An Improved BM25 for Clinical Decision Support in Precision Medicine Based on Co-word Analysis and Cuckoo Search

Zicheng Zhang¹,²

¹.School of Information Management, Nanjing University, Nanjing 210023, China
².Jiangsu Key Laboratory of Data Engineering and Knowledge Services, Nanjing 210023, China

Corresponding author E-mail: 18551701375@163.com

Abstract

Background: Retrieving gene and disease information from a very large collection of biomedical abstracts to provide doctors with clinical decision support is one of the important research directions of Precision Medicine. Method: We propose a new method for the retrieval of biomedical articles utilizing expanded word and co-word implementations and conducting Cuckoo Search to optimize parameters of the retrieval function in the final stage of the proposed method. The specific goal is to retrieve biomedical abstracts of articles addressing treatments. The method employed in this manuscript first implements the BM25 algorithm to compute the score of the abstract, then we propose a method utilizing the BM25, an improved version of BM25, to compute the scores of expanded words and co-word that lead to a composite retrieval function. Afterward, the retrieval function is optimized using Cuckoo Search. The proposed method is utilized to find both disease and gene in the abstract of the same biomedical article. By doing so, the relevance of articles would tend to increase so would the score of the biomedical article. Besides, the manuscript discusses the influence of different parameters on the retrieval algorithm and summarizes the parameters to meet various retrieval needs. Results: All data are taken from medical articles provided in the Text Retrieval Conference (TREC) utilizing Clinical Decision Support (CDS) Tracks of 2017, 2018, and 2019 in Precision Medicine. 120 standard topics are tested. Three test indicators are employed to make comparisons among the methods utilized. To conduct comparable experiments, only the BM25 algorithm and its improved version of it are utilized. The experimental results show that the proposed algorithm achieves both better results and ranking outcomes. Conclusion: The proposed algorithm, an improved version of the BM25 algorithm, utilizes both co-word implementation and Cuckoo Search and verifies that the proposed algorithm produces better results on a large number of experimental sets. On the other hand, a relatively simple query expansion method is implemented in this manuscript. As a future direction of this research, both the ontology and semantic network to expand the query vocabulary is planned to be conducted.

Keywords Clinical Decision Support, Precision Medicine, Information Retrieval, Co-word Analysis, Improved BM25, Cuckoo Search

1. Background
With the continuous development of computer technologies, that the information available on the internet has been increasing sharply leads to implementations of various means that are utilized for extracting insights in medical articles. Hence, medical treatment has also stepped into the age of big data. Managing this immense data and extracting insights from it has been a critical endeavor. If this process would be improved, the advantages that it could carry would be so beneficial that medical doctors benefit from it on a large scale. For instance, some routine decision-making tasks require significant repetition, taking time, and leading to higher costs. However, the scientific application of computerized medical information retrieval systems could effectively improve efficiency, save costs, and reduce errors. Proper use of computer technology could bring efficacy in all fields where it would be used. Therefore, the development of computer-assisted medical information retrieval systems is highly significant. In real practice, every decision made by a doctor is critical to the patient, so the doctor must constantly update his knowledge and pay attention to the latest technology and methods of clinical science. The authoritative literature and the latest research results in the medical community could be consulted via the Internet, so the medical retrieval models play a crucial role. Moreover, for medical practitioners who encounter a difficult medical problem for a certain medical record, searching the relevant biomedical literature via the internet as a reference and inspiration could provide an important way of solving difficult issues.

Information Retrieval (IR) methods for Clinical Decision Support (CDS) have been the focus of several recent types of research and assessment campaigns. Specifically, the CDS track between 2014 and 2016 Text Retrieval Conferences (TREC) [1-3] sought to assess the systems providing evidence-based information in the form of either full-text articles or abstracts from an open-access subset of the MEDLINE to clinicians in return to queries. Moreover, 2017-2019 [4-6] tracks focused on important implementations in clinical decision support that provides both useful and precise medical information to clinicians treating cancer patients. In the track, each case described the patient's disease (a type of cancer), the relevant genetic variants (which genes), and basic demographic information (age, sex). Precision Medicine (PM) in [7] is a new medical concept and model utilizing individualized medicine, develops with the rapid progress of genome sequencing technology, and the cross-application of bioinformatics and big data science.

2. Preleminaries

The objective of Information Retrieval (IR) is to retrieve related documents based on a given query. Generally, the relevancy of documents to queries is usually gauged by the score assigned by an IR model, for example, the extensively implemented classical BM25 model [8]. On the other hand, the past few decades witnessed the implementation of machine learning technology when information retrieval was a concern. The process could be classified into three varieties as follows: 1. the single document methods, 2. document pair methods, and 3. document list methods. The common single-document methods, for instance, utilizing a logistic regression method [9], deal with a feature
vector of each document as an input and an output is the relevance of each document. The document pair methods, for example, utilizing the Rank-SVM [10] or Rank-Boost [11], implement a feature vector of a pair of documents as an input, and an output is a correlation between the documents. Besides, the document list methods, for instance, utilizing the List-Net [12], or the Ada-Rank [13], or the Lambda-Mart [14], employ a set of documents associated with a query as an input and an output is a ranked list. In recent years, query expansion methods have been widely implemented in information retrieval. Singh et al. [15] suggest a query expansion method utilizing fuzzy logic. The top-ranked documents are regarded as relevant feedback documents for mining other relevant query information. The choice of different query expansion terms is determined according to their importance in the top documents in [15]. These types of methods assign each term to a different relevance score and then select the expansion term through a certain threshold.

