

Article

Combining Content-Based Filtering Methods for Building a Powerful Hybrid Recommender System to Improve Tourism in Drâa-Tafilalet Area

Khalid al Fararni ^{1,*}, Loukmane Maada ^{2,*}, Badraddine Aghoutane ², Abdelouahed Sabri ¹ , Ali Yahyaouy ¹ 
and Jamal Riffi ¹

¹ LIIAN Laboratory, Department of Informatics, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez 30000, Morocco

² IA Laboratory, Faculty of Sciences, Moulay Ismail University, P.O. Box 11201, Zitoune, Meknes 50070, Morocco

* Correspondence: khalid.alfarni@usmba.ac.ma (K.a.F.); l.maada@edu.umi.ac.ma (L.M.)

Abstract

This paper proposes a hybrid content-based recommender system aimed at enhancing personalized tourism experiences and supporting the promotion of tourism in Morocco, with particular emphasis on the Drâa-Tafilalet region. The proposed approach integrates three machine learning models—Decision Tree, k-nearest neighbors, and Support Vector Machine—to predict user ratings for historical tourism sites. Tourist attraction metadata and user-generated reviews are represented using TF-IDF vectorization, while the predictions produced by the individual models are combined through an inverse-error weighting strategy. The system is evaluated using two datasets: a subset of the Yelp dataset comprising reviews of historical buildings in the United States, and a regional Drâa-Tafilalet dataset. Experimental results, assessed using MAE and RMSE, indicate that the weighted hybrid model outperforms the individual recommendation models by achieving lower prediction errors. These findings demonstrate the potential of hybrid content-based recommendation approaches to improve the accuracy of personalized tourism recommendations, support tourist decision-making, and promote underrepresented regional destinations.

Keywords: recommendation systems; artificial intelligence; hybrid content-based filtering; machine learning; trip planning

1. Introduction

Information and communication technology (ICT) has significantly transformed the interactions between travelers and the tourism industry. The advent of online travel agencies (OTAs), such as Expedia and Booking, has fundamentally altered the process of trip planning. Online consultations of tourist information, accommodation bookings, ticket purchases for cultural or entertainment events, and interactive exchanges via social networks, forums, and blogs have become prevalent. OTAs now serve as substantial sources of tourism-related data, encompassing both reservations and customer information. Nevertheless, the abundance of available data can render travel planning a complex and time-consuming endeavor. Tourism recommender systems offer assistance to tourists by efficiently navigating data resources and identifying points of interest (POIs) most appropriately aligned with their preferences.



Academic Editor: Manoj Gupta

Received: 5 June 2026

Revised: 28 June 2026

Accepted: 2 July 2026

Published: 9 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Although recommender systems have existed for some time, organizations have adopted advanced methodologies to encourage sustained customer engagement, improve business forecasting, monitor emerging crises, and facilitate continual improvement.

Recommendation systems powered by artificial intelligence (AI) address limitations inherent in earlier models, which often relied on static methods and demanded significant manual intervention. Contemporary AI-based recommendation systems deliver personalized suggestions to travelers based on stated requirements, preferences, and constraints. These technological advancements have facilitated the discovery of new POIs, the personalization of itineraries, and the overall optimization of tourism experiences.

This research contributes to the advancement of tourism in Morocco, particularly in the Drâa-Tafilalet region, by proposing a weighted hybrid content-based recommender system that leverages machine learning to generate personalized recommendations for points of interest [1,2]. The system combines several predictive techniques to improve recommendation accuracy and provides users with curated suggestions for historical sites, events, accommodations, and tourism offers according to their preferences and requirements. By evaluating the proposed approach on both a public dataset and a regional Drâa-Tafilalet dataset, this study highlights the potential of AI-driven recommender systems to enhance tourist experience and promote underrepresented regional destinations.

The rest of the paper is organized as follows. Section 2 defines recommender systems and introduces different approaches. Section 3 covers various Machine Learning techniques widely used in tourism, along with related work. Section 4 presents our proposed solution and discusses experimental results. Section 5 concludes the study and lists potential future research.

2. Recommender Systems

According to R. Burke and L. Cui [3], a recommender system is a computational tool that can guide users to relevant products, services, or information based on their preferences or past behavior. The term Tourist Recommender System (TRS) is used in the tourism sector to refer to such systems. The TRS is developed to enhance the tourist experience through the provision of personalized information. Specifically, a recommendation engine selects points of interest (POIs) deemed most pertinent and suitable for a given visitor, based on activities that most closely align with their profile. Recommender systems are generally categorized by the data used, the algorithms implemented, and the methodologies used to generate recommendations. The primary categories include collaborative techniques, content-based methods, and hybrid approaches [4–6]. A content-based recommendation method employs item features to formulate recommendations, functioning analogously to an information retrieval system that analyzes user profiles to suggest items based on prior queries or interactions [1,2]. This approach is predicated on the principle that consumers prefer items similar to those they have previously enjoyed. Content-based methodologies infer the preferences of a target user by examining similarities between the characteristics of unrated items and those of items already evaluated. User feedback is further utilized to iteratively refine user profiles [3]. Collaborative filtering, also prevalent in the tourism domain [2,7,8], as well as for content such as music and film that lack sufficient metadata for traditional description, operates on the assumption that users with shared historical preferences will appreciate similar items in the future [9]. The two most prominent collaborative filtering strategies are memory-based and model-based approaches [10,11]. The memory-based technique compares the target user's historical records with other information in the database [10] and can be further classified into item-based and user-based [12,13]. The model-based technique establishes a predictive model for the user, employing statistical or machine learning algorithms such as decision trees,

Bayesian networks, or SVMs [11]. A hybrid approach combines multiple recommender techniques to improve system efficacy [14].

