An Improved BM25 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 massive biomedical articles to provide doctors with clinical decision support is one of the important research directions of precision medicine**. Methods:** we present a new method for biomedical article retrieval based on co-word analysis and cuckoo search. The specific goal is to retrieve biomedical articles, in the form of article abstracts, addressing relevant treatments for a given patient. The method in this paper first uses the BM25 algorithm to calculate the score of the abstract, and we designed a method based on BM25 to calculate the score of expanded words. Second, when a disease and a gene both appear in the same biomedical article, the score of the biomedical article will be increased. Finally, the cuckoo algorithm is used to optimize the parameters of the retrieval algorithm. The paper discusses the influence of different parameters on the retrieval algorithm, and summarizes the parameters to meet different retrieval needs. **Results:** all data were taken from medical articles provided in the TREC (Text Retrieval Conference) Clinical Decision Support Track 2017、 2018 and 2019 in precision medicine. 120 standard topics were tested. we chose 3 test indicators and many kinds of algorithms for experimental comparison. For the fairness of the experiment, all these selected algorithms all used the BM25 algorithm or an improved BM25 algorithm. Experimental results showed that our algorithm has achieved good results and ranking

Conclusions: we designed an improved BM25 algorithm based on co-word analysis and cuckoo search and verified the superiority of our algorithm on a large number of experimental sets. In this paper, the query expansion method is relatively simple, the next step is to consider the ontology and semantic network to expand the query vocabulary.

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

Introduction

With the continuous development of technology, the information available on the Internet is increasingly abundant, and the means of obtaining information are increasingly convenient. Medical treatment has also entered the age of big data. People can easily obtain basic medical-related knowledge from the Internet, such as symptom, treatment and disease prevention information. At the same time, many online medical question-and-answer websites have been developed, which do not require patients to visit doctors for face-to-face checkups. Instead, online questioning is used, which greatly saves manpower, material, and time and largely protects patient privacy. In addition, for some routine decision-making tasks that require significant repetition, the scientific application of computer medical information retrieval

systems can effectively improve efficiency, save costs and reduce errors. Proper use of computer technology can effectively improve the quality of clinical services and greatly reduce costs. Therefore, the development of computer-assisted medical information retrieval systems is highly significant. In actual work, every decision made by a doctor is critical to the patient, so the doctor must constantly learn and pay attention to the latest technology and methods of clinical science. The authoritative literature and the latest research results in the medical community can be consulted on the Internet, so the medical retrieval model plays a crucial role. Moreover, for medical workers who encounter a difficult medical problem for a certain medical record, searching the relevant biomedical literature on the Internet as a case reference and inspiration can provide an important way to solve difficult problems.

IR methods for CDS have been the focus of several recent studies and evaluation campaigns. Specifically, the CDS track at the 2014–2016 Text Retrieval Conference (TREC) [1-3] sought to evaluate the systems that provide evidence-based information in the form of full-text articles from an open access subset of MEDLINE to clinicians in response to medical case narratives as queries. The 2017-2019 [4-6] track focused on an important use case in clinical decision support: providing useful and precise medical information to clinicians treating cancer patients. In the track, each case describes the patient's disease (type of cancer), the relevant genetic variants (which genes), and basic demographic information (age, sex). Precision medicine [7] (PM) is a new medical concept and medical model based on individualized medicine, developed with the rapid progress of genome sequencing technology and the cross-application of bioinformatics and big data science.

Related studies

The purpose of information retrieval is to retrieve documents related to a given query. Generally, the relevancy of documents to queries is usually measured by the score given by an IR model, such as the classic BM25 model [8]. In the past few decades, machine learning technology has been applied to the field of information retrieval and has achieved good results. The earliest machine learning algorithm was learning to rank, which can be divided into three types: the single document method, document pair method, and document list method. Common single-document methods, such as logistic regression [9], use the feature vector of each document as the input and output the relevance of each document. Document pair methods, such as RankSVM [10] and RankBoost [11], use the feature vector of a pair of documents as the input and output the correlation between the documents. Document list methods such as ListNet [12], AdaRank [13], and LambdaMart [14] take a set of documents associated with a query and output a ranked list. In recent years, query expansion methods have been widely used in information retrieval. Singh et al. [15] proposed a query expansion method based on fuzzy logic. The top-ranked documents (top documents) were regarded as relevance feedback documents for mining other relevant query information. The choice of different query expansion terms was determined according to their importance in the top documents. These methods gave each term a different relevance score and then selected the expansion term through a certain threshold. Keikha et al. [16] further considered the Wikipedia corpus as a feedback set space to train the word

vector model and determined the long-term selection of the best features in the supervised and unsupervised modes. Almasri et al. [17] also used vectors to represent query words and query expansion words returned by pseudo-correlation feedback. Cosine similarity was added 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. They increased the entry weight of related documents and reduced the entry weight of non-relevant documents, but that method was very time-consuming for users to evaluate the relevance of documents. Cui et al. [19] proposed a query expansion method for Web search logs based on user interaction information. The key assumption behind this method was that the documents a user chooses to read are related to the query. The words in related documents were sorted according to their similarity with the user query, and the word with the highest similarity was 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 using UMLS query expansion. They used the MetaMap [21] program to identify medical phrases in the original query and then extended the query with phrases. Their experimental results show that query expansion using UMLS is an effective method to improve the performance of information retrieval. Li et al. [22] proposed a method of keyword-weighted network analysis to implement medical full-text recommendation. This algorithm expanded the medical acronym list by searching the full text and was evaluated by domain experts. It was verified that the algorithm works well in terms of recommendation accuracy in medical literature. Saeid et al. [23] proposed a query expansion method based on the Bayesian approach, which expanded the genes of a disease to be no less than 3, and experiments proved that the algorithm has a high precision. As a classic information retrieval algorithm, BM25 is also frequently used on TREC, such as 2017 precision medicine[34-39], 2018 precision medicine[40-43] and 2019 precision medicine[44-49]. These algorithms mainly used the original BM25 algorithm and the improved BM25 algorithm for information retrieval.

