IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 23, 2020, accepted October 11, 2020, date of publication November 6, 2020,
date of current version November 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033017

An Algorithm of Query Expansion for Chinese EMR Retrieval by Improving Expansion Term Weights and Retrieval Scores

SONGCHUN YANG[1], XIANGWEN ZHENG[1], XIANGFEI YIN[1,2], HUAJIAN MAO[1], AND DONGSHENG ZHAO[1]

[1]Academy of Military Medical Sciences, Beijing 100850, China
[2]Sansha People's Hospital, Sansha 573199, China

Corresponding author: Dongsheng Zhao (dszhao@bmi.ac.cn)

ABSTRACT Query expansion (QE) has been widely used in electronic medical record (EMR) retrieval for assisted diagnosis and clinical research. However, existing QE algorithms haven't achieved satisfactory performance in Chinese EMR retrieval, and one noticeable problem is that the weights of expansion terms and retrieval scores have unreasonable factors for lack of the solid consideration of clinical needs. Here we propose an algorithm of QE for Chinese EMR retrieval by improving expansion term weights and retrieval scores. First, the weights of expansion terms are assigned with semantic similarities, category weights and co-occurrence frequencies between expansion terms and multiple query terms. Then the retrieval scores calculated by expansion terms are limited to reduce the query drift caused by high-frequency expansion terms. Experiment results show that our method gets a 33.3% increase in the precision at top 10, a 90.4% increase in the recall, and a 13.2% increase in MAP compared with four baselines. It proves that our improvement scheme can ensure the accuracy of expansion term weights and decrease the query drift caused by QE, which substantially improves the performance of Chinese EMR retrieval.

INDEX TERMS Electronic medical record, query expansion, word2Vec, co-occurrence, BM25.

I. INTRODUCTION

Nowadays, the values of medical data have attracted the attention of medical researchers. In order to find specific information from a massive amount of EMR data in a short time, information retrieval (IR) techniques are introduced to EMR systems and improve the efficiency of clinical work [1]. The strategies of IR are divided into database retrieval and full-text search based on the structures of EMRs. For structured data, Structured Query Language (SQL) is often used to get specified information from relational databases based on keywords (query expression) in specific fields. For unstructured data, the algorithms of full-text retrieval, such as query likelihood model and BM25 [2], are used to match relevant documents. In clinical applications, there are still plenty of unstructured EMR documents in EMR systems, although they are expected to be structured by the criterion of the documentation of the medical record. Besides, the knowledge of SQL is difficult for clinical staff to master, so the rational EMR data often are converted to document structures for the convenience of retrieval in many cases [3]. Therefore, full-text retrieval has been widely used in EMR retrieval and how to improve the effectiveness of full-text retrieval has become a hot research issue.

Nevertheless, we can't ignore the differences and challenges of EMR retrieval compared with general IR tasks. Altogether, there are problems of complexity and ambiguity in the medical query [4]. On the one hand, the contents of EMRs are dominated by professional medical terms, so it's complicated for users to search EMR documents with accurate query terms. On the other hand, some EMR documents may contain non-standard medical terms, and users may get incomplete results with submitted query terms. Thus, query expansion (QE) is considered an effective way to deal with the problems. QE algorithms expand original queries with relevant terms, and more relevant documents will be searched, thus increasing the comprehensiveness of retrieval [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

