Fei [1,*] and Shyh-Wei Teng [3]

[1] School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China; www_ldx@163.com
[2] Ministry of Public Security Key Laboratory of Electronic Information Application Technology for Scene Investigation, Xi'an 710121, China
[3] Faculty of Science & Technology, Federation University Australia, Gippsland, VIC 3842, Australia; imgcsiu@163.com
* Correspondence: fgy519597702@gmail.com

Received: 1 December 2019; Accepted: 25 December 2019; Published: 3 January 2020

Abstract: Person re-identification (Re-ID) is a non-overlapping multi-camera retrieval task to match different images of the same person, and it has become a hot research topic in many fields, such as surveillance security, criminal investigation, and video analysis. As one kind of important architecture for person re-identification, Siamese networks usually adopt standard softmax loss function, and they can only obtain the global features of person images, ignoring the local features and the large margin for classification. In this paper, we design a novel symmetric Siamese network model named Siamese Multiple Granularity Network (SMGN), which can jointly learn the large margin multiple granularity features and similarity metrics for person re-identification. Firstly, two branches for global and local feature extraction are designed in the backbone of the proposed SMGN model, and the extracted features are concatenated together as multiple granularity features of person images. Then, to enhance their discriminating ability, the multiple channel weighted fusion (MCWF) loss function is constructed for the SMGN model, which includes the verification loss and identification loss of the training image pair. Extensive comparative experiments on four benchmark datasets (CUHK01, CUHK03, Market-1501 and DukeMTMC-reID) show the effectiveness of our proposed method and its performance outperforms many state-of-the-art methods.

Keywords: person re-identification; multiple granularity features; Siamese Multiple Granularity Network; multi-channel weighted fusion loss

1. Introduction

Person re-identification is a crucial task in video analytics scenarios and it received more and more attention on computer vision field [1,2]. Person re-identification, as a core technology in video analysis, aims to determine whether the objects appearing in the non-overlapping view belong to the same person. Although the researchers have made great efforts to deal with this problem, it still has challenges because of large variations in viewpoints, backgrounds, illuminations and poses. As we can see in Figure 1, there are some hard samples from baseline datasets and those difficulties usually appear in realistic camera networks.

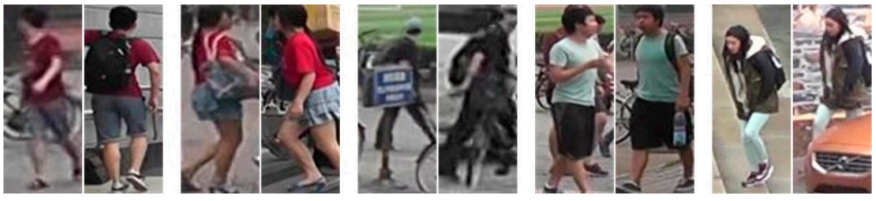
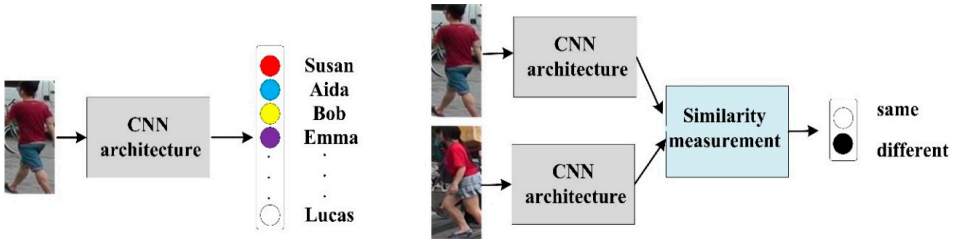


Figure 1. Example pairs of images from baseline person re-identification datasets. Every two adjacent images represent the same person. Analysis of these images suffered from much larger differences indicates person re-identification is challenging.

In order to realize person re-identification, the traditional research work mainly includes two aspects, namely feature extraction [3–6] and metric learning [7,8]. In feature extraction module, different pedestrian image descriptors are adopted to obtain discriminative information of pedestrian images. In metric learning module, there are various kind of distance metrics that are designed to find a suitable embedding space, in which the distance between similar data is pushed as close as possible while the distance between different data is pulled as far as possible.

Considering the success of deep learning in image classification problems, many researchers have applied it to person re-identification [9,10]. According to the differences in model structure, related algorithms can be divided into two categories as shown in Figure 2, namely the CNN-based identification model and Siamese based verification model. In the CNN-based identification model, the images in the training set and their labels are fed into CNN during the training processing. In order to obtain the discriminative features of pedestrian images, various loss functions are designed to take full advantage of the label information of the images, such as cross entropy loss [11], OIM (online instance matching) loss [12] etc. However, in the identification model, the problem is that it usually only uses the global information and ignores the local information of the images. In addition, the similarity metric between image pairs is not considered during model training [9–14]. Therefore, a Siamese-based verification model is proposed, which can judge whether the pedestrians in the two input images are the same person [15,16]. Compared with the identification model, the verification model constructs a loss function between the pairs of training images, and its focus is only on the similarity metric between the image pairs (that is, maximizing the similarity between positive pairs while minimizing the similarity between negative pairs as much as possible). In this case, this kind of model does not make use of the label information of the images during the training phase, which accounts for the final features of images not having the character of margin maximization for classification.



(**a**) Identification model (**b**) Verification model

Figure 2. The difference between the CNN-based identification model and the Siamese-based verification model. Identification models take one image as input and predict its identity while verification models take a pair of images as input and determine whether they belong to the same person or not.

In order to overcome the problems of the two models mentioned above during person re-identification, we fuse the two models together and design a new Siamese network model named

Siamese multiple granularity network (SMGN) in this paper. The backbone CNN of the SMGN is composed by two feature extraction branches, i.e., global and local feature extraction branches. In the proposed SMGN model, four identification loss functions and a verification loss function are designed to obtain the final multi-channel weighted fusion (MCWF) loss function. Therefore, SMGN is able to combine the advantages of identification model and verification model, and the final extracted multiple granularity features of pedestrian images have the characteristic of margin maximization for classification, namely large margin multiple granularities (LMMG) features. As a result, the algorithm based on SMGN can improve the performance of person re-identification.

The contributions of our work are threefold as follows:

- We propose a novel symmetric Siamese network model called SMGN, the backbone CNN of which is composed by two branches, i.e., a local branch and a global branch. Compared with the traditional Siamese network model, SMGN can obtain LMMG features of person images, including local features and global features, which would be of great benefit to person re-identification.
- By fusing the verification and the identification information, a new MCWF loss function is designed for the SMGN model. Compared with traditional cross entropy loss, MCWF loss function takes into account decision boundary information in identification channels, so LMMG features extracted from SMGN can be guaranteed to have the character of margin maximization for classification.
- We implement extensive experiments on four challenging person re-identification datasets (i.e., CUHK01 [17], CUHK03 [9], Market-1501 [18] and DukeMTMC-reID [19]). The experimental results show the proposed method achieves better results than the state-of-the-art methods.

The remainder of our paper is organized as follows: some related works are reviewed in Section 2. The structure of our proposed model and implementation details are presented in Section 3. Extensive comparative experiment results on four benchmark datasets are shown in Section 4, followed by conclusions drawn in Section 5.

2. Related Work

In this section, some previous works related to person re-identification are described simply.

2.1. Hand-Crafted Feature-Based Person Re-ID

The majority of traditional methods related to person re-identification pay close attention to two basic modules, i.e., feature extraction and metric learning. For feature extraction, several effective appearance cues attempt to build a robust feature representation. For example, Farenzena et al. [3] proposed symmetry-driven accumulation of local features (SDALF) to characterize pedestrian images, which are robust to image scale and illumination variations. SDALF consist of three kind of features, i.e., weighted color histograms, maximally stable color regions (MSCR) and recurrent high-structured patches (RHSP). In order to obtain discriminative features of pedestrian images, Local Maximal Occurrence representation (LOMO) is proposed by Liao et al. [4], which includes Scale Invariant Local Ternary Pattern (SILTP) descriptor and two scales of the local HSV histogram. Similarly, Yang et al. [5] utilized salient a Salient Color Name-Based Color Descriptor (SCNCD) that takes advantage of the robustness of color names to illumination to characterize pedestrian images. To further improve the performance, a Hierarchal Gaussian descriptor (GOG) was discussed in [6] that models the region as a set of multiple Gaussian distributions in which each Gaussian represents the appearance of a local patch.

For metric learning, different distance metrics have been proposed to learn a suitable metric space, in which the distance between the same pedestrian are kept as close as possible while the distance between different pedestrians are kept as far as possible. Representative metric methods include XQDA [4], KISSME [7], MLAPG [8] etc. Liao et al. [4] utilized cross-view quadratic discriminant analysis to learn a low dimensional subspace in which all the features have a character of discrimination;

meanwhile, a QDA metric is introduced. In [7], the decision on whether an image pair is similar or not is expressed as a likelihood ratio test. The pairwise difference method is adopted, and the difference space is a zero-mean Gaussian distribution. A logistic metric learning approach with the positive semi-definite (PSD) constraint and an asymmetric sample weighting strategy is derived in [8].

2.2. Deep Learned Feature-Based Person Re-ID

Previous hand-crafted descriptors and metric learning methods have made limited performance on person re-identification. Hence, many researchers tended to utilize CNN-based methods to solve person re-identification problems. Some work [20–22] shows that CNN have a great potential on image classification, object recognition, natural language processing etc. For person re-identification, Li [9] proposed a filter pairing neural network based on CNN that learn filter pairs to encode photometric transforms. Ahmed [10] proposed an enhanced deep learning framework to compute cross-input neighborhood differences and patch summary features. With the popularity of Siamese network, many works have devoted to using it to improve performance. Zheng [11] proposed a unit network that combines identification model and verification model, which learns a discriminative embedding and a similarity measurement simultaneously. Wu [13] proposed a Siamese attention structure based on joint learning spatiotemporal video representation and its similarity measurement. Chung [14] presented a two-stream convolutional neural network, in which each stream is a Siamese network. This architecture can learn spatial and temporal information separately. Benefiting from powerful deep networks, they achieved many state-of-the-art results on person re-identification.

2.3. Loss Function-Based Person Re-ID

As a supervised signal, loss functions play an important role in CNN models. For person re-identification, there are various loss functions have been proposed, such as cross entropy loss [15,23,24], binary classification loss [25,26], contrastive loss [27], center loss [28], triplet loss [29] etc. Cross entropy loss is the most popular used loss function for person re-identification, and it consider identification labels as supervised signals for reducing classification error; binary classification loss considers the deep network as a two-class model, classifying positive and negative sample from the image pair. As for contrastive loss, the Euclidean distance between two features is calculated directly by it, in order to minimize the distance between positive samples and punish the distance between negative samples when it is less than the threshold; center loss forces the similar image features into closing to their corresponding class center to reduce the intra-class variance, but it ignores pushing the distance among inter-class; Triplet loss makes the distance between positive pairs smaller than negative pairs, in other words, the distance between positive samples is pushed as close as possible while the distance between negative samples is pulled as far as possible. In addition, some loss functions based on softmax loss achieve state-of-art performance in face recognition. Liu et al. [30] proposed L-Softmax by adding angular constraints to each identity to improve the discrimination of pedestrian image features. A-Softmax [31] improves L-Softmax by normalizing the weights to learn angularly discriminative features. In addition, feature normalization is applied in [32], so that the classification results only depend on the angle between the feature vector and weight vector.

3. The Proposed Method

In this section, we first present the structure of the proposed SMGN model. Then we describe the MCWF loss function for the SMGN model. Thirdly, the training mechanism and cosine distance used in the testing phases are introduced. Finally a brief algorithm flow is concluded.

3.1. The Structure of SMGN

The overall network architecture of the proposed SMGN model is illustrated in Figure 3. It is essentially a five-channel Siamese model (including four identification channels and a verification channel), which takes a pair of person images as input. In the proposed SMGN model,

ResNet-50 is adopted as its backbone CNN because it has a competitive performance in person re-identification [10–12,16]. In order to use the local and global features to represent pedestrian images simultaneously, the subsequent part after res_conv4_1 block is divided into two independent branches in ResNet-50, namely, global and local feature extraction branches. Table 1 lists the settings of both the local and global branches. "Map Size" denotes the size of output feature maps from each branch. "Dimension" denotes the dimensionality of features for the output representations. "Feature" denotes the symbols for the output feature.

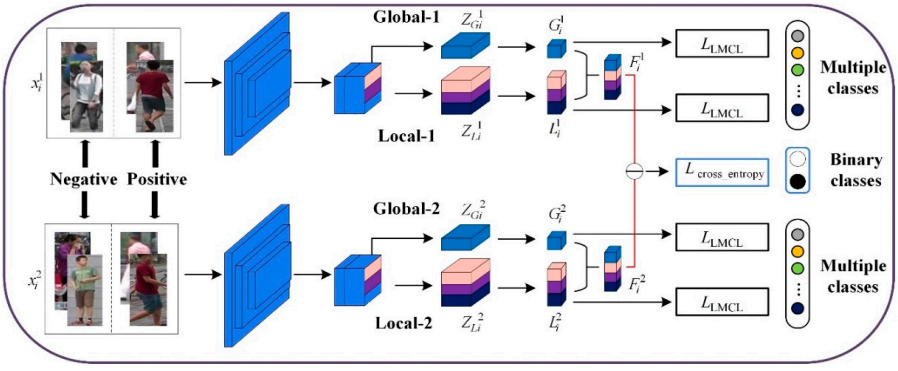


Figure 3. The framework of the proposed Siamese Multiple Granularity Network (SMGN).

Table 1. Comparison of the settings for four branches in SMGN.

Branch	Map Size	Dimension	Feature
Global-1	12×4	256	G_i^1
Global-2	12×4	256	G_i^2
Local-1	24×8	256×3	L_i^1
Local-2	24×8	256×3	L_i^2

As shown in Figure 3, "Global-1" and "Global-2" are global extraction branches while "Local-1" and "Local-2" are local extraction branches. In the global branch, down-sampling with a stride-2 convolution layer is adopted in res_conv5_1 block to address the problem that the output feature maps are sensitive to the location in the input images. After that, we perform global max-pooling (GAP) [33] operation on the corresponding output feature map. Meanwhile, batch normalization [34] and ReLU are introduced to accelerate the training and perform feature reduction respectively. In each global branch, we reduce 2048-dim features $Z_{G_i^j} \Big| j = 1, 2$ to 256-dim features $G_i^j \Big| j = 1, 2$. Different from the global branch, no down-sampling operations are adopted in the res_conv5_1 block. In this way, the appropriate areas of reception fields can be reserved for the local feature in the local feature extraction branch. Furthermore, we divide the feature maps into three uniform parts horizontally and the same following operations are conducted as the global feature extraction branch to obtain the local features of pedestrian images.

3.2. Multiple Granularity Features

During the training phase, we assume that an image pair $(x_i^1, x_i^2, l_i^1, l_i^2)$ is fed into SMGN, where x_i^1 and x_i^2 represent the first and second image in i-th image pair, and l_i^1 and l_i^2 denote the corresponding label of x_i^1 and x_i^2. The proposed SMGN can produce their descriptors from global branches and local branches. For the first image x_i^1, we can obtain its global features G_i^1 from the branch "Global-1" and its local features L_i^1 from the branch "Local-1". Similarly, we can get global features G_i^2 and local features

L_i^2 of the second image x_i^2. Finally, we concatenate global features and local features together to obtain the final representation of x_i^1 and x_i^2 through Equation (1) as follows:

$$\begin{cases} F_i^1 = [L_i^1, G_i^1] \\ F_i^2 = [L_i^2, G_i^2] \end{cases} \tag{1}$$

where F_i^1 and F_i^2 represent the multiple granularity features of the person image x_i^1 and x_i^2 respectively, which include both global information and local information from the corresponding images.

3.3. Multi-Channel Weighted Fusion Loss

To further improve the discriminability of multiple granularity features for person re-identification, we design a multi-channel weighted fusion (MCWF) loss function which include identification loss and verification loss in four identification channels and a verification channel.

3.3.1. Identification Loss

In the proposed SMGN model, there are four identification channels. For each identification channel, we introduce a new classification loss called large margin cosine loss (LMCL) [35] to make multiple granularity features have the character of margin maximization for classification.

In the traditional softmax loss function, different classes can be distinguished by maximizing the posterior probability of the ground-truth class. We assume that the i-th feature vector and its label are v_i and l_i respectively, then we can write the traditional softmax loss function as follows:

$$Loss_{\text{softmax}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{y_{l_i}}}{\sum_{j=1}^{C} e^{y_j}} \tag{2}$$

where N and C represent the number of training samples and classes respectively. Here, y_j represents the activation value of the j-th neuron in a fully connected layer with a weight vector W_j and a bias b_j. Relatively, there are C neurons in total, and the output of neurons represents the score that v_i belongs to the corresponding class. For the purpose of simplicity, we fix the bias $b_j = 0$, and then y_j can be computed by:

$$y_j = W_j^T v = \|W_j\|\|v\| \cos\theta_j \tag{3}$$

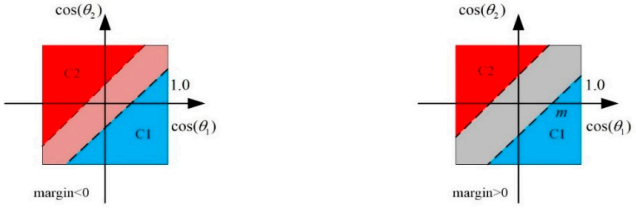
where v represents an input feature vector and θ_j is the angle between W_j and v.

In order to perform feature learning effectively, we fix $\|W_j\| = 1$ by L_2 normalization. During the testing phase, the matching score of a pair of pedestrian images is computed based on cosine similarity between the two feature vectors. This indicates that the norm of the feature vector v does not contribute to the score function. Thus, we fix $\|v\| = t$ in the training stage. Therefore, the posterior probability only depends on the cosine of the angle. To obtain a large margin classifier, we set decision boundary as follows:

$$\begin{cases} C_1 : \cos\theta_1 \geq m + \cos\theta_2 \\ C_2 : \cos\theta_2 \geq m + \cos\theta_1 \end{cases} \tag{4}$$

where $m \geq 0$ is a fixed margin parameter and it is used to better control the boundary between different classes. In Equation (4), $\cos\theta_i - m$ is smaller than $\cos\theta_i$, so that the constraint are more stringent for classification. Eventually, the modified loss enhances the discrimination of multiple granularity features by introducing an extra margin in the cosine space.

As shown in Figure 4, compared with the traditional softmax loss, there is an obvious decision boundary in large margin cosine loss. Moreover, the classification results only depend on the angle.



(**a**) Softmax loss (**b**) Large margin cosine loss

Figure 4. The decision boundary of (**a**) softmax loss and (**b**) large margin cosine loss.

Formally, the LMCL function is defined as follows:

$$Loss_{lmcl} = \frac{1}{N}\sum_{i} -\log\frac{e^{t(\cos(\theta_{l_i,i})-m)}}{e^{t(\cos(\theta_{l_i,i})-m)} + \sum_{j\neq l_i} e^{t\cos(\theta_{l_i,i})}} \tag{5}$$

In the SMGN model, the LMCL function is followed by two local branches (i.e., "Local-1" and "Local-2") and two global branches (i.e., "Global-1" and "Global-2"). Thus, we can obtain four LMCL functions, which are recorded as $Loss^1_{lmcl}$, $Loss^2_{lmcl}$, $Loss^3_{lmcl}$ and $Loss^4_{lmcl}$. Finally, we add these four LMCL functions to obtain the final identification loss function as follows:

$$Loss_{identification} = Loss^1_{lmcl} + Loss^2_{lmcl} + Loss^3_{lmcl} + Loss^4_{lmcl} \tag{6}$$

3.3.2. Verification Loss

Most previous person re-identification methods based on Siamese network regard verification process as a binary classification problem [9,27,36]. Following this idea, we adopt the widely-used cross-entropy loss function to directly compute the similarity between the extracted multiple granularity features in verification channel. For the feature pair (F_1, F_2), we compute the squared Euclidean distance as a novel feature vector in verification channel. Then the convolutional layer take the new vector as input, which is followed by a softmax output function. As a result, we can obtain a 2-dim vector (p_1, p_2) that represents the predicted probability that the two pedestrian images belong to the same person. Finally, cross-entropy loss function is formulated as follows:

$$p^s = \text{softmax}_{verification}\left((F_1 - F_2)^2 \circ \theta_{verif}\right) \tag{7}$$

$$Loss_{verification}(\theta_{verif}, s) = \sum_{i=1}^{2} p_i^s \left(\frac{1}{p_i}\right) \tag{8}$$

where s represent the target class(same/different), $\circ$ denotes a convolutional operation, p^s is the similarity score of F_1 and F_2, and the transformation is parameterized by θ_{verif}. If the predicted result indicates that the input pedestrian image pair belongs to the same person, $p_1 = 1, p_2 = 0$; otherwise, $p_1 = 0, p_2 = 1$.

3.3.3. Fusion Loss

In order to combine the advantages of verification model and identification model, two different kind of losses mentioned above are weighted fused together to formulate the MCWF loss function as follows:

$$Loss_{fusion}(\theta, s) = \lambda Loss_{verification} + Loss_{identification} \tag{9}$$

where λ is a coefficient to balance the weight of identification and verification loss function. During the training processing, the SMGN model can guarantee multiple granularities features have the

characteristic of margin maximization for classification under the constraint of the MCWF loss function. Therefore, this type of multiple granularity features extracted from the SMGN model are regarded as large margin multiple granularities (LMMG) features. As a result, the SMGN model can improve the performance of person re-identification.

3.4. Person re-Identification Based on SMGN

During the training processing of SMGN model, given a training image set with their labels $X_{train} = \left\{(x_t, l_t)\middle| t = 1, \ldots, N\right\}$, we first construct these images into many image pairs that are recorded as:

$$Pair = \{B_i | i = 1, 2, \ldots, T\} \tag{10}$$

where $B_i = (x_i^1, x_i^2, l_i^1, l_i^2, R_i)$ denotes the i-th image pair, R_i is the label that denotes whether x_i^1 and x_i^2 belong to the same person, if x_i^1 and x_i^2 represent the same person, $R_i = 1$; otherwise, $R_i = 0$. Based on the MCWF loss function and back propagation algorithm, the backbone CNN in SMGN model can be trained with *Pair*, which is recorded as Ω.

In the testing stage, given a query image x_q, its LMMG features F_q can be extracted by the backbone CNN Ω. Similarly, the LMMG features F_i^g of each gallery image in $X_{gal} = (x_1^g, x_2^g, \ldots, x_M^g)$ is also extracted by Ω. We compute the cosine distance between F_q and F_i^g as follows:

$$dist(F_q, F_i^g) = \frac{F_q \cdot F_i^g}{\|F_q\| \|F_i^g\|} = \frac{\sum_i^n F_q \times F_i^g}{\sqrt{\sum_i^n (F_q)^2} \times \sqrt{\sum_i^n (F_i^g)^2}} \tag{11}$$

where n denotes the dimension of LMMG features.

After calculating the distances between the query image x_q and each gallery image in X_{gal}, we sort these distances in ascending order to get the final ranking result. Therefore, we can calculate the corresponding right matching rates. Finally, the person re-identification procedure based on the SMGN model is summarized in Algorithm S1 (in Supplementary Materials).

4. Experiment Results

In this section, we first introduce four large-scale person re-identification databases, i.e., CUHK01, CUHK03, Market-1501 and DukeMTMC-reID. Then some experimental details are depicted, followed by some comparison with the-state-of-the-art methods on four databases. Finally, we explore the effect of the margin parameter m and the balance coefficient λ.

4.1. Datasets and Protocols

For the purpose of validating the effectiveness of the proposed model, we perform extensive experiments on four benchmark person re-identification datasets.

4.1.1. CUHK01

CUHK01 dataset is constructed by 3884 pedestrian images of 971 identities, and each identity has four images that captured by two surveillance cameras. These cameras mainly capture the front, back, left and right appearances of pedestrians. The dataset is spilt into two parts, in which 485 pedestrians are randomly selected for training and the other for testing.

4.1.2. CUHK03

CUHK03 contains 1360 people and 13,164 images captured by five non-overlapping camera pairs. Each identity is observed by two non-overlapping views and has 4.8 images under each camera on average. This dataset has two types of annotations: detector-detected (Deformable Part Model (DPM))

pedestrian bounding boxes (detected) and hand-labeled bounding boxes (labeled). All pedestrian images suffer from illumination changes, misalignment, occlusions and body part missing.

4.1.3. Market-1501

Market-1501 contains 32,668 pedestrian images of 1501 identities captured by six cameras in Tsinghua University campus. Compared with CUHK03, Market-1501 is a large scale dataset for person re-identification. In Market-1501 dataset, there are 12,396 images of 751 identities for training and 19,732 images of 750 identities for testing. All person images are detected by DPM, so some pedestrian images in Market-1501 dataset exists detection errors.

4.1.4. DukeMTMC-REID

DukeMTMC-reID is a subset of DukeMTMC that is used for multi-target tracking dataset. DukeMTMC-reID is a large scale person re-identification dataset that contains 36,411 pedestrian images of 1812 identities. The images in DukeMTMC-reID consist of 16,522 training images (from 702 people), 2228 query images (from another 702 people), and a test gallery for 17,661 images, which are captured at the Duke University campus and cropped from hand-drawn bounding boxes. The size of the images is randomly cropped, and many pedestrians are blocked.

The detail information about these datasets are summarized in Table 2. These four widely-used person re-identification datasets contain many challenges, such as misalignment, occlusions and missing body parts, low resolutions, viewpoints and background clusters. In addition, Figure 5 shows some image samples of the four datasets.

Table 2. The details of person re-identification dataset.

Dataset	Release Time	Identities	Cameras	Images	Label Method	Crop Size
CUHK01	2012	971	2	3884	Hand	160 × 60
CUHK03	2014	1467	10 (5 pairs)	13,164	Hand/DPM	Vary
Market-1501	2015	1501	6	32,217	Hand/DPM	128 × 64
DukeMTMC-reID	2017	1812	8	36,411	Hand	Vary

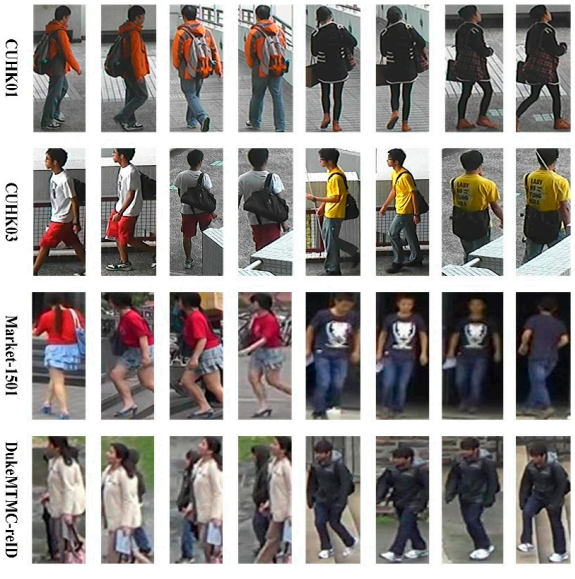


Figure 5. Some samples from CUHK01, CUHK03, Market-1501 and DukeMTMC-reID. Here, each row includes two different identities captured under different cameras.

4.1.5. Metric Protocols

As an evaluation protocol, cumulative match characteristic (CMC) is extensively applied in person re-identification to count the ranks of true matches. At the same time, we also introduce the mean average precision (mAP) for the Market-1501 and DukeMTMC-reID datasets in our experiment. These two criteria are executed under a single query setting for the four datasets. More importantly, the re-ranking method based on the k-reciprocal encoding [37] is adopted for further improvement.

4.2. Implementation Details

We use Python to implement the proposed SMGN model. Some details about data preparation, parameter settings and data augmentation are described in this section.

4.2.1. Data Preparation

For the convenience to extracting features of pedestrian images, we perform the input data preparation. Firstly, we resize all the images into 256×256. Then we utilize the resized input images to subtract the mean image. Afterwards, a random order style [11] is introduced in our paper and we set the initial ratio of positive images to negative images to improve the performance of the SMGN model. In the end, we multiple the ratio between positive and negative pairs by a factor of 1.01 every epoch until it reaches 3:1 to prevent our model from over-fitting.

4.2.2. Parameter Settings

In this experiment, we set the size of image batch to 32 for SMGN, including eight positive and eight negative image pairs. Stochastic gradient descent (SGD) is adopted to update the parameter of SMGN model. The number of training epoch is set to 1000. We set the weight decay to 5×10^{-4} and the momentum to 0.9. As for the learning rate, we set the initial learning rate to 0.001 and then set to 0.0001 for the last 10 epochs. When perform binary-class task, we randomly select negative pairs from the whole negative sample pool for each batch. For the network updating, we accumulate all the gradients produced by every image pair. In the training phase, the weight of the gradient generated by the verification loss is three times as much as the identification loss. We set the parameters $\lambda = 3$ and $m = 0.40$ empirically in all the following experiments. The validation experiment as Figures 6 and 7 illustrated.

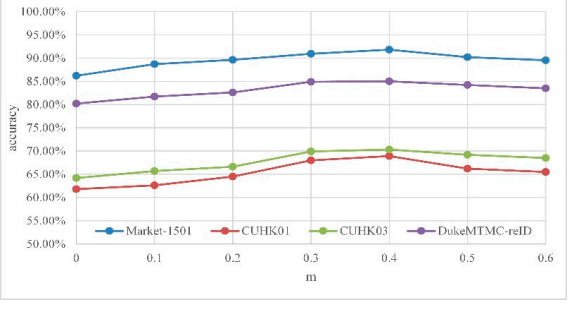


Figure 6. Rank-1 performance of SMGN with different margin parameter m on Market-1501, CUHK01, CUHK03 and DukeMTMC-reID.

