*electronics*

MDPI

*Article*

# Optimal Query Expansion Based on Hybrid Group Mean Enhanced Chimp Optimization Using Iterative Deep Learning

**Ram Kumar *, Kuldeep Narayan Tripathi and Subhash Chander Sharma**

Indian Institute of Technology Roorkee, Roorkee 247667, India; kuldeep08narayan@gmail.com (K.N.T.); subhash.sharma@pt.iitr.ac.in (S.C.S.)
* Correspondence: rkumar@pp.iitr.ac.in

**Abstract:** The internet is surrounded by uncertain information which necessitates the usage of natural language processing and soft computing techniques to extract the relevant documents. The relevant results are retrieved using the query expansion technique which is mainly formulated using the machine learning or deep learning concepts in the existing literature. This paper presents a hybrid group mean-based optimizer-enhanced chimp optimization (GMBO-ECO) algorithm for pseudo-relevance-based query expansion, whereby the actual queries are expanded with their related keywords. The hybrid GMBO-ECO algorithm mainly expands the query based on the terms that have a strong interrelationship with the actual query. To generate the word embeddings, a Word2Vec paradigm is used which learns the word association from large text corpora. The useful context in the text is identified using the improved iterative deep learning framework which determines the user's intent for the current web search. This step reduces the mismatch of the words and improves the performance of query retrieval. The weak terms are eliminated and the candidate query terms for optimal query expansion are improved via an Okapi measure and cosine similarity techniques. The proposed methodology has been compared to the state-of-the-art methods with and without a query expansion approach. Moreover, the proposed optimal query expansion technique has shown a substantial improvement in terms of a normalized discounted cumulative gain of 0.87, a mean average precision of 0.35, and a mean reciprocal rank of 0.95. The experimental results show the efficiency of the proposed methodology in retrieving the appropriate response for information retrieval. The most common applications for the proposed method are search engines.

**Keywords:** information retrieval; automatic query expansion; pseudo relevance feedback; iterative deep learning framework; enhanced chimp optimization; group mean based optimizer



**Citation:** Kumar, R.; Tripathi, K.N.; Sharma, S.C. Optimal Query Expansion Based on Hybrid Group Mean Enhanced Chimp Optimization Using Iterative Deep Learning. *Electronics* **2022**, *11*, 1556. https://doi.org/10.3390/electronics11101556

Academic Editor: José L. Abellán

Received: 10 April 2022
Accepted: 10 May 2022
Published: 12 May 2022



## 1. Introduction

Nowadays, everything is digitalized due to the development of the internet. Web pages are created and launched in societies all over the world. Therefore, there are vast quantities of data available on the internet which make the Information Retrieval (IR) process somewhat complex [1–4]. Users are not able to retrieve the correct results because of small keywords, words that are too similar to the keyword, vocabulary problems, and sometimes the fact that the user is unaware of exactly what he wants. For example, if the user searches for the word "mouse" without clarifying whether it is an animal or an electronic device. To overcome this type of vocabulary issue, various methods such as the clustering of search results, relevance feedback, and iterative query filtration are used. Almost every popular strategy adds new related terms to the initial query. The user-relevant feedback can be used to enhance the performance of the information retrieval system and search engine. Thus, Query Expansion (QE) is introduced to retrieve more accurate information [5]. QE contains three sub-divisions. They are manual, interactive, and automatic [6]. In these divisions, the user performs the Manual QE (MQE), the combined action by system and user is performed in Interactive QE (IQE), and the system alone

performs the Automatic QE (AQE). QE is not only used in IR but also in various fields such as E-commerce, medicine [7], and sports [8]. This QE can improve IR's efficacy by including relevant phrases that can aid in distinguishing between relevant and irrelevant materials [9].

To provide the easy retrieval of data, the document or webpage should be categorized based on the topic or some other metrics. However, categorization of this vast data is manually impossible. Therefore, the method text classification is introduced to overcome this issue. Text classification has a wide range of real-world applications, including spam mail filtering, document genre identification, the automatic categorization of documents or webpages based on a pre-defined label set, multimedia recommendation, etc. [10]. However, if a document or webpage has thousands of unique data then the text classification may give very poor output because of some terms that do not help for classification. These terms may mislead the classifiers [11]. The artificial neural network (ANN) technique may also be used for QE. This technique mimics the behavior of the human brain. It uses different mathematical functions to increase the ability of the model for prediction.

Formerly, text classification algorithms extracted characteristics from textual input using 'bag of words' (BOW)-based models [12]. To pick the salient characteristics, document frequency (DF), latent Dirichlet allocation (LDA), mutual information (MI), and information gain (IG) are used. These algorithms do not save the text sequence and more importantly, do not understand the context and semantics. It is worth noting that to gain a full understanding of the context of algorithmic-related metadata, the text lines are required for efficient classification. Moreover, alternative text-matching techniques such as high-order n-grams and tree kernels support the semantics and word order, though they are not able to fully capture the context and can have a significant impact on the accuracy of classification. A variety of machine learning (ML) based algorithms [13–15] has recently demonstrated significant improvements in the text sequences' retention for contextual understanding. These models' fundamental issue is extracting characteristics for each unique text item, such as sentences, documents, or phrases.

This work uses pseudo-relevance feedback for query expansion where the feedback documents are the terms in the top retrieved documents. Using query expansion, the irrelevant terms in the feedback document corpora (also known as a term pool) aims for a better understanding of the context. The addition of feedback terms enhances the query expansion result accuracy along with the efficiency of information retrieval. The performance was evaluated using various standard parameters such as precision, recall, mean average precision (MAP), mean reciprocal rank (MRR), F-measure, and normalized discounted cumulative gain (NDCG). The proposed methodology mainly uses a metaheuristic approach for automatic query expansion. The major contribution of this paper is presented as follows:

- A hybrid group mean-based optimizer-enhanced chimp optimization (GMBO-ECO) algorithm is presented for pseudo-relevance-based query expansion which finds an optimal subset of candidate terms to expand the original queries with their related keywords.
- The iterative deep learning-based methodology is used to solve the word ambiguity problem of natural language that selects the most relevant terms and eliminates the irrelevant ones.
- An experimental analysis demonstrates how the proposed methodology improves the performance of text-based information retrieval systems. The 'with' and 'without' query expansion results are evaluated using different performance metrics on the Text REtrieval Conference (TREC) and the Cross Language Evaluation Forum (CLEF) test collections.

The rest of this paper is structured accordingly. Section 2 presents the existing literary works and Section 3 describes the problem statement with the background methodologies. Section 4 presents the working of the query expansion strategy using the proposed framework and Section 5 describes the simulation experiments conducted in detail. Section 6 summarizes the paper.

## 2. Literature Survey

In recent years, various QE techniques have been presented by several researchers, including ontology-based and linguistic approaches that are used to overcome the vocabulary mismatch problem. Ontology (the knowledge structure that enumerates the ideas, attributes, and relationships between them)-based QE techniques take advantage of ontology's generalization, specialization, and other connections to extract relevant terms for QE [16]. Some of the recent query expansion approaches are explored and described in this section. Chugh et al. [17] established a spider monkey crow optimization algorithm (SMCA) with deep recurrent neural network (DRNN) for sentiment classification and information retrieval. This method was used to eliminate the stop words and unwanted data for minimizing the user's access delay. The feature extraction process was conducted by using SentiWordNet for obtaining sentiment classification. The features, such as punctuation, numerical values, SentiWordNet, hashtag, and elongated words were utilized for achieving the sentiment classification. The SMCA methodology offers accuracy, precision, recall, and an F1-score of 97.75%, 95.5%, 94.6%, and 96.7%, respectively. The main drawback of this method was the inability to predict the fake reviews and comments.

Rasheed et al. [18] proposed the query expansion method based pseudo-relevance feedback utilizing a boosting algorithm. They used a query expansion algorithm to test the five distinct hybrid approaches. The Roshni, FIRE2011, and Shahril databases were used in this study. Using this method, the accuracy rates of English, Persian and Urdu datasets were improved by 6.60%, 9.93%, and 14.02%, respectively. The query-wise test showed that the scheme detected the relevant keywords which were real for the short query. This technique requires a large running time. Jain et al. [19] explained the information retrieval of a fuzzy ontology framework utilizing semantic query expansion. The dataset was obtained from a NewsGroup20 database in the UCI Repository. The current web scheme weakness was overcome by using the fuzzy ontology based on the query envelope. MAP, precision, R-precision, and MRR were the performance metrics used in this study. The results demonstrated that their fuzzy ontology framework was more flexible and efficient than existing ontology techniques. This technique has failed to compute the concepts fusing with other ontologies such as DBpedia and Wikipedia.

David Raj, G., et al. [20] presented the patent retrieval of query expansion utilizing the modified stellar-mass black hole optimization (MSBO). MSBO enhanced the existent SBO by using the genetic operator for the new crossing over mechanism and also by using the dynamic absorption value. The CISI and 20-newsgroup datasets were used in this study. In experiments, the MSBO algorithm outperformed the existing SBO algorithm. Malik et al. [21] presented a biomedical query expansion framework to improve the actual queries by creating a secondary semantically similar term for each word in the query. In their approach, the authors integrated the word embeddings with clinical diagnosis information with the word embeddings to identify the related biometric literature connected to the specific keyword. The implicit semantics present in the text are acquired via embedding the vocabulary terms as low dimensional vectors or real-valued functions. The results show that this methodology is effective in terms of domain agnostic and domain-specific data when evaluated in terms of the Text Retrieval Conference dataset.

Safder and Hassan [22] developed a feature engineering technique based on deep learning that enhances search capability through the extraction of metadata pertaining to algorithmics. TF-IDF (term frequency-inverse document frequency) based techniques typically behave like a "bag of words" model, failing to capture either the word sequence or the semantics of the text. Here, each full-text document's semantically improved summary is created by including deep metadata text lines that are algorithmic-specific to improve the algorithm search systems with a search mechanism. Further, the bi-directional deep learning based, Long Short-Term Memory (LSTM) model classifies these text lines. The two-way LSTM model, designed with a score F1 = 0.81, outperformed the support vector machine by 9.46% when it was implemented on a database of 37,000 algorithm-specific metadata text lines which was tagged by four human experts.

