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

zicheng zhang (18551701375@163.com)

Research article

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

Posted Date: January 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-53366/v3

Version of Record: A version of this preprint was published on March 2nd, 2021. See the published version at https://doi.org/10.1186/s12911-021-01454-5.

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

Zicheng Zhang[1,2]

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 vast 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 novel article retrieval method based on expanded word and co-word analyses, also conducting Cuckoo Search to optimize parameters of the retrieval function. The main goal is to retrieve the abstracts of biomedical articles that refer to treatments. The methods mentioned in this manuscript adopt the BM25 algorithm to calculate the score of abstracts. We, however, propose an improved version of BM25 that computes the scores of expanded words and co-word leading to a composite retrieval function, which is then optimized using the Cuckoo Search. The proposed method aims to find both disease and gene information in the abstract of the same biomedical article. This is to achieve higher relevance and hence score of articles. Besides, we investigate the influence of different parameters on the retrieval algorithm and summarize how they meet various retrieval needs. **Results:** The data used in this manuscript is sourced from medical articles presented in Text Retrieval Conference (TREC):Clinical Decision Support (CDS) Tracks of 2017, 2018, and 2019 in Precision Medicine. A total of 120 topics are tested. Three indicators are employed for the comparison of utilized methods, which are selected among the ones based only on the BM25 algorithm and its improved version to conduct comparable experiments. The results showed that the proposed algorithm achieves better results.

Conclusion: The proposed method, an improved version of the BM25 algorithm, utilizes both co-word implementation and Cuckoo Search, which has been verified achieving better results on a large number of experimental sets. Besides, a relatively simple query expansion method is implemented in this manuscript. Future research will focus on ontology and semantic networks to expand the query vocabulary.

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

1. Background

With the proliferation of computer technologies, the information available on the Internet has swiftly increased leading to various implementations utilized for information extraction from medical articles. Hence, medical treatment techniques have stepped into the age of Big Data. However, managing the immense data and extracting information from them is a critical endeavor. If this process can be improved, the advantages that it could offer would be so beneficial for medical doctors. For instance, some routine decision-making tasks require significant repetition, which takes time and increases costs. However, computerized medical information retrieval systems can effectively improve efficiency, save costs, and reduce errors. Proper use of computer technology can bring efficacy to all fields where it will be used. Therefore, the development of medical information retrieval systems is crucial. In reality, every decision of a doctor is critical to the patient, so the doctor must follow the state-of-the-art techniques and keep abreast with the latest technology and methods of clinical science. The academic literature providing the latest research results in the medical community can be accessed via the Internet and the medical retrieval models play a crucial role. Furthermore, searching the relevant biomedical literature on the Internet for a reference can be highly beneficial for medical practitioners who encounter a difficult problem on a certain medical record.

Information Retrieval (IR) methods for Clinical Decision Support (CDS) have been the focus of recent 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 or abstracts from an open-access subset of MEDLINE to the clinicians in return to their queries. Furthermore, the tracks from 2017 to 2019 [4-6] focused on important implementations in clinical decision support providing both useful and precise medical information to clinicians treating cancer patients. In these, each case described the disease (a type of cancer), the relevant genetic variants (which genes), and basic demographic information (age and sex) of patients. Precision Medicine introduced in [7] is a new medical concept utilizing individualized medicine that develops with the rapid progress of genome sequencing technology and the cross-application of bioinformatics and Big Data science.

2.Preleminaries

The IR aims to retrieve related documents based on a given query. The relevancy of documents to queries is often gauged by the score assigned by an IR model, e.g., the widely-implemented BM25 model [8]. On the one hand , the past few decades witnessed the implementation of machine learning technology when information retrieval was a concern. The document ranking process could be classified into three groups as follows: i) the single document methods, ii) the document pair methods, and iii) the document list methods. The common single-document methods, such as [9] utilizing a logistic regression technique, deal with a feature vector of each document as an input, where the output is the relevance of each document. The document pair methods, e.g., the ones utilizing Rank-SVM [10] or Rank-Boost [11], implement a feature vector of a pair of documents as the input and use the correlation between the documents as the output. The document list methods, e.g., the ones proposed List-Net [12], Ada-Rank [13], or Lambda-Mart [14], employ a set of documents associated with a query as the input and a ranked list as the output. In recent years, query expansion methods have been widely implemented in information retrieval. Singh et al. [15] suggested a method based on fuzzy logic, in which the top-ranked documents were regarded as relevant feedback documents for mining query information. Furthermore, the choice of different query expansion terms was determined according to their importance. These methods often assign each term to a different relevance score and then select the expansion term based on a certain threshold.

Keikha et al. [16] considered the Wikipedia corpus as the feedback set space to train the Word Vector Model and determined the long-term selection of the best features in both supervised and unsupervised models. Almasri et al. [17] also utilized vectors to represent query words and query expansion terms returned by pseudo-correlation feedback. They added cosine similarity is to the Bag-of-Words Model, and the frequency of each word in the query term was recalculated. Rocchio et al. [18] proposed a classic correlation feedback method, which increased the entry weight of the related documents and reduced it to that of the non-relevant ones. However, one of the disadvantages of this method was to be very time-consuming for practitioners in assessing the relevance of documents.

Cui et al. [19] developed a query expansion method for web search logs utilizing the interaction information of practitioners. The key assumption behind this method was that the documents chosen by a user to read were related to the query. The new words in the related documents were sorted according to their similarity with the user query, and the new words with the highest similarity were selected as the expanded word. The candidate expanded words were extracted from the top documents, and then the

candidate expanded words were weighted and sorted by the probability generated by the language model. Aronson [20] proposed a method based on the Unified Medical Language System (UMLS) query expansion, which benefitted from the Meta-Map program [21] to identify the medical phrases in the original query and then expanded the query with new phrases. Hence, the experimental results showed that the query expansion utilizing the UMLS was an effective method to improve the performance of information retrieval.

Li et al. [22] proposed a method of keyword-weighted network analysis to implement a medical full-text recommendation, which helped to expand the medical acronym list by searching the full-text. Domain experts verified that the algorithm worked well in terms of accuracy in recommending medical literature. Saeid et al. [23] developed a query expansion method utilizing the Bayesian approach, which expanded the genes of a disease to be no less than three words. The experiments revealed that the algorithm had a higher precision value.

The literature review brings us the idea of using both query expansion and keywords to retrieve documents that are highly related to a query. Hence, this manuscript proposes a 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. This is to compute scores of the abstract, expanded words, and co-word as a composite retrieval function. Besides, when a disease and a gene both appear in the same biomedical article, the score of the article tends to increase. Finally, the Cuckoo Algorithm[28] is utilized to optimize the parameters of the proposed retrieval algorithm.

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

3. Experimental data

3.1 Data structure

The abstracts of biomedical articles are presented in XML format. 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 2017 and 2018 TREC Precision Medicine is 26,613,834, the 2019 set has 29,137,637 articles. Table 1 shows some of the statistics that are used in information retrieval, where 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 relevant to both biomedical and health sciences. The important role of the MeSH in medical information retrieval is mainly manifested with two aspects, namely accuracy and specificity. While indexers input information into the retrieval system, researchers utilize this information concerning the two aspects. The MeSH is used as the platform making the terms consistent between the index and search to achieve the best outcomes. Hence, the accurate and comprehensive usage of the MeSH has a significant impact on the results of information retrieval. In this manuscript, we utilize the MeSH database (meshb.nlm.nih.gov/MeSHonDemand) to find expansion terms or new words and their broader terms. The MeSH on Demand is utilized to expand query terms and obtain additional terms if possible.

Table 2 shows Topic 2017-1 as an example, and Table 3 presents the results of the extended words.

3.4 Age expansion

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

4. The Proposed Model

We first utilize the MeSH on Demand to find MeSH terms and additional terms that can be used in the retrieval of abstracts for any given query. Then, we construct a "wordlist" including chemical words, keywords, and MeSH headings that utilizes query expansion, thereby finding documents that are more related to query expansion. This is to increase the relevance score of documents. In the next step, we performed the co-word analysis utilizing either each separate resource, such as abstract, keywords, chemical words, or MeSh headings, or all sources 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. Then, the score of the co-word is calculated. Hence, as long as the number of documents is reduced based on the "word list", the score of the co-word tend to increase. In the last step, we compute the composite score consisting of three scores of abstract, expanded words, and co-word. Afterward, Cuckoo Search[28], an evolutionary optimization method, is applied to optimize the parameters of the proposed retrieval model.