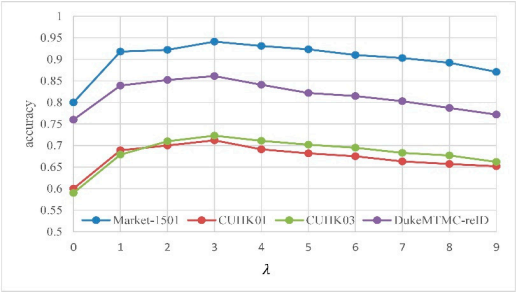


Figure 7. Rank-1 performance of SMGN with different coefficient λ on Market-1501, CUHK01, CUHK03 and DukeMTMC-reID.

4.2.3. Data Augmentation

Person re-identification datasets are composed by various images of different pedestrians, in which each pedestrian has a limited number of images. Because of this, we cannot construct adequate positive pairs to train the SMGN model. Therefore, there exists over-fitting and the performance of the Siamese network is poor.

Compared with the other datasets, CUHK01 is a small scale person re-identification dataset. To cope with over-fitting since the lack of data, data augmentation is adopted in our experiment. Specifically, all the resized pedestrian images are randomly cropped to 224×224 at first. Besides that, horizontal flipping is used on the CUHK01 dataset to implement image augmentation.

4.3. Parameter Analysis

In this section, we evaluate two important parameters, i.e., the fixed margin parameter m in Equation (5) and the balance coefficient λ in Equation (9).

4.3.1. Effect of m

The margin parameter m plays an important role in LCML. To investigate the effort of m, we conduct a comparative experiment in this part. For Figure 6, we compare the results with different margin parameter on CUHK01, CUHK03 (labeled), Market-1501 and DukeMTMC-reID. The margin parameter is used to better control the boundary between different classes. If the margin rate is too large, then the model will fail to converge. In this part, we set the range of m as $[0, 0.6]$ and for every 0.1 m increase, we do a comparison experiment once more. As shown in Figure 6, we can find that the matching performance is worst when $m = 0$ on the four person re-identification datasets. As m being increased, the accuracy of the proposed model in every dataset consistently improves and get saturated at $m = 0.40$. For convenience, the parameter m in Equation (6) is set to fixed 0.40 in the subsequent experiments. Note that λ is set to 1 in this part.

4.3.2. Effect of λ

The balance coefficient λ is to balance verification loss and identification loss. To investigate the effort of λ, we conduct a comparative experiment as Figure 7 illustrated (Note that m is set to 0.40). In this part, we set the range of λ as $[1, 9]$ and for every 1 m increase, we do a comparison experiment once more. As shown in Figure 7, we can see that the matching rates are lowest on the four datasets when $\lambda = 0$. In other words, we cannot obtain the best performance if we only use identification model. Because the identification model only makes full use of the label information of pedestrian images, which is benefit to intra-class separation. As for inter-class compactness, we assume that the verification loss equals zero if the two images belong to the same identity. So we can see that the matching degree is higher with the increase of weight coefficient λ. When λ is set to 3, we can get

the good performance on CUHK01, CUHK03 (labeled), Market-1501 and DukeMTMC-reID. In the following experiment, the parameter λ in Equation (7) is set fixed to 3 in this paper.

4.4. Performance Evaluation

We compare the proposed SMGN model with current state-of-the-art approaches on the four widely-used datasets to show our competitive performance. Comparative results in detail are given below.

4.4.1. Performance on the CUHK01 Dataset

Compared with the state-of-the-art results reported on the CUHK01 dataset, the proposed SMGN model show the best performance that are listed in Table 3. For CUHK01, we consider 486 identities for testing and the rest for training like most previous papers. As shown in Table 3, we can observe that the proposed SMGN model achieve the best rank-1 matching rate at 71.2%, which is higher 2.1% higher than the second best one NFST [38]. With the re-ranking technique in [37], we obtain a higher rank-1 rate on CUHK01.

Table 3. Comparison with the several results reported on the CUHK01 dataset using a CMC curve.

Method	Rank 1	Rank 5	Rank 10	Rank 20
FPNN [9]	27.9%	—	—	—
Deep CNN [10]	47.5%	—	—	—
KCVDCA [39]	47.8%	74.2%	83.4%	89.9%
LOMO+XQDA [4]	49.2%	75.7%	84.2%	90.8%
TCP [40]	53.7%	84.3%	91.0%	93.3%
GOG+XQDA [6]	57.8%	79.1%	86.2%	92.1%
NFST [38]	69.1%	86.9%	91.8%	95.4%
Ours	**71.2%**	**87.2%**	**90.9%**	**95.5%**
Ours+re−rank	**72.0%**	**88.1%**	**91.2%**	**96.3%**

4.4.2. Performance on the CUHK03 Dataset

The CUHK03 dataset has two types of annotations as mentioned above, i.e., labeled and detected. As we can see that the results using different methods on CUHK03 are shown in Table 4. We have the same settings as [9], that is, CUHK03 is partitioned into a training set (1160 persons), validation set (100 persons), and test set (100 persons). It is clear that the proposed SMGN outperforms the other existing methods in the case of both detected and labeled. In Table 4, we can see that the proposed algorithm achieves 70.2% at rank 1 in the case of detected boxes and 72.3% with manual bounding boxes. With the re-ranking technique described in [37], we got a better performance in both cases.

Table 4. Comparison with the several results reported on the CUHK03 dataset using a CMC curve.

Method	Detected			Labeled		
	Rank 1	Rank 5	Rank 10	Rank 1	Rank 5	Rank 10
FPNN [9]	19.9%	49.0%	64.3%	20.7%	51.7%	68.3%
DPFL [41]	40.7%	—	—	43.0%	—	—
SVDNet [42]	41.5%	—	—	40.9%	—	—
HA-CNN [43]	41.7%	—	—	44.4%	—	—
Deep CNN [10]	45.0%	75.7%	83.0%	54.7%	88.3%	93.3%
LOMO+XQDA [4]	46.3%	79.0%	88.6%	52.2%	82.3%	92.1%
MGCAM [44]	46.7%	—	—	50.1%	—	—
LOMO+MLAPG [8]	51.2%	—	—	58.0%	—	—
NFST [38]	54.7%	84.8%	94.8%	62.6%	90.1%	94.8%
PCB+RPP [45]	63.7%	80.6%	86.9%	—	—	—
GOG+XQDA [6]	65.5%	88.4%	93.7%	67.3%	91.0%	96.0%
MGN [16]	66.8%	—	—	68.0%	—	—
Ours	**70.2%**	**87.2%**	**93.9%**	**72.3%**	**89.1%**	**96.7%**
Ours+re−rank	**71.5%**	**88.3%**	**94.0%**	**73.1%**	**90.0%**	**97.1%**

4.4.3. Performance on the Market-1501dataset

We summarize the performance results on Market-1501 dataset using some state-of-the-art methods and our proposed algorithm. It can be found that the deep learning based methods (i.e., Gated SCNN [19], DPFL [42], PCB+RPP [46] etc.) obviously defeat non-deep learning based methods (i.e., BoW+kissme [28], LOMO+XQDA [4]) on the Market-1501 dataset. We can see that the proposed SMGN obtains 94.2% and 80.2% in rank-1 and mAP accuracy respectively. With the re-ranking technique [38], the proposed algorithm outperforms the second best one by a margin of 1.7% at rank-1 under the single query (SQ) setting.

4.4.4. Performance on DukeMTMC-reID

From Table 5, we can see that our algorithm on the DukeMTMC-reID dataset achieves 87.1% rank-1 matching rate and 76.0% mAP respectively, which significantly outperforms the previous state-of-the-art methods. The results on the DukeMTMC-reID dataset show that our method has a great advantage on large scale dataset. Compared with the state-of-the-art methods, our proposed method obtains competitive results on all four datasets. Especially, SMGN achieves 71.2% rank-1 accuracy for CUHK01, 70.2% rank-1 accuracy for CUHK03 (detected), 72.3% rank-1 accuracy for CUHK03 (labeled), 94.1% for Market-1501 and 86.1% for DukeMTMC-ReID without re-ranking. In addition, we visualize the top-10 ranking results on Market-1501 for some randomly-selected query pedestrian images in Figure 8. The results indicate the good performance of our model.

Table 5. Comparison with the several results reported on the Market-1501 and DukeMTMC-reID datasets using a CMC curve.

Method	Market-1501		DukeMTMC-re-ID	
	Rank-1	MAP	Rank-1	MAP
BoW+kissme [18]	39.6%	17.7%	25.1%	12.2%
LOMO+XQDA [4]	43.8%	22.2%	30.8%	17.0%
NFST [38]	55.4%	29.9%	—	—
Gated SCNN [27]	65.9%	39.6%	—	—
SVDNet [42]	82.3%	62.1%	76.7%	56.8%
MGCAM [44]	83.8%	74.3%	—	—
PSE [46]	87.7%	69.0%	79.8%	62.0%
DPFL [41]	88.6%	72.6%	79.2%	60.6%
HA-CNN [43]	91.2%	75.7%	80.5%	63.8%
Deep-Person [47]	92.3%	79.6%	80.9%	64.8%
PCB+RPP [45]	93.8%	81.6%	83.3%	69.2%
Ours	**94.1%**	**79.2%**	**86.1%**	**75.3%**
Ours+re-rank	**95.5%**	**80.3%**	**87.1%**	**76.0%**

Figure 8. Three example query images in Market-1501 test set and their corresponding top 10 ranking lists results using our method. The green boundary means true positive and red means false positive.

5. Conclusions

In this paper, we propose a novel symmetric Siamese model named SMGN for person re-identification. In order to learn multiple granularity features from global and local regions, we adopt modified ResNet-50 as the backbone network at first and use the local and global branches to extract multiple granularity features. Then a multi-channel weighted fusion (MCWF) loss function is designed to further reduce the intra-class variance while increase the inter-class variance, which consider an obvious decision boundary when classifying. Finally, we integrated SMGN and the MCWF loss function together and the large margin multiple granularities (LMMG) features can be obtained when the loss function tends to the minimum value. After waiting for SMGN to stabilize, we use the backbone network of it for testing to get the ranking lists of the target image. We validated the effectiveness of the proposed SMGN on four widely-used person re-identification datasets and the performance on those are improved comparing with many state-of-the-art methods. Our future work is to explore more robust and discriminative features of person images and investigate on how to achieve compactness of intra-class and separation of inter-class much better.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-8994/12/1/92/s1.

Author Contributions: Supervision: S.-W.T.; validation: G.-Y.F.; Writing—original draft, G.-Y.F.; writing—review and editing: D.-X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the International Cooperation and Exchange Foundation of Shaanxi, China (Grant 2017KW-013), and the Innovation & Entrepreneurship Dual Tutor Foundation of Shaanxi, China (grant nos. 2019JM-604), and the Xi'an University of Posts and Telecommunications Graduate Innovation Fund Project under grant CXJJLY2018040.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-Identification: Past, Present and Future. Available online: https://arxiv.org/abs/1610.02984 (accessed on 5 June 2019).
2. Karanam, S.; Gou, M.; Wu, Z.; Rates-Borras, A.; Camps, O.; Radke, R.J. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 523–536. [CrossRef] [PubMed]
3. Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2360–2367.
4. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
5. Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; Li, S. Salient color names for person re-identification. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Volume 8689, pp. 536–551.
6. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical Gaussian descriptor for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1363–1372.
7. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
8. Liao, S.; Li, S. Efficient PSD constrained asymmetric metric learning for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Santiago, Chile, 11–18 December 2015; pp. 3685–3693.
9. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep filter pairing neural network for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.

10. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
11. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned CNN embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 1–20. [CrossRef]
12. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3376–3385.
13. Wu, L.; Wang, Y.; Gao, J.; Li, X. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Trans. Multimed.* **2019**, *21*, 1412–1424. [CrossRef]
14. Chung, D.; Tahboub, K.; Delp, E.J. A two stream siamese convolutional neural network for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 1983–1991.
15. Yan, Y.; Ni, B.; Song, Z.; Ma, C.; Yan, Y.; Yang, X. Person re-identification via recurrent feature aggregation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 701–716.
16. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning Disriminative Features with Multiple Granularities for Person Re-Identification. Available online: https://arxiv.org/abs/1804.01438 (accessed on 25 June 2019).
17. Li, W.; Zhao, R.; Wang, X. Human re-identification with transferred metric learning. In *Lecture Notes in Computer Science, Proceedings of the Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 31–44.
18. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1116–1124.
19. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
20. He, N.; Paoletti, M.E.; Haut, J.M.; Fang, L.; Li, S.; Plaza, A.J.; Plaza, J. Feature extraction with multiscale covariance maps for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 755–769. [CrossRef]
21. Shih, Y.; Yeh, Y.; Lin, Y.; Weng, M.; Lu, Y.; Chuang, Y. Deep co-occurrence feature learning for visual object recognition. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7302–7311.
22. Gupta, U.; Chatterjee, A.; Srikanth, R.; Agrawal, P. A Sentiment-and-Semantics-Based Approach for Emotion Detection in Textual Conversations. July 2017. Available online: https://arxiv.org/abs/1707.06996 (accessed on 5 August 2019).
23. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 384–393.
24. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3980–3989.
25. Xiao, T.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1249–1258.
26. Subramaniam, A.; Chatterjee, M.; Mittal, A. Deep neural networks with inexact matching for person re-identification. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2667–2675.
27. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In Proceedings of the European Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 791–808.

28. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 499–515.
29. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. Available online: https://arxiv.org/abs/1703.07737 (accessed on 1 October 2019).
30. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-Margin Softmax Loss for Convolutional Neural Networks. Available online: https://arxiv.org/abs/1612.02295 (accessed on 10 October 2019).
31. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep hypersphere embedding for face recognition. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6738–6746.
32. Wang, F.; Cheng, J.; Liu, W.; Liu, H. Additive margin softmax for face verification. *Signal Process.* **2018**, *25*, 926–930. [CrossRef]
33. Almazán, J.; Gajic, B.; Murray, N.; Larlus, D. Re-Id Done Right: Towards Good Practices for Person Re-Identification. Available online: https://arxiv.org/abs/1801.05339 (accessed on 13 December 2019).
34. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lile, France, 6–11 July 2015; pp. 448–456.
35. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. CosFace: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5265–5274.
36. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 December 2014; pp. 34–39.
37. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3652–3661.
38. Zhang, L.; Xiang, T.; Gong, S. Learning a discriminative null space for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1239–1248.
39. Chen, Y.; Zheng, W.; Lai, J.; Yuen, P. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1661–1675. [CrossRef]
40. Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; Zheng, N. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1335–1344.
41. Chen, Y.; Zhu, X.; Gong, S. Person re-identification by deep learning multi-scale representations. In Proceedings of the Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2590–2600.
42. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. SVDNet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3800–3808.
43. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2285–2294.
44. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for personre-Identification. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1179–1188.
45. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling. In *Lecture Notes in Computer Science, Proceedings of the European Conference on Computer Vision*; Springer: Munich, Germany, 8–14 September 2018; pp. 480–496.
46. Sarfraz, M.S.; Schumann, A.; Eberle, A.; Stiefelhagen, R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 420–429.
47. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. Available online: https:/arxiv.org/abs/1711.10658 (accessed on 10 November 2019).

symmetry

MDPI

Article

Micro-Distortion Detection of Lidar Scanning Signals Based on Geometric Analysis

Shuai Liu [1,2,3], Xiang Chen [1], Ying Li [4] and Xiaochun Cheng [5,*]

1 State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang 471000, China; cs.liu.shuai@gmail.com (S.L.); cemee_xchen@163.com (X.C.)
2 College of Computer Science, Inner Mongolia University, Hohhot 010012, China
3 College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
4 College of information and communication engineering, Harbin Engineering University, Harbin 150000, China; 3188279500@hrbeu.edu.cn
5 College of Computer Science, Middlesex University, London NW4 4BT, UK
* Correspondence: x.cheng@mdx.ac.uk

Received: 20 October 2019; Accepted: 29 November 2019; Published: 3 December 2019

Abstract: When detecting micro-distortion of lidar scanning signals, current hardwires and algorithms have low compatibility, resulting in slow detection speed, high energy consumption, and poor performance against interference. A geometric statistics-based micro-distortion detection technology for lidar scanning signals was proposed. The proposed method built the overall framework of the technology, used TCD1209DG (made by TOSHIBA, Tokyo, Japan) to implement a linear array CCD (charge-coupled device) module for photoelectric conversion, signal charge storage, and transfer. Chip FPGA was used as the core component of the signal processing module for signal preprocessing of TCD1209DG output. Signal transmission units were designed with chip C8051, FT232, and RS-485 to perform lossless signal transmission between the host and any slave. The signal distortion feature matching algorithm based on geometric statistics was adopted. Micro-distortion detection of lidar scanning signals was achieved by extracting, counting, and matching the distorted signals. The correction of distorted signals was implemented with the proposed method. Experimental results showed that the proposed method had faster detection speed, lower detection energy consumption, and stronger anti-interference ability, which effectively improved micro-distortion correction.

Keywords: geometric analysis; lidar scanning signal; micro-distortion; detection technology; TCD1209DG; lossless signal transmission

1. Introduction

The recent development of wireless information has promoted the maturity of laser lidar mapping technology. As a key technology in the field of surveying and mapping, lidar scanning has received more and more attention from relevant experts and scholars [1,2]. Lidar mapping technology quickly and accurately acquires three-dimensional information of objects, making it widely used in production and life [3]. However, due to factors, such as the outdoor light, the lidar scanning signal is slightly distorted, which affects the accuracy of the acquired information [4]. Detection for micro-distortion of lidar scanning signals improves the quality of the acquired lidar scanning signal, which is of great significance for ensuring the utilization efficiency of lidar scanning signals [5]. However, existing distortion detection technologies of lidar scanning signals only detect the region where the target's distortion of three-dimensional information is severe. With the popularity of lidar scanning applications, accuracy requirements for lidar scanning are getting higher [6]. Since current complicated micro-distortion detection technologies of lidar scanning signals have poor anti-interference, micro-distortion detection technology of scanning signals has become the focus of

research in this area. With the deepening of the research content, some mature theories and applications have been produced [7].

Different experts and scholars realized the detection of micro-distortion for scanning signals by different methods in years. However, there are still some shortcomings in this research domain and need follow-up experts and scholars to study. A micro-distortion detection technology based on fiber optic gyroscope was proposed by Zheng et al. [8]. The calibrated signal structure was tested linearly by achieving a level gauge of horizontal and vertical signal distortion calibration. Micro-distortion detection was further realized by the calibration result. However, this method did not pre-process scanning signals, so its distortion detection was easily interfered by factors, such as illumination change. Lupi et al. [9] proposed a distortion detection technique for lidar scanning signals based on active panoramic vision. Lidar scanning signal information was obtained by acquiring three-dimensional coordinates of point clouds for lidar scanning signals. Then, lidar scanning signals were preprocessed, and their three-dimensional coordinates of the feature points for lidar scanning signals were determined to realize quantitative analysis. According to the analysis results, a three-dimensional model was constructed to realize distortion detection for lidar scanning signals. However, the detection process of this method was complicated, which affected high time-consuming detection. Xu et al. [10] proposed a distortion detection technique based on ASODVS (Active Stereo Omni-Directional Vision Sensing) for lidar scanning signals. The midpoint of the lidar scanning signal was determined by a Gaussian curve. The waveform of the lidar signal was smoothed by the Bezier curve. Lidar scanning signal was calibrated by ASODVS, and its distortion detection was realized by qualitative analysis. However, in the process of detecting the distortion signal designed by this method, more energy was consumed, and its detection cost was higher. In [10], it was proposed to use FPGA (Stratix, made by ALTERA, San Jose, California, USA) as the main control equipment, design a general echo signal acquisition card with 20 MHz sampling rate, filtering and hardware accumulation functions, and configurable parameters. For the design of the analog-to-digital conversion circuit and its peripheral circuit, signal conditioning circuit, level conversion circuit, RS232 interface circuit, and power circuit in the analog board card, the logic design of FPGA was given in that circuit too. For the radar scanning signal detection, the basic parameter configuration function and accumulation function were implemented, but the method had poor detection accuracy for long-distance signal detection.

Therefore, a micro-distortion detection method for lidar scanning signals with geometric statistical characteristics was proposed in this paper. The experimental study of lidar with a frequency of 500 MHz was carried out. The proposed method used TCD1209DG to model the linear array CCD (charge-coupled device), reverse optoelectronics, and store and transmit the signal charge. It improved the time-consuming and energy consumption of [9,10], respectively. The linear CCD module was designed according to LS series lidar. FPGA was used for signal preprocessing of TCD1209DG output. C8051, FT232, and RS-485 were used to reduce the loss of signal transmission. The signal TRAM publishing unit sent the distortion information of the radar scanning signal to the control host computer for micro-distortion detection of the radar scanning signal. The anti-interference ability of the method in [8] was improved. Through the experiment of micro-distortion correction, compared with the existing methods, this method had faster detection speed, lower detection energy consumption, and stronger anti-interference ability.

2. Hardware Support for the Technology

2.1. Overall Framework of the Technology

In order to realize micro-distortion detection of lidar scanning signals, technical modules, such as linear array CCD module, signal processing module, signal transmission unit, and PC control program, were designed according to LS series laser lidar. The overall framework of technology is shown in Figure 1.

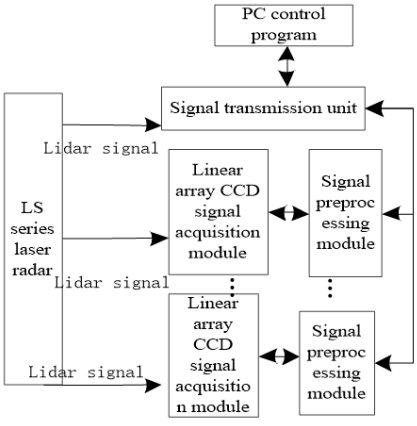


Figure 1. The overall framework of the proposed method.

According to Figure 1, each linear CCD module in the method received a word line laser emitted by the horizontal line lidar and outputted a corresponding signal through the lidar signal. When the lidar scanning signal changed slightly, its signal output from the line CCD module changed. The signal acquisition module collected the output signal and then digitized the acquired signal to determine micro-distortion information for the scanning signal. The signal transmission unit transmitted distortion information of the lidar scanning signal to the control host to realize micro-distortion detection for the lidar scanning signal. The control host could implement the setting of technical parameters and receive and process signals.

2.2. Development of Linear Array CCD Module

A charge-coupled device (CCD) [11] is a sensor that uses charge to realize signal transmission, enabling photoelectric conversion and signal charge storage and transfer. According to the working content of the linear array CCD module, chip TCD1209DG [12] with high sensitivity and high resolution is used as a chip of the linear array CCD module. The size of each photosensitive unit of tcd1209dg was 14 μm × 14 μm × 14 μm, the total length of the photosensitive array was 28.6 mm, the best working frequency was 1 MHz, and the maximum working frequency could reach 201 MHz. Because of its anti-interference advantages compared with the traditional chip, this paper chose this chip. The design of this hardware was to detect the micro-distortion signal effectively, so as to realize the detection of radar scanning signal. Using photodiode as the used pixel, the size of each pixel was set to 9.33 μm × 9.33 μm. The line CCD is shown in Figure 2.

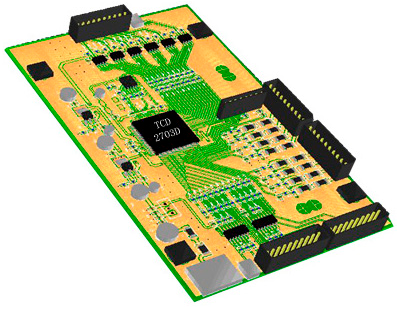
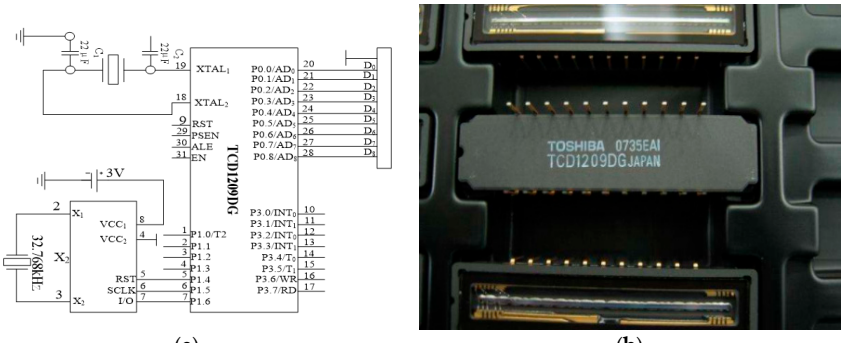


Figure 2. A general linear array CCD (charge-coupled device).

2.3. Signal Preprocessing Module

In order to realize the output of red, green, and blue light, chip TCD1209DG realized two channels of six frames of output, and its output signal size of each frame was 3984.2 × 16 *bits*. In the process of using TCD1209DG, the light entering the linear array CCD was made red by providing a red filter lens in front of it. It only processed the red light output and reduced the effect of external stray light on signal processing.

Chip TCD1209DG had five driving signals, including two-phase clock signals φ1A and φ2A, a charge conversion signal SH, a reset signal RS, and a clamp signal CP. There was a strict timing and phase relationship between the signals in TCD1209DG. Drive signal circuit diagram of TCD1209DG is shown in Figure 3a. When TCD1209DG was operating normally, the two-phase clock signal contained a high-level clamp signal and a reset signal, and the clamp signal lagged behind the reset signal. When the scanning signal was distorted, the clamp signal and the reset signal were low.



(**a**) (**b**)

Figure 3. TCD1209DG driver and its circuit. (**a**) TCD1209DG drive signal circuit, (**b**) A real TCD1209DG driver.

TCD1209DG drive signal was generated by the internal logic of FPGA and was implemented by VHDL programming [13]. Drive resulting from TCD1209DG is shown in Figure 3b.

The signal processing module mainly adjusted the signal output by TCD1209DG, digitized the processed signal, and stored the processed signal.

For signal output features of chip TCD1209DG, it was necessary to preprocess the output signal of each frame. The signal threshold processing was outputted for each frame, and the processed signal was transmitted to the signal acquisition module. The high sensitivity of TCD1209DG made the light have a greater impact on the output signal. By processing the threshold, interference of the light could be effectively reduced, and the quality of the output signal with the linear array CCD module was improved.

2.4. Signal Transmission Unit

In order to realize the lossless signal transmission for the lidar scanning signals of the control host and other slaves, a signal transmission unit was designed. The unit used FPGA as the core device of the signal processing module and adopted EP1C6Q144 [14] as the chip FPGA. After processing by FPGA, the micro-distortion information of the lidar scanning signal was stored in the off-chip SRAM.

The signal transmission unit was mainly composed of chip C8051 [15], chip FT232, and chip RS-485 [16]. In technology, the signal transmission unit host could be connected to any slave to realize communication. However, communication between slaves was impossible.

In order to realize the signal transmission unit, the connection between the USB interface with RS-232 and RS-485 interfaces was completed by using chip FT232. RS-232 interface was connected to

the RS-485 interface through chip RS-485. Using C8051F (made by Silicon Labs, Austin, Texas, USA) as an MCU, the timing between the chips was controlled to avoid communication interruption caused by bus conflicts. The specific implementation process of the signal transmission unit is shown in Figure 4.

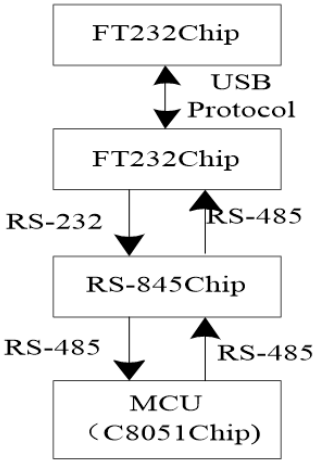


Figure 4. Composition of the signal transmission unit.

Through the above discussion, based on the requirements of the micro-distortion detection of the lidar scanning signals, the overall framework of the technology was analyzed. Designs of the linear array CCD module, the signal processing module, and the signal transmission unit for implementation determined the detection process of the micro-distortion detection technology for the lidar scanning signal.

3. Feature Matching Algorithm for Micro-Distortion Signal Based on Geometric Statistics

3.1. Frequency Matching of Micro-Distortion Based on Geometric Statistical Algorithm

The premise of realizing the detection of micro-distortion for the lidar scanning signals was the need to describe the lidar scanning signals. By accurately scanning the local properties of a signal, the frequency of the distortion signal was matched. According to the principle of lidar scanning and mapping, combined with the geometric statistical algorithm, a feature frequency band of the distorted signal was matched.

The distortion feature F_P of the scanning point band P on the segment S was introduced. A coordinate system was constructed by taking the scanning point band P as the origin and the normal [17] direction α_p as the abscissa. A tangent along the point P was equally divided into two sides. The length of S was set to L. Each micro-segment was recorded as S_i by projecting the points on S into S_i in sequence until the end of S. Using S_i statistics for a continuous segment on S, the distortion feature F_P of the scanning point band P thus obtained is:

$$F_P = \left(L, U_P^B(l,r), \left\{F'_{S_i} | i = \lfloor l/L \rfloor, \cdots, \lceil r/L \rceil\right\}\right) \tag{1}$$

In the above Equation (1), $U_P^B(l,r)$ represents the boundary farthest from the origin P projection, $\left\{F'_{S_i} | i = \lfloor l/L \rfloor, \cdots, \lceil r/L \rceil\right\}$ represents the set of all S_i distortion feature points, and $\lceil r/L \rceil - \lfloor l/L \rfloor$ represents the number of S_i associated with the scan point band.