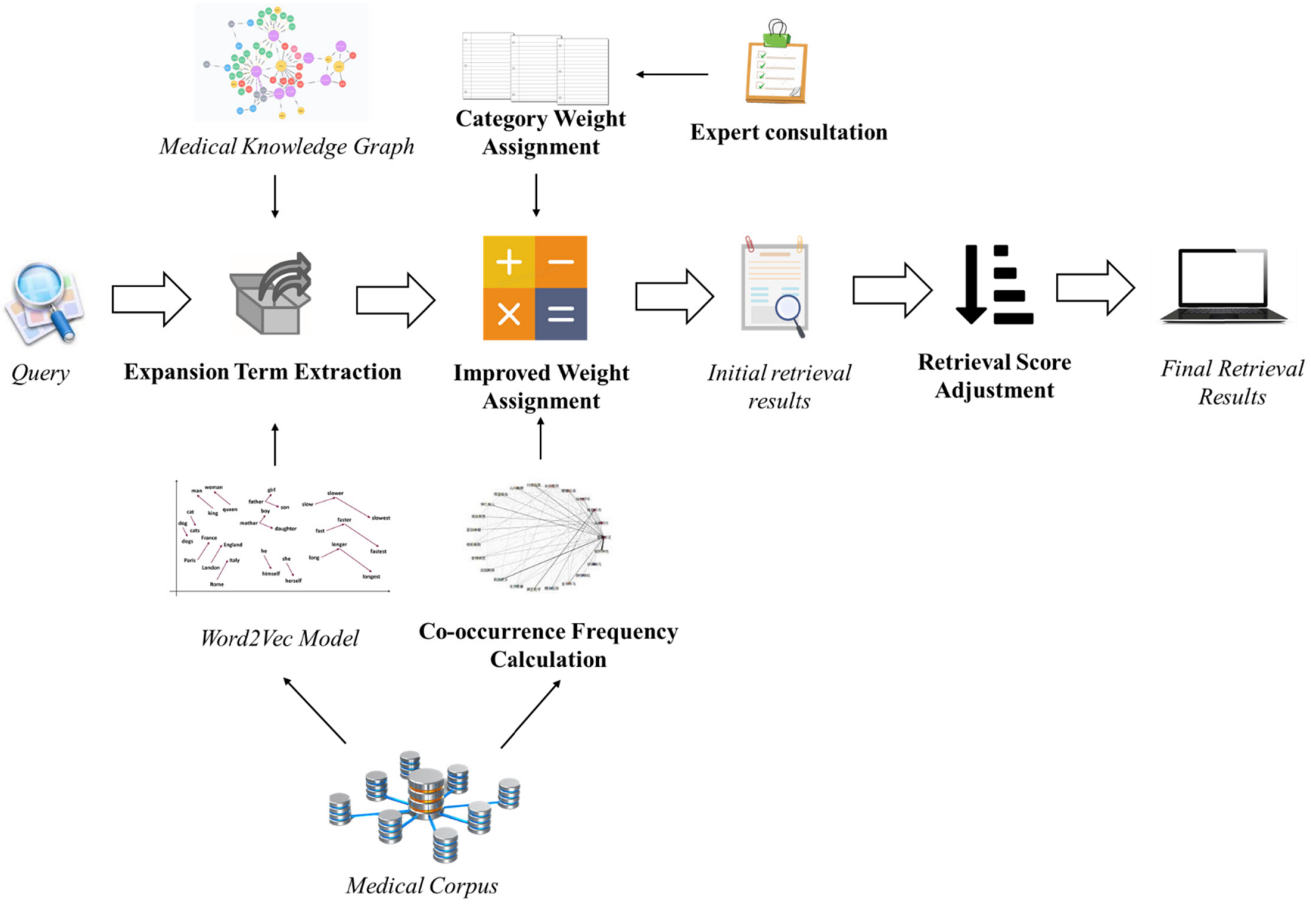

FIGURE 1. The Framework of the improved QE algorithm.

However, improper expansions for Chinese EMR retrieval may lead to the problem of query drift [6], and there are three main problems:

(1) The quality of selected expansion terms and the corresponding weights from existing algorithms haven't been high enough to achieve satisfactory performances. First, the relevance between expansion terms and multiple query terms haven't been considered [7], which may affect the performance of multi-term queries. Second, the selected expansion terms belong to different categories and should have different weights to distinguish them.

(2) The documents with query terms may get lower retrieval scores because some documents with expansion terms may get higher scores due to high term frequencies [8]. Therefore, the retrieval scores of expansion terms should be adjusted to ensure that the EMR documents with query terms and synonyms can be ranked on top.

(3) As the characteristics of Chinese text, there are two main differences in Chinese EMR retrieval compared to other languages [9], [10]. First, Chinese terms are not directly split by separators in sentences, so word segmentation is a crucial task in many tasks. Therefore, not all language models are proper for semantic similarity calculation of Chinese terms due to the characteristic, although they have achieved great performances in other languages. Second, there is a lack of national wide terminology standard in Chinese EMRs, which brings more difficulties to QE. In conclusion, the algorithm's design should consider these unique characteristics to ensure the effectiveness of QE.