This paper proposes a method based on co-word analysis to optimize the information retrieval of biomedical articles. This method first uses the BM25 algorithm to calculate the score of the abstract, and we designed a method based on BM25 to calculate the scores of expanded words. Second, when a disease and a gene both appear in the same biomedical article, the score of the biomedical article will be increased. Finally, the cuckoo algorithm is used to optimize the parameters of the retrieval algorithm

Data structure

The biomedical articles in scientific abstracts are made available in XML formats. The documents include information about the article. We chose abstracts, MeSH

headings, chemical lists and keyword lists for XML documents, as shown in figure 1.

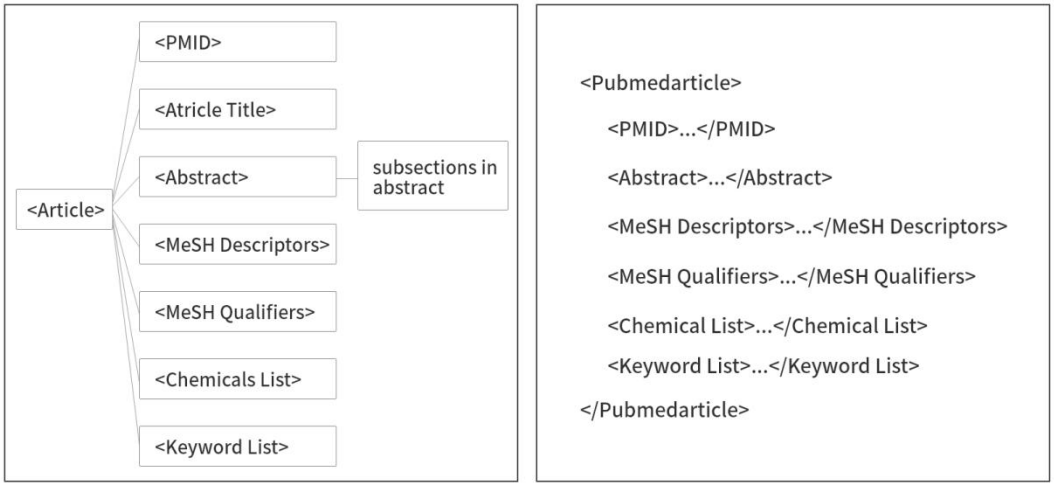

Figure 1 General structure and XML attributes of MEDLINE abstracts

Data distribution

The total number of biomedical articles in the 2017 and 2018 TREC Precision Medicine set is 26615116, and 2018 TREC Precision Medicine set is 29138919. Some other statistics are shown in table 1.

Table 1 Statistics of 2017 TREC Precision Medicine

Name	2017 and 2018	2019
abstract_mean_length	77.5	83.5
abstract_number	26613834	29137637
chemical_mean_length	3.8	3.8
chemical_number	13113093	13670358
mesh_mean_length	10.5	10.6
mesh_number	24387151	25389659
keyword_mean_length	4.1	4.4
keyword_number	4005446	5435471

In table 1, abstract_mean_length represents the average length of the abstracts after deleting stopwords; 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.

Query expansion

Medical Subject Headings (MeSH) is a controlled vocabulary developed by the U.S. National Library of Medicine. MeSH is mainly used to index, catalogue and search articles related to biomedical science and health. The important role of MeSH in medical information retrieval is mainly manifested in two aspects: accuracy and specificity. In the two processes by which the indexers enter information into the retrieval system and the searchers use the information in the system, MeSH is used as

the standard term to make the terms consistent between the index and search to achieve the best retrieval effect. MeSH is essential for the retrieval of medical articles, and the accurate and comprehensive usage of MeSH have a significant impact on the results of retrieval. In this paper, we use the MeSH database to extend MeSH (meshb.nlm.nih.gov/MeSHonDemand).

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

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

Table 3 Expanded MeSH of 2017 TREC Precision Medicine retrieval task 1

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

Age expansion

The age included in the demographic field were also expanded to the terms proposed by Kastner et al.[24]. We have readjusted the division of age, and believe that those over 18 should be adults. Our expansion of age is shown in Table 4

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

Scoring model

Score of abstract

The BM25 model [8] is a classic information retrieval model. The main idea of BM25 is to analyse the morpheme of Query Q to generate morpheme qi; then, for each search result d, calculate the correlation score of each morpheme qi and d; and finally

weight the sum of the correlation score of qi relative to d to obtain the correlation score between Q and d. The general formula of the BM25 algorithm is as follows:

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

where w_i is a weight to determine the relevance of a word to a document; we choose IDF as w_i.

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

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 the definition of IDF, for a given set of documents, the more documents containing qi, the lower the weight of qi. In other words, when many documents contain qi, the discrimination of qi is not strong, so the importance of using qi to judge relevance is weak.

The relevance score $R(q_i, d)$ of morpheme q_i to document d is defined as follows:

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

where K is defined as follows:

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

where, k_1, k_2 and b are adjustment factors, 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 will appear only once in Query; that is, qf_i = 1, so the formula can be simplified to:

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

As seen from the definition, the role of parameter b is to adjust the impact of the document length on the relevance. The larger b, the greater the impact of the document length on the relevance score, and vice versa. The longer the relative length of the document, the larger the K value and the smaller the relevance score. In summary, the correlation score formula of document $d's$ abstract can be summarized as

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

Score of expanded word

From table 1, we can see that almost all biomedical articles have abstracts and titles. The number of biomedical articles containing chemical words, MeSH headings and keywords varies widely. There are 13113093 articles containing chemical words, 2438717151 articles containing MeSH headings and 4005446 containing keywords. If the initial BM25 algorithm is used, the scores of documents containing chemical words and keywords will be too large, which may affect the final ranking results. In this section, we use an improved BM25 algorithm to calculate 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 as

$$dwl = dcl + dml + dkl \tag{7}$$

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.

Define the IDF value of the expanded word that appears in document $d's$ word list:

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

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 term frequency value of the word list is as follows:

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

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 as follows:

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

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.

Score of co-word

Co-word analysis uses 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 paper, co-word analysis is introduced into the article scoring model: if a disease and a gene co-occur across an abstract, chemical list, MeSH heading and keyword list, as shown in figure 2, or co-occur within any abstract, chemical list, MeSH heading or keyword list, as shown in figure 3, the score of the article will be increased.