3. Use of AI Techniques in Content-Based Recommender Systems

In response to evolving tourist needs, various artificial intelligence approaches have been introduced and implemented in tourism recommender systems. In their paper entitled “Artificial Intelligence in Tourism: A Review and Bibliometric Research” [15], the authors conducted a comprehensive review of 102 studies on recommender systems that use AI techniques. The majority (80.17%) of these studies adopted supervised learning approaches, utilizing 156 distinct algorithms. Conversely, 19.83% employed unsupervised learning algorithms. The most frequently employed algorithms included Decision Trees, Bayesian Classifier, Singular Value Decomposition (SVD) or Latent Factor Analysis, Ensemble Learning (i.e., combining multiple algorithms), and clustering algorithms. While these techniques have demonstrably enhanced the effectiveness of content-based tourism recommender systems [16,17], contemporary literature reveals a crucial research gap: the vast majority of existing frameworks rely on static, pre-determined blending weights or unweighted heuristic combinations. This leaves them poorly equipped to adapt to how individual algorithms perform across different text-driven features. To address this limitation, this section details the underlying mechanisms of the constituent algorithms used in our framework, actively contextualizing their specific limitations within tourism applications to justify our dynamic, performance-optimized hybridization strategy.

3.1. Decision Trees

Decision trees are predictive modeling techniques prevalent in machine learning, data mining, and statistics, and they address both regression and classification problems. These supervised learning algorithms employ tree-structured models to estimate an item's class or score using decision rules derived from training data. To determine an item's class label, processing commences at the tree's root and proceeds by comparing the item's attribute values with those at each subsequent node, following the corresponding branches [16]. P. Thiengburanathum et al. [18] advocate for a model-based destination recommender system utilizing a Decision Tree classifier to provide tourist recommendations, supported by a two-step feature selection process aimed at eliminating redundant information. The system is evaluated with extensive data collected from a case-study city to validate scalability. In our proposed framework, the decision tree's explicit rule-based predictions are balanced within a linear ensemble to add structural diversity against margin-based and instance-based models.

3.2. *k*-Nearest-Neighbor

The *k*-nearest neighbor method, which assesses the group membership of neighboring data points, serves as an established approach for data point classification. When examining a data point within a grid, the algorithm evaluates surrounding data points to determine group affiliation [19]. The proximity threshold for neighboring data points is selected to ensure a representative subset of the data. If the majority of nearby points are in group A, the likelihood increases that the examined data point also belongs to group A, and conversely for group B [20]. Typically, the *k*-nearest neighbors (*k*-NN) algorithm identifies the *k* most relevant neighbors—users or items—with the strongest correlation, to generate accurate recommendations [21]. In content-based tourism recommendation, *k*-NN excels at discovering localized metadata similarities between historical destinations. However, it scales poorly with dense feature sets and is highly sensitive to irrelevant features or noisy user reviews. We mitigate this localized bias by piping KNN outputs

into our inverse-error hybridization pipeline, allowing global margin boundaries to correct neighbor-proximity errors.

3.3. Support Vector Machine

Classification is a machine learning technique in which the goal is to predict the group (category) to which each data sample belongs [22]. This is an example of supervised learning, where the algorithm learns from labeled examples to make future predictions. An SVM classifier is a machine learning algorithm used for regression and classification. It is widely used across various fields, including image and text classification, bioinformatics, and more. This algorithm aims to determine an optimal hyperplane in a multidimensional space that maximizes the margin between examples of distinct classes. The hyperplane is determined using a set of training samples, namely the points nearest to the decision boundary, known as support vectors. Miriam Pirra et al. [22] proposed a novel approach to identifying trends in travel mode selection using the SVM classifier. The New York State dataset of tour-based demand travel is examined; it comes from the 2009 U.S. National Household Travel Survey. All trips' primary characteristics include travel, modes of transportation, and socioeconomic elements. The results show that it is possible to predict which tours people are most likely to take by non-motorized or public transportation in real-world situations where cars are the primary mode of transportation. SVMs are highly effective at navigating the sparse, high-dimensional vector spaces typical of TF-IDF text features extracted from tourism reviews. Nevertheless, they lack granular probabilistic flexibility. By coupling SVM capabilities with instance-level (KNN) and rule-based (DT) paradigms, our architecture captures both global text patterns and fine-grained user preference niches.

3.4. Ensemble Methods

Ensemble methods are learning algorithms that combine multiple classification methods (e.g., Logistic Regression, Random Forests, SVMs) and then classify new data items by averaging their predictions. Compared to single models, ensemble methods typically produce more accurate results [23]. The two simplest ensemble methods are voting and averaging. They are both easy to comprehend and implement. Voting is used to solve classification problems; it aggregates the predictions of each classifier, and the majority vote is used as the final prediction. The averaging method is used for regression and takes the average of predictions.