Keikha et al. [16] further consider the Wikipedia corpus as a feedback set space to train the Word Vector Model and determines the long-term selection of the best features in both supervised and unsupervised models. Almasri et al. [17] also utilize vectors to represent query words and query expansion terms returned by pseudo-correlation feedback. Cosine similarity is added to the Bag-of-Words Model, and the frequency of each word in the query term is recomputed. Rocchio et al. [18] suggest a classic correlation feedback method, which increases the entry weight of related documents and reduces the entry weight of non-relevant documents. However, one of the disadvantages of this method is to be very time-consuming for practitioners to assess the relevance of documents.

Cui et al. [19] propose a query expansion method for web search logs utilizing practitioners’ interaction information. The key assumption behind this method is that the documents chosen by a user to read are related to the query. The new words in related documents are sorted according to their similarity with the user query, and the new words with the highest similarity are selected as the expanded word. The candidate expanded words are extracted from the top documents, and then the candidate expanded words are weighted and sorted by the probability generated by the language model. Aronson [20] proposes a method utilizing the UMLS query expansion, which benefits from the Meta-Map program [21] to identify medical phrases in the original query and then expand the query with new phrases. Hence, the experimental results show that query expansion utilizing the UMLS is an effective method to improve the performance of information retrieval.

Li et al. [22] propose a method of keyword-weighted network analysis to implement a medical full-text recommendation, which helps expand the medical acronym list by searching the full text and is evaluated by domain experts. It is verified that the algorithm works well in terms of recommendation accuracy in the medical literature. Saeid et al. [23] propose a query expansion method utilizing the Bayesian approach,
which expands the genes of a disease to be no less than 3 words, and experiments suggest that the algorithm has higher precision value.

To retrieve documents that are highly related to a query, the literature review brings us the idea of benefitting from both query expansion and keywords that could be helpful to achieve this objective. Hence, this manuscript proposes a new method utilizing expanded words and co-word analysis as new tools to optimize the information retrieval of biomedical articles implementing the BM25 algorithm as a base method to compute scores of the abstract, expanded words, and co-word as a composite retrieval function. Besides, when disease and a gene both appear in the same biomedical article, the score of the biomedical article would tend to increase. Finally, the Cuckoo Algorithm is utilized to optimize the parameters of the retrieval algorithm.

As a classic information retrieval algorithm, the BM25 has been also frequently implemented on TREC, such as 2017, 2018, and 2019 Precision Medicine [34-49]. These algorithms mainly utilize either the original BM25 or the improved BM25 algorithms to retrieve information [37-38].

3. Data
3.1 Data structure

The biomedical articles in scientific abstracts are presented in XML formats that contain information about articles. The MeSH headings, chemical lists, and keyword lists for XML documents are selected to utilize abstracts whose displays are presented in Figure 1.

3.2 Data distribution

While the total number of biomedical articles in both the 2017 and 2018 TREC Precision Medicine set is 26,613,834, it has 29,137,637 for the 2019 TREC Precision Medicine set. Some of the statistics that are used in information retrieval are tabulated in Table 1.

Some statistics presented in Table 1 have definitions that are expressed as follows: Abstract-Mean-Length represents the average length of the abstracts after deleting stop-words; Abstract-Number represents the number of articles with abstracts; Chemical-Mean-Length represents the average length of the chemical lists; Chemical-Number represents the number of articles with a chemical list; Mesh-Mean-Length represents the average length of the MeSH headings; Mesh-Number represents the number of articles with MeSH headings; Keyword-Mean-Length represents the average length of the keyword list, and Keyword-Number represents the number of articles with keyword lists.
3.3 Query expansion

Medical Subject Headings (MeSH) is a controlled vocabulary developed by the U.S. National Library of Medicine, which is mainly utilized to index, catalog, and search articles pertinent to both biomedical and health sciences. The important role of the MeSH in medical information retrieval is mainly manifested with two aspects, which are called accuracy and specificity. While indexers enter information into the retrieval system, researchers utilize the information in the system concerning the two processes. The MeSH is utilized as the platform to make the terms consistent between the index and search to achieve the best retrieval outcomes. Hence, the accurate and comprehensive usage of the MeSH has a significant impact on the results of retrieval. In this manuscript, we utilize the MeSH database to extend the MeSH (meshb.nlm.nih.gov/MeSHonDemand) to find expansion terms or new words and their corresponding broader terms. The MeSH on Demand is utilized to expand query terms and obtain additional terms if possible.

Topic 2017-1 is taken as an example presented in Table 2. The results of the extended words are presented in Table 3.

3.4 Age expansion

The variable age included in the demographic field is also expanded to the terms or new words proposed by Kastner et al. [24]. We have readjusted the division of age and believe that those over 18 should be adults. Our expansion based on the variable age is presented in Table 4.

4. The Proposed Model

We first utilize the MeSH on Demand to find MeSH terms and additional terms that would be used in the retrieval of abstracts regarding any given query. Then, we construct a “wordlist” including chemical words, keywords, and MeSH headings utilizing query expansion. By doing so, we can find documents that are more related to query expansion. Then, the relevance score of the document would increase so would the score of it. In the next step, the co-word analysis is conducted utilizing either each separate resource of such as abstract, keywords, chemical words, or MeSh headings or utilizing them one at a time to find co-occurrence of selected words such as disease and gene in our case. While the first step deals with computing the score of abstracts based on a query and its morpheme, the second step deals with calculating the score of expansion words. In the last step, the score of the co-word is computed. Hence, as long as the number of documents is reduced to a smaller number of documents based on utilizing the “word list” to search for, the score of co-word would tend to increase. In the last step, we compute a composite score consisting of three scores of abstract, expanded words, and co-word. Afterward, Cuckoo Search, an evolutionary
optimization method, is applied to optimize the parameters of the retrieval model, which is a composite model.