To solve the issues, we propose an improved algorithm of QE for Chinese EMR retrieval, and the framework of the improved QE algorithm is shown in Fig. 1, which is composed of four steps. First, the algorithm extracts expansion terms of all query terms from a high-quality Chinese medical knowledge graph in OpenKG[1] inspiriting by the idea in [11]. Then the improved weights of expansion terms are calculated by semantic similarities, category weights and co-occurrence frequencies, and at most 20 expansion terms are reserved for QE. Third, the reserved expansion terms and the corresponding weights are added to the original query, and the reformulated query is retrieved by BM25 [12]. Finally, the retrieval scores of EMR documents are adjusted by promoting the score proportions of query terms and synonyms. Compared to other QE algorithms, our algorithm achieves

[1] http://openkg.cn/dataset/symptom-in-chinese

a balance between precision and recall, by which QE can achieve better performance.

The rest of the paper is presented as follows: In Section II, the related work and the corresponding analysis are described. Then the details of improved weight assignment and retrieval score adjustment are presented in Section III and Section IV. The experiment results and the related analyses are given in Section V. Finally, we conclude the work and propose some future research directions in Section VI.

II. RELATED WORK

Many methods have been proposed to improve the performance of QE in EMR retrieval. According to the survey, the process of QE can be divided into three steps: expansion term extraction, weight assignment and term selection, and query reformulation [5].

In order to confirm the quality of expansion terms, several sources have been used for expansion term extraction, including EMR documents in the retrieval process, hand-built knowledge bases, and query logs [13]. Nguyen and Cao [14] extract candidate terms from the documents returned by the initial search. Aronson and Rindflesch [15] extract expansion terms from UMLS, which contains many health and biomedical vocabularies. Zhu and Carterette [16] leverage PubMed query logs and extract expansion terms from similar queries. Other related works are all devoted to selecting more relevant expansion terms that are helpful to QE. It is noticed that the extracted expansion terms may still contain noisy data, so the subsequent steps are needed to filter the expansion terms of low quality.

Weight assignment and term selection ensure high-quality expansion terms for QE, and several methods have been proposed. Lee *et al.* [17] use regression models to capture linguistic and statistical properties to assign weights. Park and Croft [18] select expansion terms and assign weights by syntactic features extracted from dependency parsing results of verbose queries. Xu *et al.* [19] use supervised term ranking models and assigned weights based on learned term features. Summarizing the existing research, we can see that these methods ignore some unique characteristics of Chinese EMR retrieval, and the training datasets needed by supervised learning models are difficult to construct due to the cost of manual labeling. Most methods only consider retrieval with a single query term, which also leads to incomplete consideration for the terms and weights. Therefore, the quality of expansion terms and the corresponding weights can be improved further.

Query reformulation is the last step of QE. The query is reformulated by expansion terms to achieve better results than the original query. In EMR retrieval, researchers have been proposed several methods based on features of medical language and clinical needs. Zhu and Carterette [20] use Jensen-Shanon divergence for measuring expansion collections from different sources. Qiu and Frei [21] propose a probabilistic query expansion model and calculate the similarities between vectors of the query concept and expansion terms. Jain *et al.* [22] append expansion terms to the original query using Boolean operators and re-weigh expansion terms based on relationships in UMLS. It is observed that existing methods mostly focus on constructing reasonable expanded queries based on clinical needs. However, the existing methods of query reformulation mostly consider the improvement of queries and ignore the process of retrieval. Such may lead the problem that the documents with weak relevance or even no relevance occur on the top of the retrieval results due to expansion terms [23].

In conclusion, existing researches have noticed some problems in QE and take measures to deal with them. However, most existing QE methods for EMR retrieval still ignore some characteristics of Chinese EMR retrieval, and the performance of QE can be improved further. Therefore, our work aims to analyze the weaknesses of QE algorithms for Chinese EMR retrieval and make up for them to promote the effectiveness of retrieval.

III. IMPROVED WEIGHT ASSIGNMENT OF EXPANSION TERMS

Different expansion terms have different relevance to the corresponding query terms, so the corresponding weights should be added to retrieval formulas to distinguish the importance of expansion terms. Now many methods have been proposed for weight calculation, and the semantic similarity calculation can achieve a more satisfactory performance compared to other related methods [4], [24]. However, the direct use of semantic similarities has limitations in the QE for EMR retrieval, and we propose an improved scheme of weight calculation for expansion terms to solve the existing problems. First, a medical corpus with various types of medical documents is constructed for language model training. Second, we test three popular language models and select the most proper model for semantic similarity calculation. Third, the category weights are assigned by expert consultation. Fourth, the co-occurrence frequencies of expansion terms are calculated. Finally, the weights of expansion terms are calculated with the combination of semantic similarities, category weights and co-occurrence frequencies, and at most 20 expansion terms are reserved based on the weights.