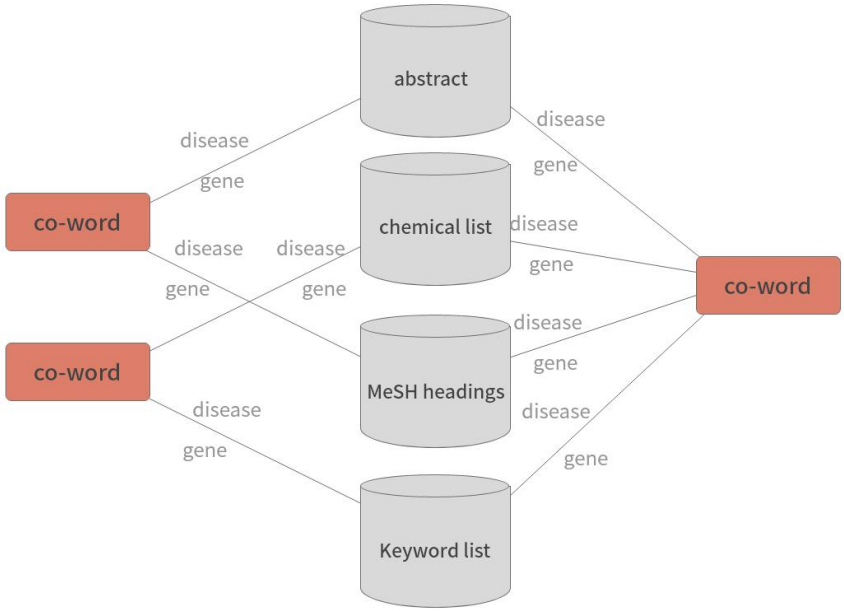

Figure 2 Cross co-word of abstract, chemical list, MeSH headings and keyword list

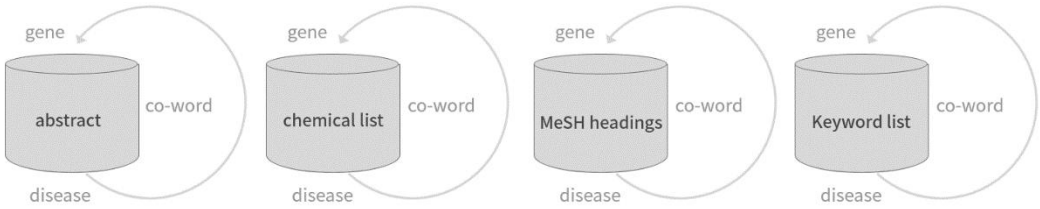

Figure 3 Co-word of abstract, chemical list, MeSH headings and keyword list

We use the IDF value as a co-word score to distinguish the importance of a gene, whose formula is defined as follows:

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

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

The co-word score is defined as follows:

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

where n represents the number of genes that co-word with a disease in Query Q, and g_i represents the morpheme of each gene in Q.

Retrieval model

We use the composite score as the final score for a document d under Query Q, and the specific formula is described as follows:

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

The system architecture for biomedical article retrieval is shown in figure 4.

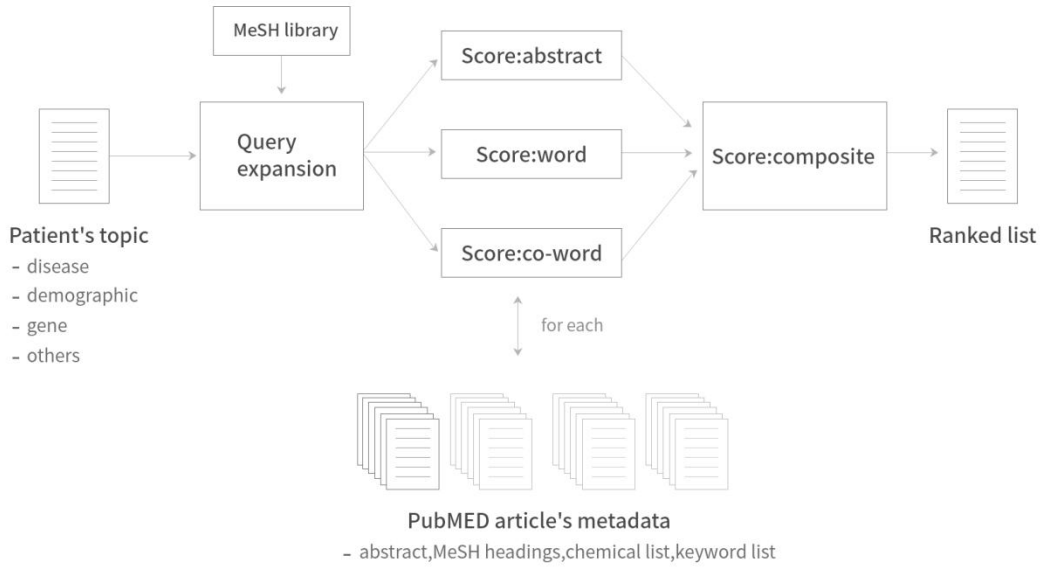

Figure 4 System architecture for biomedical article retrieval

Parameter Optimization

There are 6 parameters in the method proposed above: k_1, k_2, k_3, b_1, b_2 and α. Choosing better parameters will improve the retrieval results. Many optimization algorithms are used for function optimization,such as Genetic Algorithm (GA) [25], Simulated Annealing(SA) algorithm [26],Ant Colony(AC) algorithm [27].With the continuous development of intelligence optimization algorithms , many new swarm intelligence optimization algorithms have emerged during the recent years, such as the Cuckoo Search(CS) algorithm [28], Glow worm Swarm Optimization(GSO) algorithm [29], Particle Swarm Optimization(PSO) algorithm [30] and so on. The swarm intelligence optimization algorithm has been widely used for solving function optimization.

Cuckoo search algorithm

Cuckoo Search algorithm (CS) is a novel swarm intelligence optimization algorithm proposed by yang [28] in 2009. Maribel [31] proved that the cuckoo algorithm is more efficient than the genetic algorithm. The CS algorithm establishes idealized rules as follows:

(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.