4.1 The abstract scoring model

The BM25 [8] is a classical information retrieval model that is based on analyzing a query Q to find a morpheme qi. For each search result d, it calculates the correlation between each morpheme qi and d, and finally gives a weight to the sum of the correlation score of qi concerning d to obtain a correlation score between Q and d. The general formula of the BM25 can be expressed by:

$$Score(Q, d) = \sum_i^n W_i \times R(q_i, d) \quad (1)$$

where W_i is a weight determining the relevance of a word to a document.

The Inverse Document Frequency (IDF) is defined as:

$$IDF(q_i) = log \frac{D}{card(\{q_i | i \in d_i\})} \quad (2)$$

where D represents the total number of corpus documents, and $card(\{j|i \in d_i\})$ represents the number of documents containing morpheme qi. According to (2), the more qi contained in a document, the lower the weight of qi for a given set of documents. In other words, when several documents contain qi, the discrimination of qi 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 documenting d is defined as:

$$R(q_i, d) = \frac{f_i \times (k_1+1)}{f_i+K} \times \frac{qf_i \times (k_2+1)}{qf_i+k_2} \quad (3)$$

Where parameter K is:

$$K = k_1 \times (1 - b_1 + b_1 \times \frac{dl}{avgdl}) \quad (4)$$

where, k_1, k_2, and b are adjustment factors that are usually set according to experience, f_i is the frequency of q_i in d, qf_i is the frequency of q_i in query, dl is the length of document d, and $avgdl$ is the average length of all documents. In most cases, q_i appears only once in the query, i.e., qf_i = 1. Hence, (3) can be rewritten as:

$$R(q_i, d) = \frac{f_i \times (k_1+1)}{f_i+K} \quad (5)$$

As seen from the definition, the role of parameter b is to tune the impact of the document length on the relevance. The larger the b is, the greater the impact of the document length on the relevance score will be, and vice versa. Similarly, the longer

the relative length of the document is, the larger the K, and hence the smaller the relevance score will be. In the end, the correlation score of the abstract of the document d can be expressed as:

$$Score_{abstract}(Q,d) = \sum_i^n IDF(q_i) \times \frac{f_i \times (k_1+1)}{f_i + k_1 \times (1 - b_1 + b_1 \times \frac{dl}{avgdl})} \quad (6)$$

4.2 The expanded word score

As seen in Table 1, most biomedical articles have both abstracts and titles. The number of biomedical articles containing chemical words, MeSH headings, and keywords varies widely. Specifically, there exist 13,113,093 articles containing chemical words, 2,438,717,151 articles containing MeSH headings, and 4,005,446 articles containing keywords. As the literature suggests, direct utilization of the BM25 leads to failure when dealing with a large selection of documents [52]. In this subsection, we propose 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:

$$dwl = dcl + dml + dkl \quad (7)$$

where dcl is the length of chemical words in document d, dml is the length of MeSH headings in document d, and dkl is the length of keywords in document d.

The IDF value of the expanded word appearing in the "Word List" of document d can be given by:

$$IDF_{word}(q_i,d) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (8)$$

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

$$tf_{word}(Q,d) = \sum_i^n IDF_{word}(q_i,d) \quad (9)$$

where n represents the number of expanded words in query Q, and q_i represents the morpheme of each expanded word in query Q. The score of an expanded word in document d is defined by:

$$Score_{word}(Q,d) = \frac{tf_{word}(Q,d)\times(k_3+1)}{tf_{word}(Q,d)+k_3\times(1-b_2+b_2\times\frac{dwl}{avgdwl})} \quad (10)$$

where k_2 and b_1 are the adjustment factors that are usually set according to experience, and $avgdwl$ represents the average length of all word lists.

4.3 The co-word score

The 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, the co-word analysis is introduced into the article scoring model for the case when 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 IDF value as the co-word score to distinguish the importance of a gene, which can be formulated as:

$$IDF_{gene}(g_i,d) = log\frac{N-n(g_i)+0.5}{n(g_i)+0.5} \quad (11)$$

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

Finally, the Co-Word score is defined by:

$$Score_{co-word}(Q,d) = \sum_i^n IDF_{word}(g_i,d) \quad (12)$$

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

4.4 Retrieval Model

We utilize the composite score as the final score for document d under query Q, which can be formulated as:

$$Score_{composite}(Q,d) = Score_{abstract}(Q,d) + Score_{word}(Q,d) + \alpha \times Score_{co-word}(Q,d) \quad (13)$$

Figure 4 depicts the architecture of the biomedical article retrieval system.

5. Parameter Optimization

The proposed method has six parameters: k_1, k_2, k_3, b_1, b_2, and α, and the choice of parameters can affect the results of information retrieval. Various algorithms, e.g., the Genetic Algorithm (GA) [25], Simulated Annealing (SA) Algorithm [26], and Ant Colony (AC) Algorithm [27], have been implemented to optimize the function in use, i.e., the objective function. With the continuous effort in developing better algorithms, several new Swarm Intelligence Optimization (SIO) Algorithms have emerged during recent years, such as Cuckoo Search (CS) [28], Glow Worm Swarm Optimization (GWSO) [29], and Particle Swarm Optimization (PSO) [30]. Among them, SIO has been widely utilized.

5.1 Cuckoo Search Algorithm

CS is a SIO proposed by Yang et al. [28] in 2009. Maribel et al. [31] claimed that CS outperformed GA in terms of efficiency. Some of the idealized rules utilized by CS can be given as:

(1) Each cuckoo lays only one egg every time and selects a parasitic nest to randomly hatch its egg.

(2) The best parasitic nest will be handed down 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-given rules. The position update formula is:

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

where T is the step size ($T > 0$), $\oplus$ is the point-to-point multiplication operator, $Levy(\lambda)$ is the search path following 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, $f(x)$, $X = (x_1, .., x_d)^T$ is a d-dimensional problem

3: Generate initial population of n host nests

$$X_i \ (i = 1, 2, \ldots, n)$$

4: While ($t <$ Max Generation) or (stop criterion)

5: Get a cuckoo randomly by $Levy$ 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 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 the 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

Precision is calculated as:

$$Precision = \frac{RR}{(RR+RN)} \tag{15}$$

where RR and RN refer to relevant and irrelevant documents retrieved, respecti vely.

Then, $P@10$ is defined as the Precision when $RR + RN = 10$. Hence, the average $P@10$ can be formulated as:

$$Avg_{P@10} = \frac{\sum_{t=1}^{n} P@10(t)}{n} \tag{16}$$

where $P@10(t)$ represents the $P@10$ value of the t^{th} topic among n topics.

The Normalized Discounted Cumulative Gain(nDCG) [70] is a commonly utilized index to assess the quality of ranking in information retrieval. Let ϑ denote the relevance grade, and $gain(\vartheta)$ denote the gain associated with ϑ. Also, assume that $g_1, g_2, \ldots g_z$ are the gain values associated with the Z documents retrieved by a system in response to query q, such that $g_i = gain(\vartheta)$ if the relevance grade of the document in rank i is ϑ. Hence, the nDCG value for this system can be calculated as:

$$nDCG = \frac{DCG}{DCG_I}, \text{ where } DCG = \sum_{i=1}^{Z} \frac{g_i}{\log(i+1)} \tag{17}$$

and DCG_I denotes the DCG value for an ideal ranked list for query q.

We define the average $nDCG$ as follows:

$$Avg_{nDCG} = \frac{\sum_{t=1}^{n} nDCG(t)}{n} \tag{18}$$

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

5.3 Algorithm flow

Since k_2 has a fixed value ($k_2 = 1$), we utilize k_1, k_3, b_1, b_2, and α as input parameters. Firstly, the algorithm generates the initial population and either set the maximum number of iterations or stop the criterion. If the number of iterations reaches the maximum number or the stop criterion is met, the algorithm ends and returns the optimal solution. Otherwise, the algorithm performs a series of operations to optimize the objective function. This manuscript defines $Avg_{P@10} + Avg_{nDCG}$ as the objective

function and employs the dataset of the 2017 Precision Medicine as the training dataset to optimize the parameters. Figure 5 presents the flowchart of the proposed algorithm.

5.4. Experimental results and comparison

5.4.1. Dataset

The data used in this work were sourced from medical articles published in 2017, 2018, and 2019 TREC Precision Medicine, which can be found on 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 was formatted using the XML 2017. The assessment results of the articles were 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).

Due to the semi-structured nature of the XML format, we used MongoDB as the database for document storage and Python as the programming language. All the code can be found on the corresponding author's GitHub (https://github.com/Bruce-V/CS-BM25).