3.5. Our Innovation

Rather than adopting a static, unweighted blending rule like standard averaging, this work introduces an optimized Linear Weighted Hybridization Method driven by validation-stage performance metrics. By dynamically mapping the inverse validation error of the Decision Tree, k-NN, and SVM models to their respective operational weights, our framework automatically penalizes weaker individual predictions. This optimization ensures a robust, adaptive, and highly accurate text-driven recommendation pipeline tailored specifically for historical tourism datasets.

4. Research Method

This study seeks to answer the following questions: Which AI and machine learning techniques are most effective for predicting tourists' ratings of historical buildings in the USA and the Drâa-Tafilalet region? To what extent do building characteristics and tourist preferences influence the recommendation process? The main hypothesis is that content-based machine learning techniques can provide accurate rating predictions for historical tourism sites. Furthermore, hybrid or ensemble approaches are expected to outperform single models because they combine the strengths of several algorithms, reduce

individual model bias and variance, and capture more complex relationships between tourist preferences and historical-building features. Therefore, integrating tourist profiles with site attributes is expected to improve the accuracy, robustness, and relevance of the proposed recommendation system.

4.1. Datasets

Two datasets will be employed for training and testing the recommender system component. The first is the well-known Yelp dataset [24], and the second is the Drâa-Tafilalet dataset [1]. The Yelp Dataset is a publicly accessible dataset released by Yelp Inc. for personal, academic, and educational use. The dataset contains information on 150,346 businesses in 11 metropolitan areas, with 6,990,280 user reviews. The dataset is supplied in six JSON files: business, user, review, tip, check-in, and photo. Each file comprises one JSON object per line, giving information about a single object type. Each file comprises a single JSON object per line, providing information about a specific object type. The 'business.json' file contains data on categories, business name, location, business type, a unique identifier, and rating. The file 'user.json' contains the user's relevant metadata and their social network mapping. The file designated "review.json" comprises a review text, a rating, a date, a unique user ID, and a unique business ID. The 'checkin.json' file contains the check-in data for a given business. Due to the sheer size of this dataset, this study filters the data to consider only historical buildings. Given that the dataset contains one JSON object per line, we processed the business.json file sequentially. The inclusion criteria required the categories key to contain specific target tags, namely "Landmarks & Historical Buildings" or "History Museums". Consequently, only reviews in review.json matching the business_id of these filtered objects were retained. This filtering procedure successfully extracted 16,002 reviews corresponding to 128 unique historical buildings. Each row in the resulting dataset contains information about a historical building, including a name, city, business ID, state, longitude, latitude, rating, a review, and the historical building's attributes. Every rating is given on a range of 1 to 5.

We also consider the Drâa-Tafilalet area to test our system. It is among Morocco's most well-known tourist destinations and is characterized by numerous historical structures. The Drâa-Tafilalet dataset contains 1737 reviews, rated by 876 users, for 70 historical buildings. The data collection campaign was conducted in mid-2025, targeting public data from TripAdvisor, using customized Python scripts (Python 3.14) equipped with the BeautifulSoup library. The extraction pipeline targeted public listings explicitly categorized under historical landmarks, museums, and cultural heritage sites within the region. Once processed, these records were automatically saved into JSON files adopting the exact same line-by-line structure as the Yelp dataset to ensure seamless system compatibility.

4.2. Pre-Processing and Features Selection

The preprocessing phase consists of a series of processes designed to cleanse the dataset before processing. The objective of this phase is to enhance data quality, thereby facilitating superior results.

The block diagram (ii) in Figure 1 illustrates the pre-processing methodology employed in this research. The initial step involved filtering businesses by their classification as historical buildings. After filtering the data on businesses and their reviews and checking for missing values, we used the business and review files to create a metadata feature by combining business details (such as name, category, and reviews). Furthermore, the metadata undergoes a series of preprocessing steps, including the removal of digits, Unicode characters, and special characters; conversion of uppercase letters to lowercase; and elimination of punctuation, stop words, and stemming.

This research uses various implementations of the content-based approach to propose recommendations to users. This approach compares the features of businesses that users have previously consulted with those they haven't yet seen. To utilize the business features using the content-based method, we have transformed each business's preprocessed meta-data into a set of numeric values, called vectors. The representation of businesses in the vector space model presents two key challenges: first, how to assign appropriate weights to business metadata terms, and second, which technique to use to calculate the similarity between these feature vectors.