4.1 Scoring model for abstract

The BM25 model [8] is a classical information retrieval model whose main idea is to analyze a query \( Q \) to find morpheme \( q_i \). Then, for each search result \( d \), it calculates the correlation score of each morpheme \( q_i \) and \( d \); and finally gives a weight to the sum of the correlation score of \( q_i \) concerning to \( d \) to obtain a correlation score between \( Q \) and \( d \). The general formula of the BM25 algorithm is expressed by

\[
Score(Q, d) = \sum_{i} W_i \times R(q_i, d) 
\]

where \( W_i \) is a weight to determine the relevance of a word to a document; we define Inverse Document Frequency (IDF) as \( W_i \).

\[
IDF(q_i) = \log \frac{D}{\text{card}([j|i \in d_j])} 
\]

where \( D \) represents the total number of corpus documents, and \( \text{card}([j|i \in d_j]) \) represents the number of documents containing morpheme \( q_i \). According to the definition of the IDF, the more \( q_i \) is contained in a document, the lower the weight of \( q_i \) is for a given set of documents. In other words, when many documents contain \( q_i \), the discrimination of \( q_i \) is not so robust that the importance of utilizing \( q_i \) to judge relevance is so weak.

The relevance score \( R(q_i, d) \) of morpheme \( q_i \) to document \( d \) is defined by

\[
R(q_i, d) = \frac{f_i \times (k_1+1)}{f_i+k} \times \frac{qf_i \times (k_2+1)}{qf_i+k_2} 
\]

where \( K \) is defined as follows:

\[
K = k_3 \times (1 - b_1 + b_1 \times \frac{dl}{avgdl}) 
\]

where, \( k_1 \), \( k_2 \) and \( b \) are adjustment factors, which are usually set according to experience; \( f_i \) represents the frequency of \( q_i \) in \( d \); \( qf_i \) represents the frequency of \( q_i \) in query; \( dl \) represents the length of document \( d \); and \( avgdl \) represents the average length of all documents. In most cases, \( q_i \) would appear only once in query; that is, \( qf_i = 1 \), so the formula could be modified to:

\[
R(q_i, d) = \frac{f_i \times (k_1+1)}{f_i+k} 
\]
As seen from the definition, the role of parameter \( b \) is to tune the impact of the document length on the relevance. The larger \( b \) would be, the greater the impact of the document length on the relevance score would be, and vice versa. The longer the relative length of the document would be, the larger the \( K \) value would be. Hence, the smaller the relevance score would be. In the end, the correlation score formula of the abstract of the document \( d \) can be expressed by

\[
Score_{abstract}(Q, d) = \sum_{i}^{n} IDF(q_i) \times \frac{f_i \times (k_1 + 1)}{f_i + k_1 \times \frac{1 - b_1 \times \text{avgdl}}{b_1}}
\]  

(6)

### 4.2 Score of expanded word

As seen in Table 1, quite many biomedical articles have both abstracts and titles. The number of biomedical articles containing Chemical Words, MeSH Headings, and Keywords varies widely. There exist 13,113,093 articles containing Chemical Words, 2,438,717,151 articles containing MeSH Headings and 4,005,446 containing Keywords. As the literature suggests, utilizing directly the BM25 algorithm leads to failure when dealing with a large selection of documents [52]. In this subsection, we suggest an improved BM25 algorithm to compute the scores of expanded words. We combine Chemical Words, MeSH headings, and Keywords into a list called ‘Word List’. The length of the “Word List” in the document is defined by

\[
dwl = dcl + dml + dkl
\]

(7)

where \( dcl \) represents the length of the Chemical List in document \( d \); \( dml \) represents the length of the MeSH headings in document \( d \); and \( dkl \) represents the length of the Keyword list in document \( d \).

The IDF value of the expanded word appearing in the “Word List” of a document \( d \) is defined by

\[
IDF_{word}(q_i, d) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}
\]

(8)

where \( N \) represents the number of documents in which \( dwl \) > 0, and \( n(q_i) \) represents the number of documents containing extended morpheme \( q_i \). The frequency value of the term of the word list is defined by

\[
tf_{word}(Q, d) = \sum_{i}^{n} IDF_{word}(q_i, d)
\]

(9)

where \( n \) represents the number of expanded words in query \( Q \), and \( q_i \) represents the morpheme of each expanded word in \( Q \). The score of an expanded word in a document \( d \) is defined by
\[
Score_{\text{word}}(Q,d) = \frac{tf_{\text{word}}(Q,d) \times (k_2 + 1)}{tf_{\text{word}}(Q,d) + k_3 \times (1 - b_2 \times \frac{dwl}{avgdwl})}
\] (10)

where \(k_2\) and \(b_1\) are adjustment factors, which are usually set according to experience, and \(avgdwl\) represents the average length of all word lists.

### 4.3 Score of co-word

Co-Word analysis utilizes the co-occurrence of lexical pairs or noun phrases in an article set to determine the relationship between topics in the discipline represented by the article set. In this manuscript, Co-Word analysis is introduced into the article scoring model, when both a disease and a gene co-occur across in an abstract, Chemical List, MeSH heading, and Keyword List as presented in Figure 2 or co-occur within any abstract, Chemical List, MeSH heading, or Keyword List as presented in Figure 3.