5.4.2 Parameter setting

Table 5 presents the parameter values used in the proposed algorithm.

5.4.3 Experimental results

In Table 6, "Normal" refers to the values of empirical parameters, where CS represents the parameters trained using the 2017 dataset consisting of 1000 documents with the highest scores as a result of the selected retrieval model.

When the data of three years are compared, the optimized parameters are better than the empirical parameters. For an information retrieval system, the users desire related documents to appear earlier; hence, infNDCG and P@10 are two important indicators in assessing the performance of the information retrieval process. The parameters that

are optimized using both NDCG and P@10 would increase the weights of the word list. The word list includes extended information about age, gender, and genes, which is crucial for distinguishing the relevant literature from the irrelevant ones. In conclusion, different parameters can be utilized to meet the needs of various users.

In Figures 6, 7, and 8, RR represents the relevance in co-word documents, while RN represents all relevance except for RR. It is shown that many relevant documents contain both a disease and a gene. As a result, we define the rate of average relevant document coverage as:

$$Avg_{cov} = \frac{\sum_1^n \frac{relevance\ in\ co-word}{relevance}}{n} \quad . \tag{19}$$

The average coverage rate of 30 topics in 2017 is 52.9%, which is 74.13% for 50 topics in 2018, and 54.4% for 40 topics in 2019. These outcomes reveal that the co-word analysis has a better impact on the retrieval of relevant documents, which greatly reduces the scope of the search.

When information retrieval is a concern, NR, RR, NN, and RN respectively represent the relevant documents that are not retrieved, the irrelevant documents that are not retrieved, the relevant documents that are retrieved, and the irrelevant documents that are retrieved. Here, *Precision* is defined as Formula (15) .

Recall is defined as:

$$Recall = \frac{RR}{(RR+NR)} \tag{20}$$

and *F1-score* is defined as:

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{21}$$

As seen in Figures 9, 10, and 11, the optimized parameters are better than the empirical parameters for 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 this 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. *Precision* and *Recall* are inversely proportional to each other when the retrieval system is a concern. However, in our case examining the retrieval of biomedical articles, we are more concerned about the precision rate to alleviate the doctors' decision-making.

5.4.4 Experimental comparison

Considering the results of the models taken from 2017, 2018, and 2019 TREC Precision Medicine, three indicators called infNDCG, R-predicted, and P@10 are selected for comparison, and the experimental results are presented in Tables 7, 8, and 9.

We use three years of the TREC datasets to verify our experimental results. The selected methods either utilize the BM25 algorithm or its improved version. The experiments using the 2017 dataset showed a significant improvement in our proposed method for all indicators. For the 2018 dataset, our method performed better than similar algorithms for P@10, ranked second for infNDCG. The same result was observed for the 2019 dataset.

6. Conclusion

This manuscript proposes a BM25-based method incorporating co-word analysis to retrieve biomedical articles. We improved the BM25 algorithm and used it to compute the score of expanded words by combining the co-word score with the gene appearance weight. Then, we utilized the Cuckoo Search Algorithm to optimize parameters on the evaluation function of both P@10 and nDCG. Optimization results suggested that increasing the score weight of the "word list" could effectively improve the ranking of the related documents. The manuscript also 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, it highlights the general rules for improving the parameters of the BM25 algorithm, which were verified through numerous experiments. Since the query expansion used in this manuscript is simple, our future research will focus on adopting more linked data to investigate 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); Genetic Algorithm (GA); Simulated Annealing (SA); Ant Colony (AC);Swarm Intelligence Optimization (SIO) ;Cuckoo Search (CS);Glow Worm Swarm Optimization (GWSO);Particle Swarm Optimization (PSO)

Declarations

Ethics approval and consent to participate

On behalf of, and having obtained permission from all 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 actively involved in substantial work leading to the submitted version, 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 authors.

Consent for publication

This article uses publicly available datasets.

Availability of data and material

The data used in this work were sourced from medical articles published in 2017, 2018, and 2019 TREC Precision Medicine, which can be found on 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 was formatted using the XML 2017. The assessment results of the articles were 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).

All the code can be found on the corresponding author's 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 has been independently completed by Zicheng Zhang. The author has read and approved the final manuscript.

Acknowledgments

We would like to thank Wei Cao for their contribution to 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 Meta-Thesaurus: 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., Wilczynski 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

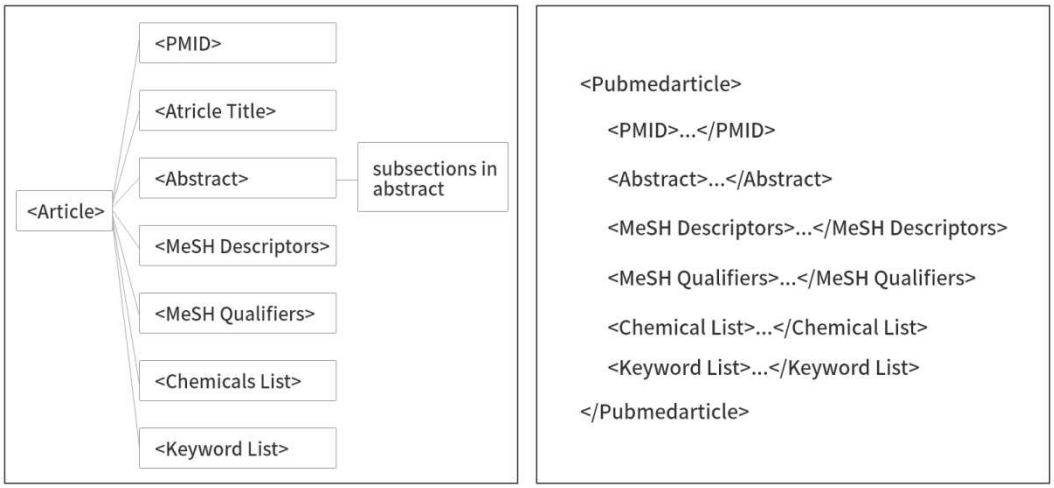

Figure 1 General structure and the XML attributes of MEDLINE abstracts

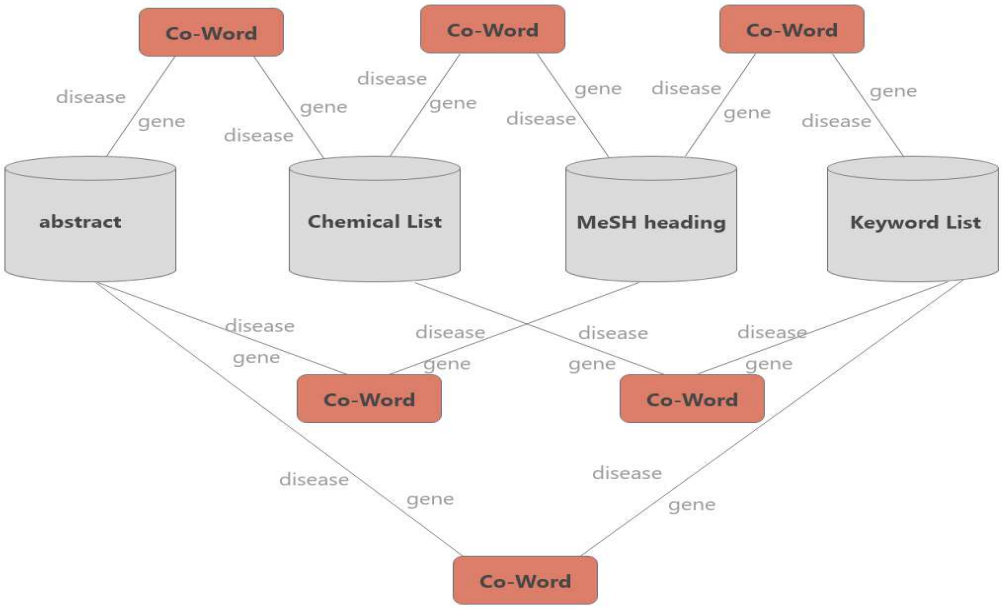

Figure 2 The cross co-word of abstract, chemical list, MeSH heading, and keyword list