ALMarwi et al. [23] presented a query expansion hybrid that employs both semantic and statistical techniques. Researchers suggested an efficient PSO (particle swarm optimization)-based weighting technique to choose the best phrases for query expansion. As a proof-of-concept, a system prototype was built and its precision tested. The test was performed on the basis of actual data. The PSO algorithm improves query expansion accuracy, according to the results of the experiments. The query reform method based on the medical term 're-weighting' is achieved by Diao et al. [24] to improve the retrieval of electronic medical record (EMR) performance, and to emphasize the relevance of medical terms. First, the approach takes medical keywords from the original query and filters them out. Then, every medical word is weighted by its self-information based on the document collection. Conclusively, the new query is built by proportionately merging the weighted medical phrases with the original query. When evaluated using the TensorFlow dataset, their strategy offered an improvement of 8%, 9.6%, and 14.2% in terms of binary preference-based measure (bpref), Precision at 10 (P@10), and MAP when compared with the existing techniques.

An approach for semantic query expansion based on ontology is presented by Khedr et al. [25] for query disambiguation in the computer programming area. Integrating the cosine similarity algorithm into the suggested model enhanced the enlarged query. The suggested technique was tested with misspelled questions and numerous confusing data, and the resulting extensions produced more relevant results when used with a search engine. The quality of the results obtained for the more detailed queries is significantly greater than for crude queries. Their approach was enabled and then evaluated by independent external testers. This method offers an average precision value of 82.2%.

Using the distribution methodology, Dahir and Qadi [26] proposed DBpedia (well-known domain-agnostic decentralized linked data repository) characteristics and Bose-Einstein statistics (Bo1). This method reconstructs candidate documents for a given query. These documents are then used to create topic models based on LDA (latent Dirichlet allocation) and the appropriate expansion words are selected. The suggested method was tested using the collection of AP datasets, and the results of the retrieval trials utilizing the distribution methodology Bo1 demonstrated considerable improvements. The LDA-linkedBo1 solution offered improvements in terms of MRR@N when compared to the association-based alternatives from DBpedia.

Recent research in the field of artificial neural networks (ANN) has changed how people retrieve relevant information from a large collection. Numerous studies have been carried out in the field of artificial intelligence modelling in different fields of study, e.g., chemical and environmental studies [27,28]. In these works, the authors used artificial intelligence modeling methods such as multiple linear regression (MLR), and multi-layer perceptron (MLP) to solve various issues associated with natural language processing [29,30].

## 3. Problem Statement and Methodologies

### *3.1. Problem Statement*

There is so much information on the internet that a user can never have enough to suit his or her needs. Hence, updates have been made to automate information retrieval systems in order to address this issue. It enables users to filter the numerous documents in the database based on their requirements. It accomplishes this by displaying only the documents that are relevant to the user. Users express their requirements to 'queries' that are used to represent IR systems. The user enters them into the system which is then compared to the documents that are stored in the database. Even though the IR system aids in the retrieval of feedback documents, one key issue it encounters is word mismatches. This is mostly due to ambiguity and contextual problems. When a user types a query containing several terms, these terms may be unrelated to the phrases that acquire the index of the files in the system's storage. One solution would be to have the user enhance their query by combining numerous expressions to the input query. However, queries should ideally be brief.

As a result, by refining the query and adding more expressions to the initial query, this problem may be addressed. Query expansion is a term used to describe such an approach. This procedure might be interactive, in which the user chooses which expressions to include in the query from a predefined list of expressions, or automated, in which the query is automatically expanded by the system. There are several methods of query expansion. This is accomplished by retrieving from the database, for each term in the query, its related expressions or synonyms. The suitable phrases are chosen based on a variety of criteria. This collection of phrases is then attached to the original query and the procedure is repeated until the user's criteria have been met.

### *3.2. Group Mean-Based Optimizer Algorithm*

The group mean-based optimizer (GMBO) algorithm [31] is also called a population-based optimization algorithm. It is designed for updating the populations in the algorithm. In GMBO, each iteration picks the two groups, e.g., bad and good groups, along with the definite number of members. The combination of two members and groups such as good and bad groups from the updated population is used to design the GMBO algorithm. The population members in the GMBO algorithm are identified using the matrix as a population matrix. The number of rows and columns in the population matrix indicates the number of members and number of problem variables, respectively. The population members in the population matrix give the solution to the optimization problem. The values of the objective functions are predicted from the below equation as,

$$f = \begin{bmatrix} f_1 & f(x_1) \\ \vdots & \vdots \\ f_j & f(x_j) \\ \vdots & \vdots \\ f_n & f(x_n) \end{bmatrix}_{n \times 1} \tag{1}$$

where the terms $f$, $f_j$ denote the vector of objective function and the value of the objective function, respectively. The bad and good groups are selected from the values of the objective function. The good group contains a definite number of members with good values of the objective function and the bad group has a definite number of members with worse values of the objective function. The simulation is performed from the selection of these two groups are derived as below,

$$f^s = \begin{bmatrix} f_1^s & Minimum(f) \\ \vdots & \vdots \\ f_j^s & \vdots \\ \vdots & \vdots \\ f_n^s & Maximum(f) \end{bmatrix}_{n \times 1} \tag{2}$$

$$H_{n_H \times M} = x_j^s \;\&\; j = 1 : \;\; n_H \tag{3}$$

$$A_{n_A \times M} = x_j^s \;\&\; j = n - n_A + 1 : \;\; n \tag{4}$$

where, $f^s$, $x^s$, $H$, $A$, $n_H$, $n_A$ and $M$ are the vector of the sorted objective function, sorted matrix population, the selected good groups, the bad groups, the number of good groups, the number of bad groups and a matrix constant (that can have a positive integer value only), respectively. When the selected good and bad groups are determined, two composite members are received. This stage performs the simulation operation by using Equations (5) and (6) which are given below,

$$PH = Mean\left(H_{n_H \times M}\right) \tag{5}$$

$$PA = Mean\left(A_{n_A \times M}\right) \tag{6}$$

where, $PA$ and $PH$ represent the composite member-based mean of the bad groups and the composite member-based mean of the good groups, respectively. Then, the population matrix is updated in three levels based on the combination of composite members and the best member of the population. At first, the good group's simulation is performed by using Equations (7) and (8) which are given below,

$$y_{j,e}^{H} = y_{j,e} + t \times \left(PH_{j,e} - y_{j,e}\right) \times Sign\left(f_j - f_{PH}\right) \tag{7}$$

$$x_j = \begin{cases} if(f_j^H < f_j),\ x_j^H \\ else,\ x_j \end{cases} \tag{8}$$

where, $y_{j,e}^{H}$, $t$, $x_j^H$ implies a new value, a random number of the interval, and the new position of a population member, respectively. $f_{PH}$, $f_j^H$ are the objective function values. In the second stage, the simulation is performed based on a composite member of the bad group which is derived below,

$$y_{j,e}^{A} = y_{j,e} + t \times \left(PA_{j,e} - y_{j,e}\right) \times Sign\left(f_j - f_{PA}\right) \tag{9}$$

$$x_j = \begin{cases} if(f_j^A < f_j),\ x_j^A \\ else,\ x_j \end{cases} \tag{10}$$

where, $y_{j,e}^{A}$, $x_j^A$ denote the new value and new position of a population member, respectively. $f_{PA}$, $f_j^A$ are the objective function value. Last, the population matrix is updated based on the best member of the population from the below-mentioned equations which are,

$$y'_{j,e} = y_{j,e} + t \times \left(y_{j,e}^{Best} - y_{j,e}\right) \tag{11}$$

$$x_j = \begin{cases} if\left(f'_j < f_j\right),\ x'_j \\ else,\ x_j \end{cases} \tag{12}$$

where, the terms $y'_{j,e}$, $x'_j$, $f'_j$ imply a new value, a new position, and an objective function value.

### *3.3. Enhanced Chimp Optimization (ECO) Algorithm*

The chimp optimization (CO) algorithm [32] is a nature-inspired metaheuristic algorithm simulated by the hunting activity of chimps. The conventional CO algorithm categorizes the chimp group into four divisions based on the actions performed to hunt the prey. These are attackers, chasers, barriers, and drivers. In this scenario, the attacker leads the population and assists other group members in catching the prey. However, the conventional algorithm has certain complications such as a lack of diversity in population and local optima issues in the searching process. To overcome these shortcomings, three mechanisms (namely a highly disruptive polynomial mutation (HDPM), Spearman's rank correlation coefficient, and Beetle antennae operator) are introduced into the CO algorithm. The implementation of these mechanisms improves the performance by eliminating the issues of the conventional CO algorithm. This improved version of the conventional CO algorithm is formulated as the enhanced chimp optimization (ECO) algorithm.

Initially, the position of chimps is updated based on the following numerical expression.

$$Z_C(x+1) = \frac{Z_a + Z_b + Z_c + Z_d}{4} \tag{13}$$

where, the terms $Z_a$, $Z_b$, $Z_c$, and $Z_d$ indicate the positions of the attacker, barrier, chaser, and driver chimps, respectively, and $x$ represents the current iteration.

3.3.1. Highly Disruptive Polynomial Mutation (HDPM)

Commonly employed mutation operators comprise polynomial mutation, random mutation as well as non-uniform mutation. When the variable is near the boundary, the mutation shows no effect in standard polynomial mutation. To increase the diversity of the population in the initialization process, the HDPM approach is employed which ultimately enhances the exploration ability of the CO algorithm efficiently. The mathematical expression of the HDPM operator is given as,

$$Z_N = Z + \lambda_c \cdot (UB - LB) \tag{14}$$

From the above equation, the term $Z_N$ depicts the young ones, $Z$ signifies the parent, $\lambda_c$ denotes coefficient, and $LB$ and $UB$ indicate the lower and upper boundaries of each search dimension, respectively. The coefficient $\lambda_c$ is determined using the following expression,

$$\lambda_c = \begin{cases} if(R \leq 0.5),\ \left[2R + \left(1 - 2R\right) \cdot (1 - \lambda_1)^{\mu_i + 1}\right]^{\frac{1}{\mu_i + 1}} - 1, \\ else,\ 1 - \left[2(1 - R) + 2(R - 0.5) \cdot (1 - \lambda_2)^{\mu_i + 1}\right]^{\frac{1}{\mu_i + 1}} \end{cases}, \tag{15}$$

where, the terms $\mu_i$ and $R$ represent the mutation index and the random number in the range [0, 1], respectively. By $\lambda_c$ equation, it reveals that this HDPM approach searches the full search space including the boundaries. Thus, it regulates the population diversity efficiently.