A. MEDICAL CORPUS CONSTRUCTION

To calculate the co-occurrence frequencies and train language models, we construct a medical corpus that contains 966883 medical documents from different sources, and the corpus sizes are shown in Table 1. The corpus contains various types of medical documents to ensure the comprehensiveness of medical information. The details of the medical corpus are described as follows:

- Medical textbooks: 43 sets of medical textbooks for clinical medicine are selected, such as Surgery Textbooks and Textbooks of Internal Medicine.
- Medical science articles from the Internet: 534853 articles are collected from authoritative public health websites.

TABLE 1. Details of Corpus sizes.

Corpus	# of words
Medical textbooks	13.93M
Medical science articles	455.83M
EMRs	456.07M
Clinical guidance and expert consensus	4.31M
Chinese wiki articles	212.81M
Total	1142.95M

TABLE 2. Details of category weights.

Model	Similarity of proper terms	Similarity of relevant terms	Similarity of irrelevant terms
CBOW	0.573	0.301	0.024
Skip-gram	0.550	0.342	0.161
GloVe	0.426	0.350	0.144
BERT	0.887	0.841	0.803

• EMRs from hospitals: 16029 EMRs extracted from the Chinese Stroke Data Center (CSDC) are used for training. CSDC has collected stroke patients' EMR since 2011 and played a crucial part in the national stroke prevention program and stroke clinical research. It is declared that the EMRs are only used for model training and algorithm evaluation. The patients' privacy data are removed before use, and the benefits of the hospital won't be harmed.

• Clinical guidance and expert consensus: 155 documents of clinical guidance and expert consensus of various diseases are collected from public online document platforms.

• Chinese wiki articles[2]: the public Chinese wiki data are added to the training corpus for the effectiveness of Word2Vec model.

B. MODEL SELECTION OF SEMANTIC SIMILARITY CALCULATION

In order to select a proper model, we test three language models—Word2Vec, GloVe, and BERT [25]–[27], which are all widely used in semantic similarity calculation, and we further compare CBOW and Skip-Gram [28]. CBOW, Skip-Gram, and Glove are trained with the medical corpus above, and an existing Chinese BERT Model[3] is selected due to experimental conditions [29]. As for the test procedure, 10 medical terms are selected as the benchmarks, and then 10 proper terms for QE, 10 improper but relevant terms that are improper but relevant, and 10 irrelevant terms are selected for every benchmark term. Finally, the average similarities of every category are calculated based on the four models, respectively, and the results are shown in Table 2.

It is noticed that CBOW and Skip-Gram can significantly reflect the difference of terms belonging to different kinds, and CBOW performs a little better than Skip-Gram. The performance of Glove is worse than CBOW and Skip-gram, especially the similarity values of proper terms. Surprisingly, the similarity values calculated by BERT are very close, which can't distinguish the terms of different kinds. By further analysis, the train of Chinese BERT is based on single characters rather than words, so the model can achieve good results in the similarity calculation of phrases or articles with supervised learning. However, it can't work well in the similarity calculation of Chinese terms with unsupervised learning. Thus, CBOW is selected as the final language model for semantic similarity calculation.

[2]http://download.wikipedia.com/zhwiki

[3]https://github.com/ymcui/Chinese-BERT-wwm

TABLE 3. Details of category weights.

Category	Weight
Synonym	0.96
Hyponym	0.60
Hypernym	0.12
Related Disease	0.11
Related Symptom	0.89
Related Drug	0.44

C. CATEGORY WEIGHT ASSIGNMENT

As mentioned above, the category weights should be assigned to enhance the effectiveness of expansion term weights. As the category weights are subjective and depend on clinical needs for EMR retrieval, we introduce the idea of expert consultation to determine the category weights. Medical experts can quantificationally estimate the importance of categories, and the categories with greater clinical importance can get higher weights.

We use the method of Delphi for reference [30] and distribute questionnaires to investigate the opinions of medical experts. We select the categories of synonyms, hyponyms, hypernyms, related diseases, related symptoms, related examinations, related operations, and related drugs, which are all common categories of expansion terms. In questionnaires, the scores are set as 3, 1, 0, −1, −3 for every category, which represent the importance of categories from high to low. First, three medical experts are invited to evaluate every category based on clinical needs for actual applications. Then we perform statistical analysis of the questionnaire results and distribute the questionnaires with first-round score statistics for reference. Finally, 20 valid questionnaires are collected, and the final category weights are assigned with the proportion of total scores given by experts to the full mark, which are shown in Table 3.