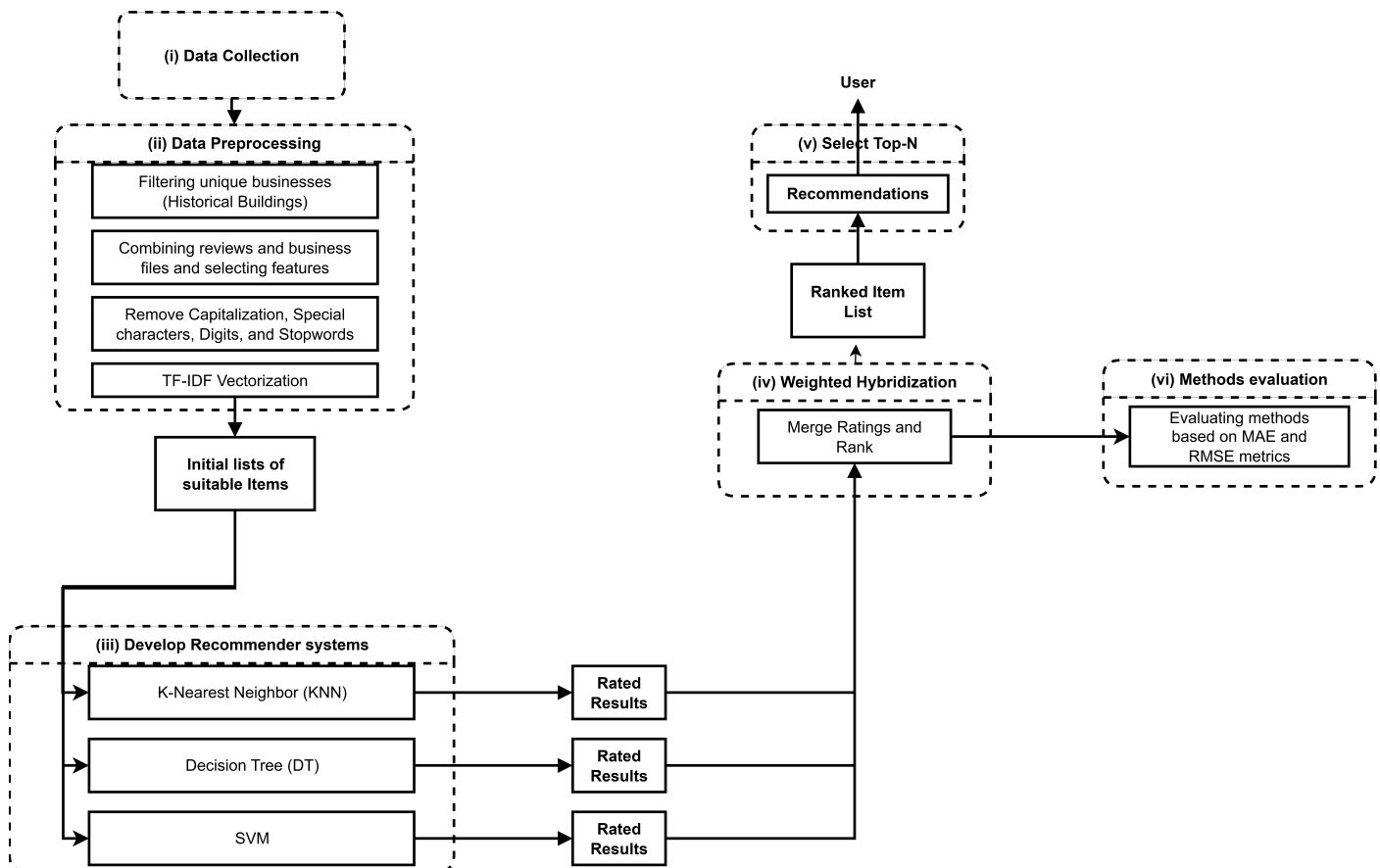


Figure 1. Hybrid Content-based Recommender System.

Term Frequency-Inverse Document Frequency (“TF-IDF”) is the most widely used term weighting scheme [25]. This is a statistical metric employed to assess a word’s significance within a document or item. The rationale for this measure is that terms that appear across many items do not distinguish between relevant and irrelevant items. However, rare and common terms for a few items can better define content similarity and the relevance of one item to another.

Formally, any business is represented as a vector of weights on terms, with each weight denoting the degree of the association between the term and that business. Let $B = \{b_1, b_2, \dots, b_n\}$ denote a set of businesses, and $T = \{t_1, t_2, \dots, t_m\}$ the vocabulary, or intersection of the set of terms that constitute the pre-processed metadata of all businesses. Each business is represented by a vector in an m -dimensional vector space. The vector for business b_j is given by $\mathbf{b}_j = \{w_{j1}, \dots, w_{jm}\}$, where w_{jk} is the weight of term t_k in business b_j . The weight of a term is calculated using the following formula:

$$w_{jk} = \text{TF}(t_k, b_j) \times \text{IDF}(t_k)$$

where:

$$TF(t_k, b_j) = \frac{\text{Frequency of occurrence of } t_k \text{ in } b_j}{\text{Total number of terms in } b_j},$$

represents the number of times the term appears in a review.

$$IDF(t_k) = \log \left(\frac{\text{Total number of businesses}}{\text{Total number of businesses with } t_k} \right),$$

is a measurement of the information provided by a term across all businesses. The output obtained from the vectorization step is an $n \times m$ (businesses/terms) matrix (Figure 2). The number of rows (n) is the number of businesses, and the number of columns (m) is the vocabulary t_k size

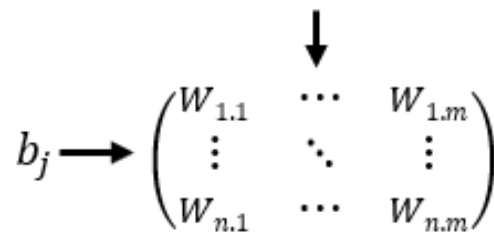


Figure 2. Vector representation of businesses.

This representation is frequently referred to as the Salton representation or matrix and was originally used in the information retrieval system SMART [26].

Once the terms have been weighted, the next step is to determine which technique to use to measure the proximity between the user-consulted item vectors and potential item vectors. Section 3 provides an overview of the techniques most commonly used in tourism recommender systems.