The cuckoo finds the nest and updates the position according to the above idealized rules. The position update formula is as follows:

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

In the formula, T stands for step size and T>0, $\oplus$ expresses a point-to-point multiplication, Levy(λ) is the search path and obey Levy distribution [32-33].The pseudo-code of the algorithm [28] as described below:

Algorithm 1 Cuckoo Search algorithm

1: **Begin**
2: Objective function,f(x), $X = (x_1,..,x_d)^T$ d is the dimension of the problem
3: Generate initial population of n host nests

$X_i(i = 1, 2, ..., n)$
4: **While (t <Max Generation) or (** stop criterion)
5: Get a cuckoo randomly by Lévy flights evaluate its quality/fitness $f(x_i)$
6: Select a nest among n (say, j) randomly.
7: **If** $f(x_i) > f(x_j)$
8: replace the nest j with a 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**

Objective function

In table information retrieval, RR represents the relevant document retrieved and RN represents the irrelevant documents retrieved. The calculation formula for Precision is as follows:

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

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

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

Where $P@10(t)$ represents the $P@10$ value of the tth topic in the n topics.

NDCG[70] is a commonly used index to evaluate the quality of ranking in information retrieval .Let ϑ denote a relevance grade and $gain(\vartheta)$ the gain as sociated with ϑ. Also, 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 $g_i = gain(\vartheta)$ if the relevance grade of the document in rank i is ϑ. Then, the nDCG value for this system can be computed as:

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

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 tth topic in the n topics.

Algorithm flow

Because k_2 is a fixed value($k_2 = 1$), we use k_1, k_3, b_1, b_2, α as input parameter. Firstly, the algorithm generates the initial population and set max generation or stop criterion. If the number of iterations reaches the max generation or the stop criterion is met, the algorithm stops and outputs the optimal solution. Otherwise, the algorithm will perform a series of optimization operations based on the value of the objective function. This article uses $Avg_{P@10} + Avg_{nDCG}$ as the objective function, and uses the 2017 precision medicine data set as the training data set to optimize the parameters. The algorithm flow is shown below:

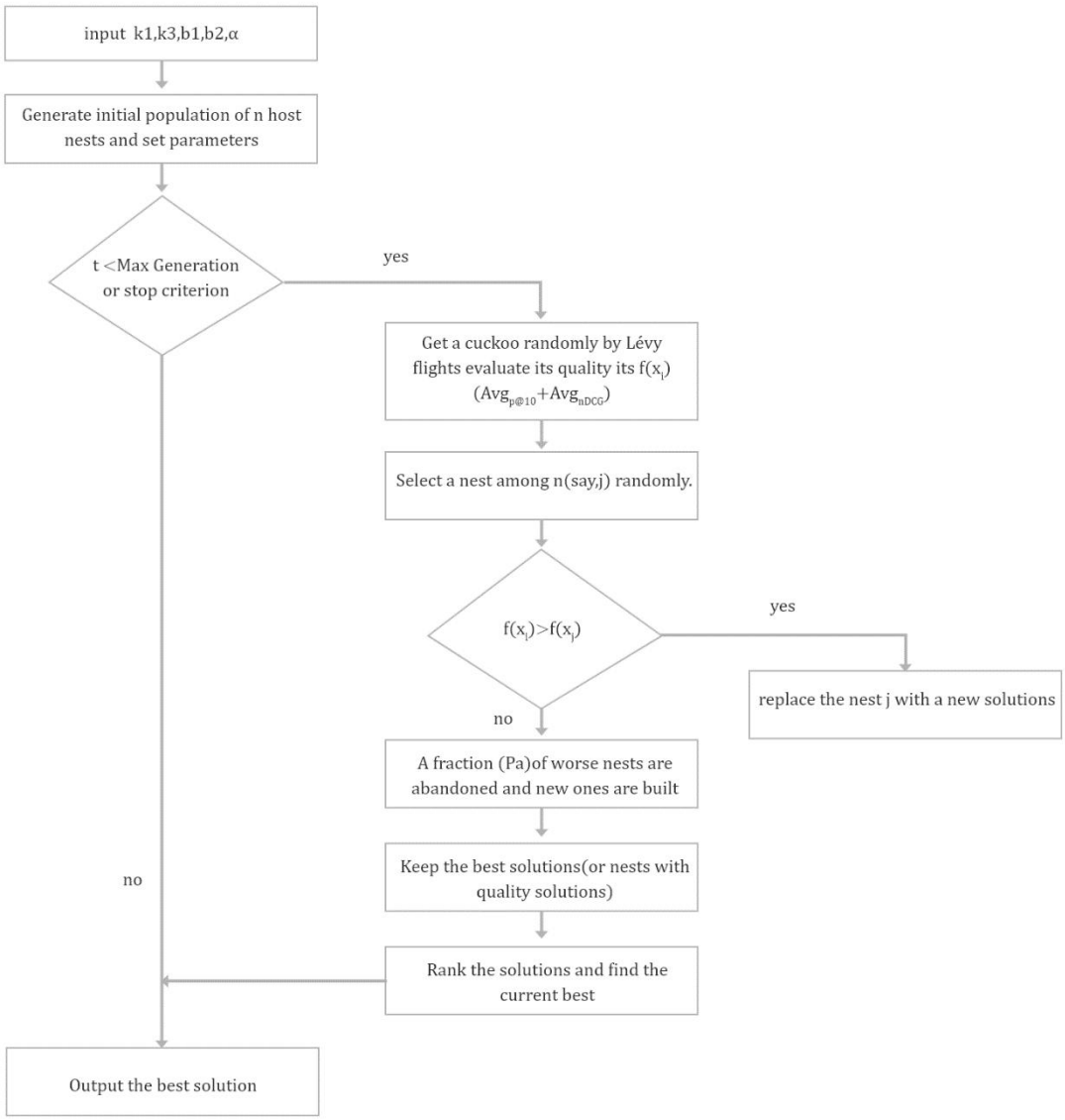

Figure 5 Algorithm flowchart

Experiments

Dataset

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-treceval-abstracts-2018-v2.txt); 2019 TREC Precision Medicine evaluation results are obtained from the website(https://trec.nist.gov/data/precmed/qrels-treceval-abstracts.2019.txt).Because of 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 experimental code can be obtained from my github(https://github.com/Bruce-V/CS-BM25)

Algorithm parameter setting

The parameter values of the algorithm in this paper are shown in table 5

Table 5 Cuckoo search algorithm parameter setting

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)

Experimental results

Table 6 Result of 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

In table 6, normal represents empirical parameters; CS represents the parameters trained on the 2017 data set using the cuckoo algorithm We choose the 1000 documents with the highest score as the result of the retrieval model.