Micro-segment S_i associated with the scanning point band was used as a segment with coordinate system information, and its distortion feature F'_{S_i} could be expressed as:

$$F'_{S_i} = (L_{S_i}, U^{\alpha}_{S_i}(\mu^{\alpha}_{S_i} - \sqrt{3}\delta^{\alpha}_{S_i}, \mu^{\alpha}_{S_i} + \sqrt{3}\delta^{\alpha}_{S_i}), C_{S_i}, U^{H}_{S_i}(t, d)) \tag{2}$$

In the above Equation (2), L_{S_i} and C_{S_i}, respectively, represent the length and overall unevenness of each micro-segment S_i after division. $U^{H}_{S_i}$ represents the vertical distance distribution range of the projected scan point band and point P [18]. $\mu^{\alpha}_{S_i}$ represents the average off-angle of the normal direction for the scanning point band with α_p in each micro-segment S_i. $\delta^{\alpha}_{S_i}$ represents the off-angle of the normal direction for the scanning point band with α_p in each micro-segment S_i. The normal direction of the micro-segment was set to satisfy consistent distribution. The associated microsegment was simplified into a distorted feature segment or arc. The range of the central angle was represented by $U^{H}_{S_i}$. S_i and $U^{H}_{S_i}$ were used to determine the position in the coordinate system. The direction of the opening was determined by $\mu^{\alpha}_{S_i}$ and C_{S_i}.

The segmentation distortion feature F_P could be seen as a special form of the associated segment distortion feature F'_{S_i}:

$$F_{S_i} = (L_{S_i}, \delta^{\alpha}_{S_i}, C_{S_i}) \Leftrightarrow F'_{S_i} = (L_{S_i}, U^{\alpha}_{S_i}(-\sqrt{3}\delta^{\alpha}_{S_i}, \sqrt{3}\delta^{\alpha}_{S_i}), C_{S_i}, \phi) \tag{3}$$

The similarity between different micro-segments S_1 and S_2 was calculated by micro-segment similarity S'_F, which is:

$$S'_F = R_L \cdot R_\delta \cdot R_H \frac{1}{1 + \left| C_{S_1}\delta^{\alpha}_{S_1} - C_{S_2}\delta^{\alpha}_{S_2} \right|} \tag{4}$$

In the above Equation (4), R_L represents the length ratio of each micro-segment in a lidar scanning signal. R_δ represents the overlap ratio of the central angle range between micro-segments. R_H represents the projection distribution ratio. According to the above equation, $S'_F \in [0, 1]$, and S'_F increases as the similarity increases.

In order to ensure the accuracy of distortion detection for the lidar scanning signal, it was necessary to accurately match the scanning point's frequency bands. For any of the scanning point bands P_1 and P_2, their coordinate systems were overlapped. Then, the matching degree of the scanning point band M_P could be expressed as:

$$M_P = (\overline{S}_F, O_B, \overline{S}'_F, O'_B) \tag{5}$$

In the above Equation (5), $\overline{S}_F$ and $\overline{S}'_F$, respectively, describe the average similarity of the corresponding micro-segment and the matching micro-segment. O_B represents overlap ratio of lidar scanning signal distortion feature span U^{B}_{P}. O'_B indicates the proportion of matching micro-segments. The average match case was described using $\overline{S}_F$ and O_B. $\overline{S}'_F$ and O'_B were used to improve the matching of scanning point bands. The highest matching degree was set to $M_{best} = (1, 1, 1, 1)$, and the matching degree could be expressed as:

$$V_{M_P} = \cos(\frac{\pi}{2} \cdot \sum_{i=1}^{4} w_i(\eta_i - 1)^2) \tag{6}$$

In the above Equation (6), $\sum_{i=1}^{4} w_i = 1$, w_i represents the weights of the distortion feature matching degree vector. η_i indicates an influencing factor of the matching. $V_{M_P} \in [0, 1]$, the larger V_{M_P}, the higher the matching. The geometrical statistical algorithm was used to match frequency bands of micro-distortion signals, and the geometrical statistical algorithm was combined according to frequency bands.

3.2. Nonlinear Micro-Distortion Signal Detection

During the process of signal scanning, the scanning signal is distorted due to the influence of illumination and scanning target chromatic aberration. Some obvious distortions can be easily detected by linear detection algorithms [3]. However, some distortion signals are nonlinear due to their distortion features, which are often overlooked by monitoring algorithms. Therefore, this paper adopted a nonlinear micro-distortion detection algorithm to detect micro-distortion signals based on the matched distortion signal frequency band. Figure 5 shows the nonlinear detection processing in this paper.

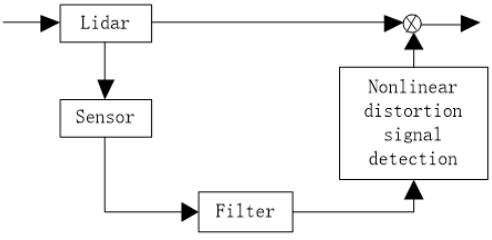


Figure 5. Nonlinear detection process of the micro-distortion signal.

The state of laser lidar scanning was fully considered. An output source of laser and the scanning signal were used as a basis for distortion detection. The analysis of the band structure for the distorted optical signal was as follows:

$$\begin{cases} q(k+1) = Aq(k) + Bh(k) + Cf(k) \\ p(k) = Dq(k) + Eh(k) \end{cases} \tag{7}$$

where $q(k)$ is the normal lidar signal scanning shape, $p(k)$ is the scanning result output, $h(k)$ is the disturbance output, $f(k)$ is the micro-distortion signal, and A, B, C, D, E are distortion dimension constant matrixes.

A lidar would have distortion problems during the scanning process. Therefore, a nonlinear state detection method was needed to detect the signal scanning with the sensor reaching the filter. Let the micro-distortion signal be $\alpha(k)$; then, its initial output signal could be expressed as:

$$p'(k) = Dq(k - \alpha(k)) + Eh(k - \alpha(k)) + Ff(k) \tag{8}$$

In the above Equation (8), $p'(k)$ is the output of dividing the micro-distortion signal set. $0 < \underline{\alpha} < \alpha(k) < \overline{\alpha}$, where $\underline{\alpha}$ is the lower bound of the micro-distortion signal, $\overline{\alpha}$ is the upper bound of the micro-distortion signal, and F is the matrix of the distortion constant for the distortion signal.

Common micro-distortion signal structures, such as the nonlinear micro-distortion state detection mechanism, could be expressed as Equation (9):

$$\begin{cases} q'(k+1) = Aq'(k) + Bg(k) + L\left(\overline{p'}(k) - p'(k)\right) \\ p'(k) = Dq'(k) \\ r(k) = M(\overline{p}(k) - p'(k)) \end{cases} \tag{9}$$

where $q'(k)$ is the control host status detection situation. $p'(k)$ is a micro-distortion output with a micro-distortion signal. $g(k)$ is the detection output. $r(k)$ is the residual signal with a micro-distortion signal. L and M are a gain matrix of the micro-distortion detection and the gain matrix of the residual signal, respectively [4].

3.3. Distortion Correction of Lidar Scanning Micro

The detected micro-distortion signal existing in the process of nonlinear micro-distortion detection was fully considered. According to the distortion distribution sequence, statistical signal features were:

$$\begin{array}{l} p(\beta(k) = 0) = R\{\beta(k)\} = \overline{\beta} \\ p(\beta(k) = 1) = 1 - R\{\beta(k)\} = 1 - \overline{\beta} \end{array} \tag{10}$$

Detection signal in the scanning and the random micro-distortion signal of lidar scanning signal were represented by $\alpha(k)$ in Equation (8) and $\beta(k)$ in Equation (10), respectively. Where $\alpha(k) = 1$ indicates the detection signal received by the control host, and $\beta(k) = 1$ indicates the lidar signal received by the control host. At this time, $\overline{p}(k) = p(k-1)$ and $g(k) = g'(k-1)$ indicate that the lidar signal had slight distortion. $\alpha(k) = 0$ and $\beta(k) = 0$ indicate that the host did not receive the detection signal and the scanning signal, respectively. At this time, $\overline{p}(k) = p(k)$ and $g(k) = g'(k)$, indicating that there was no micro-distortion signal in this frequency band.

The nonlinear state detection method was used to detect signal scanning with the sensor reaching the detection filter. According to this situation, a nonlinear micro-distortion state detection flow was designed. The micro-distortion state detection mechanism was set, the micro-distortion signal condition was analyzed, and the signal features were statistically reported to detect the micro-distortion [5].

In order to make an analysis of lidar problem more precise, the following distortion correction rules should be set:

(1) In order to improve the performance of the control host, the sensor design of the entire lidar needs to use the clock as the corrective drive, and the controller uses the event as the corrective drive.
(2) Data is scanned in a single package.
(3) The local scanning state of the micro-distortion signal is controllable.

When the lidar signal showed random micro-distortion, its output was:

$$y(k') = \alpha'(k')E'\mathrm{x}(k') \tag{11}$$

where $y(k')$ is the output of the micro-distortion detection, and E' is the matrix of the micro-distortion dimension constant.

According to the micro-distortion structure of the lidar signal, a nonlinear micro-distortion state correction mechanism, such as Equation (12), was constructed:

$$\begin{cases} x'(k'+1) = A'x'(k') + B'\overline{g}(k') + L'\left(\overline{y'}(k') - y'(k')\right) \\ y'(k') = D'g'(k') \end{cases} \tag{12}$$

where $x'(k')$ is the nonlinear control master state. $\overline{g}(k')$ is the micro-distortion signal input of the lidar. $g'(k')$ is a micro-distortion signal input without a lidar scanning signal. L', M' are the observer gain matrix and controller gain matrix for minor distortion correction, respectively [9].

The combined variable $\tau(k')$ obeyed Bernoulli distribution, and the statistical signal features were:

$$\begin{array}{l} \mathrm{Prob}(\tau(k') = 1) = R'\{\tau(k')\}\tau \\ \mathrm{Prob}(\tau(k') = 0) = 1 - R'\{\tau(k')\}1 - \tau \\ Var(\tau(k')) = R'(\tau(k') - \tau)^2 = (1-\tau)\tau = \tau^2 \end{array} \tag{13}$$

where $\alpha'(k')$ in Equation (11) and $\tau(k')$ in Equation (13) that obey Bernoulli distribution represent the distortion that occurs when the sensor transmits to the controller and the controller transmits to the actuator. $\alpha'(k') = 1$ indicates that the sensor successfully scans the signal set to the controller, and $\alpha'(k') = 0$ indicates that the sensor is distorted when transmitting the signal set to the controller.

$\tau(k') = 1$ indicates that the controller successfully scans the signal set to the actuator, and $\tau(k') = 0$ indicates that the controller is distorted when scanning the signal set to the actuator.

According to the block diagram of the micro-distortion structure containing lidar, the nonlinear state correction method was used to describe the micro-distortion signal structure of the lidar scanning signal. According to stated problems of micro-distortion signal structure, the reasonable correction rule was analyzed, and the output result when the micro-distortion had a random distortion phenomenon was calculated. Signal features under the distortion phenomenon were statistically combined with the variable $\tau(k')$ obeying Bernoulli distribution to correct the small lidar distortion.

When there was no micro-distortion of a lidar signal, the error caused by other interference factors of the control host could be ignored. At this time, the control host output result was 0, and the micro-distortion observation error was also 0. When micro-distortion of the lidar signal occurred, the output of the control host at this time was not 0, and the observation error of the micro-distortion was not 0. The magnitude of the error changed as the output was processed. Using the nonlinear feedback control law, when the micro-distortion control error increased, the micro-distortion observation error needed to be used as the residual. The entire scanning state of the micro-distortion was analyzed by observing the change in the residual. It completed the correction of micro-distortion for the lidar scanning signal.

According to the above description, these correction principles should be followed. If the absolute value of the error was less than the output result threshold, then the lidar was in a normal scanning state. If the absolute value of the error was greater than or equal to the output result threshold, then the scan signal was slightly distorted. According to this principle, the distortion correction mechanism could be designed to complete the fault-tolerant control for micro-distortion laser lidar, and the fault-tolerant phenomenon in the distortion correction process was divided and compensated.

4. Results and Analysis of Simulation Experiments

In order to test the detection time and the detection degree of the proposed geometric statistics-based micro-distortion detection technology for lidar scanning signals, the experiment was compared with methods in [8,9]. The experimental platform was built by the control host of Intel B360 i7-8700. Using Windows 2010 as the operating system, MATLAB was used to simulate the process. The detection interface of distortion signals is shown in Figure 6.

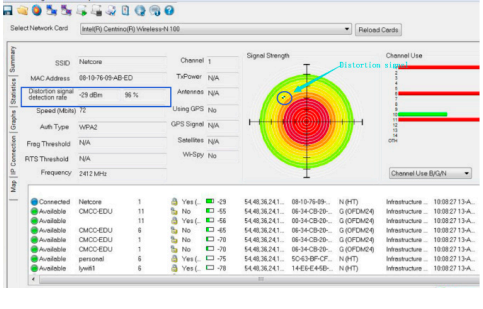


Figure 6. Detection interface of lidar signal distortion.

The proposed method and methods in [8,9] were used to detect the micro-distortion of 3000 sets for lidar scanning signals. Through experiments, time-consuming results of the three methods were recorded, as shown in Table 1.

Table 1. Comparison of time-consuming for micro-distortion detection by different methods.

Number of Experiments/Time	Proposed Method/s	Literature [8] Method/s	Literature [9] Method/s
1	2.13	2.58	3.03
2	2.16	2.64	3.09
3	2.08	2.52	2.94
4	2.09	2.53	2.95
5	2.12	2.56	3.02
6	2.15	2.63	3.07
7	2.10	2.54	3.96
8	2.14	2.60	3.05
9	2.13	2.59	3.04
10	2.11	2.26	2.99

It could be seen from Table 1 that it took less time to detect micro-distortion for the lidar scanning signal by the proposed method. It showed that its detection speed was faster, and the detection efficiency was higher. The proposed method directly matched scanning points in the process of detecting micro-distortion for lidar scanning signals. Calculation steps of the micro-distortion detection for lidar scanning signals were reduced, and its detection speed was improved. Therefore, the detection of micro-distortion for lidar scanning signals took a short time.

In order to ensure the reliability of the research and development technology, a noise signal was added through MATLAB and then connected to software Original Pro 7.0 (Website: www.Originlab.com). Using the proposed method and methods in [8,9], the micro-distortion of the lidar scanning signal was detected in a strong interference environment. The three-dimensional output of the three methods is shown in Figure 7.

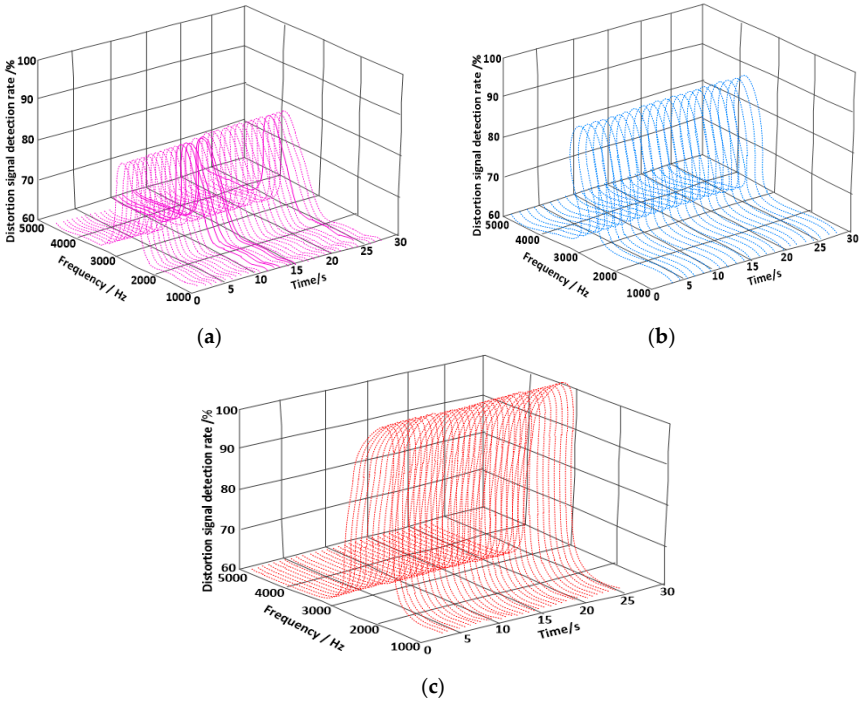


Figure 7. Comparison of detection rates for distortion signals. (**a**) Distortion detection rate of the method in [8], (**b**) Distortion detection rate of the method in [9], (**c**) Distortion detection rate of the proposed method.

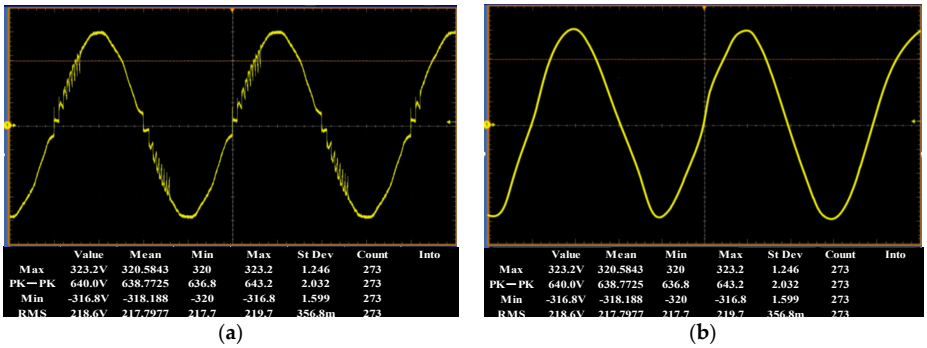
It could be seen from Figure 7 that the proposed method had a high detection rate of the distortion signal in the process of monitoring micro-distortion for the lidar scanning signal. It showed that the external factor had the least influence on the detection rate of the distortion signal, and the proposed detection method had the strongest anti-interference. In the process of detecting micro-distortion of the lidar scanning signal, the proposed method effectively reduced interference of the light and improved the anti-interference of the signal.

An experiment used the proposed method and the classical methods to compare the energy consumption of micro-distortion detection for lidar scanning signals. During the experiment, the results obtained are shown in Table 2.

Table 2. Comparison of energy consumption (nJ) for distortion signals detected by different methods.

Number of Experiments/Time	Proposed Method/nJ	Literature [8] Method/nJ	Literature [9] Method/nJ
1	235	342	318
2	241	350	326
3	237	345	321
4	240	352	328
5	236	343	317
6	235	342	315
7	238	347	323
mean			

It could be seen from Table 2 that the lidar scanning signal detected by the proposed method had the least energy consumption. It showed that the proposed method was less costly for distortion signal detection. In the process of energy consumption detection, the proposed method had fewer modules, which required less energy consumption. In order to ensure the correct performance of the distortion signal for the R&D technology, MATLAB was used to connect an oscilloscope software. The detected micro-distortion signal was corrected by the proposed method. Its correction result is shown in Figure 8.



(**a**) (**b**)

Figure 8. Comparison of micro-distortion signals before and after correction. (**a**) Minor distortion signal before correction, (**b**) Corrected micro-distortion signal.

It could be seen from Figure 8 that there were five micro-distortion signal bands in the lidar scanning signal samples selected before the correction. After the lidar scanning micro-distortion signal was corrected by the proposed method, the signal waveform before the correction was smoother, and there was no distortion phenomenon. It was proved that the proposed method had strong feasibility in the function of micro-distortion correction.

5. Conclusions

Micro-distortion detection of the lidar scanning signal could be used to improve the lidar systems. Existing micro-distortion detection technologies of lidar scanning signal have the problems of long

detection time, high energy consumption, and poor performance against interference [19,20]. To deal with these problems, a technique based on geometric statistics for micro-distortion detection of the lidar signal was proposed. This project built an overall framework for the micro-distortion detection using TCD1209DG in linear array CCD module for photoelectric conversion, signal charge storage, and transfer. FPGA chip was used for the signal preprocessing of TCD1209DG output. C8051, FT232, and RS-485 were used for signal transmission. The signal distortion features were analyzed by geometric statistics for micro-distortion detection. Experimental results showed the effectiveness of the proposed method. The following conclusions were drawn:

(1) In the process of micro-distortion monitoring of radar scanning signal, this method had a high detection rate of distortion signal compared with the methods in [8,9].
(2) The energy consumption of the radar scanning signal detected by this method was the least compared with the methods in [8,9].
(3) This method could effectively correct the distorted signal using the frequency difference formula "f = −2v/Lamda" for the change of the Doppler principle compared with the methods in [8,9].

Author Contributions: Conceptualization, S.L. and X.C. (Xiaochun Cheng); methodology, Y.L.; software, X.C. (Xiang Chen); validation, X.C. (Xiang Chen) and Y.L.; formal analysis, S.L.; investigation, X.C. (Xiaochun Cheng); resources, X.C. (Xiang Chen); data curation, X.C. (Xiang Chen); writing—original draft preparation, X.C. (Xiang Chen) and S.L.; writing—review and editing, X.C. (Xiaochun Cheng); visualization, X.C. (Xiang Chen); supervision, S.L.; project administration, S.L.; funding acquisition, S.L.

Funding: This work is supported by the Open Project Program of the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System under Grant 2019K0104B. National Natural Science Foundation of China project under Grant 61502254, Program for Yong Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region under Grant NJYT-18-B10.

Acknowledgments: We want to give our sincere gratitude for the effective work of the editorial board of journal "Symmerty", as well as the guest editors of the special section "Recent Advances in Social Data and Artificial Intelligence 2019".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aubry, A.; Maio, A.D.; Pallotta, L.A. Geometric Approach to Covariance Matrix Estimation and its Applications to Radar Problems. *IEEE Trans. Signal Process.* **2018**, *66*, 907–922. [CrossRef]
2. Muqaibel, A.H.; Abdalla, A.T.; Alkhodary, M.T.; Alawsh, S.A. Through-the-wall radar imaging exploiting Pythagorean apertures with sparse reconstruction. *Digit. Signal Process.* **2017**, *61*, 86–96. [CrossRef]
3. Lin, Y.; Wang, C.; Wan, J.; Dou, Z. A Novel Dynamic Spectrum Access Framework Based on Reinforcement Learning for Cognitive Radio Sensor Networks. *Sensors* **2016**, *16*, 86–96.
4. Rakesh, P.R.; Narayanan, G. Investigation on Zero-Sequence Signal Injection for Improved Harmonic Performance in Split-Phase Induction Motor Drives. *IEEE Trans. Ind. Electron.* **2017**, *64*, 2732–2741. [CrossRef]
5. Lin, Y.; Zhu, X.; Zheng, Z.; Dou, Z.; Zhou, R. The individual identification method of wireless device based on dimensionality reduction and machine learning. *J. Supercomput.* **2017**, *75*, 3010–3027. [CrossRef]
6. Liu, S.; Bai, W.; Zeng, N.; Wang, S. A Fast Fractal Based Compression for MRI Images. *IEEE Access* **2019**, *7*, 62412–62420. [CrossRef]
7. Piazzo, L.; Raguso, M.C.; Seu, R.; Mastrogiuseppe, M. Signal enhancement for planetary radar sounders. *Electron. Lett.* **2019**, *55*, 153–155. [CrossRef]
8. Zheng, Y.; Zhang, C.; Li, L. Influences of optical-spectrum errors on excess relative intensity noise in a fiber-optic gyroscope. *Opt. Commun.* **2018**, *410*, 504–513. [CrossRef]
9. Lupi, S.M.; Galinetto, P.; Cislaghi, M.; y Baena, A.R.; Scribante, A.; y Baena, R.R. Geometric distortion of panoramic reconstruction in third molar tilting assessments: A comprehensive evaluation. *Dentomaxillofacial Radiol.* **2018**, *47*, 20170467. [CrossRef] [PubMed]
10. Wu, T.; Lu, S.; Tang, Y. Research on panoramic point cloud data acquisition technology based on ASODVS. *J. Comput. Meas. Control* **2014**, *22*, 2284–2287.

11. Hua, X.; Cheng, Y.; Wang, H.; Qin, Y.; Li, Y. Geometric means and medians with applications to target detection. *IET Signal Process.* **2017**, *11*, 711–720. [CrossRef]
12. Gui, R.; Wang, W.Q.; Cui, C.; So, H.C. Coherent Pulsed-FDA Radar Receiver Design with Time-Variance Consideration: SINR and CRB Analysis. *IEEE Trans. Signal Process* **2017**, *66*, 200–214. [CrossRef]
13. Le, Z.; Wang, X. Super-Resolution Delay-Doppler Estimation for OFDM Passive Radar. *IEEE Trans. Signal Process.* **2017**, *65*, 2197–2210.
14. Zhang, Y.; Pan, S. Broadband Microwave Signal Processing Enabled by Polarization-Based Photonic Microwave Phase Shifters. *IEEE J. Quantum Electron.* **2018**, *54*, 1–12. [CrossRef]
15. Engels, F.; Heidenreich, P.; Zoubir, A.M.; Jondral, F.K.; Wintermantel, M. Advances in Automotive Radar: A framework on computationally efficient high-resolution frequency estimation. *IEEE Signal Process. Mag.* **2017**, *34*, 36–46. [CrossRef]
16. Wanchun, L.; Qiu, T.; Chengfeng, H.; Yingxiang, L. Location algorithms for moving target in non-coherent distributed multiple-input multiple-output radar systems. *IET Signal Process.* **2017**, *11*, 503–514. [CrossRef]
17. Pan, Z.; Liu, S.; Sangaiah, A.K.; Muhammad, K. Visual attention feature (VAF): A novel strategy for visual tracking based on cloud platform in intelligent surveillance systems. *J. Parallel Distrib. Comput.* **2018**, *120*, 182–194. [CrossRef]
18. Cheng, Z.; Liao, B.; He, Z.; Li, Y.; Li, J. Spectrally Compatible Waveform Design for MIMO Radar in the Presence of Multiple Targets. *IEEE Trans. Signal Process.* **2018**, *66*, 3543–3555. [CrossRef]
19. Duan, K.; Wang, Z.; Xie, W.; Chen, H.; Wang, Y. Sparsity-based STAP algorithm with multiple measurement vectors via sparse Bayesian learning strategy for airborne radar. *IET Signal Process.* **2017**, *11*, 544–553. [CrossRef]
20. Zhang, W.; Fu, Y.; Nie, L.; Zhao, G.; Yang, W.; Yang, J. Parameter estimation of micro-motion targets for high-range-resolution radar using high-order difference sequence. *IET Signal Process.* **2018**, *12*, 1–11. [CrossRef]

symmetry

MDPI

Article

Algorithm for Detecting Communities in Complex Networks Based on Hadoop

Mo Hai, Haifeng Li *, Zhekun Ma and Xiaomei Gao

School of Information, Central University of Finance and Economics, Beijing 100081, China; haimo@cufe.edu.cn (M.H.); mzk@cufe.edu.cn (Z.M.); gxm@cufe.edu.cn (X.G.)

* Correspondence: mydlhf@cufe.edu.cn

Received: 9 October 2019; Accepted: 6 November 2019; Published: 7 November 2019

Abstract: With the explosive growth of the scale of complex networks, the existing community detection algorithms are unable to meet the needs of rapid analysis of the community structure in complex networks. A new algorithm for detecting communities in complex networks based on the Hadoop platform (called Community Detection on Hadoop (CDOH)) is proposed in this paper. Based on the basic idea of modularity increment, our algorithm implements parallel merging and accomplishes a fast and accurate detection of the community structure in complex networks. Our extensive experimental results on three real datasets of complex networks demonstrate that the CDOH algorithm can improve the efficiency of the current memory-based community detection algorithms significantly without affecting the accuracy of the community detection.

Keywords: community detection; complex networks; Hadoop; modularity increment

1. Introduction

In the era of Web 2.0, objects are connected to each other by various technologies such as the Internet and the Internet of Things, and form a variety of complex networks such as interpersonal interaction, essay reference, transportation, and protein interaction networks. Various complex networks are widely used in sociology, management, computer science, operations, biology, and other disciplines, while their wide application prospects have attracted the interest of many researchers. For example, Watts and Strogatz [1] applied the complex network theory in the field of biology and considered the nervous system to be a complex network of large numbers of nerve cells connected by nerve fibers. Faloutsos [2] applied the method of complex network analysis to study computer networks and evaluated their stability by analyzing their robustness. Sen et al. [3] mapped the transportation network to a complex network and implemented an optimal planning and configuration of the transportation network using dynamic analysis of the complex network. Xiao et al. [4] constructed a directed and weighted complex network based on the Beijing traffic network, analyzed the traffic network load-bearing pressure, and mined the corresponding regional centers, which provided a theoretical support for optimizing urban public transport network systems. Based on the characteristic analysis of the complex network itself, Ruguo [5] proposed a method for social coordination governance and provided ideas for solving mass public events based on a characteristic analysis of complex networks.

Many studies analyzed the inherent characteristics of complex networks and discovered the relationships between node attributes and connections within networks. To discover feature of complex networks, several community detection algorithms have been proposed. The so-called "community" is a sub-network composed of a group of nodes closely connected with their internal nodes and sparsely connected with other external community nodes. The community structure is a common feature of complex networks made up of one or more communities. The accurate identification of

the community structure in complex networks play an important theoretical role for public opinion monitoring, interest recommendation, identification of the network internal structure and other related research. As a result, many researchers have studied community detection algorithms from the aspects of modularity and edge structure. For examples, Newman and Girvan [6] proposed the concept of modularity and mined the complex network community structure, Yang et al. [7] introduced a method for analyzing the edge structure and node properties allowing to improve the accuracy of the detection of the complex network community structure. The accurate identification of the community structure in complex networks have broad applications, such as influence maximization, influences discovery within a community, interest recommendation, edge intelligence empowered recommendation [8], and so on.