We utilize the Inverse Document Frequency (IDF) value as a Co-Word score to distinguish the importance of a gene whose formula is defined by

\[
IDF_{\text{gene}}(g_i, d) = \log \frac{N - n(g_i) + 0.5}{n(g_i) + 0.5}
\] (11)

where \(N\) represents the number of documents and \(n(g_i)\) represents the number of documents containing gene morpheme \(g_i\).

The Co-Word score is defined by

\[
Score_{\text{co-word}}(Q,d) = \sum_i^n IDF_{\text{word}}(g_i, d)
\] (12)

where \(n\) represents the number of genes with co-word having a disease in a query \(Q\), and \(g_i\) represents the morpheme of each gene in the \(Q\).

### 4.4 Retrieval Model

We utilize the composite score as the final score for a document \(d\) under query \(Q\), and the specific formula is expressed by

\[
Score_{\text{composite}}(Q,d) = Score_{\text{abstract}}(Q,d) + Score_{\text{word}}(Q,d) + \alpha \times Score_{\text{co-word}}(Q,d)
\] (13)

The architecture of the system for biomedical article retrieval is depicted in Figure 4.

### 5. Parameter Optimization

There exist 6 parameters in the proposed method, namely, \(k_1\), \(k_2\), \(k_3\), \(b_1\), \(b_2\) and \(\alpha\). Choosing better parameters would improve retrieval results. Many optimization algorithms are implemented to optimize the function in use such as the Genetic
Algorithm (GA) [25], Simulated Annealing (SA) Algorithm [26], Ant Colony (AC) Algorithm [27]. With the continuous effort to develop better intelligence optimization algorithms, many new Swarm Intelligence Optimization (SIO) Algorithms have emerged during recent years, such as the Cuckoo Search (CS) Algorithm [28], Glow Worm Swarm Optimization (GWSO) Algorithm [29], Particle Swarm Optimization (PSO) algorithm [30] and so on. The Swarm Intelligence Optimization (SIO) Algorithm has been widely utilized for solving function optimization.

5.1 Cuckoo Search Algorithm

Cuckoo Search Algorithm (CS) is a novel Swarm Intelligence Optimization (SIO) Algorithm proposed by Yang [28] in 2009. Maribel [31] suggests that the CS Algorithm is more efficient than the GA Algorithm. The CS Algorithm utilizes some idealized rules expressed by

1. Each cuckoo produces only one egg every time and selects a parasitic nest to hatch its egg randomly.
2. The best parasitic nest will be kept to the next generation.
3. The number of available parasitic nests is fixed and the detection probability of parasitic nest’ master is \( P_a \in (0,1) \).

The cuckoo finds the nest and updates the position according to the above-idealized rules. The position update formula is defined by

\[
X_i^{(t+1)} = X_i^{(t)} + T \oplus \text{Levy}(\lambda) \tag{14}
\]

where \( T \) stands for step size and \( T > 0 \), the operator, \( \oplus \), expresses a point-to-point multiplication, \( \text{Levy}(\lambda) \) is the search path and follows the Levy distribution [32-33]. The pseudo-code of the algorithm [28] is presented as follows:

**Algorithm 1 Cuckoo Search Algorithm**

1: Begin
2: Define the objective function, \( f(x) \), \( X = (x_1, \ldots, x_d)^T \) is a d-dimensional problem
3: Generate initial population of n host nests
   \[ X_i(i = 1, 2, \ldots, n) \]
4: While (\( t < \text{Max Generation} \)) or (stop criterion)
   5: Get a cuckoo randomly by Lévy flights evaluating its quality/fitness to \( f(x_i) \)
   6: Select a nest among n (say, j) randomly.
   7: If \( f(x_i) > f(x_j) \)
   8: Replace the nest j with new solutions
   9: end
10: A fraction \( (P_a) \) of worse nests are abandoned and new ones are built
11: Keep the best solutions (or nests with quality solutions)
12: Rank the solutions and find the current best
13: end while
14: Post-process results and visualization
15: End

5.2 Objective function
While the RR represents the relevant document retrieved, the RN represents the irrelevant documents retrieved. The calculation formula for Precision is defined by

\[
Precision = \frac{RR}{RR+RN}
\]  

(15)

Then, \(P@10\) is defined as the Precision at \(RR + RN = 10\). We define the average \(P@10\) as follows:

\[
Avg_{P@10} = \frac{\sum_{t=1}^{n} p@10(t)}{n}
\]

(16)

where \(p@10(t)\) represents the \(P@10\) value of the \(t^{th}\) topic in the \(n\) topics.

NDCG [70] is a commonly utilized index to assess the quality of ranking in information retrieval. Let \(\theta\) denote a relevance grade and \(gain(\theta)\) denote the gain associated with \(\theta\), respectively. Besides, let \(g_1, g_2, \ldots, g_z\) be the gain values associated with the \(Z\) documents retrieved by a system in response to a query \(q\), such as \(g_i = gain(\theta)\) if the relevance grade of the document in rank \(i\) is \(\theta\). Then, the nDCG value for this system is computed by

\[
\text{nDCG} = \frac{DCG}{DCG_i} \text{ where } DCG = \sum_{i=1}^{Z} \frac{g_i}{\log(i+1)}
\]

(17)

and \(DCG_i\) denotes the \(DCG\) value for an ideal ranked list for query \(q\). We define the average \(nDCG\) as follows:

\[
Avg_{nDCG} = \frac{\sum_{t=1}^{n} nDCG(t)}{n}
\]

(18)

where \(nDCG(t)\) represents the \(nDCG\) value of the \(t^{th}\) topic in the \(n\) topics.