From the comparison data of 3 years, we can see that the optimized parameters are better than empirical parameter. For a retrieval system, users want related documents to appear earlier , so infNDCG and P@10 are important indicators for evaluating information retrieval. We can see from the parameters which optimized based on NDCG and P@10 will all raise the weight of the word list. Obviously, the word list contains extended information about age, gender and genes, which is very important for distinguishing relevant literature from non-related literature. In summary, we can use different parameters to meet different user needs.

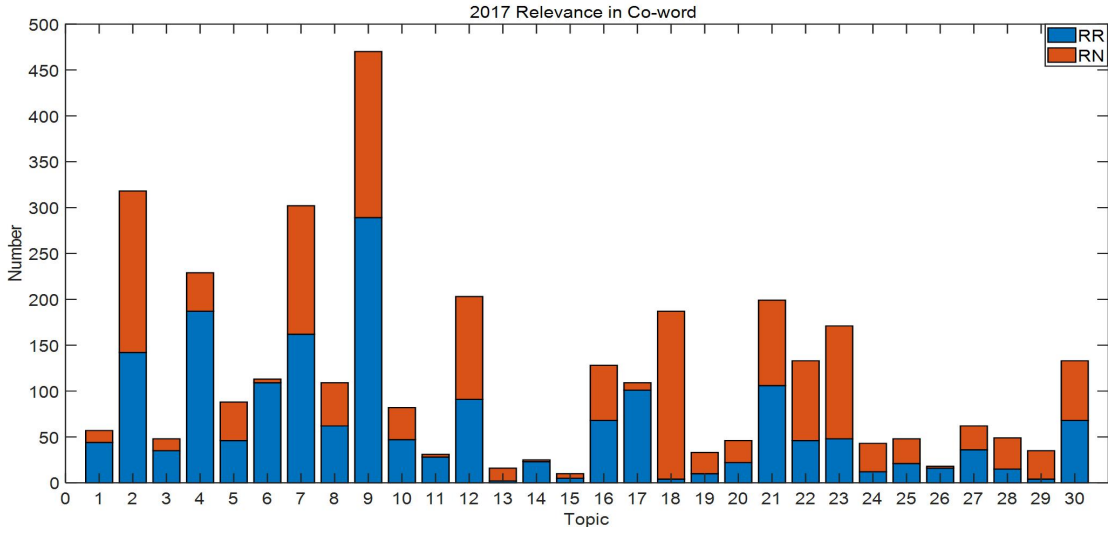

Figure 6 2017 TREC precision medicine relevance in co-word biomedical articles

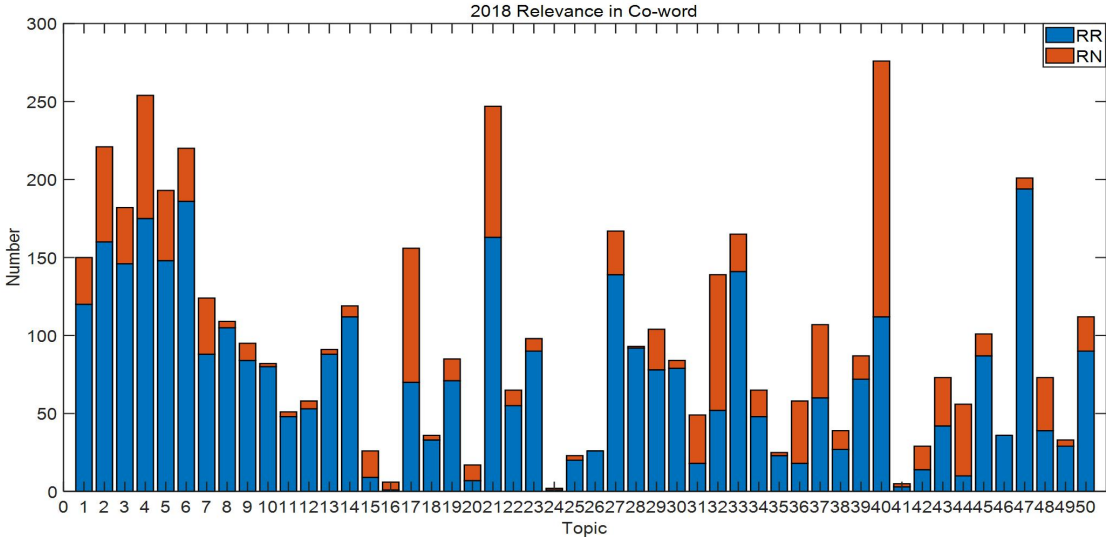

Figure 7 2018 TREC precision medicine relevance in co-word biomedical articles

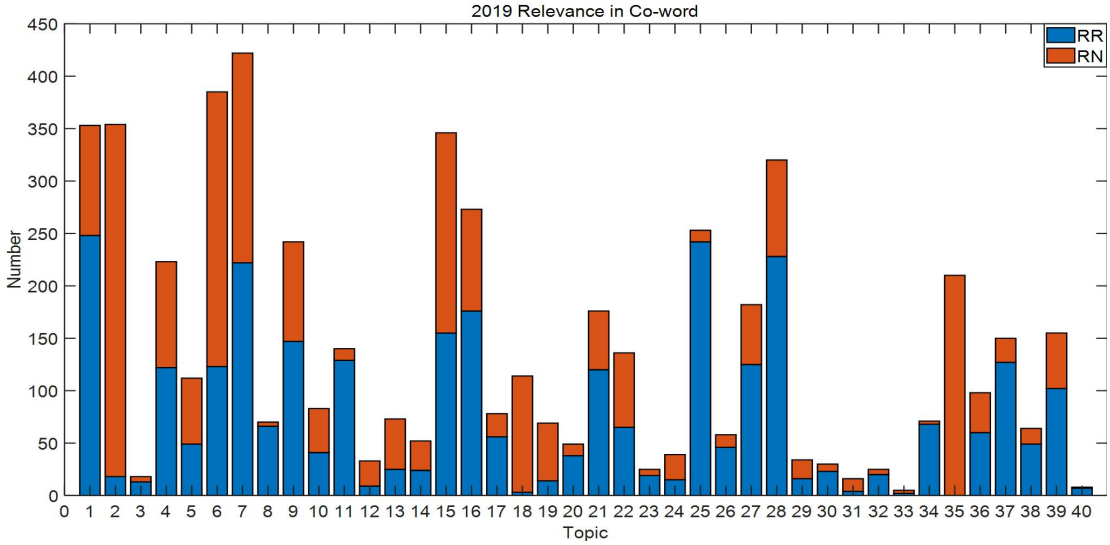

Figure 8 2019 TREC precision medicine relevance in co-word biomedical articles

As seen in figure 6、7 and 8, RR represents relevance in co-word documents; RN represents all relevance except RR. There are many relevant documents containing both a disease and a gene. We define the average relevant document coverage rate as follows:

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

The average coverage rate of 30 topics in 2017 is 52.9%、50 topics in 2018 is 74.13% and 40 topics in 2019 is 54.4%. This result shows that co-word analysis has a better influence on the retrieval of relevant documents, which greatly reduces the search scope.