However, the existing studies about complex network community detection algorithms focused on small-scale data sets and limited to the improvement of the community detection accuracy while neglecting its efficiency. At the same time, the number of nodes in complex networks demonstrates an explosive growth trend considering the advent of big data era, increasing number of network users, and exponential increase of the generated contents. At present, many social networking platforms such as WeChat, Weibo, Facebook, and Twitter, have more than 100 million on-line users and various interaction forms, including follow-ups, comments, and sharing. The large-scale complex network data sets generated by such platforms have the characteristics of node diversity, complex structure, multi-complexity fusion, which challenges the accuracy of the traditional complex network community detection algorithms. Furthermore, the traditional community detection algorithms are based on matrix iterations, which make the algorithms unable to adapt to the requirements of real-time and flexibility.

In this paper, we propose a new complex network community detection algorithm based on Hadoop framework (called Community Detection on Hadoop (CDOH)). Hadoop is a distributed system infrastructure developed by the Apache Foundation. Our contributions are as follows:

- Based on the idea of the maximum modularity, and combining the distributed characteristics of the Hadoop platform, a new modularity matrix update method is proposed and a corresponding community merging strategy is constructed to implement a fast and accurate detection and discovery of complex network community structures;
- We theoretically analyze our proposed CDOH algorithm, and show the computational cost of our algorithm can achieve $O(n)$ computational cost when we use enough parallel nodes;
- Experimental results on 3 real datasets demonstrate that CDOH significantly outperforms the traditional complex network community detection algorithm in terms of both the efficiency and accuracy of the community detection of complex networks.

The rest of our paper are organized as follows. Section 2 introduces the related works. Section 3 describes our proposed CDOH algorithm and analyzed its computational complexity. In Section 4, we show the experimental results with theoretical analysis. Section 5 concludes the paper and presents the future works.

2. Related Works

Since Newman [6,9] proposed the module optimality algorithm, the modularity-based community detection approach has been used in many network community mining algorithms such as the classic fast Newman community division algorithm [9] and CNM algorithm [10]. The fast Newman community detection algorithm is an agglomerative hierarchical clustering algorithm that starts with a state, in which each node is the sole member of n communities, and repeatedly joins communities together in pairs, choosing a joint at each step, which results in the greatest increase (or smallest decrease) in modularity. Recently, domestic researchers such as Lei et al. [11] implemented an edge community mining algorithm based on the local information of the considering network. Xiong [12] proposed a community discovery algorithm that combined the user closeness with clustering algorithms. Weiping [13] proposed the concept of new gravity of users for an accurate

community discovery; Leng [14] proposed a new network community detection algorithm based on a greedy optimization technology. Zhang et al. [15] further improved the fast Newman algorithm by introducing an improved index for the closeness centrality to classify overlapping nodes; the proposed method demonstrated a high classification accuracy in detecting overlapping communities with a time complexity of $O(n^2)$.

Blondel et al. [16] improved the modular incremental solution method by merging communities iteratively using a new calculation formula to achieve good results. Parsa et al. [17] used a probability vector model based on a single variable edge distribution algorithm, that combines an evolutionary algorithm with a community discovery method to enable the community detection; Oliverira et al. [18] used an improved Kuramoto coupled oscillator synchronization model to analyze networks from their dynamic factors and implemented a method for community discovery in complex networks. Ling Xing et al. [19] proposed a method that combines the sliding time-window method with the hierarchical encounter model based on association rules to increase the fidelity of the extracted networks by alleviating the homophily effect. Yuhui Gong et al. [20] focused on the customers' conformity behaviors in a symmetry market where customers are located in a social network. Simulation results have shown that topology structure, network size, and initial market share have significant effects on the evolution of customers conformity behaviors. Recently, Aceto et al. [21,22] and Ruoyu Wang et al. [23] applied deep learning and machine learning technologies in the research about social networking.

Recently, researchers have proposed complex network community detection algorithms based on big data platforms. Clauset [24] proposed a community-based parallel detection method based on the CNM algorithm. The basic idea of the algorithm proposed in [24] is to calculate the maximum community modularity in parallel and recognize the communities of large-scale networks by decreasing the communication overhead. The limitation of this algorithm is that it fails to run when the network scale increases and the amount of data rises to a certain level. Jinpeng [25] proposed a link community recognition algorithm based on the Hadoop platform. While this algorithm resolves the limitation of the linked community method that cannot store and process large matrices when analyzing big networks, its efficiency is still not efficient enough. Furthermore, its processing time reaches more than 5000 seconds when the scale of nodes reaches 15,000. Riedy et al. [26] used servers with multi-core processors to calculate the maximum community modularity in parallel to identify communities. However, the proposed method has strong hardware dependencies.

Moon et al. [27] proposed a parallel GN algorithm [6] based on Hadoop that can be divided into 4 stages. Each stage includes the map and reduce process. In the first stage, the tuples of all node pairs are generated; in the second and third stages, the edges with large edge betweenness values are identified and removed, respectively; in the fourth stage, the tuples are recalculated according to the new network. The experiment results demonstrated that the efficiency of the algorithm increases linearly with the increase of the number of reducers which are in charge of reduce process. Weijiang et al. [28] proposed a parallel Louvain algorithm that solved the main time-consuming problem of calculating the modularity and ergodic modularity increment in the Louvain algorithm [29]. This proposed algorithm outputs the information about all neighbors of a node in the map phase and decides the new home community of the node in the reduce phase accordingly. When computing a new community of a node, it is necessary to ensure that the neighbor's community is up-to-date, which is hard to be guaranteed in a distributed environment. Therefore, it is easy to face the problems of "community interchange" and "community ownership delay," which can be solved by resolving the associated connected graph. To solve the problem of high complexity of the fast Newman algorithm [10] in calculating the modularity of nodes. Bingzhou [30] proposed a parallel fast Newman algorithm based on Hadoop that calculates the modularity increment of each node merged with its neighbors in the map stage in parallel. In the reduce stage, the 2 nodes with the largest modularity increment are found and merged. The map and reduce processes are executed iteratively until all nodes are merged into 1 community. To deal with the problems of the fast-unfolding algorithm in processing large-scale networks. Bingzhou [30] also proposed a parallel fast-unfolding algorithm based on Hadoop and the

divide and conquer principle. First, a large-scale network is partitioned and merged separately, then the network is reconstructed according to the merging results of each partition, and finally the network is merged iteratively and reconstructed until the structure of community does not change any more. Conte et al. [31] proposed an algorithm which was able to find large k-plexes of very large graphs in just a few minutes and scale up to tens of machines with tens of cores each. Vincenzo el al. [32] proposed a novel algorithm for community detection in social networks based on game theory, and showed this algorithm outperformed other algorithms in terms of computational complexity and effectiveness. However, this algorithm cannot scale to a huge number of nodes and edges.

The traditional community detection algorithms focused on small-scale data sets and hard to scale to a large scale data sets. While parallel community detection algorithms are more scalable, they cannot achieve a good trade-off between the efficiency and accuracy. In order to overcome the shortcomings of traditional community detection algorithms and parallel community detection algorithms, we propose a new complex network community detection algorithm based on Hadoop, which effectively implements a fast and accurate detection of complex network community structure. Compared with traditional community detection algorithms, it can scale to a large scale data set. Compared with parallel community detection algorithms, it achieves a good trade-off between efficiency and accuracy.

3. Complex Network Community Detecting Algorithm Based on Hadoop

The proposed CDOH algorithm is based on the idea of the maximal modularity increment, which employs a new modularity matrix updating method and a community merging strategy.

3.1. Definitions

This section provides formal definition of the basic concepts involved in the proposed complex network community detection algorithm. The symbols and their meanings are shown in Table 1.

Table 1. Symbols and Definitions.

Symbols	Meanings
N	A complex network
V	a set of nodes
v_i	node i
E	a set of edges
e_{ij}	Denotes the connection between node v_i and node v_j, if they are connected, e_{ij} is 1; Otherwise e_{ij} is 0.
d_i	the node degree of node v_i
M	the modularity of a network
C	the set of detected network communities
c_i	a community i
l_c	the total number of edges interconnected between nodes within the community c
m	the total number of edges in the network
D_c	the sum of the node degrees of all nodes in the community c
a_c	The ratio of the sum of degrees of all nodes in the community c to the sum of degrees of all nodes in N
$\triangle M$	the modularity increment
R_{ij}	the number of connection edges between communities c_i and c_j

Definition 1. *(Complex network) A complex network is a network consisting of a series of nodes and their interconnected edges denoted as* $N = (V, E)$*. Here,* $V = \{v_i \mid i = 1, 2, \cdots, n\}$ *represents a set of nodes in a complex network, and* $E = \{e_{ij} \mid v_i, v_j \in V\}$ *represents a set of edges in a complex network, where* e_{ij} *denotes the connection between nodes* v_i *and* v_j*. If they are connected, then* $e_{ij} = 1$*; otherwise,* $e_{ij} = 0$*.*

Definition 2. *(Node degree) In a complex network $N = (V, E)$, the node degree d_i of each node v_i is defined as the number of edges connected to node v_i, which is defined by Equation (1),*

$$d_i = \sum_{v_j \in V, i \neq j} e_{ij} \tag{1}$$

Figure 1 illustrates a simple network community structure. According to Definitions 1 and 2, there are 12 nodes in the network (from v_1 to v_2), where $e_{12} = 1$, $e_{19} = 0$, and v_1 has a node degree $d_1 = 4$.

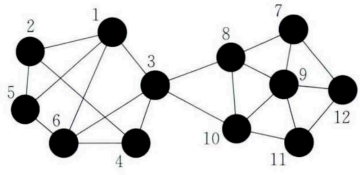


Figure 1. A Simple Network Community Structure.

Definition 3. *(Modularity) The modularity of a network M is defined by Equation (2).*

$$M = \sum_{c \in C} \left(\frac{l_c}{m} - a_c^2 \right) \tag{2}$$

Here, $C = \{c_i \mid i = 1, 2, \cdots, k\}$ denotes the detected set of network community structures, l_c denotes the total number of edges interconnected between nodes within the community c, m denotes the total number of edges in the network, and

$$a_c = \frac{D_c}{2m} \tag{3}$$

where D_c denotes the sum of the node degrees of all nodes in the community c, and D_c equals to 2 times of the sum of l_c and the total edges of connecting the community c and other external communities.

According to Equation (2), the modularity of complex networks measures the degree of closeness within the community and the degree of sparseness between the communities. The closer the internal connection of the community is and the thinner the connection between the communities is, the greater the modularity M is, and vice versa. Thus, when the modularity M of a complex network is the largest, the community detection results are optimal. However, it is quite difficult to determine directly whether M has reached its maximum. Therefore, the concept of the modularity increment $\triangle M$ proposed by Newman is adopted, where the increase or decrease in the modularity M caused by merging communities c_i and c_j, which is defined as Equation (4).

$$\triangle M = \frac{2R_{ij}}{m} - 2 \times a_i \times a_j \tag{4}$$

Here, R_{ij} denotes the number of connection edges between communities c_i and c_j in which $i \neq j$. Then the modularity M increases progressively when $\triangle M > 0$. On the contrary, if $\triangle M < 0$, the modularity M is the maximum and the process of the community detection ends.

When the number of nodes and edges in a complex network are kept the same, and different communities are merged to form a new community, the number of edges among nodes within the new community is the sum of the number of edges within the 2 merged communities and the number of edges between the 2 merged communities. Accordingly, [14] points out that when the number of nodes and edges are kept the same, the increase of the modularity between the new communities formed by merging multiple known communities and other communities can be established as Equation (5).

$$\triangle M[c_z][c_k] = \begin{cases} \triangle M[c_z][c_k] + \triangle M[c_i][c_k], < c_i, c_k > \in E, c_i \in c_z \\ \triangle M[c_z][c_k] - 2 \times a_i \times a_k, < c_i, c_k > \notin E, c_i \in c_z \end{cases} \quad (5)$$

Here, c_z denotes the new community after merging, c_k denotes the old community that does not belong to c_z, c_i denotes the old community merged to c_z, and $< c_i, c_k >$ denotes the edge set from community c_i to community c_k.

Taking the network structure in Figure 1 as an example, we can see that each node represents a community. Equation (4) can be used to calculate the modularity increment $\triangle M$ among any 2 communities and form a matrix as shown in Table 2, where the first row and column represent the community number. We focus only on the 2 same communities need to be merged, and the changes within the community need not be considered, so the diagonal of the matrix can be initialized to 0. From the values of the matrix, we can observe that communities that can be merged in this example are c_2 and c_4, c_2 and c_5, c_7 and c_{12}, c_{11} and c_{12}, where $\triangle M$ is the maximal value, that is, 0.036. Taking the community c_{13} formed by merging c_2 and c_4 as an example, the results after merging are listed in Table 3.

As can be noticed from Tables 2 and 3, the modularity increment between the community c_{13} and other communities is the sum of the modularity increment between the communities c_2, c_4, and the corresponding communities. For example, in Table 3, the modularity increment of the communities c_1 and c_{13} is 0.021, which is the sum of the modularity increment, 0.033, of c_1 and c_2, and the modularity increment, -0.012, of c_1 and c_4, as shown in Table 2.

Considering that the modularity matrix update algorithm has the characteristics of merging communities in parallel and conforms to the characteristics of parallel processing on the Hadoop platform, we select the modularity incremental update method represented by Equation (5) to construct the proposed CDOH algorithm. According to the modularity increment represented by Equation (4), we initialize the entire network, treat each node as a community, and calculate the modularity increment when merging any 2 communities. Then, we iterate consecutively to find new communities. Based on the MapReduce parallel programming model, all the 2 communities with the maximum modularity increment are identified and merged in parallel. Equation (5) is used to update the modularity increment when merging any 2 communities in parallel. The community discovery process ends when the maximum modularity increment is negative. Finally, the CDOH algorithm stores the node set V as (vId, cId), where vId denotes the node number and cId denotes the community number, and the edge set E is represented as $(s, d, \triangle M)$, where s denotes the source node of the edge, d denotes the destination node of the edge, and $\triangle M$ is the modularity increment corresponding to this edge.

Table 2. $\triangle M$ Matrix before Network Merging.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0.000	0.033	0.025	−0.012	0.033	0.029	−0.012	−0.017	−0.021	−0.017	−0.012	−0.012
2	0.033	0.000	−0.015	0.036	0.036	−0.012	−0.009	−0.012	−0.015	−0.012	−0.009	−0.009
3	0.025	−0.015	0.000	0.030	−0.015	0.025	−0.015	0.025	−0.026	0.025	−0.015	−0.015
4	−0.012	0.036	0.030	0.000	−0.009	0.033	−0.009	−0.012	−0.015	−0.012	−0.009	−0.009
5	0.033	0.036	−0.015	−0.009	0.000	0.033	−0.009	−0.012	−0.015	−0.012	−0.009	−0.009
6	0.029	−0.012	0.025	0.033	0.033	0.000	−0.012	−0.017	−0.021	−0.017	−0.012	−0.012
7	−0.012	−0.009	−0.015	−0.009	−0.009	−0.012	0.000	0.033	0.030	−0.012	−0.009	0.036
8	−0.017	−0.012	0.025	−0.012	−0.012	−0.017	0.033	0.000	0.025	0.029	−0.012	−0.012
9	−0.021	−0.015	−0.026	−0.015	−0.015	−0.021	0.030	0.025	0.000	0.025	0.030	0.030
10	−0.017	−0.012	0.025	−0.012	−0.012	−0.017	−0.012	0.029	0.025	0.000	0.033	−0.012
11	−0.012	−0.009	−0.015	−0.009	−0.009	−0.012	−0.009	−0.012	0.030	0.033	0.000	0.036
12	−0.012	−0.009	−0.015	−0.009	−0.009	−0.012	0.036	−0.012	0.030	−0.012	0.036	0.000

Table 3. $\triangle M$ Matrix after Merging c_2 and c_4.

	1	**3**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	0	0.025	0.033	0.029	−0.012	−0.017	−0.021	−0.017	−0.012	−0.012	0.021
3	0.025	0	−0.015	0.025	−0.015	0.025	−0.026	0.025	−0.015	−0.015	0.015
5	0.033	−0.015	0	0.033	−0.009	−0.012	−0.015	−0.012	−0.009	−0.009	0.027
6	0.029	0.025	0.033	0	−0.012	−0.017	−0.021	−0.017	−0.012	−0.012	0.021
7	−0.012	−0.015	−0.009	−0.012	0	0.033	0.03	−0.012	−0.009	0.036	−0.019
8	−0.017	0.025	−0.012	−0.017	0.033	0	0.025	0.029	−0.012	−0.012	−0.025
9	−0.021	−0.026	−0.015	−0.021	0.03	0.025	0	0.025	0.03	0.03	−0.031
10	−0.017	0.025	−0.012	−0.017	−0.012	0.029	0.025	0	0.033	−0.012	−0.025
11	−0.012	−0.015	−0.009	−0.012	−0.009	−0.012	0.03	0.033	0	0.036	−0.019
12	−0.012	−0.015	−0.009	−0.012	0.036	−0.012	0.03	−0.012	0.036	0	−0.019
13	0.021	0.015	0.027	0.021	−0.019	−0.025	−0.031	−0.025	−0.019	−0.019	0

3.2. The CDOH Algorithm

Based on the research framework of complex network community detection algorithm on the Hadoop platform shown in the Section 3.1. The CDOH has 4 steps, that is, first, we will initialize the parameters; second, we will find the maximum modularity increment; third, we will merge the communities and update the modularity increment; finally, we will generate the final community discovery results. Step 2 and step 3 will be repeated to find new communities until the maximum modularity increment is negative. We shown the flow charts of CDOH algorithm in Figure 2. Here, step 1 (Parameter initialization), step 2 (Finding the maximum modularity increment), and step 3 (Merging communities and updating the modularity increment) are implemented based on MapReduce parallel programming model of Hadoop.

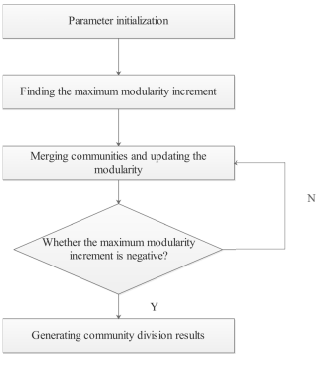


Figure 2. Flow charts of Community Detection on Hadoop (CDOH) Algorithm.

3.2.1. Parameter Initialization

The initialization phase is responsible for calculating the necessary parameters of the algorithm, which includes the total number of nodes n, total number of edges m, degree d of each node, vector a, and the modularity increment $\triangle M$ between each pair of nodes. The process is listed in Algorithm 1, the main steps of which include the following:

- First, we load the complex network data from the input file, then calculate the number of nodes n and edges m of the complex network, and broadcast the number of edges (m) to all nodes;
- Second, we calculate the degree d of each node and the vector a according to Equation (3);

- Finally, we use Equation (4) to calculate the modularity increment $\triangle M$ between each pair of nodes, and construct a new network N using this modularity increment.

Algorithm 1 Initialization of CDOH Parameters

Input:

D: Preprocessed network data;

Output:

$\triangle M$: Modularity increment;

N: Network;

```
N = networkLoad(D);
n = getVertices(N);
m = getEdges(N);
Broadcast the number of edges m to all nodes in the cluster;
for each Node i in N do
    k_i = getDegree(i);
    a_i = k_i / 2m;
for each Edge e in N do
    △M_ij = R_ij / m − 2 × a_i × a_j;
```

Here, we firstly divide the $n \times n$ matrix into multiple sub matrix, then we deploy multiple mappers, and let each mapper calculate the vector a of each node and calculate the modularity increment between each pair of nodes of each sub matrix. Each mapper works in parallel.

3.2.2. Find the Maximum Modularity Increment

After completing the modularity increment calculation, we initiate the iterative community discovery, and find multiple community pairs with the largest modularity increment, and merge them into the corresponding new communities. Taking the network shown in Figure 1 as an example. According to the $\triangle M$ matrix shown in Table 2, communities c_2 and c_4, c_2 and c_5, c_7 and c_{12}, c_{11} and c_{12} can be merged. Clearly, communities c_2, c_4, c_5 and c_7, c_{11}, c_{12} should be merged to the community c_{13} and community c_{14}, respectively.

Algorithm 2 describes the steps involved in finding the modularity increment, which has 4 steps.

- First, we compare the $\triangle M$ value of each edge e in network N, find the maximum modularity increment $max(\triangle M)$, and broadcast it to all nodes in the cluster;
- Second, we get the cartesian product T of the edge set E and node set V, $T = (s, sc, d, dc, \triangle M)$, s denotes the number of the source node, d denotes the number of destination node, sc and dc denote the community numbers of the source node and destination node respectively, and $\triangle M$ denotes the modularity increment between the source node and destination node;
- Third, we find the sub-set MC in the set T, where $\triangle M$ equals to $max(\triangle M)$;
- Finally, to organize the merged communities, we obtain the community number (i) of the source node and the community number (j) of the destination node, which represent the current communities to be merged. If i or j already belongs to a new community in C, we will get the new community to merge i and j into it, or merge i and j into another new community, whose number is $n + 1$. The final output is the community C after merging.

Algorithm 2 Find the Maximum Modularity Increment and Communities that need to be Merged

Input:

$\triangle M$: Modularity increment;

$N(E, V)$: Network;

Output:

$C = \{c_1, c_2, \cdots, c_l\}$: Communities;

$max(\triangle M)$: Maximum Modularity increment;

1: $max(\triangle M) = searchMaxDeltaM(N)$;
2: Broadcasting $\triangle M$ to all nodes in the cluster;
3: $T = E \times V$;
4: **for** each quintuple t in T **do**
5: **if** $getDeltaM(t) == max(\triangle M)$ **then**
6: $MC = insert(t)$;
7: **for** each quintuple t in MC **do**
8: $(i, j) = getCommuNum(t)$;
9: **if** $i \in C$ or $j \in C$ **then**
10: k = Get the new number of community i or j from C;
11: c_k = insert(i,j);
12: **else**
13: n = n+1;
14: c_n = insert(i, j);

Here, we find the maximum modularity increment $max(\triangle M)$ based on the MapReduce. After dividing the $n \times n$ matrix into multiple sub matrix, in the map phrase, each mapper finds the maximum modularity increment of each sub matrix and output the results to the reducer, and then in the reduce phrase, the reduce output the maximum modularity increment $max(\triangle M)$. Afterwards, we find the community pairs with the largest modularity increment based on MapReduce. Each mapper finds the community pairs with the largest modularity increment of each sub matrix in parallel.

3.2.3. Merging and Updating Communities

Merging and updating communities are the core of the proposed algorithm. Since after step 2, the community pairs with the maximum modularity increment are identified to be merged, the mapper updated the number of the communities that need to be merged and the community number of the corresponding nodes to their corresponding new community number in parallel, and the $\triangle M$ of any 2 communities are updated by the mapper in parallel.

The steps of merging and updating of communities listed in Algorithm 3 are the following.

- First, we obtain the Cartesian product T of the node set V and edge set E. Then, we look for the new community number corresponding to sc and dc in $t = (s, sc, d, dc, \triangle M)$. Let X to be the set of community numbers to be merged in this round contained by the new community of the community $t.sc$ and Y to be the set of community numbers to be merged in this round contained by the new community of the community $t.dc$;
- Second, using Equation (5), we will merge and update community i in X and community j in Y. If there is an edge connecting communities i and j, then the modularity increment between new communities X and Y should include the modularity increment between communities i and j. However, if there is no edge connecting communities i and j, the modularity increment between new communities X and Y should be reduced by the doubled product of vector value a_i of community i and vector value a_j of community j.

Algorithm 3 Merging and Updating Communities

Input:

$C = \{c_1, c_2, \cdots, c_l\}$: Communities;

N(E,V): Network;

Output:

$N(E, V)$: Updated Network;

Update the number of the communities that need to be merged and the community number of the corresponding nodes to their corresponding new community number;

$T = V \times E$;

for each quintuple t in T **do**

  $tsc = getNewCommuNum(t.sc)$;

  $tdc = getNewCommuNum(t.dc)$;

  if ($tsc \in$ C or $tdc \in C$) and $tsc \neq tdc$ **then**

    X = a set of community numbers to be merged in this round contained by the new community corresponding to $t.sc$;

    Y = a set of community numbers to be merged in this round contained by the new community corresponding to $t.dc$;

    for each community i in X and each community j in Y **do**

      if there exists at least an edge connecting i and j **then**

        $\triangle M_{XY} = \triangle M_{XY} + \triangle M_{ij}$

      else

        $\triangle M_{XY} = \triangle M_{XY} - 2 \times a_i \times a_j$

3.2.4. Generating Community Discovery Results

After the community discovery finishes, redundant data in the data set (primarily the matrix data) should be cleared, while the initial node set and their community number should be kept. Here, the node storage structure in the network is considered to be $V = (vId, cId)$, where vId denotes the node number and cId denotes the community number indicating which community each node belongs to. Algorithm 4 presents the process of generating the results of the community partitions, which has 2 steps:

- We will first traverse all nodes and keep the nodes with the same community number cId together. If cId is already in C, it means that the corresponding community of cId has already appeared. The node Ids in the community cId that have been stored in C need to be taken out, merged with the current node Id, and then stored in C; otherwise they are stored in C directly;
- Then we store the community and community's node set on the Hadoop distributed file system (HDFS) one by one. Thus, CDOH stores the final results of community discovery with a set of the tuple $(cId, vIds)$, and finishes the detection and discovery of complex network communities on Hadoop platform.

Algorithm 4 Generating Community Discovery Results

Input:

$N(E, V)$: Network;

Output:

$C = \{c_1, c_2, \cdots, c_l\}$: Communities;

1: **for** each $v = (vId, cId)$ in N **do**
2: **if** $cId \in C$ **then**
3: g = getNodeId(C, cId);
4: c = insert(g, vId);
5: C = insert(cId, c);
6: **else**
7: C = add(cId, vId);
8: **for** each community c in C **do**
9: output c;

3.3. Computational Complexity Analysis of the CDOH Algorithm

As presented before, in step 1, we let multiple mappers take charge of the initializing process of $n \times n$ sub-matrix. Supposed the matrix is divided into m matrices, and let each mapper takes charge of each sub-matrix in parallel, so the computational complexity of the initializing process of the matrix is the computational complexity of the initializing process of the sub-matrices, that is $O(\frac{n^2}{m})$. In step 2, the maximum modularity increment $max(\triangle M)$ and the community pairs with the largest modularity increment is found based on MapReduce. Again, if we divide the matrix into m sub-matrices, and let each mapper takes charge of each sub-matrix in parallel, the computational complexity of step 2 is also $O(\frac{n^2}{m})$. In step 3, the mapper updated the number of the communities that need to be merged and the community number of the corresponding nodes to their corresponding new community number in parallel, whose computational complexity is $O(1)$. After merging, the $\triangle M$ values of any 2 communities are updated by the mapper in parallel. Supposing that each mapper works on a sub-matrix, the computational complexity of updating $\triangle M$ is $O(\frac{n^2}{m})$. In step 4, all nodes are traversed and the nodes with the same community number are kept together, whose computational complexity is $O(n)$. Since step 2 and step 3 are repeated until the the maximum modularity increment $max(\triangle M)$ becomes negative, and after some iterations, the $n \times n$ matrix will shrink to a constant computing cost. As a result, our algorithm can achieve a performance that is in reverse proportion to the number of sub-matrices, which is determined by the number of nodes in the Hadoop platforms. Supposing we have n nodes to conduct the parallelly computing, we can achieve a $O(n)$ computing cost.

4. Experimental Results

4.1. Datasets and Evaluation Algorithms

To evaluate the accuracy and running time of CDOH, 3 real complex network data sets obtained from the Stanford Network Analysis Project (SNAP) were selected. The data sets contain the nodes and connection status of real complex networks and mark the communities to which the nodes belong. Table 4 gives the characteristics of the data sets used in the experiments.

Table 4. Characteristics of Datasets.

Dataset	No. of Nodes	No. of Edges	Node Average Degree	Description
Soc-Epinions	75,879	508,837	13.4118	Epinions.com Date Set
Web-NotreDame	325,729	1,497,134	9.1925	Web Graph Data Set
Soc-Pokec	1,632,803	30,622,564	37.5092	Poke Social Data Set

To evaluate our algorithm, we use 2 state-of-the-art algorithms in our experiments, that is, the traditional complex network community detection algorithm Fast Community Detection (FCD) proposed by Newman [9] and the non-overlapping community detection algorithm Non-Overlapping Community Detection Idea (OCDI) proposed by Zhang et al. [15].

All the algorithms were implemented with Java, and our algorithm was deployed on Hadoop cluster made of 3 different computers, of which 1 serving as a master node and the other 2 serving as slave nodes. The following experimental results are represented as average across the 10 runnings.