It is noticed that the weights of related examinations and related operations are not shown in Table 1 because the corresponding calculated weights are negative, which indicates that the two categories are not proper and should not be selected for QE.

D. CO-OCCURRENCE FREQUENCY CALCULATION

The concept of word co-occurrence was created in 1959 [31] and has been widely used for theme identification [32]. We assume that the expansion terms with high co-occurrence frequencies are thematically similar to the query and likely to appear together with the query terms in the same document. These terms can be considered helpful in finding more relevant EMR documents to the query. Thus, co-occurrence frequencies with multiple query terms can be combined with semantic similarities and category weights to improve the quality of weights.

The calculation formula of co-occurrence frequencies is shown as follow:

$$co\,(t, Q) = \frac{\sum_{Q} \frac{n_d(q \cap t)}{n_d(q \cup t)}}{co_{\max(T)}} \tag{1}$$

where t represents the expansion term, Q represents the set of query terms, T represents the set of expansion terms of Q, q represents a single query term in Q, $co\_max(T)$ is the maximum value of co-occurrence frequencies. In the formula, $\sum_{Q} \frac{n_d(q \cap t)}{n_d(q \cup t)}$ is the sum of co-occurrence frequencies with all query terms, and then it is divided by the maximum co-occurrence value of all expansion terms, by which the final value can avoid being too small.

Besides, the indexing technique is needed to ensure real-time response in clinical applications. Due to the huge size of the medical corpus, the retrieval system is designed to extract common medical terms from EMRs and calculate co-occurrence frequencies between every two terms during free time. Then the frequencies are stored in databases, and the system can calculate the adjusted weights of common expansion terms in real-time. If the co-occurrence frequency cannot be found in the database, the system calculates the co-occurrence frequency in EMR documents for timesaver. After the retrieval, the system re-calculates the corresponding frequency in the corpus and adds it to the database in free time.

E. FINAL WEIGHT CALCULATION AND EXPANSION TERM RESERVATION

Based on the above studies, a formula is designed to calculate the final weights, which is shown as follow:

$$w\,(t) = \sqrt{c\,(t) * \frac{sim\,(q_t, t) + co\,(t, Q)}{2}} \tag{2}$$

where $sim\,(q_t, t)$ is the semantic similarity between q_t and t, which is calculated by the CBOW model, and $c\,(t)$ is the category weight of t. In the formula, the average value is calculated to synthesize the contribution of semantic similarity and co-occurrence frequency to the weight, and the category weight is multiplied to add the importance of the corresponding category into the weight. Finally, the multiplication values are adjusted by square-root to increase the numerical values.

After the calculation, at most 20 expansion terms are reserved according to the previous study [33]. There is a worth noticing point that synonyms are semantically closer to the query than other expansion terms [34]. If the expansion terms are filtered as a whole, some synonyms may be removed, and the quality of expansion terms are decreased. Thus, we divide the expansion terms into synonyms and other expansion terms in the process of reservation. At most 20 synonyms are reserved according to the rank of weights. If the number of synonyms is less than 20, the remaining terms are extracted from other expansion terms based on the weights.

IV. RETRIEVAL SCORE ADJUSTMENT

After the weight assignment, the expansion terms and the corresponding weights are added to the query. However, most IR algorithms are based on the frequencies of query terms in documents. If used for QE, such algorithms lead to a problem that documents with high-frequency expansion terms may be ranked on top or even ahead of the documents with query terms. Therefore, the retrieval scores of expansion terms should be adjusted for the performance of QE. In this section, we first introduce the details of the adjustment scheme. Then we propose an improved algorithm of EMR retrieval based on the adjustment scheme.

A. DESIGN OF ADJUSTMENT SCHEME

Not all expansion terms cause query drift. As mentioned above, synonyms have the same semantics as the corresponding query terms. Especially if the query contains non-standard medical terms, QE with corresponding standard terms will significantly improve both precision and recall of EMR retrieval. Thus, we should take different strategies for different expansion terms.