3.3.2. Spearman's Rank Correlation Coefficient $(\rho^r)$

A non-parametric index, Spearman's rank correlation coefficient is utilized to calculate the statistic correlation among two different variables. Consider the two variables as $P_k$ and $Q_k$, and the position of these two variables are represented as ${P'}_k$ and ${Q'}_k$, respectively. Moreover, the difference among these two variables is found as $D = {P'}_k - {Q'}_k$, where $k = 1, 2, 3, \ldots, l$. The computation of Spearman's rank correlation coefficient $(\rho^r)$ is defined by the following equation,

$$\rho^r = 1 - \frac{6 \sum D_k^2}{l \cdot (l^2 - 1)} \tag{16}$$

where the term $l$ indicates series dimension and the value of $\rho^r$ lies in the interval between $-1$ and 1. If $\rho^r = 0$ there is no correlation between two variables. Moreover, the $\rho^r$ ranges [0, 1] for positive correlation and $[-1, 0]$ for negative correlations. If the value of $\rho^r$ falls outside of these ranges then it denotes a strong correlation between two variables.

3.3.3. Beetle Antennae Operator

The beetle determines the exact location of food using its two antennas. The beetle changes its flying direction with respect to concentration on food smell. If one antenna receives a strong concentration on smell, the beetle changes its movement in that direction. Inspired by this concept, the local optimum problem of the CO algorithm is prevented. The mathematical formulation of exploring the capability of the beetle in a left and right direction is given by,

$$Z_R(x) = Z(x) + D(x) \cdot \vec{g} \tag{17}$$

$$Z_L(x) = Z(x) - D(x) \cdot \vec{g} \tag{18}$$

From the above two equations, the terms $Z_R$ and $Z_L$ signify the exploring direction of the beetle between right and left, respectively. Further, $Z$ represents the actual position of the beetle, $D(x)$ depicts the distance between two antennas, and $\vec{g}$ implies a random direction vector. Furthermore, the positions are updated using the following expression,

$$D(x) = \frac{\chi(x)}{A_R} \tag{19}$$

where, the terms $A_R$ and $\chi$ signify the attenuation rate and step size, respectively. The step size $\chi$ is calculated based on the following numerical expression,

$$\chi(x) = C \cdot \chi(x-1) \tag{20}$$

In this equation, the attenuation rate of $\chi$ is represented as $C$. Moreover, the new position of the beetle is updated as,

$$Z(x+1) = Z(x) + \chi(x) \cdot \vec{g} \cdot sign(F(Z_R(x)) - F(Z_L(x))) \tag{21}$$

where, the term $sign(.)$ indicates a sign function. Thus, the ECO algorithm improves exploration capability by avoiding local optima and slow convergence problems.

## 4. Proposed Methodology

The outline of the proposed methodology is illustrated in Figure 1. Initially, the data are preprocessed using different steps such as tokenization, stop word removal, stemming, and lemmatization. Tokenization is the process of splitting complex data (pages in the document) into simple units called tokens. Two types of tokenization are used where the documents are partitioned into words and sentences. Stemming is a normalization technique that converts the tokenized words to short words (word stem, root, or base form) to remove redundancy. Stopwords are the words such as "am, is, are, was, were, etc.", which are normally used to form the sentence but have no meaning in and of themselves. These words are eliminated in the stop word removal process. The lemmatization step overcomes the complexities associated with the stemming step by eliminating the intermediate word representation which may or not result in a meaningful word. The lemmatization performs a morphological analysis on the words and returns the meanings of the words in the appropriate form. Figure 1 shows all the steps from submitting a query to returning the top results in a proper sequence. The input to the proposed model is the query and content from the dataset for which it returns the appropriate answer as the result. The Word2Vec word embedding generates a meaningful vector for each word in the top retrieved feedback. The context of the words generated by the embedding process is identified using the iterative deep learning framework. This output serves a crucial part in the candidate set generation and these terms are sent to a term pool using the Okapi measure.

The words in the top retrieved feedback are retrieved and improved using the TF-IDF and cosine similarity. In this way, the top retrieved feedback is generated which compares the query and finds the similarity to retrieve the query with the top ranks. The hybrid group mean-based optimizer-enhanced chimp optimization (GMBO-ECO) algorithm is used for query expansion which selects the weighted terms of the top retrieved feedback obtained from using the TF-IDF and cosine similarity metrics. The GMBO-ECO algorithm enhances the effectiveness of the retrieval feedback for the entire set of queries. Query expansion mainly identifies the word which is mostly related to the input query keyword. The documents are then arranged based on their word similarity to the original query. The reformulated query is given as the input to the search engine where it obtains the top-ranked feedback documents. The feedback documents are then recommended to the users based on the assumptions that the retrieved top-ranked feedback documents do not contain irrelevant information that violate the user's query.

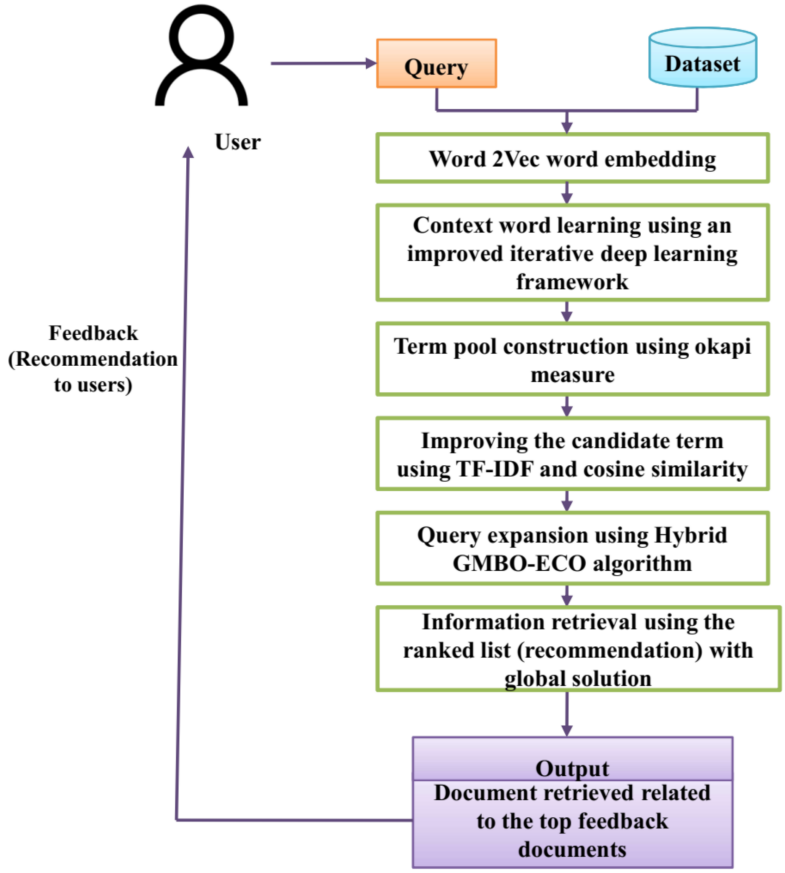


**Figure 1.** Outline of the proposed methodology.

### *4.1. Word2Vecfor Word Embedding*

Word2Vec is a vector representation paradigm for words. The similarity value may then be calculated using the Word2Vec model's word vector values and the cosine similarity formula. Several characteristics, such as the vector dimensions configuration and windows size, are employed in the building of the Word2Vec model, which is referred to as the training process [33,34]. The Content ($C$) and Query ($Q$) are the system's inputs, with the full phrase separated into individual terms:

$$\begin{aligned} Q &= l_{x1}, l_{x2}, l_{x3}, \ldots, l_{xn} \\ C &= t_{b1}, t_{b2}, t_{b3}, \ldots, t_{bm} \\ t_{b1} &= l_{t11}, l_{t12}, l_{t12}, \ldots, l_{t1c} \\ &\vdots \\ &\text{and so on} \end{aligned} \tag{22}$$

The N (i.e., $n + m$) term is combined to find the total embedded word with $m$ and $n$, where '$m$' is the words in the content and '$n$' is the total words in the query. With these N-words, the vocabulary size is retrieved $N^v$. The retrieved vocabulary size is unique. A 300-dimension-size $T^{word} \in R^{d\_size|N^v|}$ vector contains the total amount of N-words. The dimension size of this is set as a hyperparameter. Now, the individual word can be converted into the embedded vector by $M^w \times (N^v)$.

### *4.2. An Improved Iterative Deep Learning Framework for Context-Based Information Retrieval*

When searching a large dataset, context-based retrieval is critical for understanding the user's intent and generating the query result. To understand the context that exists in the query terms, we built a deep learning framework using the iterative deep convolu-

tional neural network architecture. By iterating the Deep Convolution Neural Network (DCNN) [35] and detector, a paradigm for iterative signal detection that combines the DCNN and detectors was suggested, which significantly enhanced detecting performance when there was associated noise. However, the framework's complexity increases, the performance benefit becomes marginal, and the number of iterations increases, especially when the error rate is high. If the identified contextual term is correct, no additional processing is required, reducing system complexity and overhead. As a result, an enhanced deep learning-based iterative detection framework is explained below. CRC must be applied to the decoded signal packet $\hat{Y}_d$ in each iteration. If the embedded word $\hat{Y}_d$ contains no errors, it does not participate in the following iteration. Otherwise, the embedded word must be re-entered and decoded by the detector. The estimation of the noise is calculated by the elimination and suppression of the noise. $\hat{n}(l)$ can be obtained by the following equation

$$\hat{n}(l) = z(l) - G\hat{e}(l) \tag{23}$$

where, the term $\hat{e}(l)$ represents the estimated embedded word produced by the Viterbi decoding [36]. The greater decoding and detector performance, the closer $e(l)$, and $\hat{e}(l)$ the closer $\hat{n}(l)$ is to $n(l)$. Second, $\widetilde{n}(l)$ obtains the estimated noise by inputting $\hat{n}(l)$ into the DCNN, $\widetilde{n}(l)$ which represents as more similar to $n(l)$ then $\hat{n}(l).\widetilde{z}(l)$ is the estimated embedded word obtained in the following equation

$$\widetilde{z}(l) = z(l) - \widetilde{n}(l) \tag{24}$$

$$G\hat{e}(l) = Ge(l) + k(l) \tag{25}$$

where the term $k(l) = n(l) - \widetilde{n}(l)$ represents noise in the residual output. The network's purpose is to minimize effective residual noise and obtain a more precise approximation $\widetilde{z}(l)$. Following that, $\widetilde{z}(l)$ is given back to the detector and decoded to begin the next cycle.