4.2. Analysis of Community Detection Accuracy

We used the community detection accuracy (CDA) metric to measured the accuracy of community detection. CDA is defined as the ratio of the number of nodes in the correctly identified communities to the total number of nodes in the network, which is shown in Equation (6).

$$CDA = \frac{\sum_{i=1}^{k} max\{|C_i \cap C'_j| \mid C'_j \subset C_i\}}{n}, j = 1, 2, \cdots, l \tag{6}$$

Here, $C = \{c_1, c_2, \cdots, c_k\}$ denotes the original and accurate community set, $C' = \{c'_1, c'_2, \cdots, c'_l\}$ denotes the community set identified by the community detection algorithm, $max\{|C_i \cap C'_j| \mid C'_j \subset C_i\}$ denotes the maximum number of the common nodes between all community sets and the i-th accurate community c_i, and n denotes the number of nodes. As can be seen, the larger the value is, the higher the accuracy of a community detection algorithm is and the better the quality of the resulting community is. Figure 3 shows the community discovery accuracies of the considered algorithms on the 3 different data sets.

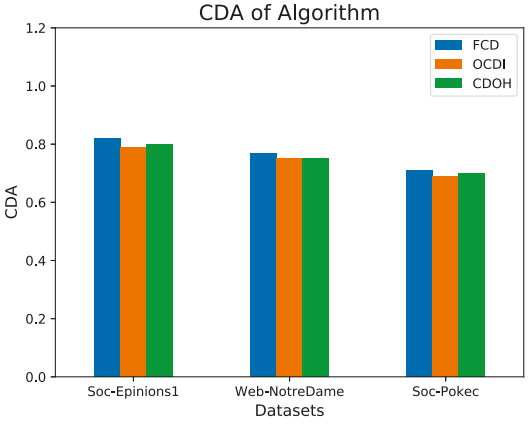


Figure 3. Comparison of the Accuracy of the Community Detection Algorithms.

It can be noticed from Figure 3 that the accuracy of the CDOH algorithm is slightly lower than that of the FCD algorithm (on average by 1.7%) and similar to that of OCDI. The reason for this is

that CDOH and OCDI have similar community merging strategies and module update principles. While multiple communities are merged at one time in the same iteration according to CDOH and OCDI, FCD only supports one-time merging of 2 communities in a single iteration, which results in the accuracy gap between FCD and the other 2 algorithms.

We also used the normalized mutual information (NMI) to evaluate our algorithm in comparison to the other 2 algorithms. NMI [33] is a standard factor which is often used to detect the difference between the results of the division and the true partition of the network. NMI can be described in Equation (7), in which $H(X)$ is the entropy of X, and $H(X|Y) = H(X,Y) - H(Y)$.

$$NMI(X,Y) = \frac{H(X) - H(X|Y) + H(Y) - H(Y|X)}{2max(H(X),H(Y))} \tag{7}$$

We can see from Figure 4 that the NMI of the 3 algorithms can reach at least 75%. Our algorithm, CDOH, has a very similar NMI score to the FCD algorithm and has a slightly higher score than OCDI. Again, we consider this is due the fact that our algorithm has similar community merging strategies and module update principles.

However, the computing cost of our algorithm is much better than that of FCD, which will be discussed in Section 4.3.

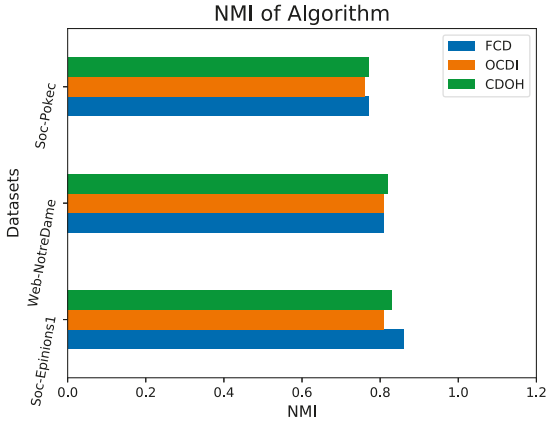


Figure 4. Comparison of the normalized mutual information (NMI) of the Community Detection Algorithms.

4.3. Analysis of Community Detection Efficiency

CDOH is a community detection algorithm based on Hadoop platform for large-scale complex networks. For processing large scale data, the run time of the algorithm is an important metric to evaluate its performance of the algorithm. Figure 5 shows the comparison of the run time of the 3 considered algorithms.

It can be noticed from Figure 5 that CDOH is highly efficient. To compared with OCDI and FCD, we can see that CDOH is about 2.1 times and 3.2 times faster, respectively, which is mainly determined by the number of slave nodes on the Hadoop platform. Compared with the traditional community detection algorithms, CDOH uses significantly less time required for community merging and modularity updating.

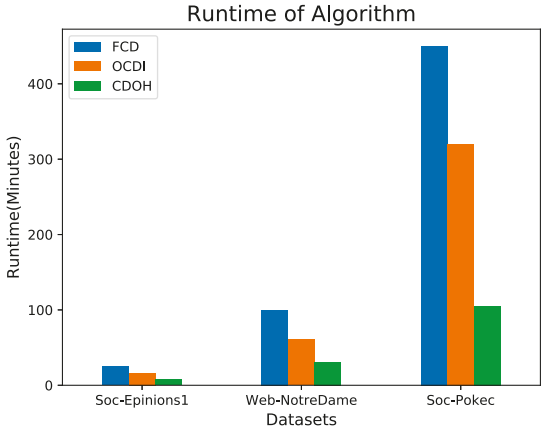


Figure 5. Comparison of the Runtime of Community Detection Algorithms.

5. Conclusions and Future Works

5.1. Conclusions

In this paper, we proposed a community detection algorithm called CDOH based on the Hadoop platform to implement accurate and fast community identification in large-scale complex networks. The algorithm was based on the modularity increment calculation method, which employed the theory of complex networks to find multiple communities satisfying certain merging conditions. The parallel merging and modularity updating of communities based on MapReduce used in the proposed algorithm reduce the number of iterations. CDOH was compared with traditional complex network algorithms using real large-scale complex networks. The experimental results evaluated the effectiveness of CDOH in large-scale network community detection.

5.2. Future Works

Our proposed CDOH algorithm is independent of the underlying big data platform. To prove its effectiveness and efficiency, we implemented the CDOH algorithm and other complex network community detection algorithms based on the Hadoop platform. However, in the Hadoop platform, the MapReduce intermediate results are first stored in disk files, and a large number of I/O operations will affect the whole calculation time; while in the Spark platform, the intermediate results are stored in memory, which avoids the performance overhead brought by I/O. In the future, we will implement the CDOH algorithm on the Spark platform and evaluate the efficiency. Furthermore, our proposed CDOH algorithm focused on static complex network community discovery, in the future, we plan to adapt the proposed algorithm to the evolving community networks.

Author Contributions: Conceptualization, M.H.; methodology, H.L.; software, Z.M.; validation, M.H. and X.G.; formal analysis, M.H.; investigation, H.L.; resources, H.L.; data curation, H.L.; writing–original draft preparation, M.H.; writing–review and editing, M.H., H.L. and Z.M.; visualization, H.L.

Funding: This research is supported by the National Natural Science Foundation of China (61100112,61309030), Beijing Higher Education Young Elite Teacher Project (YETP0987), the Top Discipline Construction Project of Central University of Finance and Economics in 2019 (Key Technologies and Application of Independent Controllable Block Chain), the Fundamental Research Funds for the Central Universities, the Education and Teaching Reform Fund of Central University of Finance and Economics in 2018(2018GRZDJG06).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440–442. [CrossRef] [PubMed]
2. Faloutsos, M.; Faloutsos, P.; Faloutsos, C. On power-law relationships of the Internet topology. *ACM SIGCOMM Comput. Commun. Rev.* **1999**, *29*, 251–262. [CrossRef]
3. Sen, P.; Manna, S.S. Clustering properties of a generalized critical Euclidean network. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2003**, *68*, 026104. [CrossRef] [PubMed]
4. Zheng, X.; Chen, J.; Shao, J.; Bie, L. Topological properties analysis of Beijing public transport network based on complex network theory. *J. Phys.* **2012**, *61*, 95–105.
5. Fan, R. Cooperative Innovation of Social Governance under the Paradigm of Complex Network Structure. *Soc. Sci. China* **2014**, *4*, 98–120.
6. Newman, M.E.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2003**, *69*, 17–32. [CrossRef]
7. Yang, J.; Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **2015**, *42*, 181–213. [CrossRef]
8. Xin, S.; Giancarlo, S.; Vincenzo, M.; Antonio, P.; Christian, E.; Chang, C. An Edge Intelligence Empowered Recommender System Enabling Cultural Heritage Applications. *IEEE Trans. Ind. Inf.* **2019**, *15*, 4266–4275.
9. Newman, M.E. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **2003**, *69*, 066133. [CrossRef]
10. Clauset, A.; Newman, M.E.; Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **2004**, *70*, 066111. [CrossRef]
11. Pan, L.; Jin, J.; Wang, C.; Xie, J. Edge Community Mining Based on Local Information in Social Networks. *J. Electron.* **2012**, *40*, 2255–2263.
12. Xiong, Z. *Community Discovery Technology and Its Application in Online Social Networks*; Central South University: Changsha, China, 2012.
13. Huang, W. *Research on Web Community Discovery Algorithms*; Beijing University of Posts and Telecommunications: Beijing, China, 2013.
14. Leng, Z. Research on network community discovery algorithm based on greedy optimization technology. *J. Electron.* **2014**, *42*, 723–729.
15. Zhang, X.; You, H.; Zhu, W.; Quiao, S.; Li, J.; Gutierrez, L.A.; Zhang, Z.; Fan, X. Overlapping community identification approach in online social networks. *Physica A Stat. Mech. Appl.* **2015**, *421*, 233–248. [CrossRef]
16. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of community hierarchies in large networks. *Comput. Res. Repos.* **2008**, abs/0803.0476.
17. Parsa, M.G.; Mozayani, N.; Esmaeili, A. An EDA-based community detection in complex networks. In Proceedings of the International Symposium on Telecommunications, Tehran, Iran, 9–11 September 2014; pp. 476–480.
18. Oliveira, J.E.M.D.; Quiles, M.G. Community Detection in Complex Networks Using Coupled Kuramoto Oscillators. In Proceedings of the International Conference on Computational Science and ITS Applications, Guimaraes, Portugal, 30 June–3 July 2014; pp. 85–90.
19. Jing-Ya, X.; Tao, L.; Lin-Tao, Y.; Davison, M. Finding College Student Social Networks by Mining the Records of Student ID Transactions. *Symmetry* **2019**, *11*, 307.
20. Yuhui, G.; Qian, Y. Evolution of Conformity Dynamics in Complex Social Networks. *Symmetry* **2019**, *11*, 299.
21. Giuseppe, A.; Domenico, C.; Antonio, M.; Antonio, P. Mobile Encrypted Traffic classification Using Deep Learning. In Proceedings of the 2018 Network Traffic Measurement and Analysis Conference (TMA), Vienna, Austria, 26–29 June 2018.
22. Giuseppe, A.; Domenico, C.; Antonio, M.; Pescapé, A. Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 445–458.
23. Ruoyu, W.; Zhen, L.; Yongming, C.; Deyu T.; Jin Y.; Zhao Y. Benchmark Data for Mobile App Traffic Research. In Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, New York, NY, USA, 5–7 November 2018.

24. Clauset, A. Finding local community structure in networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2005**, *72*, 026132. [CrossRef]
25. Li, J. *Research on Overlapping Community Discovery Algorithm Based on Hadoop Platform*; Jilin University: Changchun, China, 2014.
26. Riedy, J.; Bader, D.A.; Meyerhenke, H. Scalable Multi-threaded Community Detection in Social Networks. In Proceedings of the IEEE International Parallel and Distributed Processing Symposium Workshops & Phd Forum, Shanghai, China, 21–25 May 2012; pp. 1619–1628.
27. Moon, S.; Lee, J.G.; Kang, M. Scalable community detection from networks by computing edge betweenness on MapReduce. In Proceedings of the 2014 International Conference on Big Data and Smart Computing (BIGCOMP), Bangkok, Thailand, 15–17 January 2014; pp. 145–148.
28. Wu, W.; Li, M.; Li, G. A Parallelization of Louvain algorithm. *Comput. Digit. Eng.* **2016**, *44*, 1402–1406.
29. Blondel, V.D.; Guillaume, J.L.; Lambiotee, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *10*, P10008. [CrossRef]
30. Lai, B. *Research on Parallelization of Community Discovery Algorithm Based on Hadoop*; Jiangxi University of Science and Technology: Ganzhou, China, 2017.
31. Alessio, C.;Tiziano, D.M.; Daniele, D.S.; Grossi, R.; Marion, A.; Versari, L. *D2k: Scalable Community Detection in Massive Networks via Small-Diameter k-Plexes*; KDD 2018; ACM: New York, NY, USA, 2018; pp. 1272–1281.
32. Vincenzo, M.; Antonio, P.; Giancarlo, S. Community detection based on Game Theory. *Eng. Appl. Artif. Intell.* **2019**, *85*, 773–782.
33. Mcdaid, A.F.; Greene, D.; Hurley, N. Normalized Mutual Information to evaluate overlapping community finding algorithms. *CoRR* **2011**, abs/1110.2515.

symmetry

MDPI

Article

Centrality Metrics' Performance Comparisons on Stock Market Datasets

Jie Hua [1,*], Maolin Huang [2] and Chengshun Huang [1]

[1] Faculty of Engineering and Information Technology, Shaoyang University, Shaoyang 422000, China
[2] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
* Correspondence: Jie.Hua@alumni.uts.edu.au

Received: 13 June 2019; Accepted: 11 July 2019; Published: 15 July 2019

Abstract: The stock market is an essential sub-sector in the financial area. Both understanding and evaluating the mountains of collected stock data has become a challenge in relevant fields. Data visualisation techniques can offer a practical and engaging method to show the processed data in a meaningful way, with centrality measurements representing the significant variables in a network, through exploring the aspects of the exact definition of the metric. Here, in this study, we conducted an approach that combines data processing, graph visualisation and social network analysis methods, to develop deeper insights of complex stock data, with the ultimate aim of drawing the correct conclusions with the finalised graph models. We addressed the performance of centrality metrics methods such as betweenness, closeness, eigenvector, PageRank and weighted degree measurements, drawing comparisons between the experiments' results and the actual top 300 shares in the Australian Stock Market. The outcomes showed consistent results. Although, in our experiments, the results of the top 300 stocks from those five centrality measurements' rankings did not match the top 300 shares given by the ASX (Australian Securities Exchange) entirely, in which the weighted degree and PageRank metrics performed better than other three measurements such as betweenness, closeness and eigenvector. Potential reasons may include that we did not take into account the factor of stock's market capitalisation in the methodology. This study only considers the stock price's changing rates among every two shares and provides a relevant static pattern at this stage. Further research will include looking at cycles and symmetry in the stock market over chosen trading days, and these may assist stakeholder in grasping deep insights of those stocks.

Keywords: centrality metric; graph visualisation; visual analytics; data processing; social network; stock market

1. Introduction

Today, the importance of innovation in the business sector in achieving competitive success has become increasingly clear to most significant corporations [1], and massive financial data in relevant areas can provide the crucial information for sound business analysis, but the finance literature reveals little interest in investor decision processes or the quality of judgment [2]. In the financial sector, data is of enormous value, also generally characterised as complex and heterogeneous; once the data have been compiled, original data sets can be analysed for different purposes and stakeholders. At present, there are some contemporary common issues such as unclear relationships among multiple data entries; complex data analysis needing professional skill; data noise potentially affecting accuracy; making the wrong decision may cause the loss of revenue opportunities; uncomprehensive financial data analysis leads to low benefits deriving. Companies may not be able to meet shareholder expectations. For organisations' activities, data analysis is essential for the final decision making [3,4]. Existing data analytics methodologies usually have their pros and cons, and effective and efficient

data analysis is still a challenge that crosses multiple industry fields, which includes the financial sector. Hence, the demand to understand and make sense of a large amount of collected data is rapidly growing. However, utilising the complex data to provide stakeholders with the necessary information still remains a challenge in the financial sector. Additionally, the method of analysing relationships among multiple data entries and finding the importance of any individual data entry, is a critical factor in enabling stakeholders to grasp deep insights into their investments. One of the ultimate aims of financial data analysis is to draw the right conclusions from raw data in order to gain a competitive advantage.

The stock market is an essential sub-sector in the financial area, how to understand and make sense of the mountains of collected stock data has become a challenge in relevant fields. For instance, basic analytics in the stock market include understanding the stock price movements and the relationships among massive shares; looking at cycles and symmetry in the stock market based on counting trading days, for example, to find out if the current bear market is repeating the tops made during another period bear market. Hence, with the assistance of finalised data analytics outcomes, stakeholders may adjust investments accordingly and rationally. In addition, in this paper, we only analyse the connections among stocks, and this may help stakeholders to align their investments based on other related stocks' trends; the price prediction is not involved in our study at the stage. The raw data in this sector is usually massive and complex, it changes all the time, and the connections among stocks are complicated as well, which may lead to data analysis complexity. It arguably is that making rational judgments and decisions would require expertise and extra efforts in relevant areas; it is frequently becoming a burden of data analytics among those domains [5]. Many techniques have been applied to data analytics in the stock market, addressing multiple issues.

This study was initially encouraged by actual demand from stakeholders, creating a joint project with the financial data analytics from the business school in Western Sydney University. A general problem they found in this sector is that particular stocks are difficult to analyse for future investment due to the complexity of the stock's network. Then again, to get a better grip on all the shares, the centre/significant stocks need to be determined, factoring in the influence of essential stocks. Hereafter, this study was put on the agenda. Our proposed approach involves data processing, graph visualisation and social network analysis, and it models stock datasets into graphs based on the trading price changing rates among shares, which has not been tested in the Australian Stock Market thus far, to the best of our knowledge. The approach adopts centrality measurement concepts for importance measurement of stocks and compares the usability of those methods to find the suitable centrality metrics of the stock market data. This case study is based on raw data collected from the Australian Stock Market for over 20 years.

The significant contributions of this research are as follows:

- The provision of a supplementary method for examination of stock market data.
- The approach combines graph centrality measurements and interactive visualisation methods and applies them to real stock data.
- The method treats all selected shares as a network system other than individual entries. It brings the relationship strength into the analytics, comparing to the existing methods in the stock market analysis such as traditional charts and treemap, the connections among stocks can be identified clearly as a big picture through the finalised graph layouts. Essential features such as zoom in and out are provided as well, although the graph layout's quality is not included in this paper.
- We performed experiments on cleansed stock datasets to demonstrate that our approach is feasible and beneficial, and furthermore, the comparison of different centrality methods' performance on those datasets.
- Five centrality measurements were implemented, and their performances were compared based on the same dataset in experiments.

In the following subchapters, we briefly review related work about graph visualisation and graph centrality metrics in Section 2. Section 3 introduces the methodology we use throughout the work. Experimental results are reported in Section 4. Section 5 presents a discussion. The conclusion and future works are specified in Section 6.

2. Related Works

2.1. Graph Visualisation

In the stock market, a massive volume of raw data is extrapolated every second and said data needs to be processed into easy-to-understand forms in a reasonable time, to provide stakeholders with evidence for decision-making purposes. Many data processing methods have been applied to this sector to do so. At the final stage of financial data analytics, the finalised data has been filtered/cleansed/formatted, therefore its size might be much smaller compared to the raw data gathered, yet the finalised dataset is still large typically. For example, in our experiments, we collected around 6.4 million data entries and the completed graph model that contains the entire stock market has 1379 stocks and 11,535 edges among them, but the data is still too large to analyse, especially to non-experts. Data visualisation techniques can offer abilities for data interpretation to make raw data expressive, providing a practical and utilised way to present data in a meaningful way. Multiple visualisation tools have been adopted to visually exploit data insights in many application domains, including the stock market [6–8]. Currently, stock data analytics is still usually being conducted in traditional ways such as a spreadsheet, charts, treemaps and parallel coordinates, etc., with these methods improving the readability of finalised stock data to some extent.

A traditional stock market chart is a standard visualisation tool; it delivers representative charts and typically comes with features such as zooming, levels of details and selection etc. A chart provides a whole picture of stock market dynamics; the chart shape is considered in decision making for technical stock market analysis purposes [9]. Treemap is adopted in Smart Money—one of the most popular visualisation tools in the stock market—used by Wall-Street Magazine to show market performance [10], providing views of stock market performance and showing the changing of stock prices and the capitalisation of 500+ companies. It uses coloured rectangular tiles to represent stocks, the size represents the market capitalisation, the colour code indicates the price is decreasing/increasing, with the shading indicating the degree of change. Treemap is well-suited for presenting large hierarchical datasets where the node size feature matters. However, treemap cannot be used in decentralised networks [11], and investigating complex networks through unintuitive treemap view is difficult [12]. Parallel coordinates are used for visualising and analysing high-dimensional data. They are valid for presenting n-dimensional data, and they make it possible to explore data sets with large amounts. Although, in practice, parallel coordinates algorithms may lead to difficulty in understanding complex data, due to its sensitivity to visual clutter [13].

In addition, the node-link graph is another common visualisation technique; data with a relational structure is suitable to be modelled and visualised into node-link graphs, for a better analysis [14,15]. Hence, stock market data's relational structure makes it appropriate to apply node-link graphs methods. Here, the focus has been on how the elements are connected as a system, not just individual items [14]. Many well-developed node-link graph algorithms and tools have been built to generate graph layouts in a visually pleasing and useful way, which helps readers understand the structure and relationship patterns of the underlying graphs [16]. For example, the Visone software produces radial and spectral layouts, and integrates analysis and visualisation of networks facilitated by simple means of graphical interaction [17]. Handcock et al. created Statnet, which applies an algorithm which is called central Markov Chain Monte Carlo (MCMC) and focuses on statistical modelling of network data [18]. Rossi and Ahmed built up a web-based graph analytics platform, which is called NR, allowing users to analyse and visualise data online interactively in real-time [19]. Other similar software packages include RSiena [20], igraph [21], UCINet [22], Pajek [23], NodeXL [24,25] and Gephi [26].

Most existing visualisation tools may lose relevant data and relations among data units when they are finalising graph layouts as they place emphasis on creating graphs as abstract and straightforward as possible. Additionally, related methods adopted in the financial sector do not offer capabilities on the initiative 'unknown' relationship discovery of large-scale financial datasets, especially, to the best of our knowledge, not in the Australian stock market.

2.2. Graph Centrality Metrics

SNA (Social Networks Analysis) has experienced tremendous advances in recent years, and much research has been reported in the literature [27,28]. Centrality indices are critical metrics for network analysis. They have long been applied in SNA to provide different perceptions on the social relationships within the network, expressing the relative importance of a vertex or an edge in a network [29], and offering a detailed description of social structures [30]. Similarly, graph centrality measurements such as degree and PageRank factors etc., can provide information with importance 'ranking' in the stock market, hence, offering stakeholders a general idea on their future investments with consideration of related stocks, not only on individual shares.

Relevant studies that apply centrality metrics have been processed before. Wang et al. constructed a network to grasp the correlation structure and evolution of the world stock markets. Raw data was gathered on daily-based price indices of 57 stock markets during the 2005 to 2014 period, influence strength, betweenness centrality, and closeness centrality were adopted, and betweenness and closeness centrality metrics performed well [31]. Kazemilari et al. analysed the daily closure prices data of 70 stocks of renewable energy companies during the period from October 13, 2010 to March 4, 2015. Three centrality measures including degree, closeness and betweenness centralities were adopted to analyse the topological properties of minimum spanning trees. The outcomes showed extensive stocks within the network that played significant roles in renewable energy development in terms of market capitals. Moreover, degree, closeness and betweenness centralities provided similar results [32]. The closeness centrality was applied to measure the 2008 market crash of tensor financial network in Thailand, and it was claimed that the closeness centrality algorithm is the best tool to detect market crash [33]. Junior et al. used node strength (the sum of all values assigned to each edge that a node has) as a measure of node centrality to rank the most strong influencing nodes of related networks, including 83 stock market indices [34]. Dimitrios and Vasileios analysed stock relationships between 2007 and 2012 in the Greek Stock Market, degree, closeness, betweenness, eigenvector centralities and clustering coefficients were adopted to conclude the topology and in finding the most central shares. The outcomes showed that the Greek Market is a "shallow" market, meaning it can be affected easily by a few big investors or the economic climate [35]. Gao et al. studied the influence relations among listed stocks through generating a directed network of the Chinese stock market. They adopted in-degree, PageRank, eigenvector, authority, hub and betweenness metrics to obtain critical nodes in the influence network, and found that the in-degree, PageRank, eigenvector and authority performed well in characterising the importance of listed companies, while betweenness and hub measurements failed to do so [36]. Djauhari and Gan developed an optimal minimal spanning tree onto the daily data of closing prices of 98 stocks during the whole year of 2012, to overcome the non-optimality problem. In the experiments, degree centrality was applied to analyse network topology and determine the degree distribution [37]. Tu proposed a method based on cointegration to construct a sophisticated financial network in the Chinese stock market. In that study, directed, weighted and non-symmetric graphs were generated for showing network structure, and degree centrality, PageRank, HITS, local clustering coefficient, K-shell and strongly and weakly connected components were applied. Outcomes from the Cointegration Planar Graph (CIPG), Cointegration Threshold Network (CITN) and Partial Correlation Planar Graph (PCPG) were compared [38].

Five graph centrality metrics were proposed and applied to analyse stock market networks in this research. They are all characterized by the same monotonicity (higher statistics lead to higher centrality), symmetry (nodes' centralities only depend on their statistics and not their labels), and additivity

(statistics are processed in an additively separable manner) axioms [39]. Degree centrality is adapted to show the number of connections of a stock in the stock market; closeness centrality indicates the average length of a stock to all other shares in the network; betweenness centrality presents the number of geodesics between all pairs of shares in the network that pass through the specific capital; and PageRank is applied to relevance networks. Some metrics are more suitable than others in the stock market data analytics, yet, the stock impact cannot be measured adequately by any individual factor.

3. Methodology

To explore the relationship between data in the financial sector, a new approach was conducted, and the specific purposes of the study areas were to grasp whether significant stocks may affect others; to discover potential 'unknown' relationships among massive stock market raw data; to further explore the correctness of visualised relationship representation; and to quantitatively examine the feasibility of the approach in practice. To the best of our knowledge, few efforts that combine centrality metrics and node-link algorithms have been made to conduct visual analytics of the stock market datasets. This proposed approach combines the force-directed algorithm and five centrality metrics methods, to grasp significant stocks and show an overview of the entire structure of the Australian stock market. Although we place emphasis on the centrality metrics methods' performance in experiments, graph visualisation algorithms are included for layout representation only, see [40,41] for details regarding data processing and graph visualisation techniques involved in this work.

3.1. Data Processing and Graph Modelling

We collected all raw data in the experiments from the ASX, including 5088 stocks in the Australian stock market, which ranged from January 2, 1997 to June 30, 2017. Nearly 6.4 million data entries were gathered. In the data processing step, individual stock prices' changing rates between every two continues trading days were computed, and then the price changing rates between every two stocks during same two continues trading days were compared, hence, grasping the similarity of those two shares' price changing trends, and finding the potential connections between those two stocks. Eventually, 1379 shares and 11,535 links were generated for the graph modelling step. For details, please see [40].

In this research, each share is treated as a 'node'. Further, a relation between a stock and its connected stock can be established, and the connection is represented as an 'edge'. Hence, raw data can be transferred into undirected graph models. For example, in Figure 1, stock 'NAB' is connected to the 'CMI', and the edge weight (connection strength) (see [40,41]) is 3 in the case.