In information retrieval, NR represents the relevant document not retrieved; NN represents irrelevant documents not retrieved; RR represents the relevant document retrieved; and RN represents the irrelevant documents retrieved. The calculation formula for Precision is as follows:

$$Precision = \frac{RR}{(RR+RN)} \quad (20)$$

The calculation formula for Recall is as follows:

$$Recall = \frac{RR}{(RR+NR)} \quad (21)$$

The calculation formula for F1-score is as follows:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (22)$$

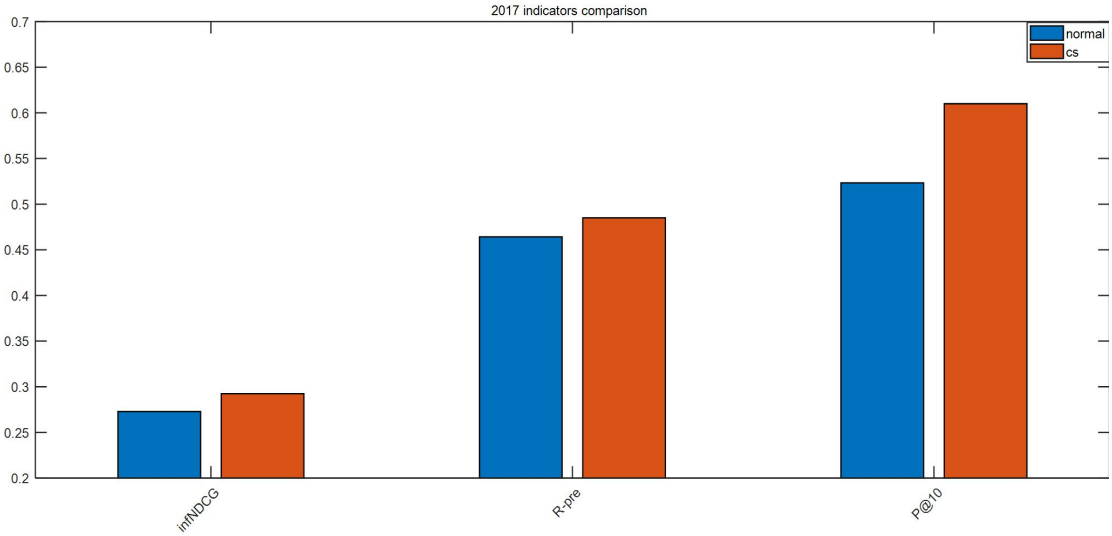

Figure 9 2017 TREC precision medicine indicators comparison

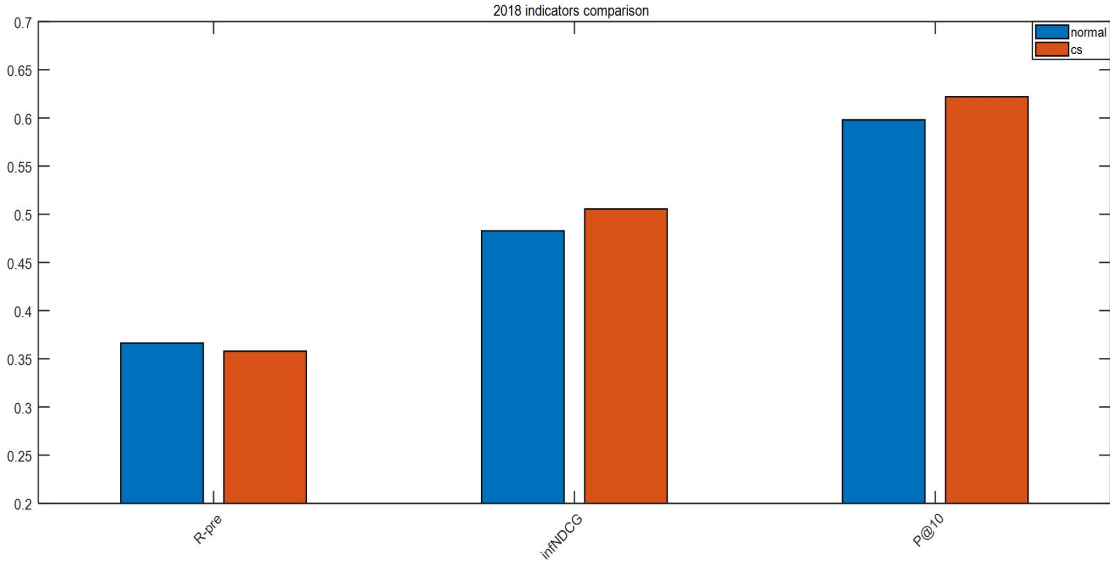

Figure 10 2018 TREC precision medicine indicators comparison

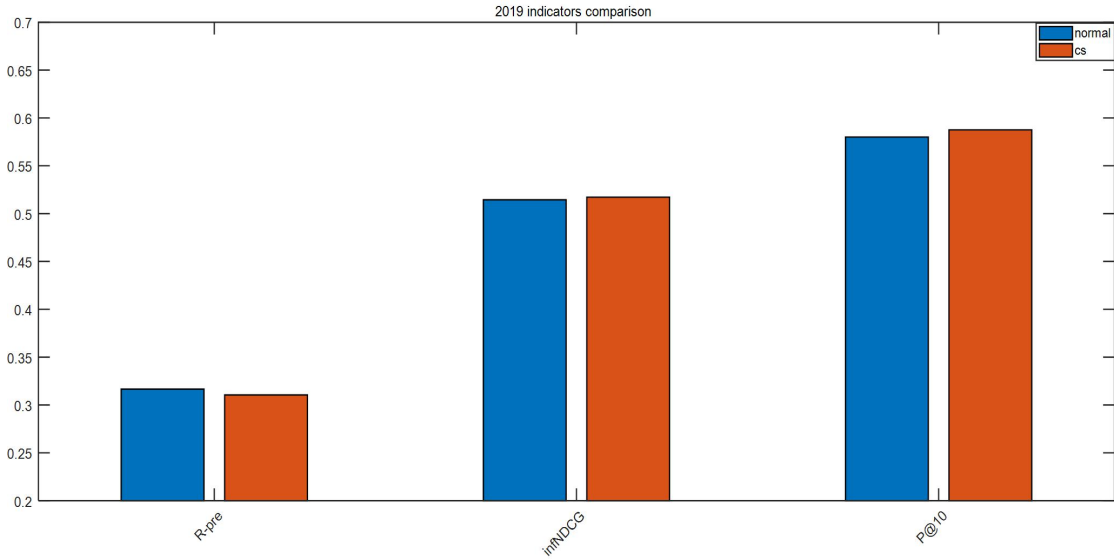

Figure 11 2019 TREC precision medicine indicators comparison

As seen in figure 9、10 and 11, the optimized parameters are better than the empirical parameters on P@10 and infNDCG. Because we used the 2017 precision medicine dataset as the training set, the optimization effect is the most obvious on the 2017 precision medicine dataset. On the test data in 2018 and 2019, we can find that P@10 and infNDCG have improved, but R-pre has declined. This is because the objective function we adopted has improved the ranking of most relevant documents. In fact, in the retrieval system, the precision and the recall are inversely proportional. In the retrieval of biomedical article, we are more concerned about the precision rate to facilitate doctors' scientific decision-making.

Experimental comparison

Compared with other literature that also selected 2017 、2018 and 2019 TREC Precision Medicine, 3 indicators (infNDCG, R-prec and P@10) were selected for comparison. The experimental results are shown in Table 7、8 and 9

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
Our 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
KlickLabs[41]	0.4432	0.2870	0.5400
UCAS[42]	0.5580	0.3654	0.5980
UTDHLTRI[43]	0.4797	0.2870	0.6160
Our 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
CincyMedIR[48]	0.4801	0.3111	0.5675
CSIROmed[49]	0.4766	0.3165	0.5825
Our Method	0.5172	0.3105	0.5875

We selected TREC datasets for 3 years to verify our experimental results. These methods all used the BM25 algorithm or an improved BM25 algorithm. In the 2017 experiments, our algorithm improved significantly in 3 indicators and the algorithm had obtained the best results among the three indicators. In the 2018 experiments, our algorithm was better than other similar algorithms on the P@10, and ranked second on infNDCG and R-pre. In the 2019 experiments, our algorithm was better than other similar algorithms on the P@10 and infNDCG, and ranked third on R-pre.

Conclusion

This article presents a method based on co-word analysis to retrieve biomedical articles. We improved the BM25 algorithm and used it to calculate the score of expanded words and combined the co-word score with the gene appearance weight. We use cuckoo algorithm to optimize parameters on the evaluation function of P @ 10 and NDCG, From the optimization results, and it can be found that increasing the score weight of word list can effectively improve the ranking of related documents.

The paper discusses the influence of different parameters on the retrieval algorithm, and summarizes the parameters to meet different retrieval needs in the future. In theory, the supervised machine learning algorithm is better in results. Although the algorithm in this paper is based on the optimization of results, it summarizes the general rules for improving the parameters of the BM25 algorithm and verifies it through a large number of experiments. Because query expansion is simple in this article, our next work will consider using more linked data to enrich the topic data.