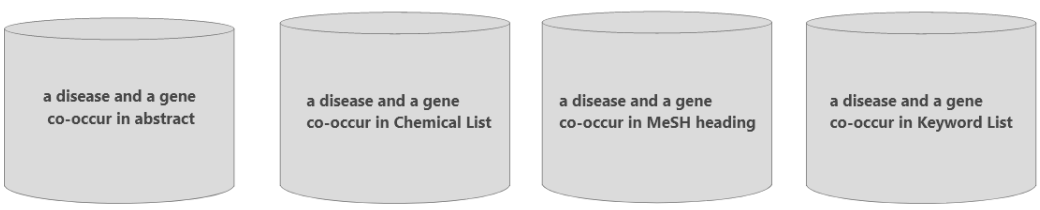

Figure 3 The co-word of abstract, Chemical List, Me-SH heading, and Keyword List

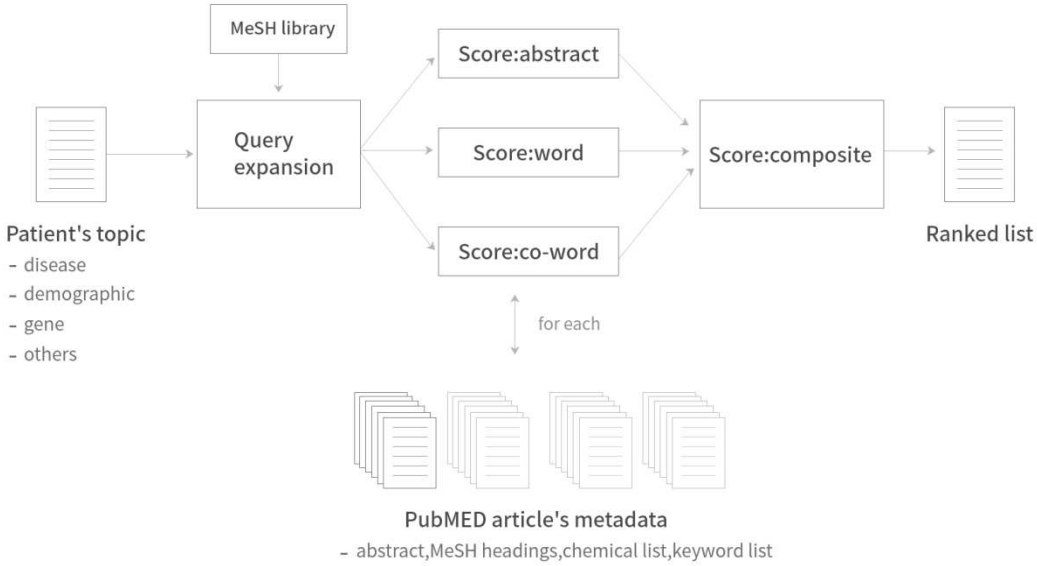

Figure 4 Architecture of the system retrieving biomedical articles

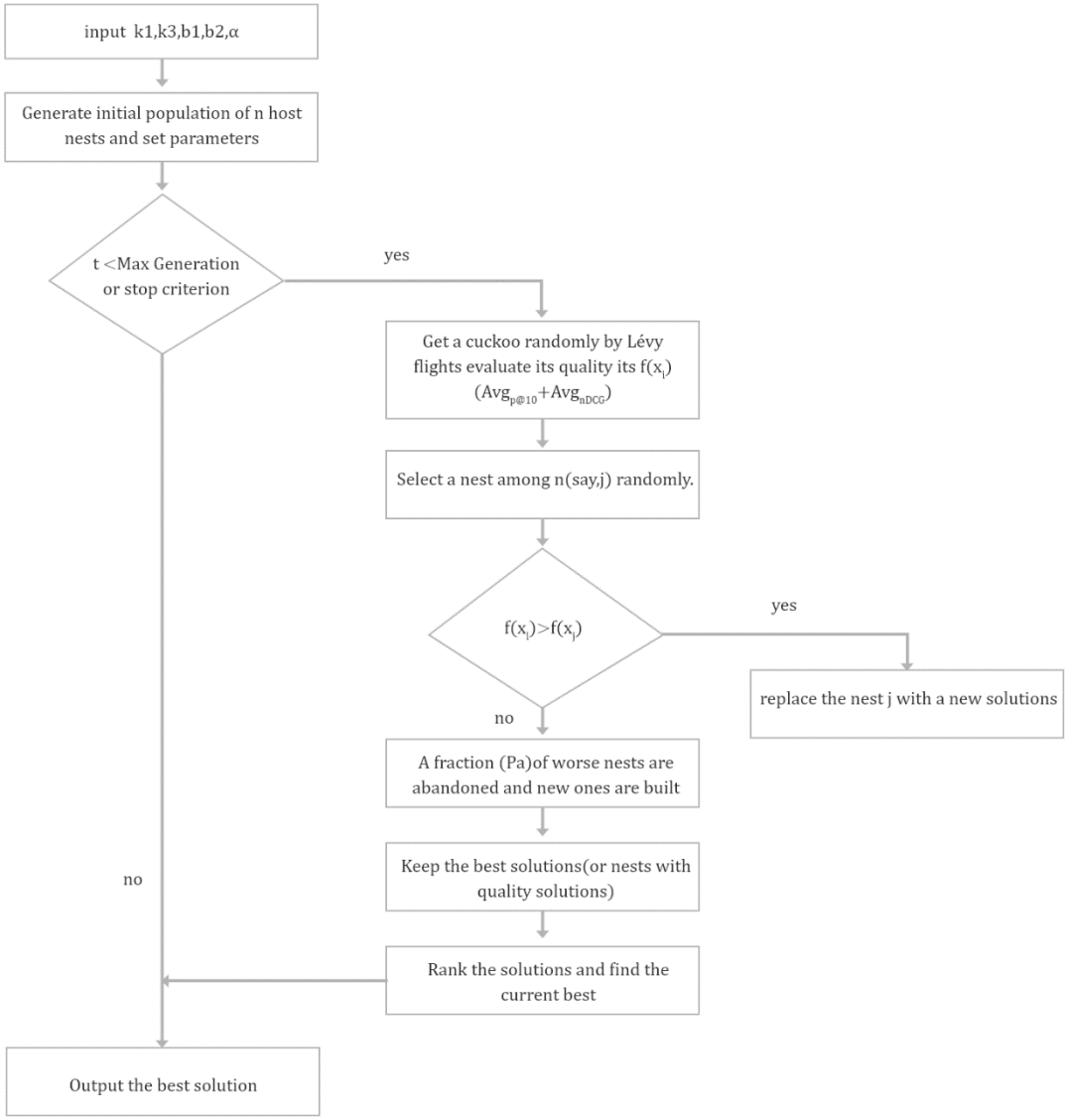

Figure 5 Flowchart of the proposed algorithm

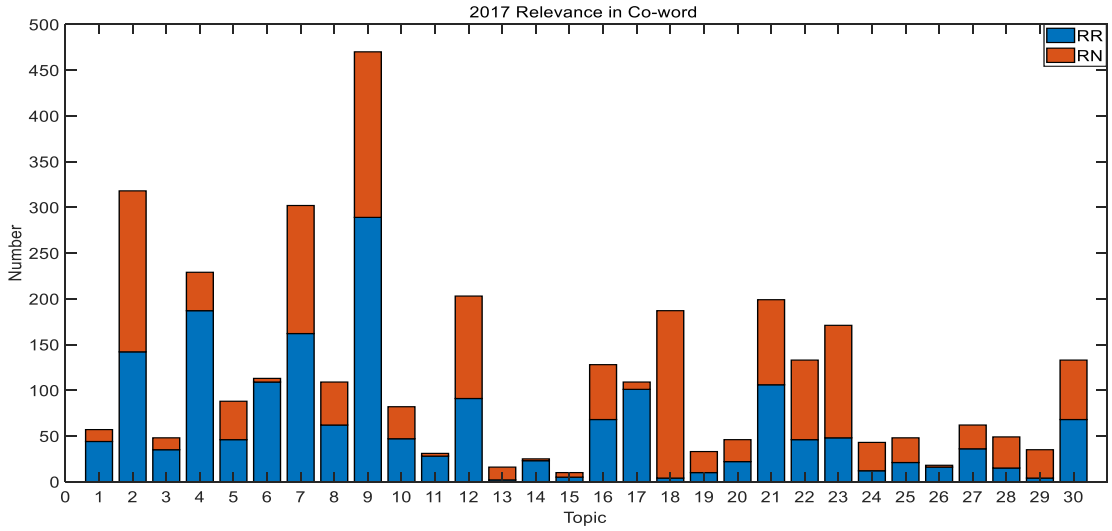

Figure 6 2017 TREC Precision Medicine relevance in co-word biomedical articles

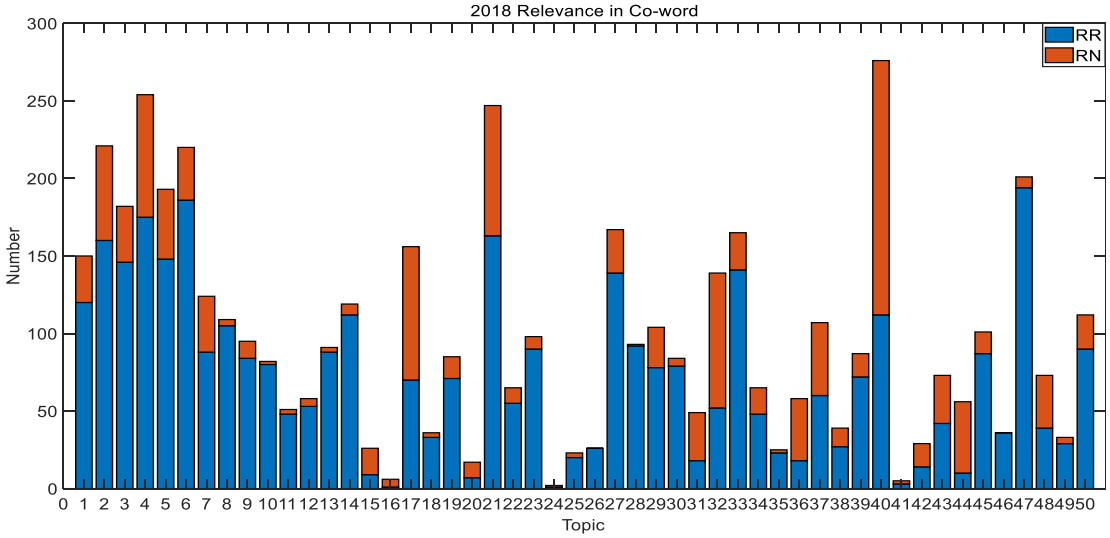

Figure 7 2018 TREC Precision Medicine relevance in co-word biomedical articles

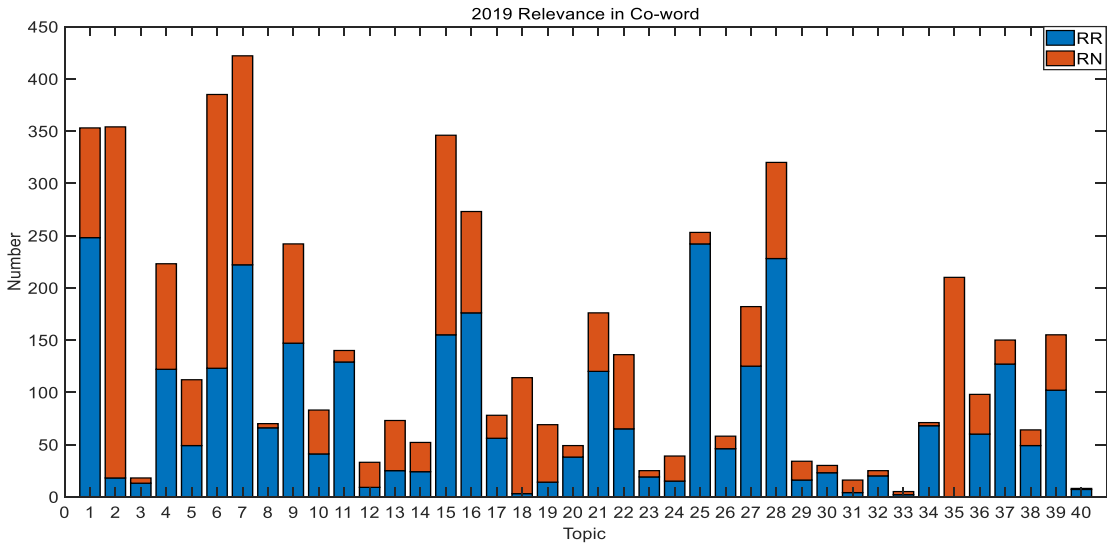

Figure 8 2019 TREC Precision Medicine relevance in co-word biomedical articles

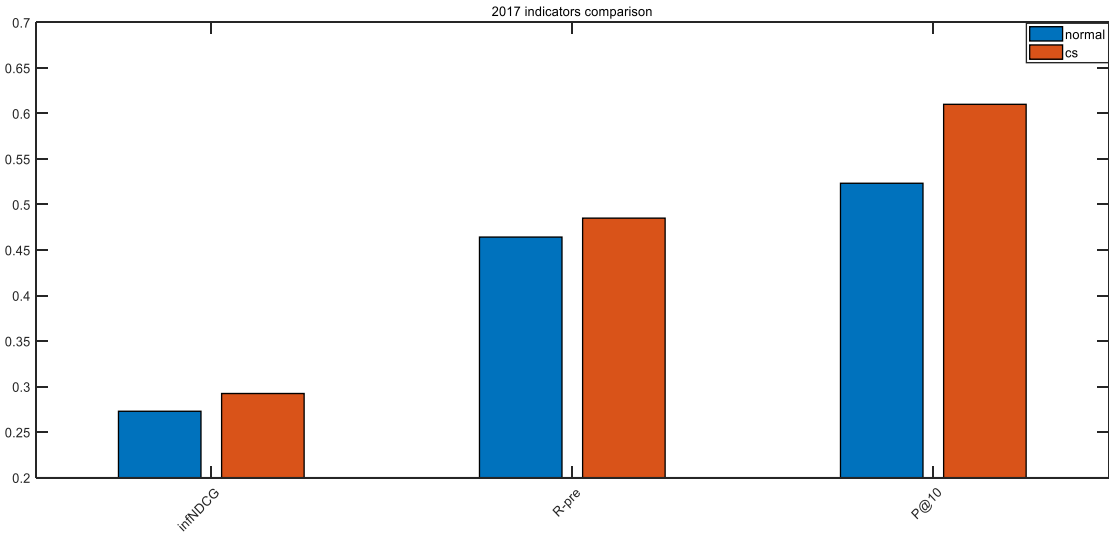

Figure 9 Comparison of 2017 TREC Precision Medicine indicators

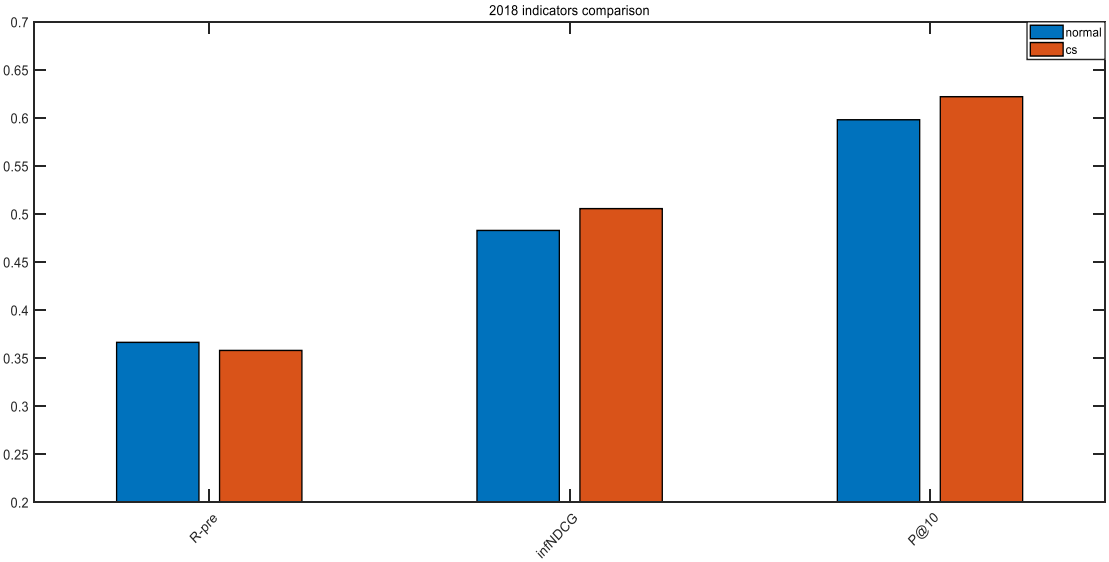

Figure 10 Comparison of 2018 TREC Precision Medicine indicators

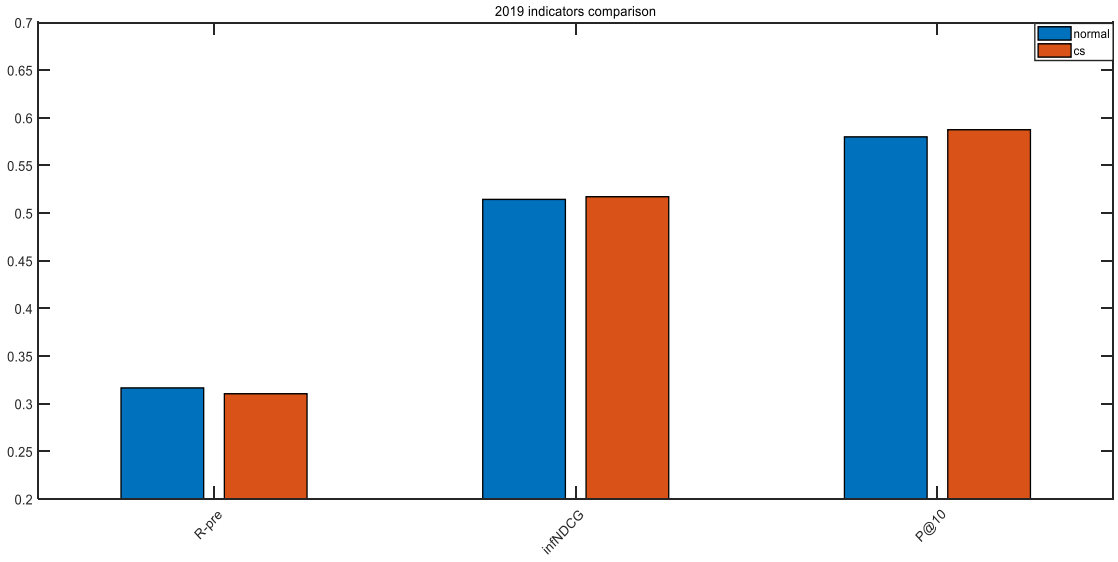

Figure 11 Comparison of 2019 TREC Precision Medicine indicators

Table 1 The statistics of the 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
Keyword-Number	4,005,446	5,435,471

Table 2 The retrieval topic description of the TREC Precision Medicine

Year	Disease	Gene	Demographic characteristics	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	Middle Aged
	Adult
Male	Human

Table 4 The expanded age of the 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
$k_1\_boundary$	Boundary of k_1	(0,100)
$k_3\_boundary$	Boundary of k_3	(0,100)
$b_1\_boundary$	Boundary of b_1	(0,1)
$b_2\_boundary$	Boundary of b_2	(0,1)
$\alpha\_boundary$	Boundary of α	(0,5)

Table 6 The results of the Cuckoo Search Algorithm

Name	k_1	k_3	b_1	b_2	α
"Normal"	1.2	1.2	0.75	0.75	1
CS	3.5	91.3	0.84	1	4

Table 7 Experimental comparison of methods published in 2017 TREC Precision Medicine

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 Experimental comparison of methods published in 2018 TREC Precision Medicine

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 Experimental comparison of methods published in 2019 TREC Precision Medicine