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml>
<key id="d0" for="node"
attr.name="color" attr.type="string">
<default> yellow </default>
</key >
<key id="d1" for="edge"
attr.name="weight" attr.type="double"/>
<graph id="G" edgedefault="undirected">
<node id="NAB" />
<node id="CMI" />
<edge id="0" source="CMI" target="NAB">
<data key="d1" > 3</data>
</edge>
</graph>
</graphml>
```

Figure 1. An XML graph model sample that contains two vertices and one edge.

3.2. Force-Directed Algorithm

Force-directed algorithms treat graph elements as a mechanical system, applying energies such as spring force onto every vertex and edge, keep working on nodes to move them to reasonable positions until the termination state is reached, for example, the energy is minimised. They usually offer aesthetically pleasing graph layouts [42,43]. The FA (ForceAtlas) layout algorithm [44] is a spatial layout method under the category of force-directed algorithms, and it addresses giving a simple shape to large real-world networks. FA2 (ForceAtlas2) offers more options and innovative optimisations; it brings good performances for a network of fewer than 100,000 nodes. FA2 is empirically observed at its best with strongly clustered networks, it excels at presenting social networks, and it takes in account the degree of the nodes in the repulsion so that the specific visual cluttering is reduced [44].

Forces involved in FA2 are the attraction force and repulsive force. Suppose there is the classical attraction force f_a between two connected nodes n_1 and n_2 depends linearly on the distance $d(n_1, n_2)$, then

$$f_a(n_1, n_2) = d(n_1, n_2). \tag{1}$$

FA2 brings poorly connected nodes closer to very connected nodes, and it tweaks the repulsion force so that poorly connected nodes and very connected nodes repulse less. The repulsive force f_r is proportional to the product of the degrees plus one (deg+1) of the two nodes n_1 and n_2, and the coefficient k_r is pre-defined by the settings, then

$$f_r(n_1, n_2) = k_r \frac{(\deg(n_1) + 1)(\deg(n_2) + 1)}{d(n_1,\ n_2)}. \tag{2}$$

Given a graph $G = (V, E)$, the combined force applied to vertex v is represented as:

$$F(v) = \sum_{(u,v)\in \mathrm{E}} f_a(u, v) + \sum_{(u,v)\in \mathrm{V}*\mathrm{V}} f_r(u, v). \tag{3}$$

3.3. Centrality Metrics

To grasp a more comprehensive impression on all the stock data collected, we conducted a study to determine which stocks are at the centre of all stocks; therefore, graph centrality metrics exist to discover relationships between stocks. Here, the centrality measurements were characterized by the same axioms of monotonicity, symmetry and additivity, and symmetry guarantees that a centrality measure does not depend on the identity of a node. Note that in most of the unweighted networks, the edges are treated equally, which is not the instance in this study. Each connection in the stock market network may have different underlying significances in network structures and functions, and centrality metrics can be influenced by taking into account the weightings that are applied to them [45,46]. For different kinds of network flows, various centrality measures should be used [47]. More specifically, five metrics were adopted in our experiments to examine the stock's network. Here, a stock network is a labelled undirected weighted graph $G = (V, E, w)$, in which V presents the set of vertices, E is the set of edges and w is the weight function (see [40]).

Eigenvector centrality is an expansion of the degree centrality metric. It considers the importance of the nodes connected to the current node, which means not all vertices are equivalent. The eigenvector centrality attributes a value that represents the connection intensity among nodes, a higher value indicates a more critical node, and a node that has few but essential linkers is still with high eigenvector centrality [48–50]. In this study, the eigenvector centrality concept is adopted in undirected graphs as a ranking measure to analyse the importance of stocks; it measures the extent to which a stock interacts with other shares in the stock market. Let A show an $(n \times n)$ similarity matrix, λ be the largest eigenvalue of A and x the corresponding eigenvector, the eigenvector centrality x_i of node i is defined

as the i^{th} entry in the normalized eigenvector belonging to the largest eigenvalue of A, $N(i)$ are the node i's neighbouring nodes, consider a particular node i with its adjacent nodes $N(i)$, then

$$x_i = \sum_{j \in N(i)} A_{ij} x_j. \tag{4}$$

PageRank defines a link analysis method to evaluate a user's influence, so that not only the immediate information flow is incorporated, but the information flow after that would also be considered. A node here acts more critical when it is linked from other nodes that play essential roles, or it is highly connected. It assigns probability distributions to each node, indicating the importance of the node through determining the probability of being at that node throughout the random walk [51,52]. The PageRank centrality is also applied in undirected graphs here. At each node in an undirected graph, the next node is selected with probability from the set of successors of the present node. Or else, when a node has no successors, the next node is chosen from all nodes, and nodes with higher measurement are more likely to be determined.

The weighted degree centrality takes into consideration the weights of ties, and this has been the preferred measure for analysing weighted networks [46,53]. In this research, a number of connections pointing to or emerging from a stock in the graph, and edge weight is included. Suppose w is the weighted adjacency matrix, V is the nodes added, if node i is connected to node j then w_{ij} is greater than 0, and the value indicates the weight of the tie. Then the weighted degree x_i of a node i is the sum of weight values that from all nodes connect to it, the weight is defined as:

$$x_i = \sum_{j \in V} w_{ij}. \tag{5}$$

Betweenness centrality is a primary measure in SNA, and it expresses the importance of elements (vertex/edge) involved in a network, evaluating traffic in communication networks, and also identifies critical intersections in road networks [29]. Betweenness centrality is generally observed as a measure of others' dependence on a given node, and therefore as a measure of potential control. A node comes with high betweenness centrality if it lies between many other nodes concerning their shortest path, and this node controls the flow of information between many other nodes. The vertex with the highest betweenness value is on the closest link among all nodes in a network [54,55].

Closeness centrality is designed as the reciprocal of the sum of the length of the shortest paths between the selected node and all other nodes in a network. Accordingly, the more central a node is, the closer it is to all other nodes. Closeness centrality is usually construed either as a measure of access efficiency or of independence from potential control by intermediaries. The node with the highest closeness metrics can reach every other node in a network on a short path, and it has the power of access [54–56].

3.4. Procedure

Based on the raw data generated from data processing and graph modelling steps, the proposed approach applies five centrality metrics onto finalised graph models; then, compares the metrics results with top 300 stocks from ASX to find out which central metrics algorithms are suitable for analysis of the stock market data. The steps included in the workflow of this study are shown below, as well as in Figure 2.

(1) Collecting raw data from ASX;
(2) Data filtering and formatting, such as removing stocks not existing at present and stocks' existence are less than one year;
(3) Computing individual stocks' price changing rates;
(4) Cross comparing stocks' price changing rate and finding similarities between every two stocks;

(5) Generating graph models based on stock data processed (every stock is a node, and connections among stocks are edges, the relationship strength is presented via edge weight);
(6) Finalising experiments:

- Applying the FA2 algorithm to the graph models and generating graph layouts;
- Using five centrality metrics methods to the graph models and getting five groups of top 300 rankings, comparing results to the top 300 stocks from ASX;

(7) Outcomes and discussion.

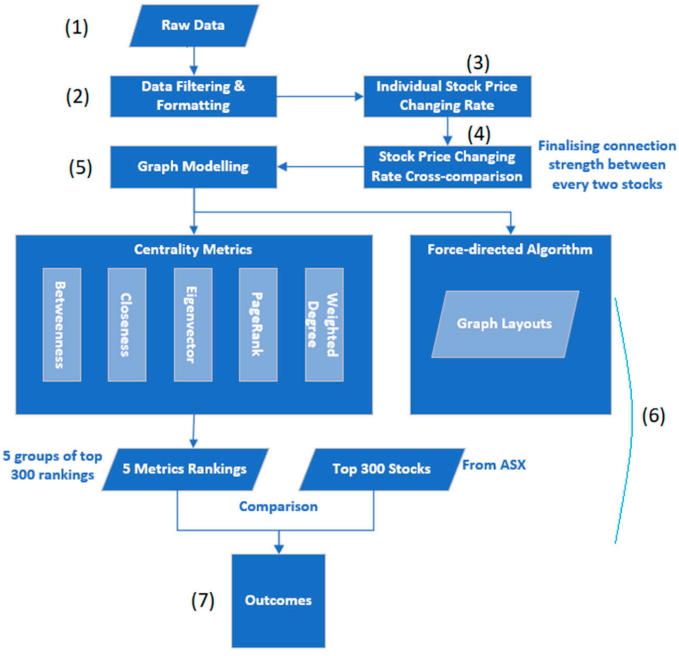


Figure 2. The workflow of this study.

3.5. Tools and Technology Summary

A summary containing details of evaluations and relevant technologies is shown in Table 1.

Table 1. Tools and technology summary in this study.

Evaluation	Tools and Technology
Data formatting/filtering/cleansing	Individual stock price changing rate calculation Cross-comparisons of stock price changing rates
Graph model generation	XML/graphml
Graph layout generation	Gephi/ForceAtlas 2
Centrality measurements	Betweenness/Closeness/Eigenvector/PageRank/Weighted Degree

4. Results

Eventually, 1379 stocks were kept after data processing steps, as well as 11,535 edges for presenting relationships among shares. Table 2 shows all the comparison results. For example, the top 300 shares from the weighted degree centrality match 64.13% of the top 300 stocks from the S&P/ASX 300 (XKO) on January, 2019. The S&P/ASX 300 (XKO) Index provides information to Australia's large-, mid-

and small-cap equities, and this index consists of all S&P/ASX 200 companies plus 100 smaller-cap companies that have market capitalisations' above $100 million (AUD). Investors typically use the ASX 300 as a benchmark for superannuation portfolios and managed funds due to its exposure to smaller companies. On the other hand, PageRank metrics match 61.56%, and closeness and eigenvector centrality only reach 50.57% and 48.74%. In this case, the weighted degree and PageRank tend to have a 'better' similarity in practice, and all other three metrics present worse.

Table 2. Similarity rates between five centrality measurements results and top 300 stocks from ASX on selected dates.

	Similarity Rate (%)				
Metrics / ASX Date	**Betweenness**	**Closeness**	**Eigenvector**	**PageRank**	**Weighted Degree**
January, 2019	57.17	50.57	48.74	61.56	64.13
December, 2018	56.98	50.41	48.58	61.00	63.56
July, 2018	55.72	49.53	47.71	59.73	62.64
February, 2018	54.94	48.76	46.57	59.31	61.86
December, 2017	55.19	48.98	46.78	59.58	62.13
July, 2017	54.46	48.94	46.73	59.25	62.19
January, 2017	52.29	47.84	45.98	57.11	61.56
August, 2016	50.16	45.72	45.35	55.39	60.59

In Table 3, due to the limitation of paper length, an example of only the top 20 stocks of giving centrality measurements are given (stock information details such as stock name and sector etc. can be referred from https://www.asx.com.au/). In the experimental evaluation the top 300 rankings were applied.

Table 3. Top 20 stocks of five centrality measurements (please refer to www.asx.com.au for stock details).

	Betweenness	**Closeness**	**Eigenvector**	**PageRank**	**Weighted Degree**
1	SCP	MIG	BXB	RAC	BXB
2	IAM	VGL	WOW	WOR	TLS
3	WOR	CQG	FXL	WMC	WOW
4	IAF	SGC	WBC	WAM	FXL
5	RCB	PGS	CHC	STW	WBC
6	CNU	SAS	DWS	SLF	MFF
7	MPL	ICI	TAH	TLS	AUB
8	CAT	CCE	EHL	SMI	WAM
9	OML	SRO	STO	WOW	CHC
10	BIR	LI3	MFF	WBC	SOL
11	RAC	TRA	SHL	TNE	TAH
12	SMI	ITL	AUB	MAY	RHC
13	QFY	CIA	SGP	MRD	RIO
14	WAM	KSM	SOL	BXB	DWS
15	INA	SUM	WPL	SUN	WPL
16	PLG	TMT	CUV	AUB	STO
17	TLS	AGC	TGG	RHC	SHL
18	FCT	GTT	RIO	SEK	SGP
19	SLF	AJN	WTP	WTP	WTP
20	HUB	AJO	CFE	MIG	WES

In addition, the similarity rates of the top 300 shares from giving centrality metrics are shown in Table 4. PageRank and weighted degree metrics are subject to provide similar results, and their outcomes are 81.33% matching; the others all differ a lot.

Table 4. Similarity rates between five centrality measurements results (%).

	Betweenness	Closeness	Eigenvector	PageRank	Weighted Degree
Betweenness	100	56.67	39	68.67	62.33
Closeness	56.67	100	47.67	54	56.33
Eigenvector	39	47.67	100	54	65
PageRank	68.67	54	54	100	81.33
Weighted Degree	62.33	56.33	65	81.33	100

Figure 3 offers an overview of the entire Australian stock market network based on shares cleansed in this study, which includes 1379 stocks and 11,535 connections among them. Some peripheral vertices may not be shown in Figure 3 due to the paper layout size limit. Figure 3 applies the weighted degree centrality measurement for computing the ranking of each share, the darker colour and larger node name indicate significant (higher ranking) stocks, and the thicker edge width represents stronger connections between two shares.

Figure 3. An overview of the Australian stock market ($|E| = 1379$; $|V| = 11{,}535$; weighted degree metrics applied).

5. Discussion

From the results in Table 2, it can be found out that centrality measurements' rankings do not match the actual top 300 stocks perfectly. The best similarity percentages are 64.13% (weighted degree metrics on January, 2019) and 61.56% (PageRank metrics on January, 2019), and the worst is 45.35% (eigenvector metrics on August, 2016). The possible reasons may include:

- The S&P/ASX 300 (XKO) provides comprehensive information on the Australian share market. It has a diverse mix of large-, mid- and small-cap shares, yet, the index only accounts for the partially Australian equity market. For example, it accounted for 89% of the Australian equity market in March 2018; it did not involve all the stocks at that time. Additionally, the index takes into account the market capitalisation of each stock, and the connections among stocks are not included;
- The ASX 300 consists of all S&P/ASX 200 companies plus 100 smaller-cap companies. ASX 200 indicates a company's contribution relative to its total market value. In addition the ASX 200 is also float-adjusted, which presents that the total numerical contribution to the index of a stock is comparable to the stock's value at the float of the stock [57]. The calculation of the ASX 200 starts with a sum of the market capitalisation of the constituent stocks, and it is intended to reflect changes in share price, not only market capitalisation, and only consider the other 100 smaller-cap

stocks with market capitalisations' above $100 million (AUD). On the other hand, our experiments rely on the price changing rates between every two stocks on the same continuous trading days. Hence, the relationship of stocks was emphasised; the capitalisation was not taken into account at this stage; this may lead to incorrect results.

- The top 300 stocks vary on a different period; it considers present market capitalisations of each share. For example, the weighted degree metrics' similarity with the real top 300 shares on January, 2019 was 64.13%, and the value was 60.69% on August, 2016, that makes a 3.44% difference between those two months. The proposed approach focuses on the historical data and ignores the short-term data changing; this may also cause unmatched outcomes.

Hence, due to different calculations between ASX 300 and our methodology, the results are not perfectly matched, albeit expectedly so. Additionally, from Table 1, the weighted degree and PageRank measurements perform well, closeness and eigenvector centrality metrics hardly reach 50% similarity, however betweenness metrics, on the other hand, have average performance among the five methods. What is more, most top shares in Table 2 belong to the financial and mining sector, for example, in the weighted degree metrics ranking category, there are 35% in the economic sector (FXL, WBC, MFF, AUB, WAM, CHC, SOL); 20% are in the mining/energy area (RIO, WES, WPL, STO); the health/medical, construction and retail industries own others. These results are in line with people's common sense of the Australian market, but further discovery of the stock sector distribution has not been done at the stage.

In the stock market, there are drawbacks of traditional data analysis methods. For example, stock charts can only present selected shares in limited amounts (see Figure 4, it shows a case that only contains five stocks BXB, TLS, WOW, FXL, WBC, stock details can be referred from https://www.asx.com.au/); treemap's capability is restricted in a complex network, and the way of representing parallel coordinates' makes it hard to understand and show complicated datasets. Comparing to those existing works, this research combines data processing, graph visualisation and centrality metrics methods, calculates edge weights based on stocks' price changing rates, provides not only the structure of the Australian stock market network (Figure 3), but also offers a feasible way to do the 'importance' analytics among massive stocks (top rankings). Hence, the methodology could deliver significant shares and relations among them that previous studies may lack.

Figure 4. An example of a traditional chart for showing five stocks' price changing trends (from https://www.investing.com/charts/real-time-stocks-charts).

6. Conclusions

The existing approaches apply proper centrality measurements to analyse the relationship among complex stock markets and grasp the importance of stocks. However, details of the network generation and centrality metrics comparisons are still lacking in most cases. Our methodology offers a clear procedure of raw data processing, generates graph models and compares centrality measurements' performance based on the S&P/ASX 300 (XKO) in practice. Hence the results are more convincing theoretically.

Based on raw data collected between 1997 and 2017 from ASX, we conducted an approach that applied betweenness, closeness, eigenvector, PageRank and weighted degree centrality measurements on graph models generated, furthermore, we compared the top 300 ranking stocks from five measurements and the ASX top 300, revealing that PageRank and weighted degree centrality measurements performed well. As far as we know, this study is the first work which has adopted the methodology above in the Australian stock market. The experimental results were computed based on historical data gathered in a 20 years period, and only considered the connection among stocks, but in practice, a new company with high capitalisation, which does not have relations to other shares, can still be in top 300 at ASX. Hence, the differences between our approach and practical market ranking were reported.

The present approach provides a static pattern that offers stakeholders relationship descriptions of all stocks, and each share's centrality metrics changes over time which has not been taken into account at the stage. On the other hand, the epdf (Equal-Error Probability Density Function) can identify the network topology quickly [58]; and probability distribution of centrality would offer a powerful supplementary aspect to the proposed approach, which can draw dynamic changing patterns of selected essential shares to study the symmetry of the stock market network, and grasp deep insights of those shares in the Australian stock market. Hence, the trend analytics would be more convincing; the probability distribution of centrality will be included in our future work. In addition, the present stock capitalisation, as well as the increasing rate of the stock capitalisation, will be taken into account as node weights to enhance the approach's feasibility.

Author Contributions: Conceptualization, J.H.; methodology, J.H. and M.H.; software, J.H.; validation, J.H.; formal analysis, C.H.; investigation, J.H.; resources, J.H.; data curation, J.H.; writing—original draft preparation, J.H.; writing—review and editing, J.H., M.H. and C.H.; visualisation, J.H.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bereznoi, A. Business model innovation in corporate competitive strategy. *Probl. Econ. Transit.* **2015**, *57*, 14–33. [CrossRef]
2. De Bondt, W.F.; Thaler, R.H. Financial decision-making in markets and firms: A behavioral perspective. *Handb. Oper. Res. Manag. Sci.* **1995**, *9*, 385–410.
3. McAfee, A.; Brynjolfsson, E.; Davenport, T.H.; Patil, D.J.; Barton, D. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68. [PubMed]
4. Larson, D.; Chang, V. A review and future direction of agile, business intelligence, analytics and data science. *Int. J. Inf. Manag.* **2016**, *36*, 700–710. [CrossRef]
5. Gärling, T.; Kirchler, E.; Lewis, A.; Van Raaij, F. Psychology, financial decision making, and financial crises. *Psychol. Sci. Public Interest* **2009**, *10*, 1–47. [CrossRef]
6. Bikakis, N.; Sellis, T. Exploration and visualization in the web of big linked data: A survey of the state of the art. *arXiv preprint* **2016**, arXiv:1601.08059.

7. Zhang, L.; Stoffel, A.; Behrisch, M.; Mittelstadt, S.; Schreck, T.; Pompl, R.; Weber, S.; Last, H.; Keim, D. Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems. In Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), Seattle, WA, USA, 14–19 October 2012.
8. Parsons, P.; Sedig, K.; Didandeh, A.; Khosravi, A. Interactivity in Visual Analytics: Use of Conceptual Frameworks to Support Human-Centered Design of a Decision-Support Tool. In Proceedings of the 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2015; pp. 1138–1147.
9. Šimunić, K. Visualization of Stock Market Charts. In Proceedings of the 11-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2003, Plzen, Czech Republic, 3–7 February 2003.
10. Johnson, J.E.; Morris, J.K.; de Brun, S.C.; Gunther, N.L. Visible Market, Inc., Dynamic Visual Statistical Data Display and Method for Limited Display Device. U.S. Patent 8,972,295, 3 March 2015.
11. Kolomeets, M.; Chechulin, A.; Kotenko, I.; Strecker, M. Voronoi Maps for Planar Sensor Networks Visualization. In *Communications in Computer and Information Science, Proceedings of the International Symposium on Mobile Internet Security, Jeju Island, Korea, 19–22 October 2017*; Springer: Singapore, 2017.
12. Blue, R.; Dunne, C.; Fuchs, A.; King, K.; Schulman, A. Visualizing real-time network resource usage. In *Lecture Notes in Computer Science, Proceedings of the International Workshop on Visualization for Computer Security, Cambridge, MA, USA, 15 September 2008*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 119–135.
13. Johansson, J.; Forsell, C. Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. *IEEE Trans. Vis. Comput. Graph.* **2015**, *22*, 579–588. [CrossRef]
14. Card, S.K.; Nation, D.A. Xerox Corp. System and method for browsing hierarchically based node-link structures based on an estimated degree of interest. U.S. Patent 7,392,488, 24 June 2008.
15. Huang, W.; Hong, S.H.; Eades, P. Effects of crossing angles. In Proceedings of the 2008 IEEE Pacific Visualization Symposium, Kyoto, Japan, 5–7 March 2008; pp. 41–46.
16. Brandes, U.; Wagner, D. Analysis and visualisation of social networks. In *Graph Drawing Software*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 321–340.
17. Baur, M.; Benkert, M.; Brandes, U.; Cornelsen, S.; Gaertler, M.; Köpf, B.; Lerner, J.; Wagner, D. Visone Software for visual social network analysis. In *Lecture Notes in Computer Science, Proceedings of the International Symposium on Graph Drawing, Vienna, Austria, 23–26 September 2001*; Springer: Berlin/Heidelberg, Germany, 2001.
18. Handcock, M.S.; Hunter, D.R.; Butts, C.T.; Goodreau, S.M.; Morris, M. statnet: Software Tools for the Representation, Visualisation, Analysis and Simulation of Network Data. *J. Stat. Softw.* **2005**, *14*, 1548–7660. [CrossRef]
19. Rossi, R.; Ahmed, N. The network data repository with interactive graph analytics and visualization. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
20. Ripley, R.; Boitmanis, K.; Snijders, T.A.B. RSiena: Siena Simulation Investigation for Empirical Network Analysis. R package version 1.1-232. 2013. Available online: http://CRAN.R-project.org/package=RSiena (accessed on 5 April 2019).
21. Csárdi, G.; Nepusz, T. The igraph software package for complex network research. *Inter. J. Comp. Syst.* **2006**, *1695*, 1–9.
22. Borgatti, S.; Everett, M.G.; Freeman, L.C. *UCINET 6.0 Version 1.00*; Analytic Technologies: Natick, MA, USA, 1999.
23. Batagelj, V.; Andrej, M. Pajek—Analysis and visualization of large networks. In *Lecture Notes in Computer Science, Proceedings of the International Symposium on Graph Drawing, Vienna, Austria, 23–26 September 2001*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 477–478.
24. Bonsignore, E.M.; Dunne, C.; Rotman, D.; Smith, M.; Capone, T.; Hansen, D.L.; Shneiderman, B. First steps to NetViz Nirvana: Evaluating social network analysis with NodeXL. In Proceedings of the 2009 International conference on computational science and engineering, Vancouver, BC, Canada, 29–31 August 2009; pp. 332–339.
25. Hansen, D.L.; Shneiderman, B.; Smith, M.A. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*; Morgan Kaufmann: San Mateo, CA, USA, 2010.
26. Bastian, M.; Heymann, S.; Jacomy, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2009.

27. Scott, J.; Carrington, P.J. *The SAGE Handbook of Social Network Analysis*; SAGE Publications: Thousand Oaks, CA, USA, 2011.
28. Aggarwal, C.C. An introduction to social network data analytics. In *Social Network Data Analytics*; Springer: Boston, MA, USA, 2011; pp. 1–15.
29. Riondato, M.; Kornaropoulos, E.M. Fast approximation of betweenness centrality through sampling. *Data Min. Knowl. Discov.* **2016**, *30*, 438–475. [CrossRef]
30. Mika, P.; Gangemi, A. Descriptions of social relations. *Benefits* **2016**, *1*, 14.
31. Wang, G.J.; Xie, C.; Stanley, H.E. Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks. *Comput. Econ.* **2018**, *51*, 607–635. [CrossRef]
32. Kazemilari, M.; Mardani, A.; Streimikiene, D.; Zavadskas, E.K. An overview of renewable energy companies in stock exchange: Evidence from minimal spanning tree approach. *Renew. Energy* **2017**, *102*, 107–117. [CrossRef]
33. Kanjamapornkul, K.; Pinčák, R.; Bartoš, E. The study of Thai stock market across the 2008 financial crisis. *Phys. A Stat. Mech. Its Appl.* **2016**, *462*, 117–133. [CrossRef]
34. Junior, L.S.; Mullokandov, A.; Kenett, D.Y. Dependency Relations among International Stock Market Indices. *J. Risk Financ. Manag.* **2015**, *8*, 227–265. [CrossRef]
35. Dimitrios, K.; Vasileios, O. A network analysis of the Greek stock market. *Procedia Econ. Financ.* **2015**, *33*, 340–349. [CrossRef]
36. Gao, Y.C.; Zeng, Y.; Cai, S.M. Influence network in the Chinese stock market. *J. Stat. Mech. Theory Exp.* **2015**, *2015*, P03017. [CrossRef]
37. Djauhari, M.A.; Gan, S.L. Optimality problem of network topology in stocks market analysis. *Phys. A Stat. Mech. Its Appl.* **2015**, *419*, 108–114. [CrossRef]
38. Tu, C. Cointegration-based financial networks study in Chinese stock market. *Phys. A Stat. Mech. Its Appl.* **2014**, *402*, 245–254. [CrossRef]
39. Bloch, F.; Jackson, M.O.; Tebaldi, P. Centrality measures in networks. *arXiv* **2017**, arXiv:1608.05845. [CrossRef]
40. Hua, J.; Huang, M.; Zreika, M.; Wang, G. Applying Data Visualization Techniques for Stock Relationship Analysis. *Filomat* **2018**, *32*, 1931–1936. [CrossRef]
41. Zreika, M.; Hua, J.; Wang, G. Applying Data Processing Method for Relationship Discovery in the Stock Market. In *Recent Developments in Data Science and Business Analytics*; Springer: Cham, Switzerland, 2018; pp. 247–253.
42. Gansner, E.R.; North, S.C. Improved force-directed layouts. In Proceedings of the Graph Drawing 98, Montreal, QC, Canada, 13–15 August 1998; pp. 364–373.
43. Kobourov, S.G. Force-directed drawing algorithms. In *Handbook of Graph Drawing and Visualization*; CRC Press: Boca Raton, FL, USA, 2013; pp. 383–408.
44. Jacomy, M.; Heymann, S.; Venturini, T.; Bastian, M. Force Atlas 2, a Graph Layout Algorithm for Handy Network Visualisation. 2011. Available online: https://medialab.sciencespo.fr/publications/Jacomy_Heymann_Venturini-Force_Atlas2.pdf (accessed on 5 April 2019).
45. Friedkin, N.E. Theoretical foundations for centrality measures. *Am. J. Sociol.* **1991**, *96*, 1478–1504. [CrossRef]
46. Wang, J.; Hou, X.; Li, K.; Ding, Y. A novel weight neighborhood centrality algorithm for identifying influential spreaders in complex networks. *Phys. A Stat. Mech. Its Appl.* **2017**, *475*, 88–105. [CrossRef]
47. Borgatti, S. Social Network Analysis. Available online: http://www.analytictech.com/mb109/slides/networks.pdf (accessed on 5 April 2019).
48. Bonacich, P. Some unique properties of eigenvector centrality. *Soc. Netw.* **2007**, *29*, 555–564. [CrossRef]
49. sci.unich.it. Eigenvector Centrality. Available online: https://www.sci.unich.it/~{}francesc/teaching/network/eigenvector.html (accessed on 2 April 2019).
50. Lohmann, G.; Margulies, D.S.; Horstmann, A.; Pleger, B.; Lepsien, J.; Goldhahn, D.; Schloegl, H.; Stumvoll, M.; Villringer, A.; Turner, R. Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PloS ONE* **2010**, *5*, e10232. [CrossRef] [PubMed]
51. Langville, A.N.; Meyer, C.D. *Google's PageRank and beyond: The science of search engine rankings*; Princeton University Press: Princeton, NJ, USA, 2011.
52. Page, L.; Brin, S.; Motwani, R.; Rajeev, M.; Terry, W. *The PageRank Citation Ranking: Bringing Order to the Web*; Stanford InfoLab Publication Server: Stanford, CA, USA, 1998.

53. Tang, X.; Wang, J.; Zhong, J.; Pan, Y. Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **2014**, *11*, 407–418. [CrossRef] [PubMed]
54. Brandes, U.; Borgatti, S.P.; Freeman, L.C. Maintaining the duality of closeness and betweenness centrality. *Soc. Netw.* **2016**, *44*, 153–159. [CrossRef]
55. Krebs, V. Power in Networks. Available online: http://www.orgnet.com (accessed on 28 March 2019).
56. Hua, J.; Huang, M.; Huang, W.; Zhao, C. Applying Graph Centrality Metrics in Visual Analytics of Scientific Standard Datasets. *Symmetry* **2019**, *11*, 30. [CrossRef]
57. Spindices, Index Mathematics Methodology. Standard & Poor's. Available online: https://us.spindices.com/documents/methodologies/methodology-index-math.pdf (accessed on 28 March 2019).
58. Li, J.; Covertino, M. Optimal Microbiome Networks: Macroecology and Criticality. *Entrop* **2019**, *21*, 506. [CrossRef]

symmetry

MDPI

Review

A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources

Chengyu Sun, Liang Hu, Shuai Li, Tuohang Li, Hongtu Li and Ling Chi *

College of Computer Science and Technology, Jilin Unversity, Chaoyang District, Changchun 130012, China; cysun20@mails.jlu.edu.cn (C.S.); hul@jlu.edu.cn (L.H.); shuaili18@mails.jlu.edu.cn (S.L.); lith19@mails.jlu.edu.cn (T.L.); lihongtu@jlu.edu.cn (H.L.)

* Correspondence: chiling@jlu.edu.cn

Received: 22 October 2020; Accepted: 9 November 2020; Published: 12 November 2020

Abstract: An essential part of a text generation task is to extract critical information from the text. People usually obtain critical information in the text via manual extraction; however, the asymmetry between the ability to process information manually and the speed of information growth makes it impossible. This problem can be solved by automatic keyphrase extraction. In this paper, the mainstream unsupervised methods to extract keyphrases are summarized, and we analyze in detail the reasons for the differences in the performance of methods then provided some solutions.

Keywords: keyphrase extraction; unsupervised method; feature selection

1. Introduction