4.3. System Design

The global project [1] aims to build a hybrid recommender system that combines different methods of recommendation (Content-Based, Collaborative, Context-aware, and Demographic Filtering) to suggest the most convenient PoIs (i.e., hotels, restaurants, transport, heritage, activities, etc.) in a particular destination to the visitor, based on his past choices and his constructed profile. In this study, we chose to implement the content-based component to predict ratings for a specific type of tourist visiting historical buildings in the USA and the Drâa-Tafilalet area. To this end, a linear vector of features was used to represent each item. The proposed Tourism Recommender Engine uses a hybrid content-based methodology, calculating distances between each item the user has rated and potential items using various machine learning techniques. It then combines the results to produce an accurate estimate of the item's ratings.

The hybrid content-based component that must be constructed is illustrated by the block diagram in Figure 1. Initially, a data preprocessing stage is conducted before the exploitation of the data (ii). This stage provides an initial selection of items. The content-based approach uses TF-IDF to encode the features of these items into weighted term vectors.

The similarity between item feature vectors can be measured using various methods. Three methods employed and endorsed by numerous existing content-based tourism recommender systems have been selected for consideration. These methods include Support Vector Machines (SVMs), K nearest neighbors (KNN), and Decision trees. The aforementioned three methods are executed in parallel (iii). Based on the results of each similarity method, the prediction rating for each item is calculated (e.g., using the mean, the mul-

tidimensional vote vector length, etc.). Furthermore, the hybrid method (iv) employs a weighting strategy that combines the three rated outcomes using the weights of each method to produce a hybrid predicted rating. The result is a ranked list of recommendations. Ultimately, the predicted hybrid rating is used to generate a ranked list of recommendations, which is then presented to the tourist as a final Top-N list (v).

4.4. Performance Evaluation

To evaluate the performance of this system, two well-established metrics are employed: the mean absolute error (MAE) and the root mean squared error (RMSE) [27]. They aim to test the similarity between users' actual evaluations and the predicted values.

The MAE (Mean Absolute Error) [27] of a model on a test set is the mean of the absolute differences between the actual values for all instances in the test set and the predicted values. It is formally represented as follows:

$$MAE = \frac{\sum_{(u,i) \in T} |r_{u,i} - \hat{r}_{u,i}|}{|T|}$$

where, $r_{u,i}$ is the actual rating that user u has given to item i , $\hat{r}_{u,i}$ is the corresponding estimated rating, and $|T|$ is the number of ratings used in the test. The recommendations produced have higher accuracy when the MAE is lower.

We have also used the RMSE (Root Mean Squared Error) metric [27], which is frequently used to evaluate recommender system model performance. It shows the deviation between predicted and observed true values using Euclidean distance, and it's defined by:

$$RMSE = \sqrt{\frac{\sum_{(u,i) \in T} (r_{u,i} - \hat{r}_{u,i})^2}{|T|}}$$

where, $r_{u,i}$ is the actual rating given to item i by user u , $\hat{r}_{u,i}$ is the corresponding predicted rating, and $|T|$ is the number of predicted values. Compared to the other metrics, RMSE is more sensitive to errors, imposing a more severe penalty (the difference is squared before it is summed).

4.5. Results and Discussion

The implemented recommender system used a hybrid Content-based Filtering method, which suggests recommendations based on the user's previous items. For each test user, a masked hold-out validation strategy was executed by partitioning the historical data into a 60% training set, 10% validation set, and 30% test set. Specifically, 30% of the user's seen historical buildings were selected uniformly at random and converted into unseen items to serve as the definitive ground-truth test set. The remaining 70% was utilized for model training (60%) and inverse-error weight optimization (10%). Finally, the framework evaluates the results by generating recommendations for these hidden items and calculating predictive accuracy using the evaluation metrics described above.

To determine the most suitable similarity technique for the Hybrid Content-based Filtering system, experiments were conducted using three techniques commonly employed in Content-based Tourism Recommender Systems: Decision Tree (DT), KNN, and SVM. The evaluation was performed on two datasets, namely the Yelp dataset and the Drâa-Tafilalet dataset. The accuracy of these techniques was assessed using two standard evaluation metrics: mean absolute error (MAE) and root mean square error (RMSE). The MAE results are illustrated in Figure 3 and summarized in Table 1, while the RMSE results are presented in Figure 4 and Table 2.

Table 1. MAE comparison of the proposed hybrid content-based recommender system.

Dataset	KNN	Decision Tree	SVM	Weighted Hybridization Method
Yelp Dataset	0.054	0.056	0.047	0.030
Drâa-Taifalat Dataset	0.046	0.061	0.039	0.037
Average	0.050	0.058	0.043	0.034