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

Figures

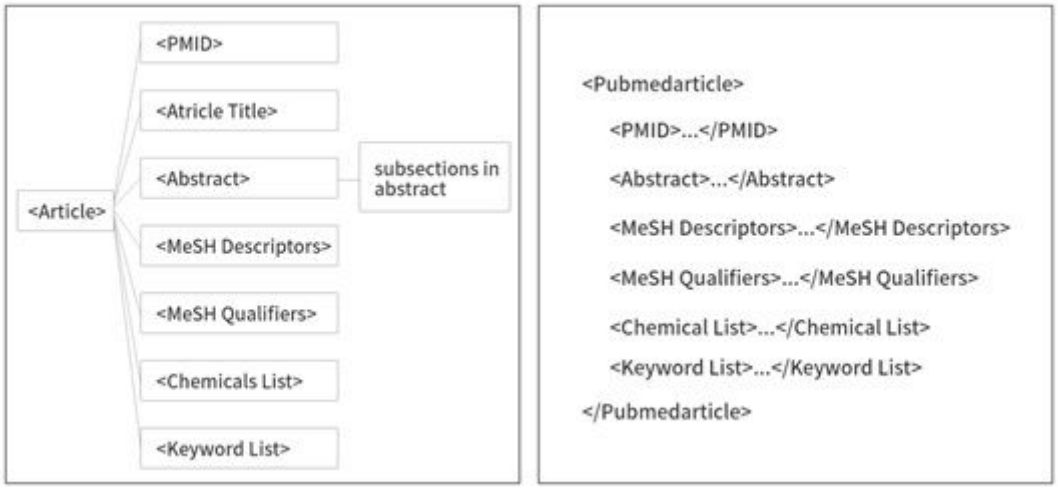

Figure 1

General structure and the XML attributes of MEDLINE abstracts

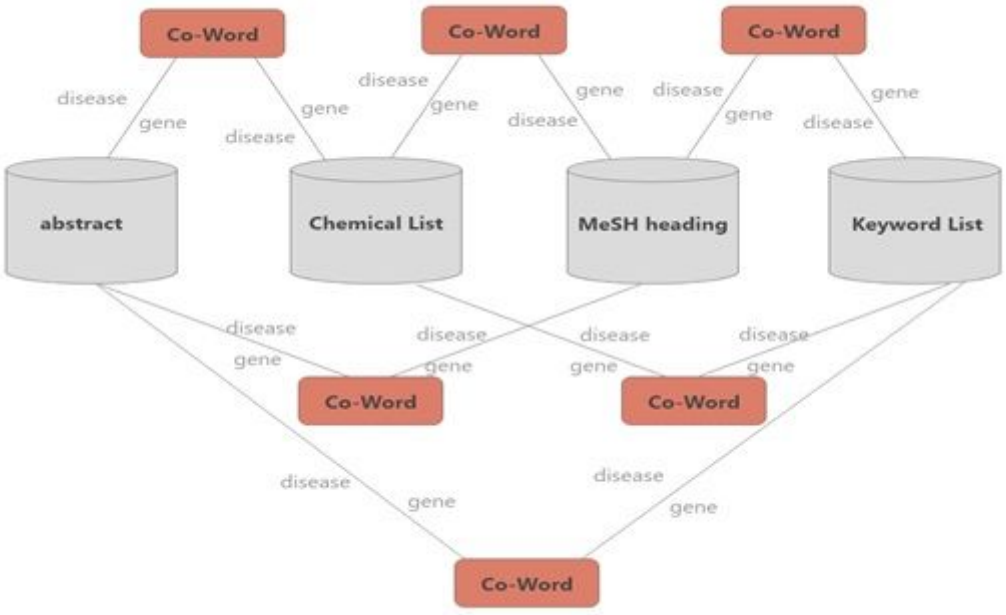

Figure 2

The cross co-word of abstract, chemical list, MeSH heading, and keyword list

Figure 3

The co-word of abstract, Chemical List, Me-SH heading, and Keyword List

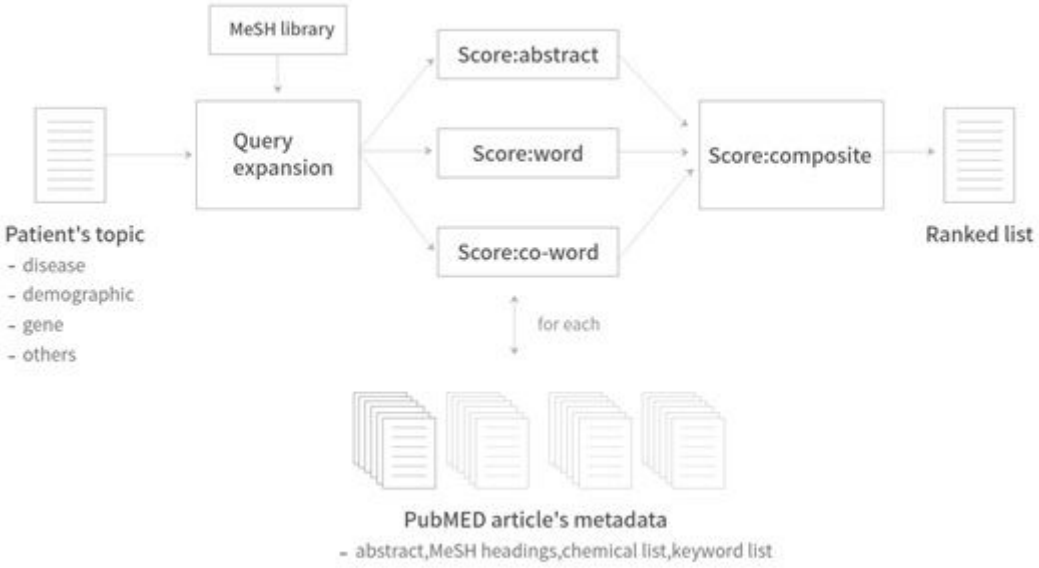