Under the background of the continuous development of the information age, the content based on words grows exponentially, making it more challenging to manage this large-scale information. This information could be processed manually in the past. However, now, it is impossible because of the asymmetry between the amount of data and the ability to process information manually, which exemplifies the efforts to handle the current data scales, thereby promoting the development of automatic key sentence and keyphrase extraction methods that use the mighty computing power of computers to replace the manual labor [1]. Keyphrase extraction and key sentence extraction are two important subtasks in the text generation task [2–5]. Among them, the key sentence extraction task separates the most important part of a text and combines it in a specific way to an abstract that can express the text's main content while retaining the readability [6]. The main task of keyphrase extraction is to identify a single word or phrase representing the text's main content [7]. The extracted results are called keyphrases, the most common of which include the keyword in the abstracts of academic papers, representing the core content that the author wants to express. As the concise expression of an article's main idea, keyphrase makes the information easy to be managed, classified, and retrieved [8]. At present, keyphrase extraction is widely used in many fields, such as natural language processing (NLP), information retrieval (IR) [9–12], opinion mining [13–15], document indexing [16], and document classification [17].

Keyphrase extraction is divided into supervised methods and unsupervised methods based on the training set. The difference between them is whether there is a labeled training set in the learning process. Among them, the supervised method [18] transforms the keyphrase extraction task into a classification problem [19,20] or regression problem [21]. It trains the model on the labeled training set and uses the trained model to determine whether a candidate word in a text is a keyphrase. For example, KEA (Automatic keyphrase extraction) [19] determines whether a candidate word is a keyphrase by calculating the TF-IDF (Term Frequency–Inverse Document Frequency) [22] value of each candidate word and the location where it first appears in the text and inputs these two values into the Naive Bayes classifier. Generally, the supervised method is superior to the unsupervised

method [23]. However, compared with the past, the explosive growth of all kinds of information makes the types and quantity of information increase significantly, and the supervised method requires many labeled training sets, thus it require large amounts of manual labor [24]. Moreover, there are no labeled datasets that can serve as references in many fields, especially in some languages that are not well known by human beings, such as the translation tasks of hieroglyphs and cuneiform characters, which makes unsupervised methods without human intervention essential.

Based on the features of the unsupervised keyphrase extraction methods selected by researchers, unsupervised methods can be divided into the statistics-based method, graph-based method, topic-based method, language model-based method, and these methods can be classified into two schools: the linguistic school and the statistical school. The first school mainly extracts keyphrases by analyzing texts using linguistic methods, among which the most common method is to analyze the topic distribution of articles, such as KeyCluster [25] and CommunityCluster [26]. The statistical school mainly analyzes an article's probability features such as KP-Miner [27] and YAKE [28] based on TF-IDF, TextRank [29], or SingleRank [30]. The linguistic school and statistical school have been influencing and promoting each other. As time has passed, researchers have proposed new methods to cross-utilize the two schools' knowledge, such as TopicRank based on clustering (linguistic school) and graphs (statistical school).

In the above discussion, we divide keyphrase extraction into the linguistic school and the statistical school. We continue this classification method to divide commonly used metrics, features that affect keyphrase extraction, and mainstream unsupervised keyphrase extraction methods, making the structure and development path of the entire field look clear.

This paper aims to introduce the mainstream unsupervised learning methods, which are reflected in [26,31], but we have done other work as follows:

- Based on the characteristics of different methods, combined with the human language habit, the reasons for the performance differences between methods are analyzed in detail (Section 5.1).
- In keyphrase extraction, the characteristics of the datasets directly affect the performance of the methods, so we analyze how different datasets affect method performance (Section 5.2).
- We analyze the reasons for the limitations of the keyphrase methods and propose corresponding solutions, which will help the following researchers to explore further (Section 6).

The remainder of this paper is organized as follows. Section 2 introduces some preliminary knowledge, including the datasets (Section 2.1) and evaluation metrics (Section 2.2) that are commonly used in the automatic keyphrase extraction field, the features affecting keyphrase extraction (Section 2.3), and how to use these features for keyphrase extraction (Section 2.4). Section 3 mainly introduces several types of unsupervised methods for keyphrase extraction (Section 3.1), which are divided into statistics-based methods (Section 3.2), graph-based methods (Section 3.3), topic-based methods (Section 3.4), and language model-based methods (Section 3.5). Section 4 is the experimental results, and we analyze the reasons for the differences in performance of methods based on human language habits in Section 5. In Section 6, we analyze the limitations of keyphrase extraction methods and provided some solutions. Finally, this paper is summarized in Section 7.

2. Datasets, Evaluation Metrics and Features

2.1. What Datasets Are There in the Keyphrase Extraction Field?

An unsupervised keyphrase extraction system can be applied to many datasets for testing, such as the full-text of a paper, the abstract of a paper, news, web page, and email. In this paper, the names, types, number of texts (Docs), contributors, number of tokens per text, language, and annotation (annotation for gold keyphrases are performed by authors (A), readers (R), editors (E), or professional indexers (I)) of multiple datasets are sorted out in detail, as shown in Table 1.

Table 1. Evaluation datasets grouped by their type.

Type	Dataset	Contributor	Tokens/Doc	Docs	Keyphrases/Doc	Language	Annotation
Full-text papers	ACM	Krapivin et al. [32]	2∼12k	2304	6	English	A
	Citeulike-180	Medelyan et al. [33]		181	5	English	A+R
	CSTR	Witten et al. [19]		630	-	English	A
	SemEval-2010	Kim et al. [23]		283	15	English	A+R
	NUS	Nguyen and Kan [34]		211	11	English	A+R
	PubMed	Schutz [35]		1320	5	English	A
Papers abstracts	Inspec	Hulth [36]	100∼200	2000	10	English	I
	KDD	Gollapalli et al. [37]		755	4	English	A
	WWW	Gollapalli et al. [37]		1330	5	English	A
	TALN	Boudin [38]		641	4	English	A
	TermLTH-Eval	Bougouin [39]		400	12	English	I
News	DUC-2001	Wan and Xiao [40]	300∼850	308	10	English	R
	500N-KPCrowd	Marujo et al. [41]		500	46	English	R
	110-PT-BN-KP	Marujo et al. [42]		110	28	Portuguese	R
	Wikinews	Bougouin et al. [7]		100	10	French	R
Web pages	Blogs	Grineva et al. [43]	500∼1k	252	8	English	R
	-	Hammouda et al. [44]		312	-	English	-

2.2. What Are the Evaluation Metrics in the Keyphrase Extraction Field?

It is not an easy task to design an evaluation metric that can reflect an algorithm's advantages and disadvantages. Since an evaluation metric may only evaluate one aspect of the algorithm, multiple metrics can more precisely and comprehensively evaluate an algorithm. For example, researchers usually use precision, recall, and F-score to evaluate a method from multiple perspectives. In this section, some standard evaluation metrics are introduced and divided into statistics-based and linguistics-based ones.

2.2.1. Statistics-Based Metrics

Statistics-based evaluation metrics analyze the performance of a method by calculating the proportion of the number of various keyphrases, such as the number of extracted keyphrases, correct keyphrases, wrong keyphrases, and manually assigned keyphrases. Standard statistics-based metrics include precision, recall, and F1-score.

Precision:

It represents the number of real keyphrases in the extracted keyphrases, reflecting the accuracy of the keyphrases output by the algorithm.

$$precison = \frac{tp}{tp+fp} = \frac{correct\ keyphrases}{extracted\ keyphrases} \tag{1}$$

Here, tp represents true positives, i.e., the number of keyphrases that are correctly extracted, and fp represents false positives, i.e., the number of keyphrases that are incorrectly extracted.

Recall:

It represents the number of extracted keyphrases among the real keyphrases, reflecting the comprehensiveness of the keyphrases output by the algorithm.

$$recall = \frac{tp}{tp+fn} = \frac{correctly\ matched\ keyphrases}{assigned\ keyphrases} \tag{2}$$

Here, fn represents false negatives, which are the keyphrases that are not correctly extracted.

Fα-score:

The precision and recall interact with each other. In an ideal situation, they are both high, but, in general, when precision is high, recall is low, and vice versa. The F-score is formed by combining precision and recall.

$$F\alpha\text{-}score = \frac{(\alpha^2 + 1) \cdot precision \cdot recall}{\alpha^2 \cdot precision + recall} \tag{3}$$

When $\alpha = 1$, it is the $F1\text{-}score$.

2.2.2. Linguistics-Based Metrics

The above evaluation metrics are based on the assumption that keyphrases are mutually independent, but, based on human language habits, we hope that the more essential keyphrases should be ranked higher.

The following three evaluation metrics can reflect the order features between the keyphrases output by an algorithm.

Mean Reciprocal Rank (MRR):

In MRR [45], $rank_d$ is denoted as the rank of the first correct keyphrase with all extracted keyphrases, D is the document set for keyphrase extraction, and d is a specific document.

$$MRR = \frac{\sum\limits_{d \in D} \frac{1}{rank_d}}{|D|} \tag{4}$$

Mean Average Precision (MAP):

The MAP takes the ordering of a particular returned list of keyphrases into account. The average precision (AP) is defined as follows:

$$AP = \frac{\sum\limits_{\mathrm{n}=1}^{|N|} P(n) gd(n)}{|LN|} \tag{5}$$

where $|N|$ is the length of the list, $|LN|$ is the number of relevant items, $P(n)$ is the precision, and $gd(n)$ equals one if the nth item is gold keyphrase and 0 otherwise. By averaging AP over a set of n documents, the Mean Average Precision (MAP) is defined as follows:

$$MAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{6}$$

where AP_i is the average precision of the extracted keyphrases list.

Binary Preference Measure (Bpref):

The $Bpref$ [46] represents the number of correct keyphrases in front of incorrect keyphrases extracted by the algorithm. Its definition is as follows:

$$Bpref = \frac{1}{C} \sum_{c \in C} 1 - \frac{|I|}{M} \tag{7}$$

where C represents the number of correct keyphrases, M represents the number of all extracted keyphrases, and I represents the number of correct keyphrases in front of incorrect keyphrases.

We organize all the formulas in Table 2.

Table 2. Formulas for all evaluation metrics.

Evaluation Metrcis	precision	$precison = \frac{tp}{tp+fp} = \frac{thenumberofcorrectkeyphrase}{thenumberofextractedkeyphrase}$
	recall	$recall = \frac{tp}{tp+fn} = \frac{the\ number\ of\ correctly\ matched\ keyphrase}{the\ number\ of\ assigned\ keyphrase}$
	F-socre	$F\alpha - score = \frac{(\alpha^2+1)\cdot precision \cdot recall}{\alpha^2 \cdot precision + recall}$
	MRR	$MRR = \frac{\sum_{d\in D} \frac{1}{rank_d}}{\lvert D\rvert}$
	AP	$AP = \frac{\sum_{n=1}^{\lvert N\rvert} P(n)gd(n)}{\lvert LN\rvert}$
	MAP	$MAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$
	Bpref	$Bpref = \frac{1}{C}\sum_{c\in C} 1 - \frac{\lvert I\rvert}{M}$

2.3. What Are the Features that Affect Keyphrase Extraction?

Many features affect keyphrase extraction methods performance, and these features are divided into linguistic-based features and statistical-based features.

2.3.1. Linguistic-Based Features

Topic distribution:

The locations where keyphrases appear are often not fixed in different text types and are affected by the distribution of topics since human language habits determine new keyphrases will appear whenever a new topic appears [8]. In academic papers and scientific articles, there is only one topic in the whole text, and so the keyphrases usually appear at the beginning and the end of a text [33]. However, most texts contain multiple topics, such as news and web pages, so new keyphrases will appear when the topic changes. To extract keyphrases from multi-topics articles, researchers have introduced clustering methods, such as Latent Dirichlet Allocation (LDA) [47], KeyCluster [25], the Topical PageRank (TPR) [24], CommunityCluster [26], and the topic-sensitive Topical PageRank (tsTPR) [48]. These methods are described in detail in Section 3.

Topic correlation:

For texts such as academic papers, the text's keyphrases are typically related to the others, so the correlation between texts can be used when extracting keyphrases [40]. However, this observation does not necessarily hold for emails or chats because there are no restrictions on the topics discussed between people, so it is difficult to use the relationship between the texts to extract keyphrases, and further increase the difficulty of the keyphrase extraction task.

2.3.2. Statistical-Based Features

Keyphrase density:

The concept of keyphrase density is proposed and defined as the ratio of the frequency of a keyphrase's occurrence to the total number of words in a text. To improve the algorithm's performance, we need to preprocess the document before calculating the keyphrase density, that is, delete the function words and restore the remaining words to their root patterns. The keyphrase density is

usually related to the document's length, while the average length of the document in different datasets is often different. For example, there are 300 documents in the DUC-2001 dataset, with an average of 847 words in each document, and its keyphrase density is 0.56%. There are 1330 documents in the WWW dataset, with an average of 163 words in each document, which keyphrase density of 0.87%. According to the relevant experimental experience, the longer the document length is, the more difficult it is to extract keyphrases [26]. Therefore, the higher the keyphrase density in a document is, the easier it is to extract keyphrases because a lower keyphrase density means that there are relatively few keyphrases, and the document is relatively long, which makes it more difficult to extract real keyphrases.

Lexical density:

The lexical density is used to express the structure and complexity of human language [48]. The definition of a lexical density is the ratio of the number of lexical words, which are simply nouns, adjectives, verbs, and adverbs in the document, to the total number of words in the document [49]. Lexical words give a text its meaning and provide information regarding what the text is about. In the keyphrase extraction task, lexical words are usually used as candidate keyphrase, so, when there are more lexical words in a text (larger lexical density), we need to select the real keywords from more candidate words, which increases the difficulty.

Keyphrase density and lexical density are used to reflect the features of datasets. Their difference is that keyphrase density reflects the frequency of the keyphrases, while the lexical density reflects the richness of text semantics.

Structural features:

The difficulty of keyphrase extraction will be reduced by their fixed structures. In texts with fixed formats, such as scientific research papers, which generally include an abstract, introduction, related work, experiment, and conclusion, keyphrases often appear at fixed positions such as the abstract and conclusion [19,33]. Simultaneously, it is more challenging to extract keyphrases in texts without a fixed format, such as news, blogs, and email [23].

2.4. How to Use These Features for Keyphrase Extraction?

In Section 2.3, the features that affect keyphrase extraction are introduced, and we show how researchers use these features to complete the keyphrase extraction task. The explanation is divided into linguistic-based and statistical-based sections.

2.4.1. Linguistic-Based

Topic distribution:

As mentioned above, the topic distribution has an impact on the difficulty of keyphrase extraction. Researchers expect to extract keyphrases that can cover all topics of a given document; therefore, they take the topic distribution into account and generally use Latent Dirichlet Allocation (LDA) [47] or a clustering method to detect the topic distribution. Tthese methods are described in detail in Section 3.

Syntactic features:

It can be seen that keyphrases are generally composed of lexical words; therefore, grammar patterns can be set to filter candidate words, such as nouns or adjectives plus nouns [50]; thus, in the keyphrase extraction task, the first step is to delete the non-lexical word and select keyphrase from the remaining words.

2.4.2. Statistical-Based

Frequency of words:

Generally, if a lexical word appears more frequently in a text and less frequently in another text, it can better represent the document's critical information. Based on this finding, researchers proposed the TF-IDF [22], where TF is the term frequency, representing the frequency of a candidate word in a document, and IDF is the inverse document frequency, representing the frequency of the candidate word in other documents.

Distance of words:

Generally, if a word appears at the top of a document, it is more likely to be a keyphrase [34]. Based on this finding, researchers took the location information as a feature defined as the distance of the first occurrence of a word in a document, and the length of documents is usually used to regularize each word's location information.

Structural features:

As stated in Section 2.3.2, for texts with a fixed format, such as scientific research papers, if a word often appears in abstract and introduction, it is more likely to be a keyphrase.

3. Unsupervised Keyphrase Extraction Methods

This section introduces the classification of unsupervised keyphrase extraction methods (Section 3.1).

We introduce various types of unsupervised methods following the chronological order of the publication of papers and show how these research works are optimized from generation to generation (Sections 3.2–3.5).

3.1. Classification of Unsupervised Keyphrase Extraction Methods

The mainstream unsupervised keyphrase extraction methods are divided into four categories: statistics-based method, graph-based method, topic-based method, and language model-based method. The methods covered in this article are summarized in Figure 1.

The mainstream unsupervised methods usually preprocess documents when performing keyphrase extraction task. Because the keyphrases are the lexical words (nouns, adjectives, verbs, and adverbs), deleting other words except lexical words in the document is necessary. Since some words have different forms but have similar meanings (such as play and playing), they are restored to their root forms. Finally, the remaining words are treated as candidate keyphrases, where the real keyphrases are extracted. The unsupervised method described below does not introduce this step.

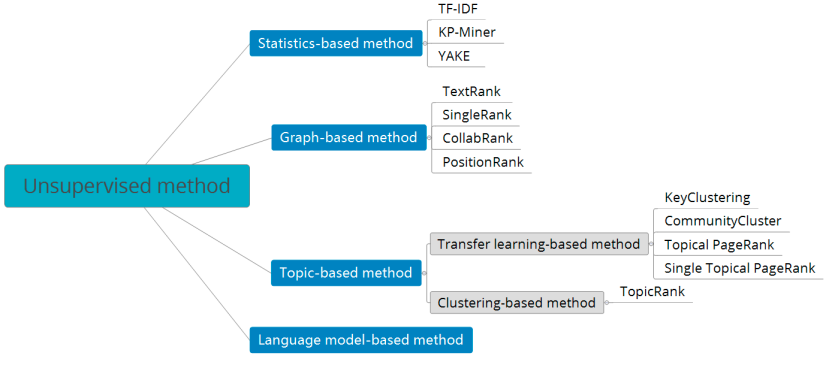


Figure 1. Summary of unsupervised methods.

The symbols in the unsupervised keyphrase extraction method in this paper are described in Table 3.

Table 3. Symbols used in this paper.

S (W/V)	The score for a word W or node V
WE (Vk, Vm)	Edge weight of node Vk and node Vm
α	damping factor
NB (V)	the neighboring node of node V

3.2. Statistics-Based Methods

TF-IDF:

The TF-IDF is a common baseline method in the keyphrase extraction field, in which the Term Frequency (TF) represents the frequency of a word in a document. To prevent the frequency of words in a long document from being too high, the TF usually uses the document length to normalize the value, that is, $TF = \frac{TN}{DL}$, in which TN represents how many times the word T appears in a specific document D, and DL represents the length of document D. The Inverse Document Frequency (IDF) represents how many documents the word T has appeared in. The main idea of the TF-IDF is that, when the frequency of T in a document is very high (that is, TF is very large) while other documents containing T are very few (that is, IDF is huge), it indicates that T has a good ability to distinguish keyphrases. Among them, $IDF = \log(\frac{DN}{DC+1})$, where DN represents the total number of documents and DC represents the number of documents containing the word T.

KP-miner:

The TF-IDF is generally only used as a statistical method applied by other unsupervised keyphrase extraction methods to calculate keyphrases' importance. For example, El-Beltagy and Rafea proposed the KP-miner [27] in 2009. This method is a typical unsupervised keyphrase extraction method using the TF-IDF, divided into three steps. The first step is to select the candidate words from documents, the second step is to calculate the candidate words' score, and the third step is to select the candidate word with the highest score as the final keyphrase. KP-miner introduced two new statistical features in the candidate word selection stage. (i) The least allowable seen frequency (lasf) factor means that only words that appear more than n times in a document can be regarded as candidate words. (ii) CutOff is based on the fact that, if a word appears after a given threshold position in a long document, it will not be a keyphrase, which means the word appearing after CutOff will be filtered out. Finally, the final keyphrases are selected by combining the candidate words' positions and the TF-IDF score. Experiments show that the efficiency of the algorithm is higher than Extractor [51] and KEA.

YAKE:

Campos et al. proposed YAKE [28] in 2018, as a typical unsupervised keyphrase extraction method using the TF-IDF. The difference between YAKE and KP-miner is that it uses candidate word locations or TF-IDF information and introduces a new feature set, which contains five features. The Word Casing (WC) reflects the cases of the candidate words. The Word Position (WP) reflects the position of a word, which means the more often the word is in the front of the document, the greater its value. The Word Frequency (WF) reflects that the higher is the frequency of a word in a document, the greater is its value. The Word Relatedness to Context (WRC) indicates the number of different words appearing on both sides of a candidate word. The Word DifSentence (WD) indicates the frequency of a candidate word in different sentences. The five values are combined to calculate $S(w)$, as shown in the following formula.

$$S(w) = \frac{WR*WP}{WC+\frac{WF}{WRC}+\frac{WD}{WR}} \tag{8}$$

Finally, the final $S(kw)$ of each candidate word is calculated by using the 3-gram model, as shown in Formula (7).

$$S(kw) = \frac{\prod_{w \in kw} S(w)}{TF(kw) * (1 + \sum_{w \in kw} S(w))} \quad (9)$$

where kw represents the candidate word and TF represents the frequency of the keyphrase. The smaller is $S(kw)$, the more likely kw is to be a keyphrase.

3.3. Graph-Based Methods

The keyphrase extraction task is transformed into a graph sorting problem using a graph-based algorithm based on the basic assumption that more connections mean more important candidate words. The idea originated from PageRank [52] of Google with a basic idea of voting or recommendations, which means the graph's edges are considered votes. The more votes a node gets, the higher its score, and the more critical it is. Specifically, PageRank generates a directed graph containing all pages with a single page as a node. If there is a link pointing to B in web page A, node A in the graph has an edge pointing to B, regarded as A "voting" for B. The more votes a node receives the higher its score is and the higher its web page ranking. Moreover, the voting of high score nodes will contribute higher scores to the voted nodes [53]. Combining PageRank and word embedding [54], the performance on Chinese and English datasets exceeds TF-IDF and PositionRank.

TextRank:

Based on the idea of PageRank, Mihalcea and Tarau proposed TextRank [29] in 2004, which is the first algorithm to use PageRank for keyphrase extraction. The first thing TextRank does for a document is to delete the function words in the document. Only certain words with fixed parts of speech (such as adjectives and names) can be candidate words. The algorithm then links the selected candidate words according to the co-occurrence relationships between words to generate a directed and powerless graph. The initial score of each node is 1. If two words are within the window of w (w takes a random value from 2 to 20), the two words are connected by lines in the graph. Next, PageRank is run to calculate each node's final score, where the score of node VK is determined by the Formula (8). Finally, the document's consecutive candidate words will be connected into multi-word keyphrases, where the score is the sum of the scores of each candidate word, and the top-ranked candidate words are taken as the keyphrases.

$$S(Vk) = (1 - \alpha) + \alpha \sum_{m \in NB(Vk)} \frac{1}{|NB(Vm)|} S(Vm) \quad (10)$$

To prevent TextRank from encountering a dead cycle in the recursive computations, it introduces a damping factor (α). NB (vi) represents the neighboring node set of node vi.

SingleRank:

In view of the fact that the graphs constructed by TextRank are unweighted graphs and the weights of the edges can reflect the strength of the semantic relationship between the two nodes, using the weighted graph may be better in the keyphrase extraction task. Based on this assumption, Wan and Xiao proposed SingleRank [30] in 2008, which added weights on the basis of the TextRank between nodes appearing in the window of w at the same time, and the weight value was determined by the number of times the two words appeared in the window of w at the same time. The final score of nodes is determined by Formula (9), where $C(Vj, Vm)$ represents the number of times that nodeVj and node Vm appear together in a document.

$$S(Vk) = (1 - \alpha) + \alpha \sum_{m \in NB(Vk)} \frac{C(Vj, Vm)}{\sum_{Vk \in NB(Vm)} C(Vm, Vj)} S(Vm) \quad (11)$$

ExpandRank:

In 2008, based on SingleRank, Wan and Xiao proposed ExpandRank [30], which takes the neighboring documents in the same dataset into account to provide the background knowledge when extracting keyphrases from a specific document. Specifically, ExpandRank first uses vectors to represent the documents in the dataset. Next, it calculates the k neighboring documents similar to the extracted document d0 to form a $k + 1$ document set D. Then, it builds a global graph to assist in extracting the keyphrases by using D, where the edge weight WE (Vk, Vm) between nodes Vk and Vm in the global graph is determined by Formula (10). Sim (d0,di) represents the similarity of documents d0 and di and Fdi (Vk,Vm) represents the number of times that nodes Vk and Vm appear in document di at the same time. The efficiency of ExpandRank is not significantly better than that of SingleRank.

$$WE(Vk, Vm) = \sum_{di \in D} sim(d0, di) \cdot Fdi(Vk, Vm) \tag{12}$$

PositionRank:

Florescu et al. proposed PositionRank [55] in 2017, which introduces location information based on SingleRank according to the idea that the earlier the candidate words appear in a document, the more important they are. As shown in Formula (11), where each item in vector P represents the normalized location information of a candidate word, the final score of each candidate word can be calculated by bringing the location information of each node into Formulas (12) and (13), where pk is the kth element in P, that is, the ratio of the position of the kth candidate word to the sum of positions of all candidate words; w is the weight of the edge; and $adj(v)$ is the adjacent node of v.

$$P = [\frac{p1}{p1 + p2 + ... + pn}, \frac{p2}{p1 + p2 + ... + pn}, ..., \frac{pn}{p1 + p2 + ... + pn}] \tag{13}$$

$$S(Vk) = (1 - \alpha)pk + a \cdot \sum_{vm \in adj(Vk)} \frac{Wmk}{O(Vm)} S(Vm) \tag{14}$$

$$O(Vm) = \sum_{vi \in adj(Vm)} Wmi \tag{15}$$

Graph-based algorithms have some disadvantages. As far as multi-topics documents (such as news) are concerned, human language habits determine that a new topic will have corresponding new keyphrases. However, in graph-based methods, all candidate words (node) are uniformly sorted, and the node with the highest score is taken as the keyphrase. This does not completely guarantee that the keyphrases output by the algorithm can cover all topics, and it may cause the phenomenon that all the keyphrases describe the same topic [24], which is improved by topic-based methods.

3.4. Topic-Based Methods

Topic-based methods can be further divided into transfer learning-based methods and clustering-based methods.

3.4.1. Transfer Learning-Based Methods

Applying the knowledge acquired from one problem to another different but related problem is the primary motivation of transfer learning [56]. Common knowledge in keyphrase extraction includes Wikipedia [33] and citation networks [37]. Because some background knowledge is needed to classify candidate words in topic-based methods, transfer learning is widely used. The following introduces several mainstream transfer learning-based methods.

KeyCluster:

Applying the knowledge acquired from one problem to another different but related problem is the primary motivation of transfer learning [56]. Common knowledge in keyphrase extraction includes Wikipedia [33] and citation networks [37]. In 2009, Liu et al. proposed KeyCluster [25], divided into four steps. As with other methods, the first step is to preprocess the document, delete the function words, and use the remaining words as candidate words. The second step is to use the Wikipedia-based method to calculate the semantic relationships of candidate words. The Wikipedia-based method regards each word as a vector with each item being the TF-IDF value in Wikipedia. The correlation between the two words can be measured by comparing the vector representations of the two words. The third step is to group the candidate words based on these semantic relationships and find each group's exemplar. The fourth step is to extract the final keyphrases from the exemplar. The experimental results show that the performance of KeyCluster is better than TextRank, and the extracted keyphrases cover the whole document.

CommunityCluster:

In 2009, Grineva et al. proposed CommunityCluster [26] based on the assumption that the words related to the same topic are generally aggregated into a subgraph (or community), and the most connected subgraph generally corresponds to a theme of a document. CommunityCluster uses Girvan–Newman network analysis to detect communities and uses all words in the most closely connected communities as keyphrases. According to the experimental results, CommunityCluster is superior to the baseline system, such as TF-IDF, Yahoo!, and Wikify! [57], in precision and recall.

Topical PageRank (TPR):

In 2010, Liu et al. proposed TPR [24], which uses Wikipedia articles as resources to train the potential Dirichlet Distribution (LDA) and uses the trained LDA model to calculate the topic distribution of documents. Then, it uses PageRank for each topic, as shown in Formula (15), to calculate the topic-specific importance scores. Finally, it combines these scores to calculate the candidate words' total score and selects the top-ranked word as keyphrases. TPR, similar to KeyCluster, ensures that the extracted keyphrases cover the entire document. According to the experimental results, TPR is better than the baseline methods, such as TF-IDF and PageRank, in precision, recall, F-score, Bpref, MRR, and MR.

$$St(Vk) = \alpha \sum_{m:Vm \to Vk} \frac{Wmk}{O(Vm)} St(Vm) + (1-\alpha) pt(Vk) \quad (16)$$

Here, t represents a topic and pt represents the LDA distribution of t.