5.3 Algorithm flow

Since \(k_2\) is a fixed value, namely, \(k_2 = 1\), we utilize \(k_1, k_3, b_1, b_2, \alpha\) as input parameters. Firstly, the algorithm generates the initial population and set the maximum number of iterations or stop criterion. If the number of iterations reaches the max generation or the stop criterion is met, the algorithm stops and gives the optimal solution. Otherwise, the algorithm will perform a series of optimization operations based on the value of the objective function. This manuscript utilizes \(Avg_{P@10} + Avg_{nDCG}\) as the objective function, and employs the data set of the 2017 Precision
Medicine as the training data set to optimize the parameters. The flow chart of the algorithm is presented in Figure 5.

5.4. Experiments

5.4.1. Dataset

All the data are taken from medical articles provided in 2017, 2018, and 2019 TREC Precision Medicine whose web pages could be found in (http://www.trec-cds.org/2017.html), (http://www.trec-cds.org/2018.html), and (http://www.trec-cds.org/2019.html), respectively. Each article is formatted utilizing the XML 2017. The assessment results of the 2017 TREC Precision Medicine, of the 2018 TREC Precision Medicine, and of the 2019 TREC Precision Medicine are obtained from (https://trec.nist.gov/data/precmed/qrels-final-abstracts.txt), (https://trec.nist.gov/data/precmed/qrels-treceval-abstracts-2018-v2.txt), and (https://trec.nist.gov/data/precmed/qrels-treceval-abstracts.2019.txt), respectively.

Due to the semi-structured nature of the XML format, we employed MongoDB as the database for document storage and python as the programming language. All the experimental code could be obtained from the author’s GitHub (https://github.com/Bruce-V/CS-BM25).

5.4.2 Algorithm for parameter setting

The parameter values of the algorithm in this manuscript are presented in Table 5.

5.4.3 Experimental results

In Table 6, while “Normal” represents values of the empirical parameters; the CS represents the parameters trained utilizing the 2017 data set implementing the Cuckoo Search algorithm using 1000 documents having the highest scores as the result of the retrieval model that is selected.

When 3 years of data are compared, the optimized parameters are better than the empirical parameters. For a retrieval system, users want related documents to appear earlier, so infNDCG and P@10 are two important indicators to assess information retrieval. The parameters that are optimized utilizing both NDCG and P@10 would increase the weights of the word list. The word list contains extended information about age, gender, and genes, which is very important for distinguishing relevant literature from non-related literature. In conclusion, different parameters could be utilized to meet the needs of various users.

As seen in Figures 6, 7 and 8, while the RR represents relevance in co-word documents; the RN represents all relevance except for the RR. Many relevant documents contain both a disease and a gene. We define the average relevant document coverage rate as follows:
The average coverage rate of 30 topics in 2017 is 52.9%, that of 50 topics in 2018 is 74.13%, and that of 40 topics in 2019 is 54.4%. These outcomes suggest that the Co-Word Analysis has a better impact on the retrieval of relevant documents, which greatly reduces the search scope.

When information retrieval is a concern, the NR, RR, NN, and RN represent the relevant documents that are not retrieved, the irrelevant documents that are not retrieved, the relevant documents that are retrieved, the irrelevant documents that are retrieved, respectively. Precision is defined by

\[ Precision = \frac{RR}{(RR+RN)} \]  

(20)

The Recall is defined by

\[ Recall = \frac{RR}{(RR+NR)} \]  

(21)

The F1-score is defined by

\[ F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \]  

(22)

As seen in Figures 9, 10 and 11, the optimized parameters are better than the empirical parameters on both P@10 and infNDCG. Since we utilize the 2017 Precision Medicine dataset as the training set, the optimization effect is the most obvious on the 2017 precision medicine dataset. When the test data in 2018 and 2019 are a concern, both P@10 and infNDCG have improved, but R-predicted has declined. This happens since the adopted objective function has improved the ranking of the most relevant documents. The Precision and the Recall are inversely proportional when the retrieval system is a concern. In our case examining the retrieval of biomedical articles, we are more concerned about the precision rate to facilitate doctors’ scientific decision-making.

5.4.4 Experimental comparison

Compared with the results in the literature based on the selected 2017, 2018, and 2019 TREC Precision Medicine, 3 indicators, namely infNDCG, R-predicted, and P@10, are
selected for comparison. The experimental results are presented in Tables 7, 8, and 9, respectively.

We select 3 years of the TREC datasets to verify our experimental results. These methods all utilize the BM25 algorithm or an improved BM25 algorithm. In the 2017 experiments, the significant improvement of the proposed method could be observed in three indicators. In the 2018 experiments, the proposed method is better than other similar algorithms when the P@10, ranked second on infNDCG, and R-predicted are a concern. In the 2019 experiments, the proposed method is better than other similar algorithms when the P@10, ranked second on infNDCG, and R-predicted are a concern.

6. Conclusion
This manuscript proposes a method based on BM25 taking into account Co-Word Analysis to retrieve biomedical articles. We improved the BM25 algorithm and employed it to compute the score of expanded words by combining the co-word score with the gene appearance weight. Then, we utilize the Cuckoo Search Algorithm to optimize parameters on the evaluation function of both P @ 10 and NDCG. Optimization results suggest that increasing the score weight of the “word list” can effectively improve the ranking of related documents. Besides, the manuscript discusses the influence of different parameters on the retrieval algorithm and presents the parameters to meet different retrieval needs in the future. Although the proposed algorithm in this manuscript is based on the improved version of BM25 utilizing expanded words and co-word analysis and the optimization of parameters of the retrieval function, it summarizes the general rules for improving the parameters of the BM25 algorithm and verifies it through many experiments. Since the query expansion is simple in this manuscript, our future research will focus on utilizing more linked data to conduct the research utilizing the topic data.