Code availability

All the experimental code can be obtained from my github(https://github.com/Bruce-V/CS-BM25)

Consent for publication

This article uses 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-treceval-abstracts-2018-v2.txt); 2019 TREC Precision Medicine evaluation results are obtained from the website(https://trec.nist.gov/data/precmed/qrels-treceval-abstracts.2019.txt)

Ethics approval

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 in relation to patient or subject protection, or animal experimentation. I testify to the accuracy of the above on behalf of all the authors.

Conflicts of interest

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

Authors' contributions

This article was independently completed by zicheng zhang. All authors read and approved the final manuscript.

Acknowledgements

We appreciated wei cao for data collection

References

1. M.S. Simpson, E.M. Voorhees, W. Hersh, Overview of the TREC 2014 Clinical Decision Support Track, in: Proceedings of Text Retrieval Conference (TREC), 2014.
2. K. Roberts, M.S. Simpson, E.M. Voorhees, W.R. Hersh, Overview of the TREC 2015 clinical decision support track, in: Proceedings of Text Retrieval Conference (TREC), 2015.
3. K. Roberts, D. Demner-Fushman, E.M. Voorhees, W.R. Hersh, Overview of the TREC 2016 clinical decision support track, in: Proceedings of Text Retrieval Conference (TREC), 2016.
4. K. Roberts, D. Demner-Fushman, E.M. Voorhees, W.R. Hersh, S. Bedrick, A.J. Lazar, S. Pant, Overview of the TREC 2017 precision medicine track, in: Proceedings of Text Retrieval Conference (TREC), 2017.
5. K. Roberts, D. Demner-Fushman, E.M. Voorhees, W.R. Hersh, S. Bedrick, A.J. Lazar, Overview of the TREC 2018 precision medicine track, in: Proceedings of Text Retrieval Conference (TREC), 2018.
6. K. Roberts, D. Demner-Fushman, E.M. Voorhees, W.R. Hersh, S. Bedrick, A.J. Lazar, Overview of the TREC 2019 precision medicine track, in: Proceedings of Text Retrieval Conference (TREC), 2019.
7. Francis S Collins, Harold Varmus. A New Initiative on Precision Medicine. New England Journal of Medicine, 2015, 372(9):793-795.
8. S.E.Robertson,S.Walker,M.Hancock-Beaulieu,M.Gatford,and A.Payne,"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,1994:222-231
10. Joachims T.Optimizing search engines using clickthrough data.Proceedings of the eight ACM SIGKDD international conference on Knowledge discovery and data mining.ACM,2002:133-142
11. Freund Y,lyer R,Schapire R E,et al.An efficient boosting algorithm for combining preferences.Journal of machine learning research,2003,4(9):933-969.
12. Cao Z,Qin T,Liu T Y,et al.Learning to rank:from pairwise approach to listwise approach.Proceedings of the 24th international conference on Machine learning.ACM,2007:129-136
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,2007:391-398
14. Burges C J C.From ranknet to lambdarank to lambdamart:An overview.Learning,2010,11(23-581):81
15. Singh J, Prasad M, Prasad O K, et al. A novel fuzzy logic model for pseudo-relevance feedback-based query expansion. International Journal of Fuzzy Systems, 2016, 18(6): 980-989.
16. Keikha A, Ensan F, Bagheri E. Query expansion using pseudo rel-evance feedback on Wikipedia. Journal of Intelligent Information Systems, 2018, 50(3): 455-478.