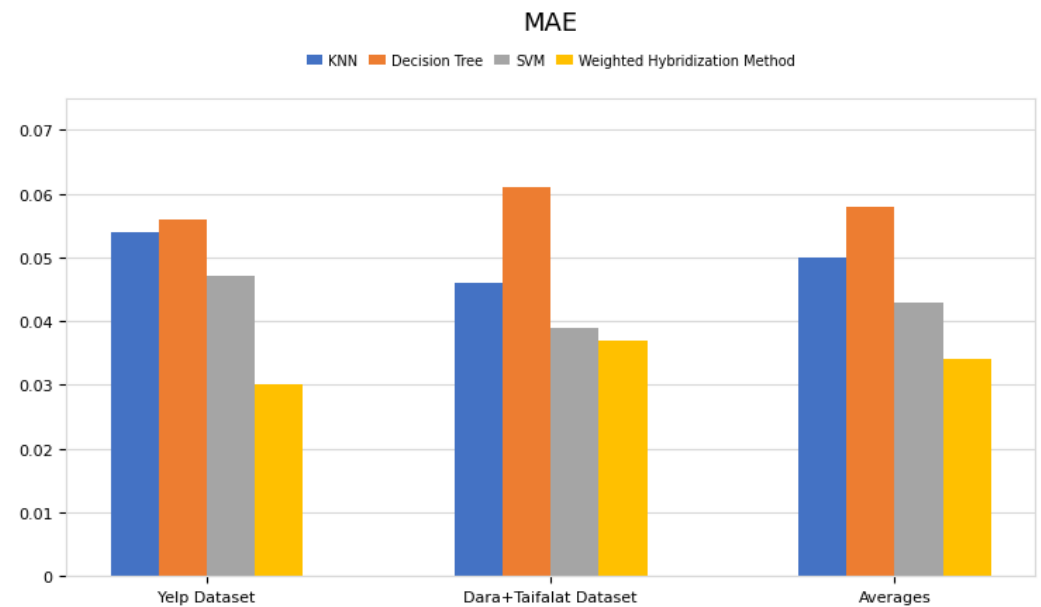
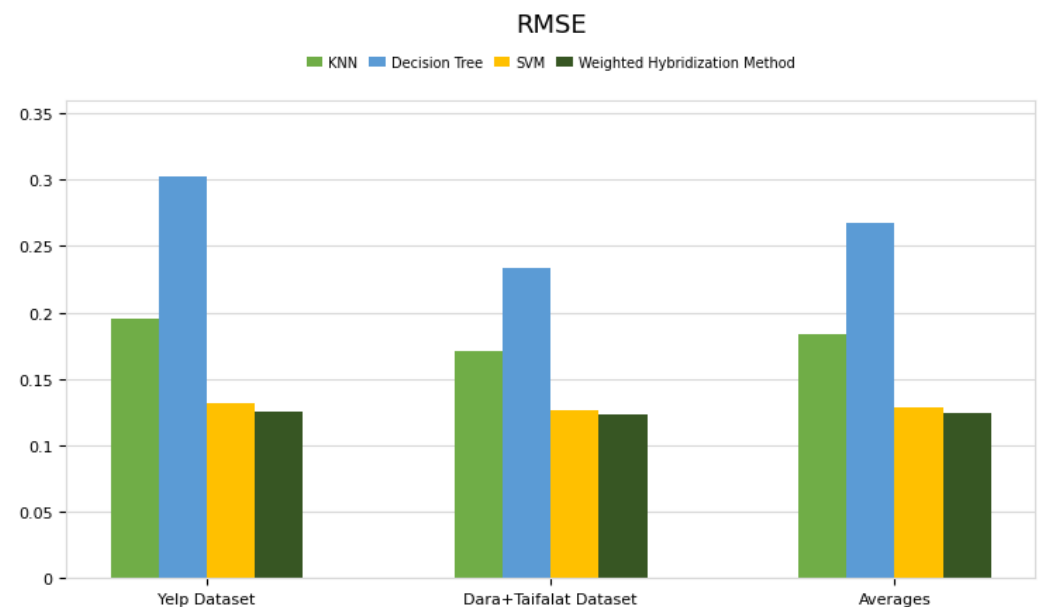
**Figure 3.** Method comparison based on the MAE metric.**Figure 4.** Method comparison based on the RMSE metric.

Table 2. RMSE comparison of machine learning models for the hybrid content-based recommender system.

Dataset	KNN	Decision Tree	SVM	Weighted Hybridization Method
Yelp Dataset	0.196	0.302	0.132	0.126
Dara+Taifat Dataset	0.172	0.233	0.127	0.123
Averages	0.184	0.267	0.130	0.125

The weighted hybrid method combines the predicted ratings generated by the three recommender techniques, namely kNN, SVM, and Decision Tree, using a linear weighted formula:

$$\widehat{r}_{\text{hybrid}} = \alpha \widehat{r}_{\text{DT}} + \beta \widehat{r}_{\text{kNN}} + \gamma \widehat{r}_{\text{SVM}}$$

where $\widehat{r}_{\text{DT}}$, $\widehat{r}_{\text{kNN}}$, and $\widehat{r}_{\text{SVM}}$ represent the predicted ratings produced by the Decision Tree, kNN, and SVM models, respectively. The coefficients α , β , and γ represent the weights assigned to each model. These weights are determined according to the predictive performance of each model on the validation set. Since lower error values indicate better predictive accuracy, models with lower MAE and RMSE values are assigned higher weights.