Figure 4

Architecture of the system retrieving biomedical articles

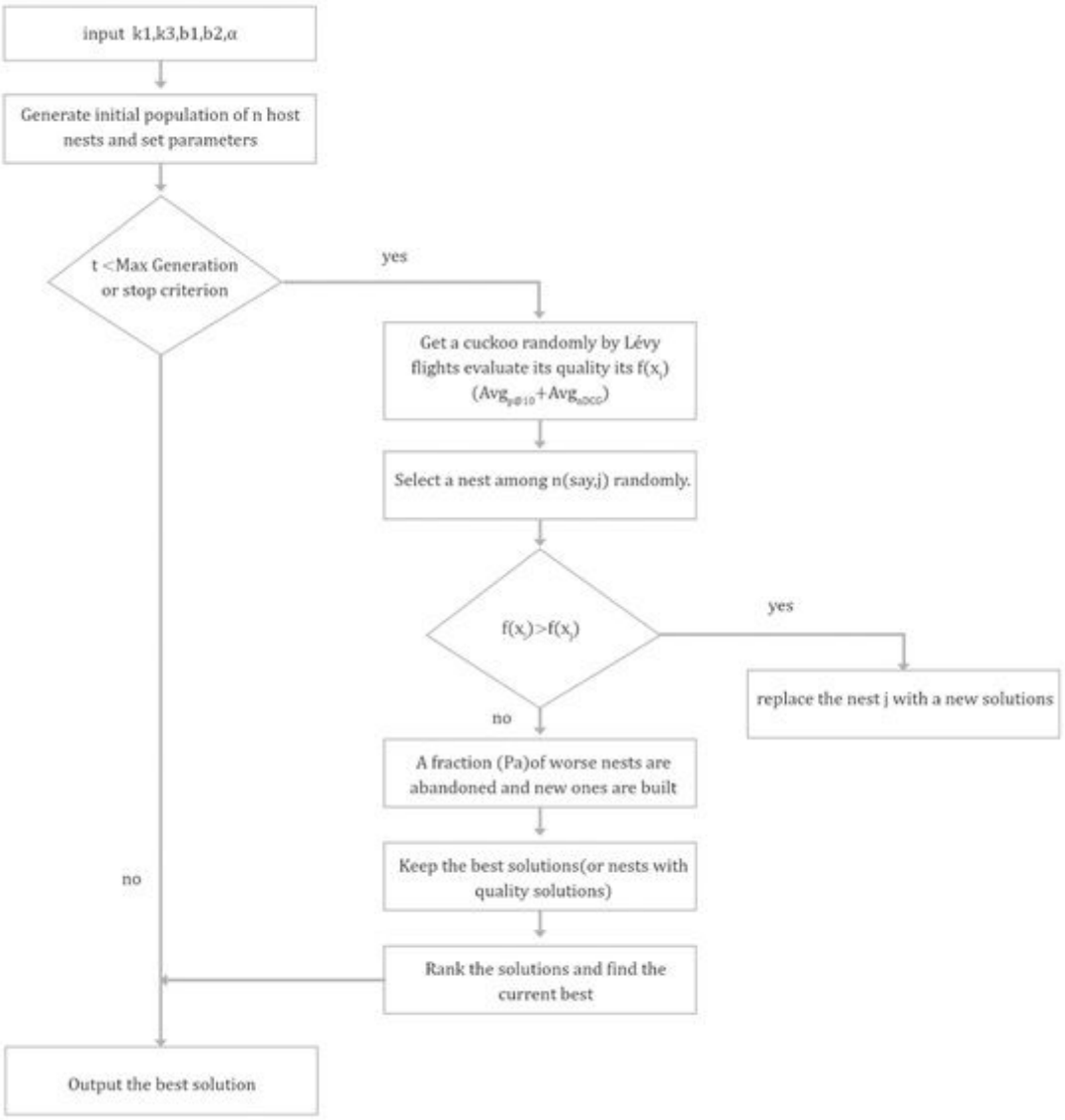

Figure 5

Flowchart of the proposed algorithm

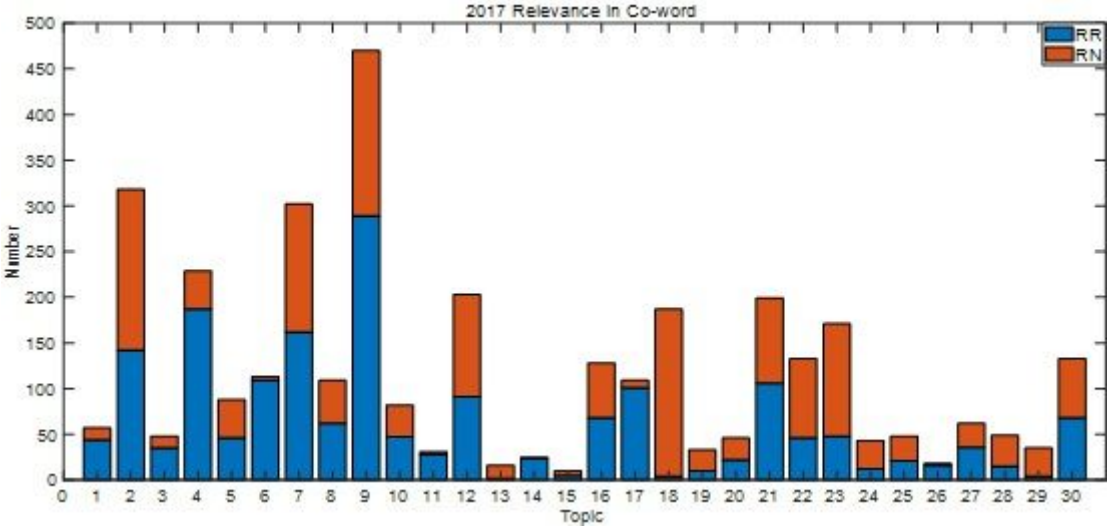

Figure 6

2017 TREC Precision Medicine relevance in co-word biomedical articles

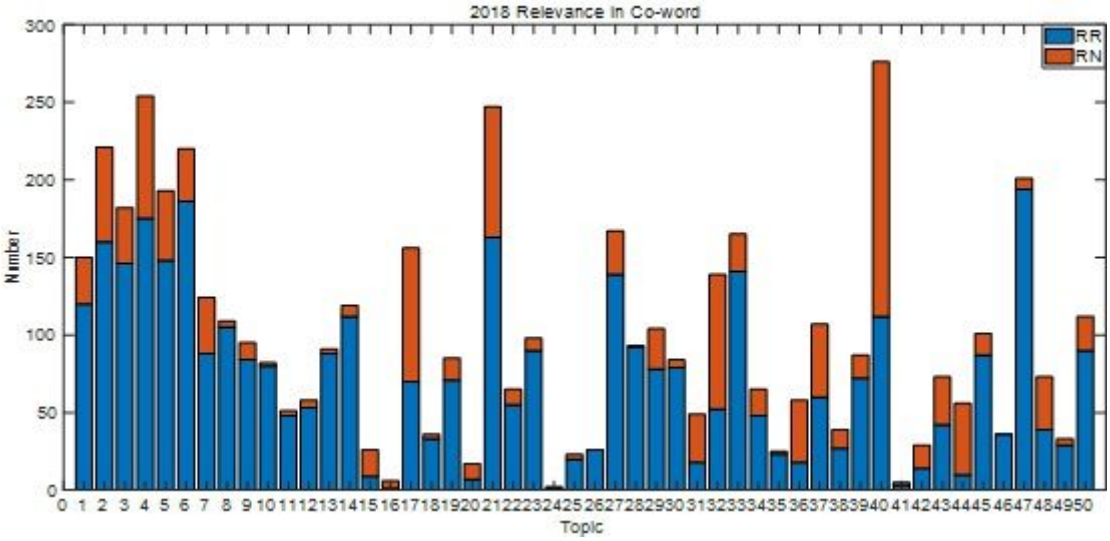

Figure 7

2018 TREC Precision Medicine relevance in co-word biomedical articles