A DCNN is built with an $L$ convolutional layer. This networks' every layer consists of $M_j$ convolution kernels and input data, where $j \in [1, 2 \ldots, L]$. The previous layers' output is fed as an input to the current layer. Here the convolutional layer is represented as $kernel_{(J,m)}$, $m \in [1, 2 \ldots, M_J]$. For input data, first, the zero-padding operation is performed before the convolution operation is performed. By this technique, even after the convolutional operation is performed, the same input data size is maintained. So, the feature map output of the 1D convolutional operation performed using a convolutional kernel size $K_j$ is achieved by setting the Rectified Linear Unit (ReLU) as an activation function. The initial input layer is set as a zero layer that receives $\hat{n}(l)$ then acts as the first layers' input after the zero-padding operation. The $M_J$ convolutional kernel is created by the first layer, and every kernel is referred to as $kernel_{(J,m)}$ of $K_j$ size. The convolution operation is executed at first between padded input data and kernels, followed by the addition of bias. Then, the result of the convolution is applied to the ReLU activation function to achieve the $M_1$ feature maps, which act as the first layers' output and the second layers' input. This procedure continues until the network reaches its last layer. As the network's final output, the final layer outputs $\widetilde{n}(l)$, whose size is the same as $\hat{n}(l)$. Generally, DCNN is represented as, $\{L;\ K_1,\ K_2, \ldots K_L; M_1, M_2 \ldots, M_L\}$.

In a deep convolutional neural network (DCNN), the network metrics updates will affect the loss function, and the training direction is also determined by the loss function. The loss function is used to train the networks. If $\widetilde{x}(o)$ is near to $x(o)$ then the value of $s(o)$ is 0. Then, the Gaussian distribution deviates $s(o)$ and the Gaussian distribution is not followed $s(o)$. Hence the loss function is a required design. Not only does the non-Gaussian distribution repress the noise, it also deals efficiently. Initially, the candidate's probability is

required to build and compute the exact probabilities for every candidate in the *oth* symbol. The probabilities for the actual candidates are represented by $r_j(o)$ which is expressed as

$$r_j(o) = \frac{\left\| z(o) - \widetilde{x}(o) - It_j \right\|^2}{\sum_{k=1}^{|\Theta|} \left\| z(o) - \widetilde{x}(o) - It_k \right\|^2} \tag{26}$$

Considering the derivation $r_j(o)$ for $t_j$ and it is expressed as

$$r_j(o) = \frac{e^{-\| z(o) - \widetilde{x}(o) - It_j \|^2}}{\sum_{k=1}^{|\Theta|} e^{-\|z(o) - \widetilde{x}(o) - It_j\|_2}} \tag{27}$$

The signal identification issue is the multi-category issue utilized as the one-hot method, so the candidate probabilities are rewritten as

$$R(o) = \left[ r_1(o), r_2(o) \ldots\ldots r_{1\Theta|}(o) \right] \tag{28}$$

where $\sum_{j=1}^{|\Theta|} r_j(o) = 1$. The distribution of the true symbol is represented by $q_j(o)$ and the one-hot method is utilized for the representation of categorical data where $q_j(o)$ = 1, then $t_j$ is 0 or correct otherwise. The true symbol possibility is denoted by

$$q(o) = \left[ q_1(o), q_2(o), \ldots\ldots q_{|\Theta|}(o) \right] \tag{29}$$

where, $\sum_{j=1}^{|\Theta|} q_j(o) = 1$, the difference between the probability distribution $q_j(o)$ and the evaluated probability distribution is $r_j(o)$, the cross-entropy is utilized for modifying it. Thus, the cross-entropy among $R(o)$ and $Q(o)$ is expressed as;

$$I(Q, R) = \frac{1}{O} \sum_{o=1}^{O} \sum_{j=1}^{|\Theta|} q_j(o) \log_2 \frac{1}{r_j(o)} \tag{30}$$

For the enhancement of the decoding performance and detection, $I(Q, R)$ cross-entropy will be reduced. Particularly, the true distribution $Q$ and the computed distribution $R$ are equal, $I(Q, R)$ which is exceedingly near to zero. Then the loss function is expressed as

$$g_f = \sum_{o=1}^{O} \sum_{j=1}^{|\Theta|} \frac{q_j(o)}{O} \log_2 \frac{\sum_{k=1}^{|\Theta|} e^{-\|z(o) - \widetilde{x}(o) - It_k\|2}}{e^{-\|z(o) - \widetilde{x}(o) - It_j\|_2}} \tag{31}$$

The first $|\Theta|$ is the group of user's actual queries which represent the one-hot encoding and the second $|\Theta|$ is the group of all top feedback documents. The loss function assisted to enhance the performance of detection by decreasing the noise and repressing the difference among the $s(o)$ residual noise and the Gaussian distribution. The Adam optimizer is used for reducing the loss functions. For noise suppression in the loss function, the mean squared error (MSE) is used. The loss $O$ at time intervals is acquired by

$$loss = \frac{1}{O} \sum_{o=1}^{O} \left\| x(o) - \widetilde{x}(o) \right\| \tag{32}$$

$N_s$ antennas, receiver and the loss function MSE is expressed as

$$g_n = \frac{\sum_{o=1}^{O} \sum_{j=1}^{N_S} \left\| x_j(o) - \widetilde{x}_j(o) \right\|}{N_S O} \tag{33}$$

The $j$th elements of estimated noise and true noise in the $o$th time slot are represented by $\widetilde{x}_j(o)$ and $x_j(o)$ The loss function MSE is utilized for suppressing the $s(o)$ residual noise. The Gaussian distribution deviates from the residual noise and the non-Gaussian distribution effect is not handled effectively in MSE. The iterative detection is utilized for enhancing the performance of detection. The cyclic redundancy check (CRC) is used to identify if the user relevance feedback comprises an error or not for controlling the iteration of identification that can be used to minimize the computation difficulty. The objective of the network is to minimize the $s(o)$ residual noise and retrieve the $x(o)$ noise. Because of limitations, the conventional detector performance that obtains the $\hat{x}(o)$ estimated noise initially, deviates the noise $x(o)$. After the network metrics training, the network output is obtained through the input $\hat{x}(o)$ within the network that approximates the $x(o)$. The words present in the query determine the importance of the results retrieved. The iterative deep learning architecture assigns a positive value to terms of greater relevance and a negative value to those that are unrelated to the query. These values are then taken for candidate terms generation and term pool construction by applying different similarity measures before query expansion.

### *4.3. Term Pool Construction Using Okapi Ranking Measure*

For the construction of the term pool (retrieval feedback documents), initially, top $N$ documents are retrieved by means of the improved iterative deep learning framework. Here, the Okapi ranking measure is utilized as a matching function for the document retrieval process. The numerical expression of the Okapi measure is defined based on the following expression.

$$OKAPI(q, d_x) = \sum_{t \in q} \omega \frac{(C_1 + 1)\ T_f}{C + T_f} \times \frac{(C_3 + 1)\ Q_{Tf}}{C_3 + Q_{Tf}} \tag{34}$$

The query with $t$ words is represented as $q$; the constant parameters $C_1 = 1.2$ and $C_3 = 7.0$. Moreover, $T_f$ and $Q_{Tf}$ indicate term frequency in $d_x$ document and term frequency in query $q$.

$$C = C_1(1 - k) + \left(k \cdot \frac{DL}{DL_A}\right) \tag{35}$$

From the above equation, $k$ signifies the constant parameter, $DL$ implies document length, and $DL_A$ depicts average document length.

$$\omega = \log \frac{(L - N + 0.5)}{(N + 0.5)} \tag{36}$$

where, the term $L$ and $N$ depict a total number of documents and document numbers comprising of terms. Each document is sorted based on the Okapi ranking function. Moreover, the individual terms of the whole document are chosen based on the co-occurrence of query terms. The co-occurrence is evaluated by means of the Jaccard coefficient $J_{C\_}Co$ which is formulated in the following expression,

$$J_{C\_}Co(T_a, T_b) = \frac{D_{ab}}{D_a + D_b - D_{ab}} \tag{37}$$

From the above equation, co-occurrence terms are represented as $T_a$ and $T_b$; document numbers in which the terms tend to occur are depicted as $D_a$ and $D_b$; document numbers of $T_a$ and $T_b$ are depicted as $D_{ab}$. This coefficient is employed to estimate similarities among document terms and query terms. The co-occurrence degree of the candidate term $C_a$ with respect to the query term is defined by the inclusion of inverse document frequency (IDF) and the application of normalization using the below equation.

$$Co\_d(C_a, T_a) = \log_{10}(Co(C_a, T_a) + 1) * \left(IDF(C_a) / \log_{10}(d)\right) \tag{38}$$

$$IDF(C_a) = \log_{10}\left(\frac{L}{L_C}\right) \tag{39}$$

where, the terms $L$, $L_C$ and $d$ indicate total documents in the corpus, the number of documents containing candidate terms, and top-ranked document numbers. In order to achieve better $C_a$ from the whole query, the co-occurrence degree is incorporated with each individual query which is illustrated in the below expression as,

$$q = F(C_a, q) = \prod_{T_a \; in \; q} \left(\chi + Co\_d(C_a, T_a)^{IDF(T_a)}\right) \tag{40}$$

The above equation offers the suitability score for sorting the terms that are interrelated with the overall query.

*4.4. Improving Query Terms*

For a good combination of extension terms, the improving query term approach is employed. The user's input is used to add a specific weight to the retrieved documents and query keywords. The major step of this method is to estimate the weights in each term. The query term present in the term pool is obtained from the vector $t = T_1,\ T_2,\ T_3,\ \ldots\ldots$. The total weight of each query is given below,

$$\sigma_{T_j} = \sum_{N}^{i=1} \sigma_{T_j}^{e_i} \tag{41}$$

where, $\sigma_{T_j}$, $N$, $\sigma_{T_j}^{e_i}$ is the mean total weight, the number of selected documents from the retrieved documents, and the weight of documents. The $\sigma_{T_j}^{e_i}$ is only valid when $e_i$ is applied according to the requirement of the user. The weight of the document term is defined as below,

$$\sigma_{T_j}^{e_i} = \eta_{T_j}^{e_i} \times \gamma_{e_i} \tag{42}$$

The traditional weighting model (TF-IDF) is utilized for calculating the weight of all query terms based upon the document of $e_i$. The increasing value of $\eta_{T_j}^{e_i}$ is directly proportional to the number of times $T_j$ appears in the documents and the decreasing value depends on the number of documents present in the collection. The high value of $\eta_{T_j}^{e_i}$ is computed from the term $T_j$ which has a low frequency in document collection and high frequency from the document. To examine the obtained primary query documents, the retrieved documents are assumed as a perfect match for the user. Each document was measured by using the distance between two retrieved documents. The similarity between the top document and another document in the retrieved document set is calculated by,

$$\gamma_{e_i} = \frac{\sum_{K=1,\ K\neq 1} sim\cos\left(\overrightarrow{e_M},\ \overrightarrow{e_i}\right)}{N-1} \tag{43}$$

where, $\overrightarrow{e_M}$, $\overrightarrow{e_i}$ denotes Euclidean vectors in the retrieved set and $sim$ cos is the cosine similarity.