Abbreviations

Text Retrieval Conference (TREC); Clinical Decision Support (CDS); Information Retrieval (IR); Support Vector Machines(SVM); Unified Medical Language System(UMLS); Normalized Discounted Cumulative Gain(NDCG)

Declarations

Ethics approval and consent to participate

On behalf of, and having obtained permission from all the authors, I declare that: the material has not been published in whole or in part elsewhere; the paper is not currently being considered for publication elsewhere; all authors have been personally and actively involved in substantive work leading to the report, and will hold themselves jointly and individually responsible for its content; all relevant ethical safeguards have been met concerning patient or subject protection, or animal experimentation. I testify to the accuracy of the above on behalf of all the authors.
Consent for publication

This article utilizes public datasets

Availability of data and material

All the data were taken from medical articles provided in 2017 TREC Precision Medicine (http://www.trec-cds.org/2017.html); 2018 TREC Precision Medicine (http://www.trec-cds.org/2018.html) and 2019 TREC Precision Medicine (http://www.trec-cds.org/2019.html). Each article was formatted using XML. 2017 TREC Precision Medicine evaluation results are obtained from the website(https://trec.nist.gov/data/precmed/qrels-final-abstracts.txt); 2018 TREC Precision Medicine evaluation results are obtained from the website(https://trec.nist.gov/data/precmed/qrels-treeceval-abstracts-2018-v2.txt); 2019 TREC Precision Medicine evaluation results are obtained from the website (https://trec.nist.gov/data/precmed/qrels-treeceval-abstracts.2019.txt). All the experimental code can be obtained from my GitHub (https://github.com/Bruce-V/CS-BM25)

Competing interests

All authors declare that: (i) no support, financial or otherwise, has been received from any organization that may have an interest in the submitted work; and (ii) there are no other relationships or activities that could appear to have influenced the submitted work.

Funding

None

Authors’ contributions

This article was independently completed by Zicheng Zhang. All authors read and approved the final manuscript.

Acknowledgments

We appreciated Wei Cao for data collection

References

1. Simpson M.S., Voorhees E.M., Hersh W., Overview of the TREC 2014 Clinical Decision Support Track, in Proceedings of Text Retrieval Conference (TREC), 2014.
2. Roberts K., Simpson M.S., Voorhees E.M., Hersh W.R., Overview of the TREC 2015 clinical decision support track, in Proceedings of Text Retrieval Conference (TREC), 2015.
3. Roberts K., Demner-Fushman D., Voorhees E.M., Hersh W.R., Overview of the TREC 2016 clinical decision support track, in Proceedings of Text Retrieval Conference (TREC), 2016.
4. Roberts K., Demner-Fushman D., Voorhees E.M., Hersh W.R., Bedrick S., Lazar A.J., Pant S., Overview of the TREC 2017 precision medicine track, in Proceedings of Text Retrieval Conference (TREC), 2017.
5. Roberts K., Demner-Fushman D., Voorhees E.M., Hersh W.R., Bedrick S., Lazar S.J., Overview of the TREC 2018 precision medicine track, in Proceedings of Text Retrieval Conference (TREC), 2018.
6. Roberts K., Demner-Fushman D., Voorhees E.M., Hersh W.R., Bedrick S., Lazar S.J., Overview of the TREC 2019 precision medicine track, in Proceedings of Text Retrieval Conference (TREC), 2019.
7. Collins F.S., Varmus H., A New Initiative on Precision Medicine. New England Journal of Medicine, 372 (9), 793-795, 2015.
8. Robertson S.E., Walker S., Hancock-Beaulieu M., Gatford M., Payne A., "Okapi at TREC-4," in TREC, 1995.
9. Gey F. C., Inferring probability of relevance using the method of logistic regression, SIGIR’94, Springer, London, 222-231, 1994.
10. Joachims T., Optimizing search engines using clickthrough data, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 133-142, 2002.
11. Freund Y., Layer R., Schapire R.E., An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 4 (9) 933-969, 2003.
12. Cao Z., Qin T., Liu T.Y., Learning to rank: from pairwise approach to listwise approach, Proceedings of the 24th International Conference on Machine Learning, ACM, 129-136, 2007.
13. Xu J., Li H., Adarank: A boosting algorithm for information retrieval, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 391-398, 2007.
14. Burges C. J. C., From ranknet to lambdarank to lambdamart: An Overview. Learning, 11, 523-581, 81, 2010.
15. Singh J., Prasad M., Prasad O.K., A novel fuzzy logic model for pseudo-relevance feedback-based query expansion. International Journal of Fuzzy Systems, 18 (6), 980-989, 2016.
16. Keikha A., Ensan F., Bagheri E. Query expansion using pseudo relevance feedback on Wikipedia, Journal of Intelligent Information Systems, 50 (3): 455-478, 2018.
17. Almasri M., Berrut C., Chevallet J. P., A comparison of deep learning-based query expansion with pseudo-relevance feedback and mutual information, Proceedings
of European Conference on Information Retrieval Padua, ECIR Press, 709-715, 2016.

18. Abacha A. B., Khelifi S., LIST at TREC 2015 Clinical Decision Support Track: Question Analysis and Unsupervised Result Fusion. TREC, 2015.

19. Cui H., Wen J. R., Nie J.Y., Probabilistic query expansion using query logs. Proceedings of the 11th International Conference on World Wide Web. ACM, 325-332, 2002.

20. Aronson A. R., Rindflesch T. C., Query expansion using the UMLS Meta Thesaurus: Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 485, 1997.