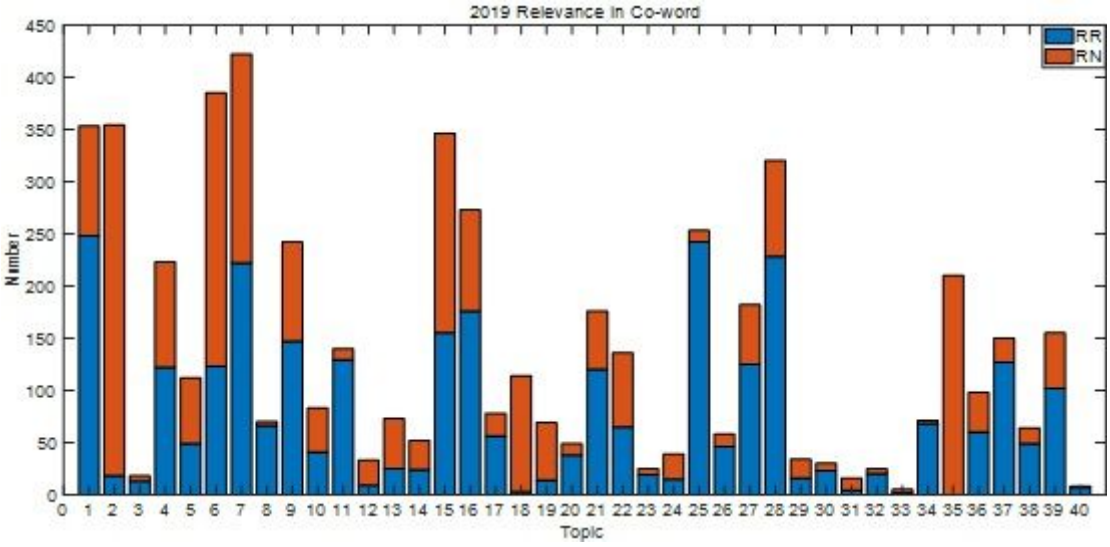

Figure 8

2019 TREC Precision Medicine relevance in co-word biomedical articles

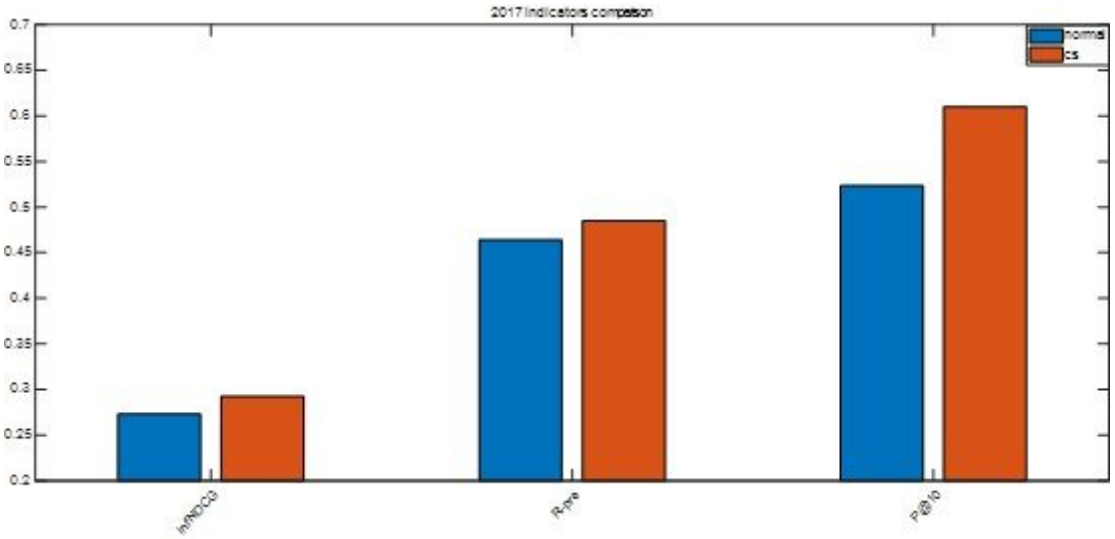

Figure 9

Comparison of 2017 TREC Precision Medicine indicators

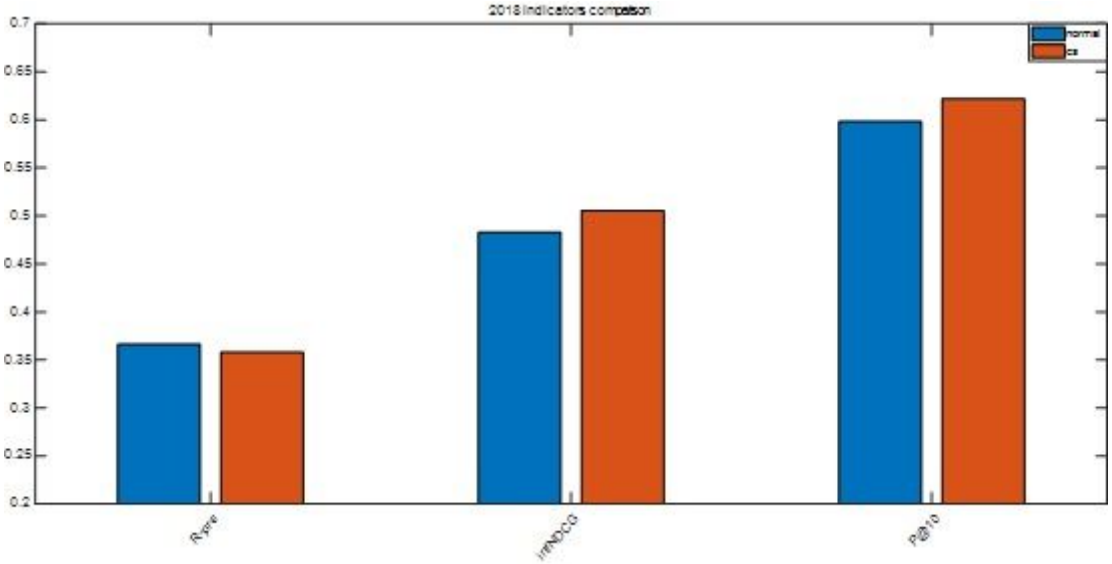

Figure 10

Comparison of 2018 TREC Precision Medicine indicators

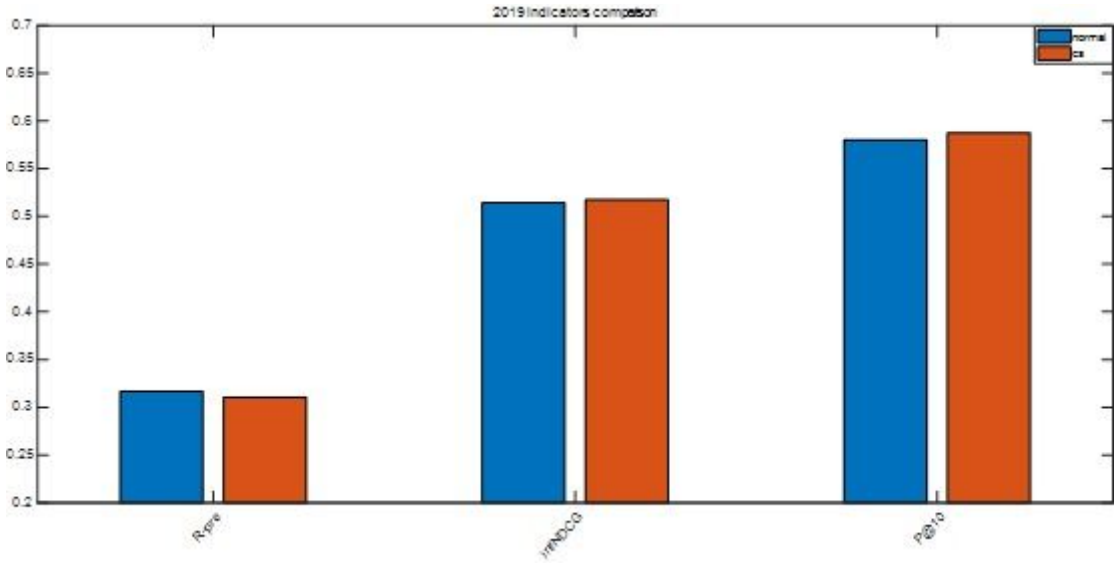

Figure 11

Comparison of 2019 TREC Precision Medicine indicators