At last, the total weight after applying log normalization method is shown in the below equation,

$$\sigma_{T_j} = \log\left(1 + \sum_{i=1}^{N} f_{T_j}^{e_i} \times jef_{T_j} \times \gamma_{e_i}\right) \tag{44}$$

The value one is added to the logarithmic function for avoiding the zero values. The $\sigma_{T_j}$ is related to the occurrence of the top documents $t$ term and the inverse document frequency of $T$ terms in the whole collection.

### *4.5. Query Expansion Using the Hybrid GMBO-ECO Algorithm*

The investigation revealed that the group mean-based optimizer (GMBO) algorithm has the ability to obtain globally optimal solutions but it affects the movement of individuals which results in fast convergence problems. Moreover, the GMBO algorithm also suffers from local optima problems. Therefore, to overcome such difficulties, the enhanced chimp optimization (ECO) algorithm is hybridized to achieve optimal solutions. The hybridized flow diagram of GMBO along with the ECO algorithm is delineated in Figure 2. The fitness function is determined by the query's appropriateness or goodness in obtaining relevant documents, as assessed by recall or precision. As a fitness function, the recall of the retrieved result is employed.

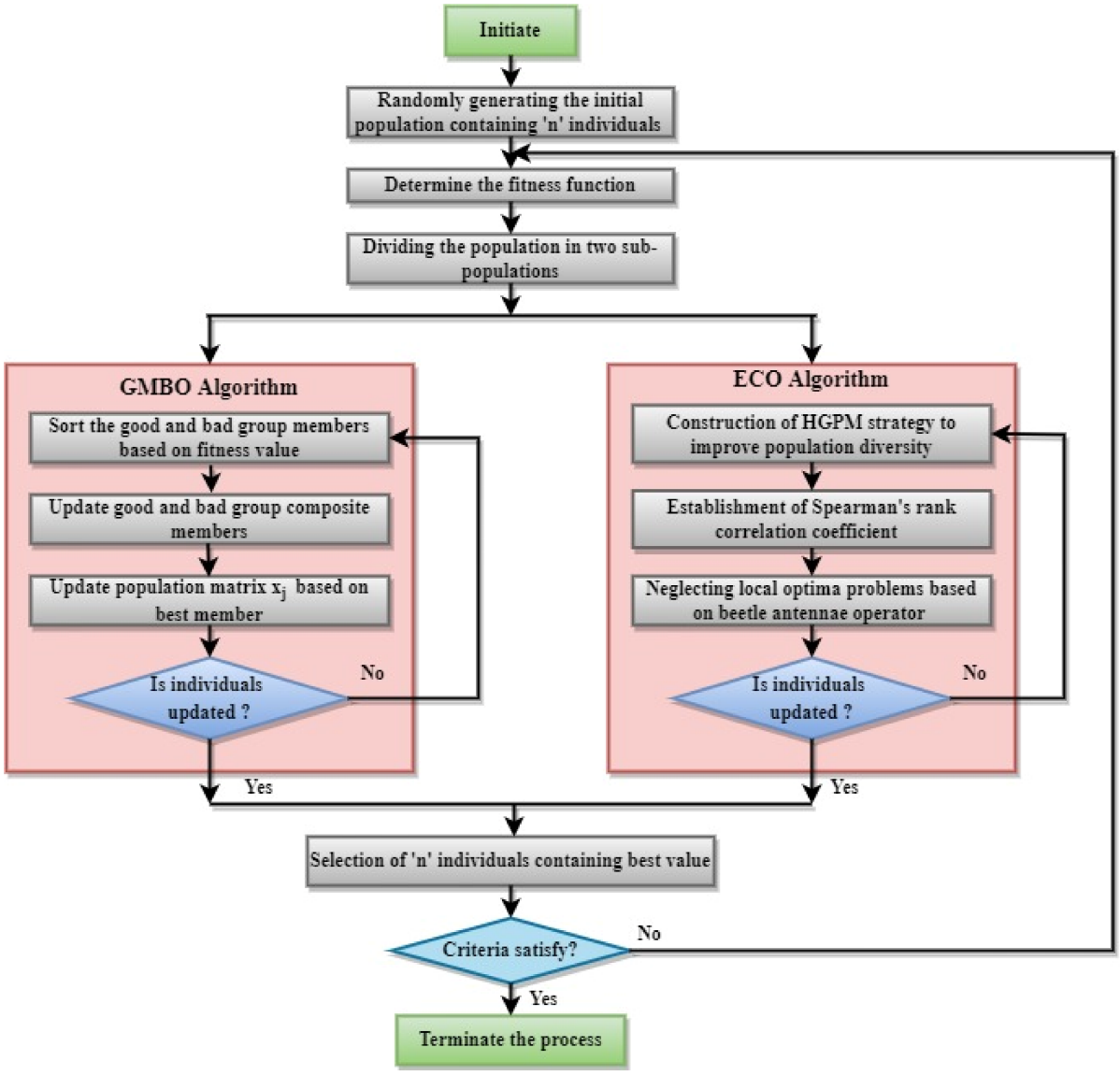


**Figure 2.** Hybrid GMBO-ECO algorithm.

Highly disruptive polynomial mutation (HDMP) Spearman's rank correlation coefficient and Beetle antennae operators are applied for individuals in the populations. The pseudocode of the Hybrid GMBO-ECO algorithm for query expansion is presented in Algorithm 1. In general, query expansion refers to the expansion of a query using the most commonly connected words (i.e., the word containing high fitness) that are close to the query keyword. Here, a particular query is entered by the user and the queries

are expanded by identifying the relevant words using the proposed technique. Hence unambiguous solutions are achieved by assigning a fitness value, thereby determining the most frequent concepts or words for a specified query. The query expansion method uses the weighted candidate set terms as well as global knowledge to deal with the uncertainty present in-between the relationships. The matching processes in retrieving information are carried out by employing a similarity measure which is stated below.

$$S\ (Q,\ D) = \frac{\sum_j Q_j D_j}{\sqrt{\sum_j Q_j^2}\ \sqrt{\sum_j D_j^2}} \tag{45}$$

From the above equation, $S$ signifies the similarity measure and $Q_j D_j$ denotes the query and the document. During the process of query expansion, the queries are expanded by choosing the most relevant word for a particular query. By applying the expanded query to the matching process, $(Q,\ D)$ is converted into $(Q',\ D')$. Thus, the similarity measure $S_{New}\ (Q',\ D')$ becomes,

$$S_{New}\ (Q',\ D') = \frac{\sum_j Q'_j D'_j}{\sqrt{\sum_j Q'^2_j}\ \sqrt{\sum_j D'^2_j}} \tag{46}$$

Then, the document is ranked based on the similarity of words and as the queries expand, the document is also refined. The top retrieved documents values are mainly ranked using the Equation (49) for query expansion term selection. An example of the query keywords and the expanded terms is presented in Table 1.

**Algorithm 1:** Pseudocode of the hybrid GMBO-ECO algorithm.

**Input:** Query term, fitness function, and candidate_term_set (terms in the top retrieved feedback document)
**Output:** Optimal solution
Initialize population, Maximum_generation = 1, and end_criterion = optimal solution
Candidate_term_set = terms selected by iterative deep learning model and improved with co-occurrence score
While (query_expansion (Query, candidate_term_set)) do
While (iteration<Maximum_generation) or (end_criterion met) do
Select a solution randomly using the GMBO algorithm
Apply the TF-IDF function and cosine similarity and save it in the index file
Compute the fitness value of the current solution
If the fitness value of the current solution is higher than the previous solution then
Replace the new solution
End If
The worst individual in the population is eliminated and the new ones are built using the different operations of the ECO algorithm
Rank the current best solution by leaving the best solutions obtained
End
Output the result and terminate the process
End

**Table 1.** Expanded keywords for queries using the hybrid GMBO-ECO algorithm.

| Query | Keyword | Expanded Terms |
|---|---|---|
| 1 | Examination | Test exam viva questionnaire assessment |
| 2 | Tiger | Carnivore feline lynx jaguar cheetah |
| 3 | Plant | Flora Vegetation Tree wood Shrub |
| 4 | Dance | Breakdance Twirl disco rock party |

## 5. Experimental Result and Discussion

### *5.1. Hardware and Softwareplatform*

The hardware specification of the PC on which we performed the experiments includes an Intel(R) Core(TM) i5-8265U 1.60 GHz CPU and 8 GB primary memory (RAM). The Windows 10 operating system was installed on the PC. For the simulation of results, Terrier, an open-source search engine that comprises a massive set of documents for text retrieval [37] was used. The terrier is an open-source platform for text retrieval. It provides standard indexing and retrieval functionalities and serves as an efficient platform for large-scale retrieval applications. The terrier code is written in Java and developed by the information retrieval group from Glasgow University. The proposed methodology is tested on the standard TREC and CLEF test collections. The python bindings of the terrier are called Pyterrier which allows the experiments to be held simply and explicitly using the Google Colaboratory, or "Colab" and Jupiter notebooks along with the benefits obtained from Terrier.

### *5.2. Parameter Setting*

The comparison is mainly held by using a diverse number of top feedback documents in the range 5–50 for both the ECO-based GMBO and the existing algorithms to select the optimal number of feedback documents. The optimal number of top feedback documents is set as 15 since the proposed model offers improved performance for the top 15 documents. In the query term enhancement step, the query expansion terms are selected in the range 15, 20, 30, 40, and 65 and these are the candidate terms ranked based on the similarity value. The proposed methodology offered optimal performance for a total of 40 terms and these 40 terms were selected for query reformulation based on the user's input search term. The parameters of the ECO and GMBO algorithm are listed in Table 2.

**Table 2.** Parameter settings for the proposed method.

| Techniques | Parameters | Ranges |
|---|---|---|
| Enhanced chimp optimization (ECO) algorithm | Size of population | 30 |
| | Maximum number of iterations | 200 |
| | Rank correlation coefficient | $[-1, 1]$ |
| | Attenuation rate $(A_R)$ | 2 |
| | $C$ | 0.95 |
| Group mean-based optimization (GMBO) algorithm | Population size | 30 |
| | Total iterations | 200 |
| | Random number | $[0, -1]$ |

### *5.3. Simulation Measures*

The performance metrics such as precision, *MAP*, *Recall*, *F-measure*, *NDCG*, and *MRR* were evaluated to determine the efficiency of the proposed method.