17. Almasri M, Berrut C, Chevallet J P. A comparison of deep learn-ing based query expansion with pseudo-relevance feedback and mutual information. Proceedings of European Conference on Information Retrieval . Padua: ECIR Press, 2016: 709-715.
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, et al. Probabilistic query expansion using querylogs.Proceedings of the 11th international conference on World Wide Web.ACM, 2002: 325-332
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, 1997: 485.
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, 2001: 17.
22. Li Shuqing,Sun Ying,Soergel Dagobert. Automatic Decision Support for Clinical Diagnostic Literature Using Link Analysis in a Weighted Keyword Network. Journal of medical systems,2017,42(2).
23. Saeid Balaneshinkordan,Alexander Kotov. 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,2019,98.
24. Monika Kastner, Nancy L. Wilczynski, Cindy Walker-Dilks, Kathleen Ann McKibbon, and Brian Haynes. 2006. Age-specific search strategies for medline. Journal of Medical Internet Research 8, 4 (2006), 1–10.
25. Holland JH ,Adaptation in Natural and Artificial Systems . Ann Arbor , Michigan ,USA:The University of Michigan Press,1975.
26. Kirkpatrick S, Gelatt Jr CD and Vecchi M P, “Optimization by Simulated Annealing,” Science, vol. 220, no. 4598, pp. 671 -680，1983.
27. Dorigo M and Gambardella LM, “A Study of Some Properties of Ant-Q,” in Proc of the 44th Int 'l Conf on Parallet Problem Solving from Nature,1996,pp. 656-665.
28. Yang XS and Deb S, “Cuckoo search via le'vy flights,” in World congress on Nature & biologically inspired computing, 2009, pp. 210－214.
29. Krishnand KN and Ghose D, “Detection of multiple source location using a glowworm metaphor with applications to collective robotics,” In Proc of IEEE Swarm Intelligence Symposium, 2005 , pp. 84-91.
30. Kenney J and Eberhart R, “Particle swarm optimization,” in Proceedings of IEEE Conference on Neural Networks,1995.
31. Maribel Guerrero, Oscar Castillo and Mario García Valdez, “Cuckoo Search via Lévy Flights and a Comparison with Genetic Algorithms,” Fuzzy Logic Augmentation of Nature-Inspired Optimization Metaheuristics, vol. 574, pp.91-103,2015.

32. Pavlyukevich I, “L ' evy flights, non-local search and simulated annealing,” Computational Physics, vol. 226,pp.1830-1844,2007
33. Pavlyukevich I, “Cooling down L 'evy flights, ” J. Phys.A: Math. Theor., vol. 40, pp. 12299-12313,2007.
34. Yanshan Wang,Ravikumar Komandur-Elayavilli,Majid Rastegar-Mojarad.Leveraging both Structured and Unstructured Data for Precision Information Retrieval.Proceedings of Text Retrieval Conference (TREC), 2017.
35. Canjia Li,Ben He,Yingfei Sun.UCAS at TREC-2017 Precision Medicine Track Proceedings of Text Retrieval Conference (TREC), 2017.
36. Seung-Hyeon Jo and Kyung-Soon Lee.CBNU at TREC 2017 Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2017.
37. Yue Wang and Hui Fang.Combining Term-based and Concept-based Representation for Clinical Retrieval.Proceedings of Text Retrieval Conference (TREC), 2017.
38. Yuan ling,Sadid A.Hasan,Michele Filannino.A Hybrid Approach to Precision Medicine-related Biomedical Article Retrieval and Clinical Trial Matching.Proceedings of Text Retrieval Conference (TREC), 2017.
39. Jiho Noh and Ramakanth Kavuluru.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. Prakrit Baruah,Riya Dulepet,Kyle Qian.Brown University at TREC Precision Medicine 2018.Proceedings of Text Retrieval Conference (TREC), 2018.
41. Lediona Nishani,Maheedhar Kolla,Gaurav Baruah.KlickLabs at TREC 2018 Precision Medicine track.Proceedings of Text Retrieval Conference (TREC), 2018.
42. Zhi Zheng,Canjia Li,Ben He.UCAS at TREC-2018 Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2018.
43. Stuart J.Taylor,Travis R.Goodwin,Sanda M.Harabagiu.UTD HLTRI at TREC 2018:Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2018.
44. Seung-Hyeon Jo and Kyung-Soon Lee.CBNU at TREC 2019 Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2019.
45. Qi Zheng,Yong Li,Jiaying Hu.ECNU-ICA team at TREC 2019 Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2019.
46. Giorgio Maria Di Nunzio,Stefano Marchesin,Maristella Agosti.Exploring how to Combine Query Reformulations for Precision Medicine.Proceedings of Text Retrieval Conference (TREC), 2019.
47. Artur Cie´slewicz,Jakub Dutkiewicz,Czes law J¸edrzejek.Poznan Contribution to TREC-PM 2019.Proceedings of Text Retrieval Conference (TREC), 2019.
48. Danny T.Y Wu,PhD,MSI,Wu-Chen Su.Retrieving Scientific Abstracts using Venue-and Concept-based Approaches:CincyMedIR at TREC 2019 Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2019.

49. Maciej Rybinski,Sarvnaz Karimi,Cecile Paris.CSIRO at 2019 TREC Precision Medicine Track.Proceedings of Text Retrieval Conference (TREC), 2019.