21. Aronson A. R. Effective mapping of biomedical text to the UMLS MetaThesaurus: the MetaMap program. Proceedings of the AMIA Symposium, American Medical Informatics Association, 17, 2001.

22. Shuqing L., Ying S., Dagobert S., Automatic decision support for clinical diagnostic literature using link analysis in a weighted keyword network, Journal of Medical Systems, 42:27, 2018.

23. Balaneshinkordan S., Kotov A.. Bayesian approach to incorporating different types of biomedical knowledge bases into information retrieval systems for clinical decision support in precision medicine, Journal of Biomedical Informatics, 98,2019.

24. Kastner M., Wileczynski N.L., Walker-Dilks C., Ann McKibbon K., Haynes B., Age-specific search strategies for MedLine. Journal of Medical Internet Research 8, 4, 1–10, 2006.

25. Holland J.H., Adaptation in Natural and Artificial Systems, Ann Arbor, Michigan.

26. Kirkpatrick S., Gelatt Jr. C.D., and Vecchi M. P., Optimization by Simulated Annealing, Science, Vol. 220, No. 4598, pp. 671 -680, 1983.

27. Dorigo M. Gambardella L.M., A Study of Some Properties of Ant-Q, in Proc.of the 44th Int' l. Conf on Parallel Problem Solving from Nature, 656-665, 1996.

28. Yang X.S., Deb S., Cuckoo search via levy flights, in World Congress on Nature & Biologically Inspired Computing, 210–214, 2009.

29. Krishnand K.N., Ghose D., Detection of multiple source locations using a glowworm metaphor with applications to collective robotics, In Proc. of IEEE Swarm Intelligence Symposium, 84-91, 2005.

30. Kenney J., Eberhart R., Particle swarm optimization, in Proceedings of IEEE Conference on Neural Networks, 1995.

31. Guerrero M., Castillo O., Valdez M.G., Cuckoo Search via Lévy Flights and a Comparison with Genetic Algorithms, Fuzzy Logic Augmentation of Nature-Inspired Optimization Metaheuristics, Vol. 574, 91-103, 2015.

32. Pavlyukevich I., Levy flights, non-local search, and simulated annealing, Computational Physics, Vol. 226, 1830-1844, 2007.

33. Pavlyukevich I., Cooling down Levy flights, J. Phys.A: Math. Theor., vol. 40, 12299-12313, 2007.
34. Wang Y., Komandur-Elayavilli R., Rastegar-Mojarad M., Leveraging both structured and unstructured data for Precision Information Retrieval. Proceedings of Text Retrieval Conference (TREC), 2017.

35. Li C., He B., Sun Y., UCAS at TREC-2017 Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2017.

36. Jo S-H., Lee K-S., CBNU at TREC 2017 Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2017.

37. Wang Y., Fang H., Combining Term-based and Concept-based Representation for Clinical Retrieval. Proceedings of Text Retrieval Conference (TREC), 2017.

38. Ling Y., Hasan S.A., Filannino M., A hybrid approach to Precision Medicine-related biomedical article retrieval and clinical trial matching. Proceedings of Text Retrieval Conference (TREC), 2017.

39. Noh J., Kavuluru R., Team UKNLP at TREC 2017 Precision Medicine Track: A Knowledge-Based IR System with Tuned Query-Time Boosting. Proceedings of Text Retrieval Conference (TREC), 2017.

40. Baruah P., Dulepet R., Kyle Qian. Brown University at TREC Precision Medicine 2018. Proceedings of Text Retrieval Conference (TREC), 2018.

41. Nishani L., Kolla M., Baruah G., Klick Labs at TREC 2018 Precision Medicine track. Proceedings of Text Retrieval Conference (TREC), 2018.

42. Zheng Z., Li C., He B., UCAS at TREC-2018 Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2018.

43. Taylor S.J., Goodwin T.R., Harabagiu S.B, UTD HLTRI at TREC 2018: Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2018.

44. Jo S-H, Lee K-S., CBNU at TREC 2019 Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2019.

45. Zheng Q., Li Y., Hu J., ECNU-ICA team at TREC 2019 Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2019.

46. Di Nunzio G.M., Marchesin S., Agosti M, Exploring how to combine query reformulations for Precision Medicine, Proceedings of Text Retrieval Conference (TREC), 2019.

47. Cieslewicz A., Dutkiewicz J., Jedrzejek C.L, Poznan Contribution to TREC-PM 2019, Proceedings of Text Retrieval Conference (TREC), 2019.

48. Wu D.T.Y., Su W-C., Retrieving Scientific Abstracts using Venue-and Concept-based Approaches: CincyMedIR at TREC 2019 Precision Medicine Track, Proceedings of Text Retrieval Conference (TREC), 2019.

49. Rybinski M., Karimi S., Paris C., CSIRO at 2019 TREC Precision Medicine Track. Proceedings of Text Retrieval Conference (TREC), 2019.

50. Yang X-S., Deb S., Cuckoo Search via Lévy flights 2009 World Congress on Nature Biologically Inspired Computing, NaBIC, IEEE, 210–214, 2009.