*Precision (P):*

Precision is defined as the ratio of the total number of the retrieved document which is relevant ($T_{RD}$) to the number of retrieved documents ($T_D$ ) from the query and is expressed as,

$$P = \frac{T_{RD}}{T_D} \tag{47}$$

*Mean average precision (MAP):*

The arithmetic mean of the precision score of the retrieved document over a group of $k$ queries is termed as the mean average precision. The $MAP$ is calculated as,

$$MAP = \frac{1}{k}\sum_{k} AP_k \tag{48}$$

*Recall (R):*

The recall is defined as the ratio of the total number of the retrieved documents which is relevant ($T_{RD}$) to the total number of the relevant documents ($T_R$ ) in the database and is expressed as,

$$R = \frac{T_{RD}}{T_R} \tag{49}$$

*F-measure:*

The harmonic mean of precision ($P$) and recall ($R$) is considered as the *F-measure* and is defined as,

$$F - measure = \frac{2PR}{P+R} \tag{50}$$

*Normalized Discounted Cumulative Gain:*

The ratio of discounted cumulative gain ($DCG$) to the ideal discounted cumulative gain ($IDCG$) is defined as the normalized discounted cumulative gain. $NDCG$ is expressed as,

$$NDCG = \frac{DCG}{IDCG} \tag{51}$$

*Mean reciprocal rank (MRR):*

Mean reciprocal rank is calculated as,

$$MRR = \frac{\sum_{d=1}^{k} \frac{1}{RANKd}}{k} \tag{52}$$

### *5.4. Performance Analysis*

The performance of the proposed method is determined by comparing with various existing methods such as the Spider Monkey Crow Optimization algorithm with a deep recurrent neural network (SMCA-DRNN), Modified Stellarmass Blackhole Optimization (MSBO), the Fuzzy Ontology Framework (FOF) and the Boosting Algorithm (BA) with query and without query expansion. The proposed method has achieved better efficiency in terms of precision with and without query expansion and it is summarized in Table 3.

**Table 3.** Comparative analysis of methods with and without query expansion.

| Methods | Without Query | | | With Query | | |
|---|---|---|---|---|---|---|
| | **P@10** | **P@20** | **P@30** | **P@10** | **P@20** | **P@30** |
| SMCA-DRNN | 0.3572 | 0.3269 | 0.3147 | 0.3873 | 0.3771 | 0.3548 |
| MSBO | 0.3571 | 0.3374 | 0.3206 | 0.4547 | 0.4286 | 0.4155 |
| FOF | 0.3904 | 0.3808 | 0.3619 | 0.4510 | 0.4023 | 0.3985 |
| BA | 0.3829 | 0.3627 | 0.3495 | 0.3976 | 0.3918 | 0.3529 |
| Proposed Method | **0.4434** | **0.4105** | **0.3864** | **0.4568** | **0.4437** | **0.4123** |

In the retrieval experiments, if we increase the number of retrieved documents then the precision decreases. From the precision results given in Table 3, this important tradeoff can be observed clearly. In the category of without query expansion results the highest precision was reported with the minimum number of retrieved documents i.e., 10. (P@10 mean precision at level of 10 documents). Similarly, in the category of expanded query precision results the best precision value was achieved at 10 retrieved documents.

Figure 3 depicts the comparative analysis of the mean average precision of the proposed method without query expansion. The proposed method is compared with SMCA-DRNN, MSBO, FOF and BA without query expansion. The proposed method obtains a higher MAP rate than other existing methods.

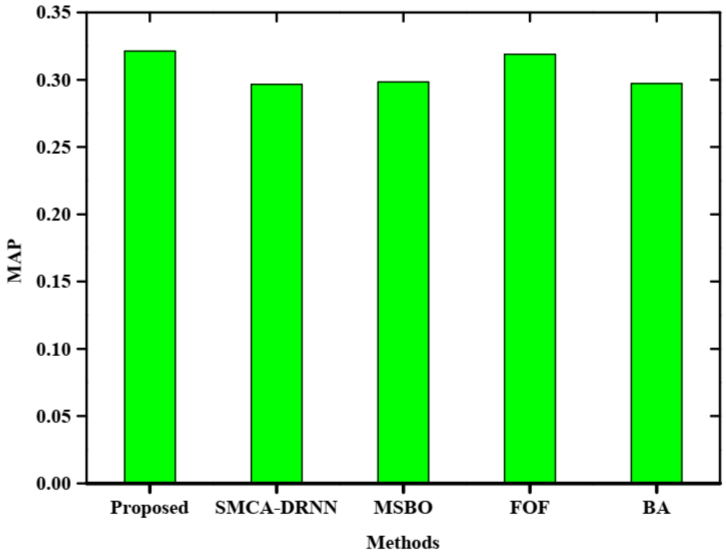


**Figure 3.** Comparative analysis of MAP without query expansion.

Figure 4 represents the comparative analysis of the mean average precision of the proposed method with query expansion. The X-axis specifies the various methods and the Y-axis specifies MAP. The proposed method obtains a higher MAP rate than SMCA-DRNN, MSBO, FOF and BA, respectively. The proposed method attained a 0.35 MAP rate.

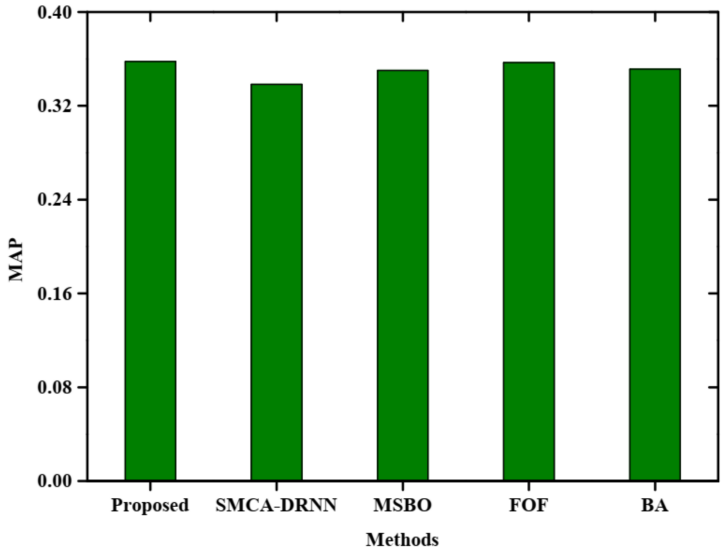


**Figure 4.** Comparative analysis of MAP with query expansion.

Figure 5 shows the comparative analysis of the Precision–Recall curve without query expansion. The proposed method is compared with existing methods such as SMCA-DRNN, MSBO, FOF and BA, respectively. The precision rate is depleted when the recall rate increases. However, the figure shows an improvement in information retrieval and it is mainly achieved due to the usage of the proposed GMBO-ECO algorithm.

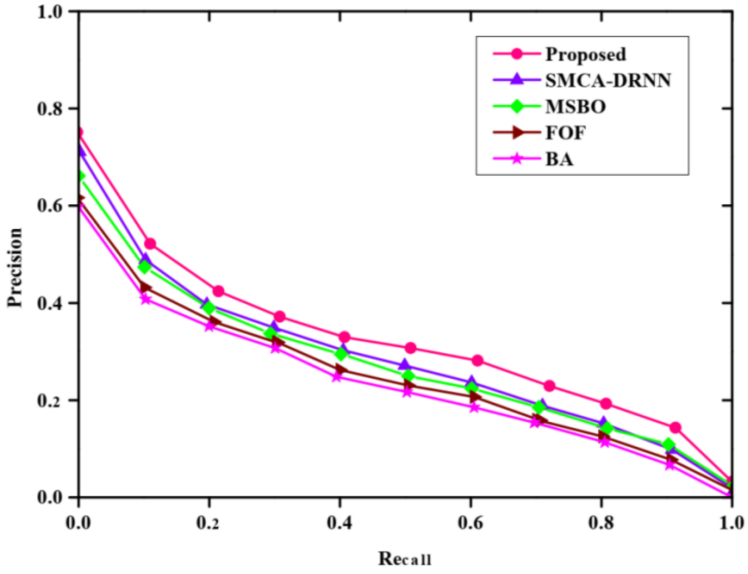


**Figure 5.** Comparative analysis of precision and recall without query expansion.

Figure 6 shows the comparative analysis of the Precision–Recall curve with query expansion. The proposed method is compared with existing methods such as SMCA-DRNN, MSBO, FOF and BA, respectively. The precision rate gets depleted when the recall rate increases. However, the figure shows an improvement in information retrieval and it is mainly achieved due to the usage of the proposed GMBO-ECO algorithm. Further, Figures 4 and 5 shows that accurate technique selection is more important for improving query extension.

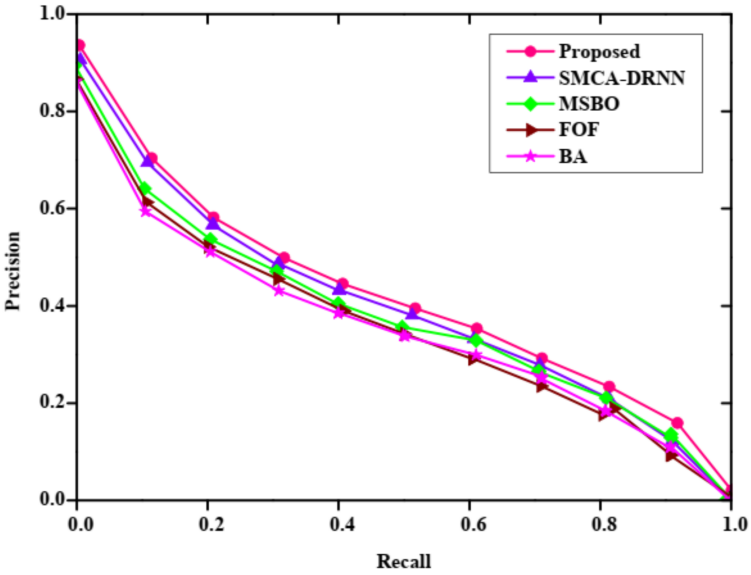


**Figure 6.** Comparative analysis of precision and recall with query expansion.

Figure 7 describes the F-measure of without query expansion and the X, Y axis represent the number of queries and F-measure, respectively. In this plot, there are five methods used such as the proposed, SMCA-DRNN, MSBO, FOF, and BA. From this plot, the proposed method has a high F-measure value when compared with other methods. The BA method has a very low F-measure value.