To compute the weight of each model, an inverse-error weighting strategy is used. For a given model i , its weight w_i is calculated as follows:

$$w_i = \frac{\frac{1}{e_i}}{\sum_{j=1}^n \frac{1}{e_j}}$$

where e_i denotes the prediction error of model i , and n is the total number of models combined in the hybrid system ($n = 3$). In our implementation, the inverse-error optimization was explicitly evaluated using the RMSE metric on the validation set. Given validation error scores of $e_{\text{DT}} = 0.268$, $e_{\text{kNN}} = 0.184$, and $e_{\text{SVM}} = 0.130$, the corresponding individual inverse errors $\left(\frac{1}{e_i}\right)$ are calculated as 3.73, 5.43, and 7.69, respectively, yielding cumulative sum of 16.85. Normalizing these intermediate values guarantees that the system configuration fulfills the strict total weight constraint:

$$\alpha + \beta + \gamma = 1$$

By dividing each individual inverse error by the cumulative sum, we obtain the final exact empirical coefficients utilized in this study to generate the “Weighted Hybridization Method” results shown in Figures 3 and 4: $\alpha = 0.221$, $\beta = 0.322$, and $\gamma = 0.456$. This weighting strategy allows the hybrid recommender system to give greater importance to the best-performing models while still preserving the contribution of each technique. As a result, the final predicted rating is expected to be more robust and accurate than the prediction produced by any individual model.

As illustrated in Figures 3 and 4, the weighted hybridization method exhibits the lowest mean absolute error (MAE) among the three algorithms, namely DT, kNN, and SVM, when evaluated on the two datasets. The mean absolute error (MAE) of the weighted method was 0.034, while the support vector machine (SVM) yielded a close value of 0.038. The mean absolute error (MAE) for KNN was 0.050, while DT exhibited the highest value at 0.058. Furthermore, across both datasets, the weighted hybridization method yields superior results compared to the remaining three techniques. The average root mean square error (RMSE) for this method is 0.124, whereas the respective values for the decision tree (DT), k-nearest neighbor (kNN), and support vector machine (SVM) techniques are 0.268, 0.184, and 0.130, respectively.

The results demonstrate that our proposed recommender approach, based on hybrid content-based methods, achieved the lowest mean absolute error (MAE) and root mean square error (RMSE) among all methods. In light of these findings, we selected the weighted method as a technique of similarity to our content-based component.

5. Conclusions

This article presents the findings of a study conducted using a hybrid content-based methodology. The Drâa-Tafilalet area was selected as the focus of the study to promote its attractions to travelers. The recommendation engine uses a hybrid content-based approach that combines three techniques: the Decision Tree, kNN, and SVM. The experimental findings, evaluated using RMSE and MAE metrics, demonstrate that the Hybrid Method with weighting methodology exhibits superior prediction accuracy compared to the recommendation method when performed independently. The content-based approach effectively addresses the cold-start problem for new items; however, it still poses a challenge for new users, as there are no rated items from active users to serve as a point of comparison, hindering the ability to generate results. Future research will develop the remaining components using collaborative filtering and demographic-based methodologies, subsequently integrating them into the Big Data architecture framework of the Hybrid Tourism Recommendation System [1]. The objective is to augment the database with additional points of interest (POIs) and recommend a trip itinerary based on the selected top-rated sites and activities. This novel system will be designed as a trip-planning service to promote travel to Morocco, with a particular focus on the Drâa-Tafilalet region.

Author Contributions: Conceptualization, K.a.F. and L.M.; methodology, K.a.F., L.M., B.A., A.S., A.Y. and J.R.; software, K.a.F. and L.M.; validation, K.a.F., L.M., B.A., A.S., A.Y. and J.R.; formal analysis, K.a.F. and L.M.; investigation, K.a.F. and L.M.; resources, K.a.F. and L.M.; data curation, K.a.F. and L.M.; writing—original draft preparation, K.a.F. and L.M.; writing—review and editing, K.a.F., L.M., B.A., A.S., A.Y. and J.R.; visualization, K.a.F. and L.M.; supervision, B.A., A.S., A.Y. and J.R.; project administration, K.a.F., L.M., B.A., A.S., A.Y. and J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Yelp dataset used in this study is publicly available from Yelp Inc. for personal, academic, and educational purposes. In this work, a subset of the Yelp dataset containing 16,002 reviews related to 128 historical buildings in the USA was used for training the recommender system component. The Drâa-Tafilalet dataset used for testing contains 1737 reviews from 876 users for 70 historical buildings in the Drâa-Tafilalet region of Morocco. This dataset is private and is not publicly available due to privacy and ethical restrictions related to the collected user-review data. The processed data supporting the findings of this study may be made available from the corresponding author upon reasonable request, subject to applicable privacy restrictions.

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

References

1. Al Fararni, K.; Nafis, F.; Aghoutane, B.; Yahyaouy, A.; Riffi, J.; Sabri, A. Hybrid Recommender System for Tourism Based on Big Data and AI: A Conceptual Framework. *Big Data Min. Anal.* **2021**, *4*, 47–55. [[CrossRef](#)]
2. Al Fararni, K.; Aghoutane, B.; Yahyaouy, A.; Riffi, J.; Sabri, A. Tourism Recommender Systems: An Overview. *Int. J. Cloud Comput.* **2021**, *10*, 603–612. [[CrossRef](#)]
3. Burke, R. Hybrid Recommender Systems: Survey and Experiments. *User Model. User-Adapt. Interact.* **2002**, *12*, 331–370. [[CrossRef](#)]