51. Yang X-S., Deb S., Multi-objective cuckoo search for design optimization, Comput. Oper. Res. 40 (6) 1616–1624, 2013.

52. Trotman A., Choosing document structure weights, Information Processing, and Management 41 243–264, 2005.
Figures and Tables

**Figure 1** General structure and the XML attributes of the MEDLINE abstracts

**Figure 2** Cross Co-Word of abstract, Chemical List, MeSH headings and Keyword List

**Figure 3** Co-Word of abstract, Chemical List, MeSH headings and Keyword List
Figure 4 The architecture of the system for biomedical article retrieval
Figure 5 The Flowchart of the Algorithm
Figure 6 2017 TREC precision medicine relevance in co-word biomedical articles

Figure 7 2018 TREC precision medicine relevance in co-word biomedical articles

Figure 8 2019 TREC precision medicine relevance in co-word biomedical articles
Table 1 Some of the Statistics in TREC Precision Medicine covering the period of 2017-2019

| Name                  | 2017 and 2018   | 2019           |
|-----------------------|-----------------|----------------|
| Abstract-Mean-Length  | 77.5            | 83.5           |
| Abstract-Number       | 26,613,834      | 29,137,637     |
| Chemical-Mean-Length  | 3.8             | 3.8            |
| Chemical-Number       | 13,113,093      | 13,670,358     |
| Mesh-Mean-Length      | 10.5            | 10.6           |
| Mesh-Number           | 24,387,151      | 25,389,659     |
| Keyword-Mean-Length   | 4.1             | 4.4            |
### Table 2 The description of TREC Precision Medicine retrieval topic

| Year  | disease          | gene             | demographic       | other    |
|-------|------------------|------------------|-------------------|----------|
| 2017-1| Liposarcoma      | CDK4 Amplification| 38-year-old male  | GERD     |
| 2018-1| Melanoma         | BRAF (V600E)     | 64-year-old male  | None     |
| 2019-1| Melanoma         | BRAF (E586K)     | 64-year-old female| None     |

### Table 3 The expanded MeSH of 2017 TREC Precision Medicine retrieval task 1

| Search word         | Expanded word                                                                 |
|---------------------|-------------------------------------------------------------------------------|
| Liposarcoma         | Myxoid                         |
| CDK4 Amplification  | Cyclin-Dependent Kinase 4         |
|                     | Proto-Onkogene Proteins c-mdm2       |
| 38-year-old Male    | Middle Aged Adult                |
| Male                | Human                            |

### Table 4 Expanded Age of TREC Precision Medicine

| Term      | Range                        |
|-----------|------------------------------|
| Fetus     | Fetus                        |
| Newborn   | Birth to 1 month             |
| Infant    | > 1 month to < 24 months     |
| Preschool | 2 years to < 6 years         |
| Child     | 6 years to < 13 years        |
| Adolescent| 13 years to < 19 years       |
| Young     | 19 years to < 35 years       |
| Middle age| 35 years to < 60 years       |
| Aged      | 60 years to < 80 years       |
| Aged 80   | ≥ 80 years                   |
| Adult     | ≥ 18 years                   |

### Table 5 The parameter settings of the Cuckoo Search Algorithm

| Parameter   | Description          | Value          |
|-------------|----------------------|----------------|
| n           | population number    | 40             |
| T           | step size            | 1              |
| Max_Generation | Max Generation   | 500            |
| k1_boundary | Boundary of k1       | (0,100)        |
| k3_boundary | Boundary of k3       | (0,100)        |
| b1_boundary | Boundary of b1       | (0,1)          |
| b2_boundary | Boundary of b2       | (0,1)          |
| α_boundary  | Boundary of α        | (0,5)          |
**Table 6** Results of Cuckoo Search Algorithm

| Name      | $k_1$ | $k_3$ | $b_1$ | $b_2$ | $\alpha$ |
|-----------|-------|-------|-------|-------|----------|
| “Normal”  | 1.2   | 1.2   | 0.75  | 0.75  | 1        |
| CS        | 3.5   | 91.3  | 0.84  | 1     | 4        |

**Table 7** 2017 TREC Precision Medicine Experimental comparison

| Methods                | InfNDCG | R-Prec | P@10  |
|------------------------|---------|--------|-------|
| MayoNLPTeam [34]       | 0.2864  | 0.1698 | 0.3931|
| UCAS [35]              | 0.3271  | 0.2227 | 0.4276|
| Cbnu [36]              | 0.3218  | 0.2287 | 0.4614|
| Udel-Fang [37]         | 0.3879  | 0.2503 | 0.5067|
| Prna-Mit [38]          | 0.4070  | 0.2620 | 0.5300|
| UKNLP [39]             | 0.3852  | 0.2518 | 0.5533|
| Proposed Method        | 0.4850  | 0.2924 | 0.6100|

**Table 8** 2018 TREC Precision Medicine Experimental comparison

| Methods                | InfNDCG | R-Prec | P@10  |
|------------------------|---------|--------|-------|
| Brown [40]             | 0.4000  | 0.2350 | 0.4980|
| Klick-Labs [41]        | 0.4432  | 0.2870 | 0.5400|
| UCAS [42]              | 0.5580  | 0.3654 | 0.5980|
| UTDHLTRI [43]          | 0.4797  | 0.2870 | 0.6160|
| Proposed Method        | 0.5055  | 0.3579 | 0.6220|

**Table 9** 2019 TREC Precision Medicine Experimental comparison

| Methods                | InfNDCG | R-Prec | P@10  |
|------------------------|---------|--------|-------|
| Brown [44]             | 0.4052  | 0.2527 | 0.4625|
| ECNU-ICA [45]          | 0.4432  | 0.2870 | 0.5400|
| Ims-Unipd [46]         | 0.4750  | 0.3000 | 0.5450|
| Poznan [47]            | 0.4800  | 0.3100 | 0.5500|
| Cincy-MedIR [48]       | 0.4801  | 0.3111 | 0.5675|
| CSIR-Omed [49]         | 0.4766  | 0.3165 | 0.5825|
| Proposed Method        | 0.5172  | 0.3105 | 0.5875|