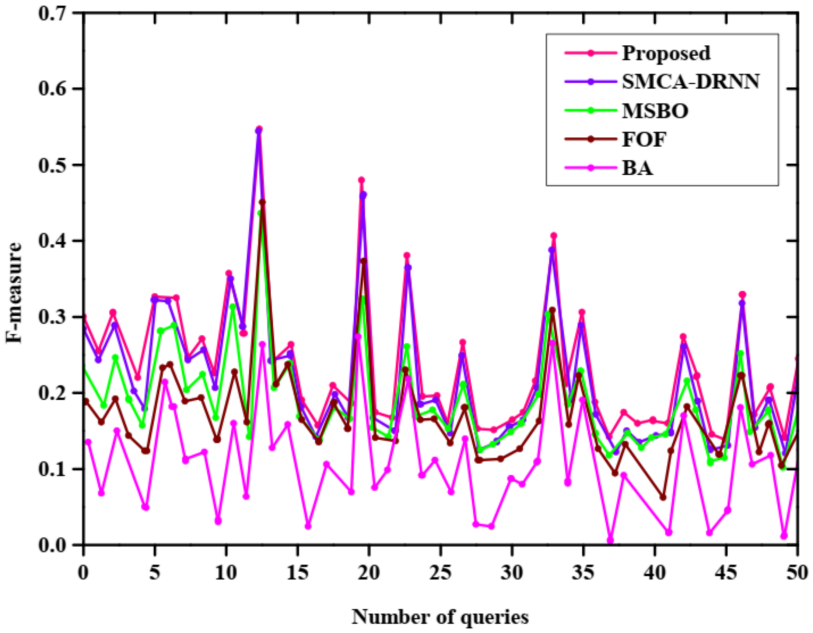


**Figure 7.** F-measure without query expansion.

Figure 8 represents the F-measure of with query expansion and X, Y denotes the F-measure and number of queries, respectively. To predict the best F-measure value, five methods are employed that are BA, SMCA-DRNN, FOF, MSBO, and the proposed method. The proposed method has a high F-measure value compared with other methods and the BA method has a very low F-measure value.

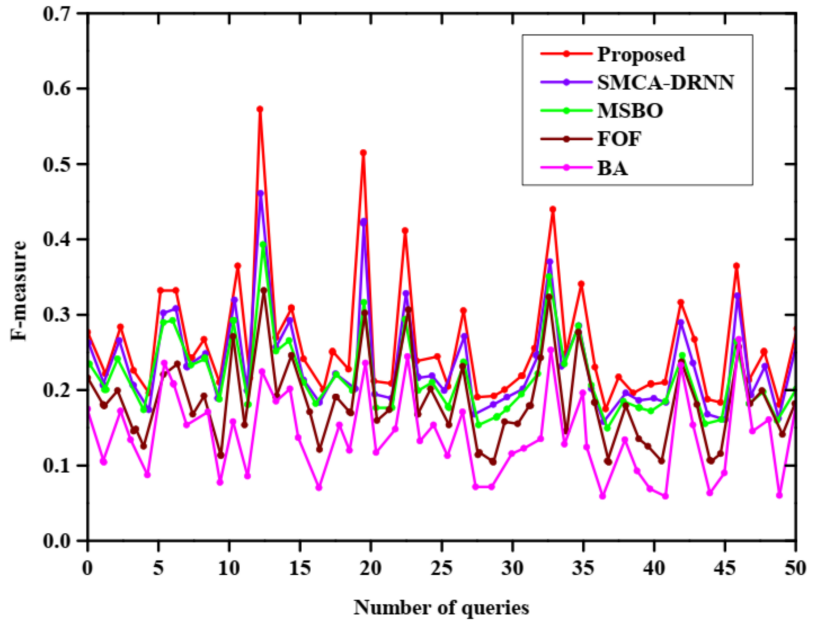


**Figure 8.** F-measure with query expansion.

Figure 9 shows the mean reciprocal rank (MRR) of without query expansion and X, Y axis represents techniques and MRR, respectively. Here, five techniques are used to determine the best MRR value, such as SMCA-DRNN, FOF, MSBO, BA, and the proposed method. From this graph, the proposed technique has a high mean reciprocal rank of 0.89 and the BA technique has a very low mean reciprocal rank compared with other techniques.

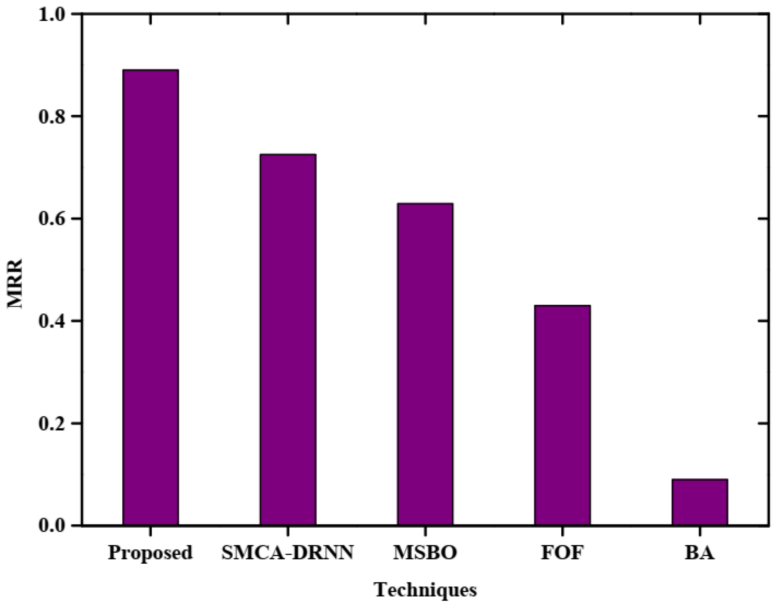


**Figure 9.** Mean reciprocal rank without query expansion.

Figure 10 represents the mean reciprocal rank (MRR) of with query expansion and the X, Y axis denotes the techniques and MRR, respectively. Five techniques such as the proposed, BA, MSBO, SMCA-DRNN, and FOF were utilized for examining the high MRR rate. This plot shows the proposed technique has a high MRR rate of 0.95 and the BA technique has a low MRR rate compared with other techniques. Figure 11 represents the comparative analysis of the normalised discounted cumulative gain (NDCG) of the proposed method without query. The graph is plotted between various techniques and NDCG. The proposed method is compared with SMCA-DRNN, MSBO, FOF and BA without query. The proposed method obtains a higher NDCG than other existing methods. The obtained normalized discounted cumulative gain without query of the proposed method is 0.82.

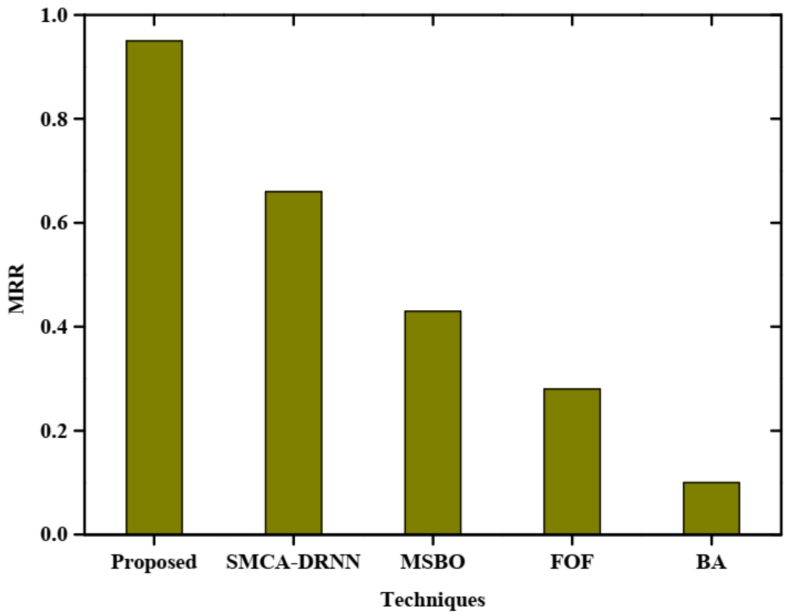


**Figure 10.** Mean reciprocal rank with query expansion.

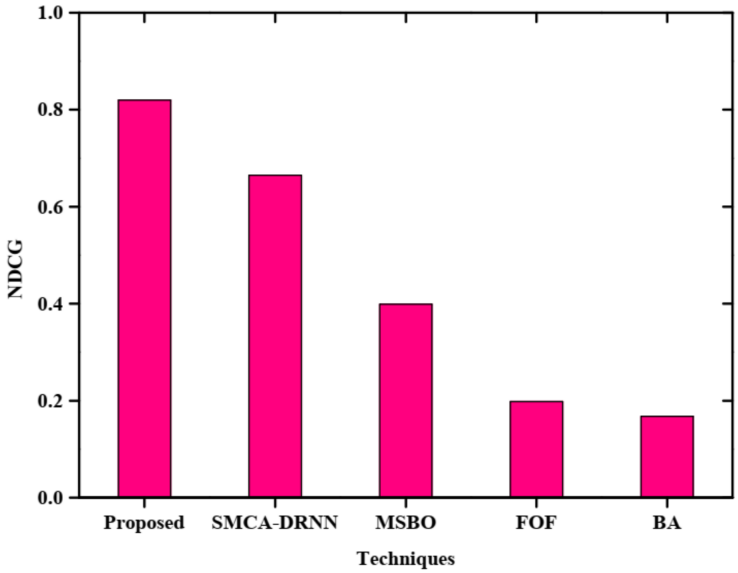


**Figure 11.** Comparative analysis of NDCG without query expansion.

Figure 12 represents the comparative analysis of the normalised discounted cumulative gain (NDCG) of the proposed method with query. The graph is plotted between various techniques and NDCG. The proposed method is compared with existing methods such as SMCA-DRNN, MSBO, FOF and BA, respectively. The proposed method obtains a higher NDCG than other existing methods with query. The obtained normalized discounted cumulative gain with query of the proposed method is 0.87.