It is worth mentioning that the introduction of LDA makes each topic have different weights when using TPR, and topics with low weights may not output related keyphrases, which is more in line with human language habits. For example, when we write an article on natural language processing, we may use 20% of the content to describe human language habits, 70% of the content to write about how a computer deals with human language, and 10% to write other things, and this 10% may not be needed to extract keyphrases, which is a feature that KeyCluster does not have.

Single Topical PageRank:

Because the TPR needs to run PageRank once for each topic, its running efficiency is reduced. Based on this weakness of TPR, Sterckx et al. improved TPR in 2015 and proposed Single Topical PageRank (Single TPR) [58]. Single TPR only needs to run PageRank once for a document, which significantly improves the running efficiency on the premise of accuracy, especially when dealing with large datasets.

The topic-based method includes the use of transfer learning and the use of hierarchical aggregative clustering to complete the keyphrase extraction task.

3.4.2. Clustering-Based Methods

TopicRank:

In 2013, Bougouin et al. proposed topicRank [22] that is similar to TextRank using candidate words as graph nodes, as TopicRank uses topics as graph nodes. Specifically, TopicRank first uses hierarchical agglomerative clustering [33] to divide the document into multiple topics, uses PageRank to score each topic, then selects the first candidate word from each top-ranked topic, and finally uses all the selected candidate words as the keyphrases. According to the experimental results, the method makes the extracted keyphrases cover all topics, and the performance is better than TF-IDF, SingleRank, and TextRank in precision, recall, and F-score.

3.5. Language Model-Based Methods

Based on Kullback–Leibler (KL) divergence that can measure the loss of two language models, Tomokiya et al. used two kinds of datasets with different functions, foreground corpus and background corpus, to assist in keyphrase extraction [59]. The foreground corpus is the dataset for keyphrase extraction, while the background corpus provides background knowledge. Similar to TF-IDF, this method reflects each keyphrase's unique extent by using background knowledge and introduces two new features, namely phraseness and informativeness. Phraseness represents the extent to which a word sequence can be used as a phrase, while informativeness represents the extent to which the phrase can express a document's central idea. This method uses the n-gram model to learn these two features in the foreground corpus and the background corpus. The phraseness and informativeness determine the final scores of the candidate words.

In the above three types of methods (statistics-based methods, graph-based methods, and topic-based methods), each method often contains more than one idea. For example, TPR uses two ideas of topic and graph. This connection is described in detail in Figure 2.

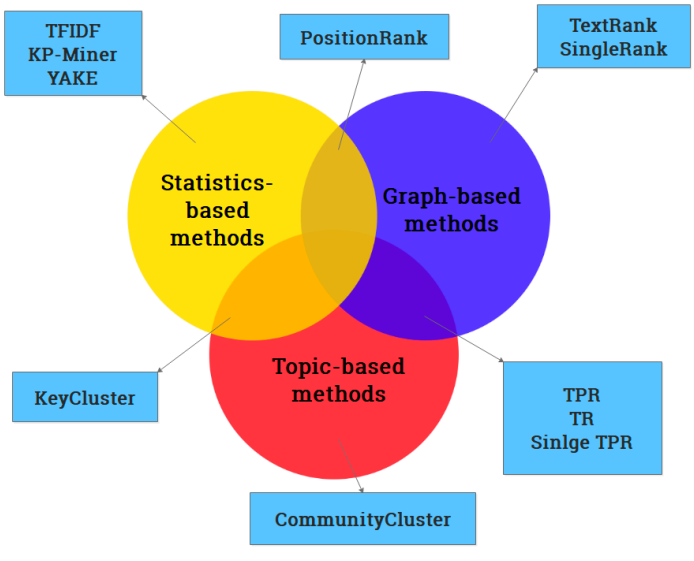


Figure 2. All methods are classified according to the technology applied. The overlapping part represents that the method uses multiple technologies.

4. The State of the Art

The performances of the mainstream unsupervised methods in the keyphrase extraction field are tested and analyzed in this section, including the statistics-based method TF-IDF; the graph-based

methods TextRank, SingleRank, and PositionRank; and the topic-based methods TPR and TR. Each algorithm outputs ten keyphrases. The experimental datasets selected are mainly from the abstracts of academic papers (KDD, WWW, and Nguyen) and news (DUC-2001), which are benchmark datasets in this field, as detailed in Section 2.1. The experimental results are shown in Table 4.

Table 4. Scores achieved on various datasets (P, R, and F1 are the abbreviations of precision, recall, and F1-score, respectively).

Dataset	Method	P%	R%	F1%	Dataset	Method	P%	R%	F1%
DUC	TF-IDF	9.4	12.4	10.6	KDD	TF-IDF	9.2	23.3	12.9
	TextRank	11.1	14.1	12.2		TextRank	6.1	13.6	7.9
	SingleRank	21.5	27.4	23.8		SingleRank	7.1	16.1	9.2
	PositionRank	18.9	24.8	21.2		**PositionRank**	**9.6**	**24.3**	**13.4**
	TR	18.2	23.3	20.2		TR	6.3	14.1	8.2
	TPR	**22.3**	**28.2**	**24.6**		TPR	7.2	16.5	9.5
Nguyen	TF-IDF	10.6	26.2	14.8	WWW	TF-IDF	9.7	21.6	13.2
	TextRank	6.7	16.6	9.3		TextRank	6.6	13.3	8.1
	SingleRank	8.4	21.2	11.8		SingleRank	7.6	15.4	9.5
	PositionRank	**11.3**	**28.1**	**15.7**		**PositionRank**	**11.1**	**23.7**	**14.4**
	TR	8.1	19.6	11.1		TR	7.3	14.1	8.9
	TPR	8.7	21.8	12.1		TPR	7.9	16.1	9.9

In terms of the abstracts of academic papers (KDD, WWW, and Nguyen), it can be found using the precision, recall, and F1-score that PositionRank has the best performance, followed by TF-IDF, TPR, SingleRank, TR, and TextRank, among which TPR and SingleRank are almost the same, as shown in Figure 3.

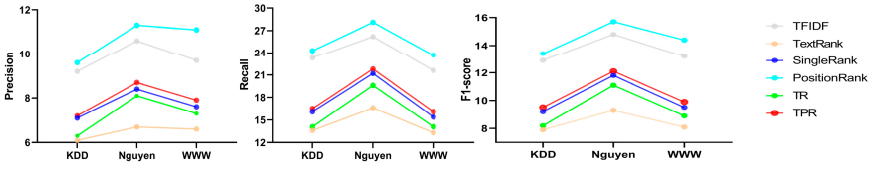


Figure 3. The performance of state of the art evaluated on KDD, Nguyen, and WWW, which are all paper abstract datasets (single topic). Overall, the performance of PositionRank and TF-IDF is higher than other methods (see Section 5.1 for more details). Precision, recall, and F1-score are introduced in Section 2.2.

5. Analysis

Although some researchers have introduced the performance differences of each method in their work, as far as we know, they have not pointed out why [26]; thus, in this section, we analyze the performance of each keyphrase extraction method from two perspectives. First, from each method's characteristics, we analyze why their performance is different (Section 5.1). Secondly, we start from the characteristics of the datasets themselves and show how different datasets affect the performance of the methods (Section 5.2).

5.1. The Performance of the Methods

Overall, the performance of PositionRank and TF-IDF is higher than other methods. We infer that, compared with other methods, these two methods use statistical data to reflect the document's

important content better. The performance of the PositionRank is higher than that of TF-IDF because PositionRank not only uses statistical data but also uses a graph method, which further reflects the structure of the document.

In the graph-based method, SingleRank adds the weight of the edge in constructing the graph based on TextRank, and the weight of the edge can reflect the strength of the semantic relationship between the two words so that it can increase the performance. Based on SingleRank, PositionRank takes the position information of words into account, which is in line with human language habits that important words often appear in the front of the article, which improves the algorithm's accuracy.

In the topic-based methods, TR uses the given document itself (single document) for keyphrase extraction, while TPR trains the LDA model using Wikipedia to learn the relevant knowledge and then uses the trained model to extract keyphrases. It is evident that Wikipedia resources are much greater than a single document, so TPR has better performance than TP.

5.2. The Impacts of Dataset on Performance

While using the news dataset (DUC-2001) to test each method's performance, it is found that the methods show different performance compared with the paper's abstract dataset, as shown in Figure 4.

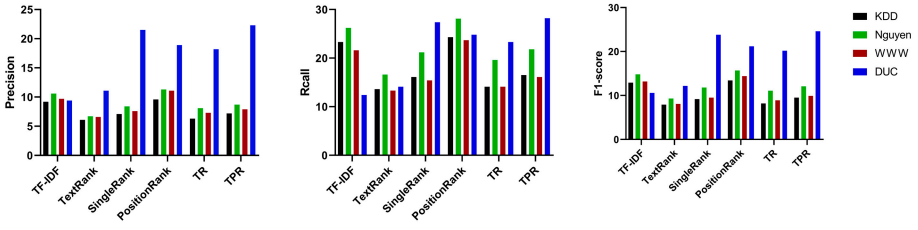


Figure 4. The performance of state of the art evaluated on KDD, Nguyen, and WWW, which are all paper abstract datasets (single topic), and DUC-2001, which is a news dataset (multiple topics). The topic-based method (TR,TPR) greatly improves the performance on multi-topic datasets (see Section 5.2 for more details). Precision, recall and F1-score are introduced in Section 2.2.

Th precision and F1-score change slightly for the TF-IDF method, but the recall is significantly reduced. For example, the precision, recall, and F1-score of TF-IDF on the KDD, Nguyen, and WWW datasets are 9.2–10.6% (9.8% on average), 21.6–26.2% (23.7% on average), and 12.9–14.8% (13.6% on average), respectively. The precision, recall, and F1-score on DUC-2001 are 9.4%, 12.4%, and 10.6%, respectively, which is decreased by 4%, 40%, and 22%, respectively. Similarly, in the graph-based methods, TextRank and PositionRank have significantly higher precisions and F1-scores, but there is little change in the recall. Specifically, the two methods, respectively, experienced 71% and 77% higher precisions, 45% and 46% higher F1-scores, and 3% and 2% lower recall. SingleRank significantly improved in all evaluation indexes, including its precision, recall, and F1-score increasing by 180%, 56%, and 133%, respectively. In the topic-based methods, th e TP and TPR have significantly improved in all evaluation indexes. Specifically, TR had a 152%, 47%, and 115% higher precision, recall, and F1-score, respectively, and TPR had a 182%, 56%, and 146% higher precision, recall, and F1-score, respectively.

The following conclusions can be drawn. Because news articles usually contain more topics than paper abstracts and the recall can reflect the comprehensiveness of a method, as described in Section 2.2, the recall of TF-IDF on multi-topic news datasets is reduced, which shows that this method cannot effectively cover all topics of an article, and the extracted keyphrases are not comprehensive enough. Further analysis shows that this is because the TF-IDF holds that, if a keyphrase appears more often in a document and less frequently in other documents, it will get a higher score. However, the numbers of keyphrases corresponding to each topic in multi-topic documents are often different (related to the topic); therefore, the candidate words in important topics may cover all the final keyphrases,

while the keyphrases corresponding to unimportant topics are ignored. As a result, this method cannot guarantee the comprehensiveness of the keyphrase extraction. Based on this, it can be concluded that the TF-IDF is more suitable for the keyphrase extraction of single topic documents.

In the graph-based methods, TextRank calculates each candidate word's score using PageRank, which reflects the semantic relationships between words. Therefore, the precision and F1-score are improved for the multi-topic datasets. However, similar to the TF-IDF, TextRank cannot cover all topics well and guarantee the comprehensiveness of the extracted keyphrases, and so the recall does not change much. For SingleRank, it adds weight to the edges based on TextRank, strengthening the semantic connection between words. Therefore, the precision, F1-score, and recall have all been greatly improved. PositionRank considers the first occurrence of words based on SingleRank, but it has lower performance because the essential words in the abstract will appear in front of the article, but the news does not have this feature. It can be concluded that TextRank and PositionRank are more suitable for the keyphrase extraction of single topic documents, while SingleRank can be used for single documents and multi-topic documents.

The topic-based methods (TR and TPR) have a considerable advantage in the news datasets because the main idea of the TPR is to divide a document into several topics through LDA, calculate the score of each keyphrase corresponding to the related topics, and select the final keyphrase based on these scores. Moreover, TR divides the document into multiple topics using hierarchical agglomerative clustering, uses PageRank to score each topic, and finally selects the top topics' final keyphrases. TPR and TR have well reflected the semantic associations between topics in documents; therefore, their performances have been greatly improved, and there is no doubt that these two methods are more suitable for the keyphrase extraction of multi-topic documents.

6. Limitation of Keyphrase Extraction Methods

The limitation of the keyphrase extraction task makes various unsupervised methods unable to complete the task well. One is the impact of the "gold standard" problem on evaluation (Section 6.1), and the second is due to the habit of artificially annotating the datasets, resulting in the algorithm being unable to extract keyphrase that is consistent with the artificial annotation label (Section 6.2). Based on these two limitations, we discuss the possible ways to solve these problems and provide some new features that may improve the keyphrase extraction method's performance in Section 6.3.

6.1. The Impact of Gold Standard on Evaluation

The precision, recall, and F-score have a common disadvantage that each extracted keyphrase is considered correct only when it is entirely consistent with the gold standard keyphrase, which will cause two problems. The first is that the extracted keyphrase has the same root but different expression forms with the gold standard keyphrase. For example, if the gold standard keyphrase is "concept of wealth", the algorithm will not use the "conception of wealth" as a keyphrase, which is not what we want. This phenomenon is called the exact match problem, which can be solved using a stemmer. The keyphrases can be reduced to their root form and then compared with the stemmer. The second is that the extracted keyphrase and gold standard keyphrase have the same semantics but different words. For example, the gold standard keyphrase is "data processing", while the algorithm will not use "data handling" as the keyphrase. At present, there is no right solution.

6.2. The Impact of Manually Assigned Labels on Evaluation

The current keyphrase extraction method takes the words or phrases in an article as the final keyphrases. However, some keyphrases that are manually assigned are not the words appearing in the original document (it may be a summary of the semantics of content in the original document), and the keyphrase extraction method cannot extract such keyphrases, which is the limitation of the method. On some datasets, the error caused by this problem can reach 52–73%. For example, Figure 5 is a document from WWW, which contains five keyphrases: link analysis, newsgroup, social network,

text mining, and web mining. Among them, only "newsgroup" has appeared in the original document, while the other keyphrases are summary expressions of the meaning of the content. The blue sentence in the figure is the sentence expressing the keyphrases, the blue italics and bold words are the keyphrases for which their original words did not appear in the document, and the red italics and bold words are the keyphrases for which original words appear in the document.

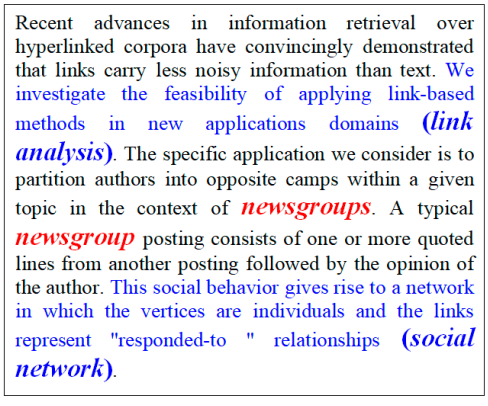


Figure 5. The impact of manually assigned labels on evaluation. Only "newsgroups" can be extracted by the algorithm.

6.3. Our Recommendations

For the impact of the gold standard on evaluation, we need to introduce external knowledge to determine whether the keyphrases extracted by a method and the gold keyphrase have the same semantics. For example, we can use external resources to train Word2vec [54], a commonly used model in the NLP field, to assist in the task of keyphrase extraction. Specifically, we can use the trained Word2vec model to convert each extracted keyphrase into an embedding vector, and then compare the similarity with the gold keyphrase. If the similarity is higher than a certain threshold, the extraction can be considered successful.

Regarding the influence of manually assigned labels on evaluation, on the one hand, through observation of the datasets, we find that the original words can directly express many keyphrases in the document without artificial summary. Thus, we can start from this aspect when constructing the dataset and select the words that have appeared in the article as the gold keyphrase. On the other hand, similar to solving the gold standard problem, we can also introduce external knowledge into the method. For example, we can use Word2vec to understand the meaning of the sentence by constructing the embedding vector and then find the words that are most relevant to the meaning of the sentence as keyphrases.

For these recommendations, we have done some experiments for future researchers to study further. We use Word2vec of gensim (a Python library) to try to solve the "gold standard" and "manually assigned labels" problems.

We use PositionRank to test on the WWW dataset. Unlike the usual evaluation method, the extracted keyphrases and gold keyphrases will not be restored to the root form with a stemmer, nor will they be directly matched with strings. We use the trained Word2vec model to convert the extracted keyphrase and gold keyphrase into vectors, and then calculate the Euclidean distance or cosine similarity of the two vectors. If the Euclidean distance is less than a certain threshold or the cosine similarity is higher than a certain threshold, we manually compare whether the words corresponding to the two vectors have similar semantics. If so, the extracted keyphrases and gold keyphrases are considered to match.

The experimental results show that using Word2vec to evaluate the performance of the method can better reflect the similarity of extracted keyphrases than the conventional evaluation metrics (precision, recall, and F1-score), and it can also help extraction method to extract keyphrases with different from but similar semantics with gold keyphrases.

However, we also encounter some difficulties in setting a threshold (Euclidean distance or cosine similarity) to determine whether an extracted keyphrase can be considered as a gold keyphrase because the threshold is an empirical parameter that requires much labor to compare the semantic similarity between the extracted keyphrases and the gold keyphrases. We will continue to study this idea further.

7. Conclusions and Future Directions

In this paper, the unsupervised learning methods in the field of keyphrase extraction are summarized, and the performance of each method on different datasets and the reasons for the performance difference are analyzed in detail, to help future researchers understand mainstream solutions in the field of keyphrase extraction from multiple perspectives.

The reasons for the limitations of the keyphrase extraction field are pointed ("gold standard" and "manually assigned labels"), and our recommendations to solve these problems are proposed.

To help researchers further improve the performance of the method in the keyphrase extraction task, we introduce some new features that may be helpful.

The relative position of words: It is found that many methods, such as PositionRank and YAKE, have introduced the position information of the word, that is, the positions of the word in the full document. However, for long texts with multiple topics, keyphrase will appear with the appearance of new topics, so some keyphrase will appear in topics that are located later, that is, the keyphrase will be relatively far away from the beginning of the article and the previous location-based features would not give these words high weight. Based on this discovery, instead of using an entire document as a reference, we can use each paragraph as an independent unit to calculate the position of the word from the beginning of the paragraph.

The role of conjunctions: It is worth mentioning that, if the second clause of two comma-connected clauses starts with a conjunction, there is likely to be some semantic relationship between the two sentences [60]. Based on this discovery, paying attention to the function and position of conjunctions in the keyphrase extraction process may help to improve the performance.

Design method for different types of datasets: In Section 5.2, we show how the datasets affect the performance of methods; thus, in future work, researchers should design different methods based on the characteristics of the datasets.

Author Contributions: Conceptualization, L.C. and C.S.; methodology, C.S.; software, C.S.; validation, C.S., L.C.; formal analysis, C.S.; investigation, T.L.; resources, T.L.; data curation, T.L.; writing—original draft preparation, C.S.; writing—review and editing, S.L.; visualization, S.L.; supervision, L.C.; project administration, H.L.; funding acquisition, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Plan of China under Grant No. 2017YFA0604500, and by National Sci-Tech Support Plan of China under Grant No. 2014BAH02F00, and by National Natural Science Foundation of China under Grant No. 61701190, and by Youth Science Foundation of Jilin Province of China under Grant No. 20160520011JH & 20180520021JH, and by Youth Sci-Tech Innovation Leader and Team Project of Jilin Province of China under Grant No. 20170519017JH, and by Key Technology Innovation Cooperation Project of Government and University for the whole Industry Demonstration under Grant No. SXGJSF2017-4, and by Key scientific and technological R&D Plan of Jilin Province of China under Grant No. 20180201103GX and by Project of Jilin Province Development and Reform Commission under Grant No. 2019FGWTZC001.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Babar, S.A.; Patil, P.D. Improving Performance of Text Summarization. *Procedia Comput. Sci.* **2015**, *46*, 354–363. [CrossRef]
2. Welleck, S.; Brantley, K.; Daumé, H., III; Cho, K. Non-Monotonic Sequential Text Generation. *arXiv* **2019**, arXiv:1902.02192.
3. Welleck, S.; Kulikov, I.; Roller, S.; Dinan, E.; Cho, K.; Weston, J. Neural Text Generation with Unlikelihood Training. *arXiv* **2019**, arXiv:1908.04319.
4. Puduppully, R.; Dong, L.; Lapata, M. Data-to-Text Generation with Content Selection and Planning. *arXiv* **2019**, arXiv:1809.00582.
5. Shen, S.; Fried, D.; Andreas, J.; Klein, D. Pragmatically Informative Text Generation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4060–4067.
6. Mallett, D.; Elding, J.; Nascimento, M.A. Information-content based sentence extraction for text summarization. In Proceedings of the International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, 5–7 April 2004; IEEE: Las Vegas, NV, USA, 2004; Volume 2, pp. 214–218 .
7. Bougouin, A.; Boudin, F.; Daille, B. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In Proceedings of the IJCNLP, Nagoya, Japan, 14–18 October 2013.
8. Liu, Z.; Liang, C.; Sun, M. Topical Word Trigger Model for Keyphrase Extraction. Available online: https://www.aclweb.org/anthology/C12-1105.pdf (accessed on 12 June 2020).
9. Azad, H.K.; Deepak, A. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.* **2019**, *56*, 1698–1735. [CrossRef]
10. Guo, J. A Deep Look into Neural Ranking Models for Information Retrieval. *arXiv* **2019**, arXiv:1903.06902.
11. Gutierrez, C.E.; Alsharif, M.R. A Tweets Mining Approach to Detection of Critical Events Characteristics using Random Forest. *Int. J. Next Gener. Comput.* **2014**, *5*, 167–176.
12. Gutierrez, C.E.; Alsharif, M.R.; He, C.; Khosravy, M.; Villa, R.; Yamashita, K.; Miyagi, H. Uncover news dynamic by principal component analysis. *ICIC Express Lett.* **2016**, *7*, 1245–1250.
13. Dave, K.; Lawrence, S.; Pennock, D. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the WWW'03, Budapest, Hungary, 20–24 May 2003.
14. Hemmatian, F.; Sohrabi, M.K. A survey on classification techniques for opinion mining and sentiment analysis. *Artif. Intell. Rev.* **2019**, *52*, 1495–1545. [CrossRef]
15. Asghar, M.Z.; Khan, A.; Zahra, S.R.; Ahmad, S.; Kundi, F.M. Aspect-based opinion mining framework using heuristic patterns. *Cluster. Comput.* **2019**, *22*, 7181–7199. [CrossRef]
16. Frank, E.; Paynter, G.W.; Witten, I.; Gutwin, C.; Nevill-Manning, C. Domain-Specific Keyphrase Extraction. In Proceedings of the IJCAI, Stockholm, Sweden, 31 July–6 August 1999.
17. Hulth, A.; Megyesi, B.B. A Study on Automatically Extracted Keywords in Text Categorization. Available online: https://www.aclweb.org/anthology/P06-1068.pdf (accessed on 15 June 2020).
18. Turney, P.D. *Learning to Extract Keyphrases from Text*; Technical Report; National Research Council, Institute for Information Technolog: Ottawa, ON, Canada, 2002.
19. Witten, I.H.; Paynter, G.W.; Frank, E.; Gutwin, C.; Nevill-Manning, C.G. KEA: Practical Automatic Keyphrase Extraction. In Proceedings of the Fourth ACM Conference on Digital Libraries, Berkeley, CA, USA, 11–14 August 1999.
20. Wang, R.; Wang, G. Web Text Categorization Based on Statistical Merging Algorithm in Big Data. *Environ. Int. J. Ambient. Comput. Intell.* **2019**, *10*, 17–32. [CrossRef]
21. Gutierrez, C.E.; Alsharif, M.R.; Khosravy, M.; Yamashita, K.; Miyagi, H.; Villa, R. Main Large Data Set Features Detection by a Linear Predictor Model. Available online: https://aip.scitation.org/doi/abs/10.1063/1.4897836 (accessed on 20 June 2020).
22. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
23. Kim, S.N.; Kan, M.Y. Re-examining automatic keyphrase extraction approaches in scientific articles. In Proceedings of the Workshop on Multiword Expressions Identification, Interpretation, Disambiguation and Applications—MWE'09, Singapore, 25–27 November 2009; p. 9.

24. Liu, Z.; Huang, W.; Heng, Y.; Sun, M. *Automatic Keyphrase Extraction via Topic Decomposition. Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Singapore, 2010.
25. Liu, Z.; Li, P.; Zheng, Y.; Sun, M. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing—EMNLP'09, Singapore, 6–7 August 2009; Association for Computational Linguistics: Singapore, 2009; Volume 1, p. 257.
26. Hasan, K.S.; Ng, V. Automatic Keyphrase Extraction: A Survey of the State of the Art. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 23–25 June 2014; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 1262–1273.
27. El-Beltagy, S.R.; Rafea, A. KP-Miner: A keyphrase extraction system for English and Arabic documents. *Inf. Syst.* **2009**, *34*, 132–144. [CrossRef]
28. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.M.; Nunes, C.; Jatowt, A. YAKE! Collection-Independent Automatic Keyword Extractor. In *Advances in Information Retrieval*; Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 10772, pp. 806–810, ISBN 978-3-319-76940-0.
29. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, 25–26 July 2004.
30. Wan, X.; Xiao, J. CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In Proceedings of the 22nd International Conference on Computational Linguistics—COLING'08, Manchester, UK, 18–22 August 2008; Association for Computational Linguistics: Manchester, UK, 2008; Volume 1, pp. 969–976.
31. Papagiannopoulou, E.; Tsoumakas, G. A review of keyphrase extraction. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, e1339. [CrossRef]
32. Krapivin, M.; Autayeu, A.; Marchese, M. *Large Dataset for Keyphrases Extraction*; Technical Report DISI-09-055; DISI, University of Trento: Trento, Italy, 2009.
33. Medelyan, O.; Frank, E.; Witten, I.H.Human-competitive tagging using automatic keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing—EMNLP'09, Singapore, 6–7 August 2009; Association for Computational Linguistics: Singapore, 2009; Volume 3, p. 1318.
34. Nguyen, T.D.; Kan, M.-Y. Keyphrase Extraction in Scientific Publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*; Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4822, pp. 317–326, ISBN 978-3-540-77093-0.
35. Schutz, A. Keyphrase Extraction from Single Documents in the Open Domain Exploiting Linguistic and Statistical Methods. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.394.5372&rep=rep1&type=pdf (accessed on 30 June 2020).
36. Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan, 11–12 July 2003; Volume 10, pp. 216–223.
37. Gollapalli; Sujatha, D.; Cornelia, C. Extracting Keyphrases from Research Papers Using Citation Networks. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 1629–1635.
38. Jiang, M.; Chen, Y.; Liu, M.; Rosenbloom, S.T.; Mani, S.; Denny, J.C.; Xu, H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 601–606. [CrossRef] [PubMed]
39. Bougouin, A.; Barreaux, S.; Romary, L.; Boudin, F.; Daille, B. TermITH-Eval: A French Standard-Based Resource for Keyphrase Extraction Evaluation. In Proceedings of the Language Resources and Evaluation Conference (LREC), Portorož, Slovenia, 23–28 May 2016.
40. Wan, X.; Xiao, J. Single Document Keyphrase Extraction Using Neighborhood Knowledge. Available online: https://www.aaai.org/Papers/AAAI/2008/AAAI08-136.pdf (accessed on 25 June 2020).
41. Marujo, L.; Gershman, A.; Carbonell, J.; Frederking, R.; Neto, J.P. Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 23–25 May 2012.

42. Marujo, L.; Viveiros, M.; Neto, J.P. Keyphrase Cloud Generation of Broadcast News. In Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy, 27–31 August 2011.
43. Grineva, M.; Grinev, M.; Lizorkin, D. Extracting Key Terms from Noisy and Multitheme Documents. Available online: https://dl.acm.org/doi/abs/10.1145/1526709.1526798 (accessed on 26 June 2020).
44. Hammouda, K.M.; Matute, D.N.; Kamel, M.S. CorePhrase: Keyphrase Extraction for Document Clustering. In *Machine Learning and Data Mining in Pattern Recognition*; Perner, P., Imiya, A., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3587, pp. 265–274, ISBN 978-3-540-26923-6.
45. Voorhees, E.M.The TREC-8 question answering track report. In Proceedings of the Eighth Text Retrieval Conference, TREC 1999, Gaithersburg, MD, USA, 17–19 November 1999.
46. Buckley, C.;Voorhees, E.M. Retrieval Evaluation with Incomplete Information. Available online: https://dl.acm.org/doi/abs/10.1145/1008992.1009000 (accessed on 18 June 2020).
47. Blei, D.M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
48. Halliday, M.A.K.; Matthiessen, C.M.I.M. *An Introduction to Functional Grammar*, 3rd ed.; Distributed in the United States of America by Oxford University Press: Oxford, MS, USA, 2004; ISBN 978-0-340-76167-0.
49. Johansson, V. Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective. Working Papers. Available online: https://www.semanticscholar.org/paper/Lexical-diversity-and-lexical-density-in-speech-and-Johansson/f0ec9ed698d5195220f80b732e30261eafbe1ad8?p2df (accessed on 10 June 2020).
50. Yih, W.; Goodman, J.; Carvalho, V.R. Finding advertising keywords on web pages. In Proceedings of the 15th International Conference on World Wide Web—WWW'06, Edinburgh, UK, 23–26 May 2006; p. 213.
51. Turney, P.D. Learning Algorithms for Keyphrase Extraction. *Inf. Retr.* **2000**, *2*, 303–336. [CrossRef]
52. Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Available online: http://ilpubs.stanford.edu:8090/422/ (accessed on 5 June 2020).
53. Wang, H.; Ye, J.; Yu, Z.; Wang, J.; Mao, C. Unsupervised Keyword Extraction Methods Based on a Word Graph Network. *Int. J. Ambient. Comput. Intell.* **2020**, *11*, 68–79. [CrossRef]
54. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. Available online: https://ui.adsabs.harvard.edu/abs/2013arXiv1301.3781M/abstract (accessed on 1 June 2020).
55. Florescu, C.; Caragea, C. PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1105–1115.
56. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
57. Mihalcea, R.; Csomai, A. Wikify!: Linking documents to encyclopedic knowledge. In Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management—CIKM'07, Lisbon, Portugal, 6–10 November 2007; p. 233.
58. Sterckx, L.; Demeester, T.; Deleu, J.,; Develder, C. Topical word importance for fast keyphrase extraction. Presented at the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, 18–22 May 2015; pp. 121–122. [CrossRef]
59. Tomokiyo, T.; Hurst, M. A language model approach to keyphrase extraction. In Proceedings of the ACL 2003 Workshop on Multiword Expressions Analysis, Acquisition and Treatment, Sapporo, Japan, 12 July 2003; Volume 18, pp. 33–40.
60. Jernite, Y.; Bowman, S.R.; Sontag, D. Discourse-Based Objectives for Fast Unsupervised Sentence Representation Learning. *arXiv* **2017**, arXiv:1705.00557.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Symmetry Editorial Office
E-mail: symmetry@mdpi.com
www.mdpi.com/journal/symmetry

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-0365-4022-1