4. Adomavicius, G.; Tuzhilin, A. 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. [\[CrossRef\]](#)
5. Thorat, B.P.; Goudar, R.; Barve, S. Survey on Collaborative Filtering, Content-Based Filtering, and Hybrid Recommendation Systems. *Int. J. Comput. Appl.* **2015**, *110*, 31–36. [\[CrossRef\]](#)
6. Al Fararni, K.; Aghoutane, B.; Riffi, J.; Sabri, A.; Yahyaouy, A. Comparative Study on Approaches of Recommendation Systems. In *Embedded Systems and Artificial Intelligence*; Bhateja, V., Satapathy, S., Satori, H., Eds.; Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; Volume 1076. [\[CrossRef\]](#)
7. Adil, S.; Thangaraj, S.J.J. Collaborative Content Filtering-Based Tourism Places Recommendation System Compared with a Hybrid Recommender System. *AIP Conf. Proc.* **2023**, *2822*, 020014. [\[CrossRef\]](#)
8. Mu, Z.; Yi, C.; Xiaohong, Z.; Junyong, L. Study on the Recommendation Technology for Tourism Information Service. In *Proceedings of the 2009 Second International Symposium on Computational Intelligence and Design*; IEEE: New York, NY, USA, 2009; Volume 1, pp. 410–415.
9. Yurchak, I.; Hryhlevych, M. Recommender System Based on Collaborative Filtering. *Comput.-Integr. Technol. Educ. Sci. Prod.* **2023**, *53*, 78–85. [\[CrossRef\]](#)
10. Delgado, J.; Ishii, N. Memory-Based Weighted Majority Prediction. In *Proceedings of the SIGIR Workshop Recommender Systems*; Citeseer: State College, PA, USA, 1999.
11. Aggarwal, C.C. Model-Based Collaborative Filtering. In *Recommender Systems*; Springer: Cham, Switzerland, 2016. [\[CrossRef\]](#)
12. Moscato, V.; Picariello, A.; Rinaldi, A.M. Towards a User-Based Recommendation Strategy for Digital Ecosystems. *Knowl.-Based Syst.* **2013**, *37*, 165–175. [\[CrossRef\]](#)
13. Musa, J.M.; Zhihong, X. Item-Based Collaborative Filtering Approach in Movie Recommendation System Using Different Similarity Measures. In *Proceedings of the 2020 6th International Conference on Computer and Technology Applications*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 31–34.
14. Chen, J.-H.; Chao, K.-M.; Shah, N. Hybrid Recommendation System for Tourism. In *Proceedings of the 2013 IEEE 10th International Conference on e-Business Engineering*; IEEE: New York, NY, USA, 2013; pp. 156–161.
15. Kirtil, I.G.; Aşkun, V. Artificial Intelligence in Tourism: A Review and Bibliometrics Research. *Adv. Hosp. Tour. Res.* **2021**, *9*, 205–233. [\[CrossRef\]](#)
16. Patel, H.; Prajapati, P. Study and Analysis of Decision Tree-Based Classification Algorithms. *Int. J. Comput. Sci. Eng.* **2018**, *6*, 74–78. [\[CrossRef\]](#)
17. Nayak, N.R. Application of Naive Bayes Classifier for Information Extraction. *OSF Prepr.* **2020**, z7q2e. [\[CrossRef\]](#)
18. Thiengburanatham, P.; Cang, S.; Yu, H. A Decision Tree-Based Destination Recommendation System for Tourists. In *Proceedings of the IEEE 21st International Conference on Automation and Computing*; IEEE: New York, NY, USA, 2015.
19. Perdana, A.; Hermawan, A.; Avianto, D. Analyze Important Features of PIMA Indian Database for Diabetes Prediction Using KNN. *J. Sisfokom Sist. Inf. Dan Komput.* **2023**, *12*, 70–75. [\[CrossRef\]](#)
20. Prasad, B.V.V.S.; Gupta, S.; Borah, N.; Dineshkumar, R.; Lautre, H.K.; Mouleswararao, B. Predicting Diabetes with Multivariate Analysis: An Innovative KNN-Based Classifier Approach. *Prev. Med.* **2023**, *174*, 107619. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Tech-AI-Math. K-Nearest Neighbors (KNN) in Depth. *Medium*, 2023. Available online: <https://ai.plainenglish.io/k-nearest-neighbors-knn-769bd39514c6> (accessed on 31 May 2026).
22. Pirra, M.; Diana, M. A Study of Tour-Based Mode Choice Based on a Support Vector Machine Classifier. *Transp. Plan. Technol.* **2019**, *42*, 23–36. [\[CrossRef\]](#)
23. Schlar, A.; Tsikinovsky, A.; Rokach, L.; Meisels, A.; Friedman Antwarg, L. Ensemble Methods for Improving the Performance of Neighborhood-Based Collaborative Filtering. In *Proceedings of the ACM Conference*; Association for Computing Machinery: New York, NY, USA, 2009; pp. 261–264. [\[CrossRef\]](#)
24. Yelp Open Dataset. Available online: <https://www.yelp.com/dataset> (accessed on 31 May 2026).
25. Bafna, P.; Pramod, D.; Vaidya, A. Document Clustering: TF-IDF Approach. In *Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques, Chennai, India, 3–5 March 2016*; IEEE: New York, NY, USA, 2016; pp. 61–66. [\[CrossRef\]](#)
26. Salton, G.; Lesk, M.E. The SMART Automatic Document Retrieval Systems and Illustration. *Commun. ACM* **1965**, *8*, 391–398. [\[CrossRef\]](#)
27. Chai, T.; Draxler, R.R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Arguments against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.