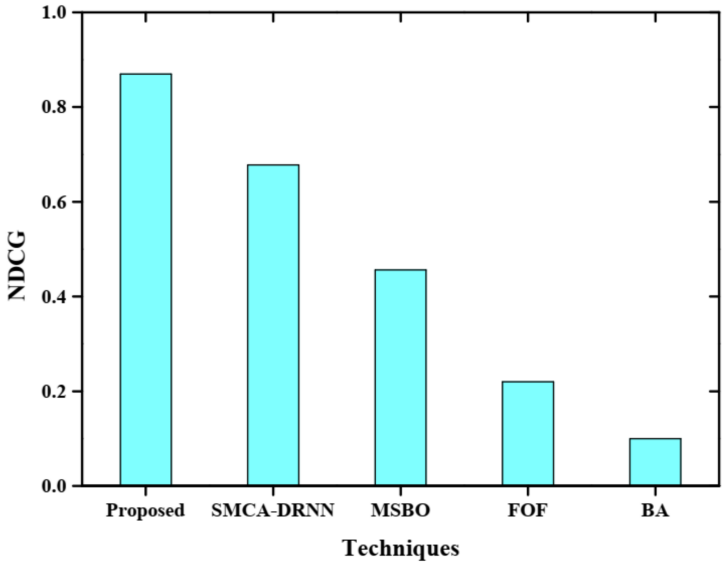


**Figure 12.** Comparative analysis of NDCG with query expansion.

## 6. Conclusions

This paper presents the hybrid group mean-based optimizer-enhanced chimp optimization (GMBO-ECO) algorithm for automatic query expansion from a retrieval feedback corpus which selects the weighted terms derived from the TF-IDF and cosine similarity metrics that are mostly linked to the input query keyword. The query and content from the dataset are fed into the proposed model, and this model delivers a suitable response as a result. For each word in the text, the Word2Vec word embedding creates a meaningful vector. The iterative deep learning framework is used to determine the context of the words created by the embedding process. The Okapi measure is used to transfer these

terms to a term pool which plays an important role in the candidate set creation. The TF-IDF and cosine similarity are used to retrieve and refine the terms in the candidate set. TREC and CLEFtest sets are primarily used to implement the proposed methodology. The proposed methodology enhances the performance of the information retrieval system, and it also returns more relevant results to the users. The efficiency of the proposed method is determined by evaluating various metrics such as precision, MAP, Recall, F-measure, NDCG, and MRR. The proposed method achieved a higher MAP, NDCG, MRR, F-measure, precision, and recall rate by comparing various methods such as SMCA-DRNN, MSBO, FOF, and BA with and without query expansion. A normalized discounted cumulative gain of 0.87 was obtained using the proposed method which is relatively higher than the existing techniques. The proposed technique has a high MRR rate of 0.95 and a high mean reciprocal rank of 0.89. The Hybrid GMBO-ECO retrieval feedback technique not only works well for query expansion but also considerably increases system accuracy. In the future, further improvement in the retrieval performance may be achieved by varying the number of query expansion terms using the Kullback Leibler QE approach. This method also adds similar terms to the original user query, such as the pseudo relevance feedback.

**Author Contributions:** Conceptualization, R.K.; methodology, R.K.; software, R.K.; validation, K.N.T.; formal analysis, R.K.; investigation, K.N.T.; resources, S.C.S.; data curation, R.K.; writing—original draft preparation, R.K.; writing—review and editing, R.K. and K.N.T.; visualization, R.K. and K.N.T.; supervision, S.C.S.; project administration, S.C.S. 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:** Publicly available datasets from Text REtrieval Conference (TREC) and Cross-Language Evaluation Forum (CLEF) test collections were analyzed in this study.

**Conflicts of Interest:** The authors declare that there are no conflict of interest regarding the publication of this paper.

## References

1. Kumar, R.; Sharma, S.C. Information retrieval system: An overview, issues, and challenges. *Int. J. Technol. Diffus. (IJTD)* **2018**, *9*, 1–10.
2. Liu, H.; Zheng, C.; Li, D.; Zhang, Z.; Lin, K.; Shen, X.; Xiong, N.N.; Wang, J. Multi-perspective social recommendation method with graph representation learning. *Neurocomputing* **2021**, *468*, 469–481.
3. Liu, H.; Nie, H.; Zhang, Z.; Li, Y.-F. Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction. *Neurocomputing* **2020**, *433*, 310–322.
4. Liu, T.; Liu, H.; Li, Y.-F.; Chen, Z.; Zhang, Z.; Liu, S. Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing. *IEEE Trans. Ind. Inform.* **2019**, *16*, 544–554.
5. Azad, H.K.; Deepak, A. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.* **2019**, *56*, 1698–1735.
6. Nasir, J.A.; Varlamis, I.; Ishfaq, S. A knowledge-based semantic framework for query expansion. *Inf. Process. Manag.* **2019**, *56*, 1605–1617.
7. Saleh, S.; Pecina, P. Term selection for query expansion in medical cross-lingual information retrieval. In *Advances in Information Retrieva, Proceedings of the 41st European Conference on IR Research*; Clogne, Germany, 14–18 April 2019, Springer: Cham, Switzerland, 2019; pp. 507–522.
8. Sharma, D.K.; Pamula, R.; Chauhan, D.S. A hybrid evolutionary algorithm based automatic query expansion for enhancing document retrieval system. *J. Ambient Intell. Humaniz. Comput.* **2019**, 1–20. [CrossRef]
9. Azad, H.K.; Deepak, A. A new approach for query expansion using Wikipedia and WordNet. *Inf. Sci.* **2019**, *492*, 147–163.
10. Sahin, D.O.; Kural, O.E.; Kilic, E.; Karabina, A. A text classification application: Poet detection from poetry. *arXiv* **2018**, arXiv:1810.11414.
11. Deng, X.; Li, Y.; Weng, J.; Zhang, J. Feature selection for text classification: A review. *Multimedia Tools Appl.* **2019**, *78*, 3797–3816.
12. Yan, D.; Li, K.; Gu, S.; Yang, L. Network-based bag-of-words model for text classification. *IEEE Access* **2020**, *8*, 82641–82652.
13. Liu, H.; Zheng, C.; Li, D.; Shen, X.; Lin, K.; Wang, J.; Zhang, Z.; Zhang, Z.; Xiong, N.N. EDMF: Efficient deep matrix factorization with review feature learning for industrial recommender system. *IEEE Trans. Ind. Inform.* **2021**, *18*, 4361–4371.

14. Liu, H.; Liu, T.; Zhang, Z.; Sangaiah, A.K.; Yang, B.; Li, Y.F. ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human-machine Interaction. *IEEE Trans. Ind. Inform.* **2022**, 1–11. [CrossRef]
15. Li, Z.; Liu, H.; Zhang, Z.; Liu, T.; Xiong, N.N. Learning knowledge graph embedding with heterogeneous relation attention networks. *IEEE Trans. Neural Networks Learn. Syst.* **2021**, 1–13. [CrossRef]
16. Raza, M.A.; Mokhtar, R.; Ahmad, N.; Pasha, M.; Pasha, U.; Rahmah, M.; Noraziah, A. A taxonomy and survey of semantic approaches for query expansion. *IEEE Access* **2019**, *7*, 17823–17833.
17. Chugh, A.; Sharma, V.K.; Kumar, S.; Nayyar, A.; Qureshi, B.; Bhatia, M.K.; Jain, C. Spider monkey crow optimization algorithm with deep learning for sentiment classification and information retrieval. *IEEE Access* **2021**, *9*, 24249–24262.
18. Rasheed, I.; Banka, H.; Khan, H.M. Pseudo-relevance feedback-based query expansion using a boosting algorithm. *Artif. Intell. Rev.* **2021**, *54*, 6101–6124.
19. Jain, S.; Seeja, K.R.; Jindal, R. A fuzzy ontology framework in information retrieval using semantic query expansion. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100009.
20. Raj, G.D.; Mukherjee, S.; Uma, G.V.; Jasmine, R.L.; Balamurugan, R. Query expansion for patent retrieval using a modified stellar-mass black hole optimization. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 4841–4853.
21. Malik, S.; Shoaib, U.; Bukhari, S.A.C.; El Sayed, H.; Khan, M.A. A hybrid query expansion framework for the optimal retrieval of the biomedical literature. *Smart Health* **2021**, *23*, 100247.
22. Safder, I.; Hassan, S.-U. Bibliometric-enhanced information retrieval: A novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics* **2019**, *119*, 257–277.
23. Almarwi, H.; Ghurab, M.; Al-Baltah, I. A hybrid semantic query expansion approach for Arabic information retrieval. *J. Big Data* **2020**, *7*, 39.
24. Diao, L.; Yan, H.; Li, F.; Song, S.; Lei, G.; Wang, F. The research of query expansion based on medical terms reweighting in medical information retrieval. *EURASIP J. Wirel. Commun. Netw.* **2018**, 105. [CrossRef]
25. Khedr, M.A.; El-Licy, F.A.; Salah, A. Ontology based semantic query expansion for searching queries in programming domain. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 449–455.
26. Dahir, S.; El Qadi, A. A query expansion method based on topic modeling and DBpedia features. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100043.
27. Saffariha, M.; Jahani, A.; Potter, D. Seed germination prediction of Salvia limbata under ecological stresses in protected areas: An artificial intelligence modeling approach. *BMC Ecol.* **2020**, *20*, 48.
28. Saffariha, M.; Jahani, A.; Jahani, R.; Latif, S. Prediction of hypericin content in *Hypericum perforatum* L. in different ecological habitat using artificial neural networks. *Plant Methods* **2021**, *17*, 10.
29. Jahani, A.; Saffariha, M. Modeling of trees failure under windstorm in harvested Hyrcanian forests using machine learning techniques. *Sci. Rep.* **2021**, *11*, 1124.
30. Jahani, A.; Saffariha, M. Human activities impact prediction in vegetation diversity of Lar National Park in Iran using artificial neural network model. *Integr. Environ. Assess. Manag.* **2021**, *17*, 42–52.
31. Dehghani, M.; Montazeri, Z.; Hubálovský, Š. GMBO: Group mean-based optimizer for solving various optimization problems. *Mathematics* **2021**, *9*, 1190.
32. Jia, H.; Sun, K.; Zhang, W.; Leng, X. An enhanced chimp optimization algorithm for continuous optimization domains. *Complex Intell. Syst.* **2021**, *8*, 65–82.
33. Jatnika, D.; Bijaksana, M.A.; Suryani, A.A. Word2Vec model analysis for semantic similarities in English words. *Procedia Comput. Sci.* **2019**, *157*, 160–167.
34. Lilian, J.F.; Sundarakantham, K.; Shalinie, S.M. QeCSO: Design of hybrid cuckoo search based query expansion model for efficient information retrieval. *Sādhanā* **2021**, *46*, 181.
35. Wang, Z.; Zhou, W.; Chen, L.; Zhou, F.; Zhu, F.; Fan, L. An adaptive deep learning-based UAV receiver design for coded MIMO with correlated noise. *Phys. Commun.* **2021**, *47*, 101365.
36. Raviv, T.; Schwartz, A.; Be'Ery, Y. Deep ensemble of weighted viterbi decoders for tail-biting convolutional codes. *Entropy* **2021**, *23*, 93.
37. Macdonald, C.; Tonellotto, N. Declarative experimentation in information retrieval using PyTerrier. In Proceedings of the ACM SIGIR on International Conference on Theory of Information Retrieval, Stavanger, Norway, 14–17 September 2020; pp. 161–168.