The expansion terms are divided into synonyms and other expansion terms, then the synonyms are combined with the query terms. For a convenient description in the paper, the synonyms and query terms are denoted as QS terms, and other expansion terms are denoted as Oth terms. With the analysis of users' retrieval intents, the importance of QS terms should be higher than Oth terms, so the score proportions of QS terms should also be adjusted higher. The adjustment scheme contains three factors:

(1) If the documents contain no Oth terms, the retrieval score is unadjusted.

(2) If the documents contain only Oth terms, the retrieval scores should be adjusted lower than the scores of QS terms in all retrieved documents.

(3) The retrieval scores of Oth terms should be lower than the scores of QS terms.

B. DESIGN OF IMPROVED RETRIEVAL ALGORITHM

As mentioned in the description of the algorithm framework, BM25 is selected as the baseline algorithm for our improvement. The formulas are described as follows:

$$score\,(d, Q) = \sum_{i=1}^{n} IDF\,(q_i) * R(q_i|d) \tag{3}$$

$$IDF\,(q_i) = log\frac{N - n\,(q_i) + 0.5}{n\,(q_i) + 0.5} \tag{4}$$

$$R(q_i \mid \mathrm{d}) = \frac{f_i(k_1+1)}{f_i+K} = \frac{f_i(k_1+1)}{f_i+k_1\left(1-b+b\frac{dl}{avgdl}\right)} \quad (5)$$

where $IDF(q_i)$ isthe inverse frequency of q_i, $R(q_i|\mathrm{d})$ is the retrieval score of q_i in d, N is the total number of documents in the index, $n(q_i)$ is the number of documents that contain q_i, f_i is the term frequency of q_i in d, k_1 and b are adjustment factors, dl is the length of d, $avgdl$ is the average length of all documents in the index. k_1 is set as 1.5, and b is set as 0.75 based on experience.

According to the demand of QE and the adjustment scheme, the expansion terms and corresponding weights are added to the calculation of retrieval scores. Then the idea of sigmoid functions [35] is introduced for the score limitation to Oth terms. The formulas of the improved algorithm are shown as follows:

$$score(QS|d) = \sum_{t\in QS} IDF(t) * w(t) * R(t \mid d) \quad (6)$$

$$score(Oth|d) = \sum_{t\in Oth} IDF(t) * w(t) * R(t \mid d) \quad (7)$$

$$H(t) = \begin{cases} score(QS|d) & t > 0 \\ min_{d'\epsilon D}\left(score\left(QS|d'\right)\right) & t = 0 \end{cases} \quad (8)$$

$$S(x) = \begin{cases} 1/1+e^{-x} & x > 0 \\ 0 & x = 0 \end{cases} \quad (9)$$

$$score'(Oth|d) = H(score(QS|d)) * S(score(Oth|d)) \quad (10)$$

$$score(Q+E|d) = score(QS|d) + score'(Oth|d) \quad (11)$$

where $w(t)$ is the weight of t and $w(t) = 1$ for all query terms, $score(QS|d)$ is the retrieval score of QS terms in d, $score(Oth|d)$ is the original retrieval score of Oth terms in d, $H(t)$ is the upper bound value of adjustment and $min_{d'\epsilon D}\left(score\left(QS|d'\right)\right)$ is the minimum of $score\left(QS|d'\right)$ in D, $S(x)$ is the constraint function for adjustment, $score'(Oth|d)$ is the adjusted retrieval score of Oth terms, $score(Q+E|d)$ is the final retrieval score of d.

In the above functions, $S(x)$ is the modified sigmoid function, which is a type of squashing function and can limit the output to a range between 0 and 1. The input is the original retrieval score of Oth terms, and the output is the proportion of score adjustment. As the original score is in $[0, +\infty)$ and the output should be set 0 when the original score is 0, the sigmoid function is transformed accordingly. It is noticed that the output values are in (0.5, 1) when the original score is not 0, which is shown in Fig. 2. Thus, the function ensures that the adjusted scores are limited and not too small meanwhile.

$H(t)$ defines the upper bound values according to the adjustment scheme. The input is the retrieval score of QS terms in a retrieved document, and the function is designed as a piecewise function so that it can return the corresponding upper bound value, which is cooperated with $S(x)$ for adjustment later.

The process of final score calculation is divided into three steps. First, the original scores of QS terms and Oth terms are calculated with the addition of the corresponding weights. Second, the two scores are correspondingly input into $H(t)$ and $S(x)$ for the calculation of adjusted retrieval score of Oth terms. Finally, the adjusted score is added with the retrieval score of QS terms, and the result is the final retrieval score of the retrieved document.

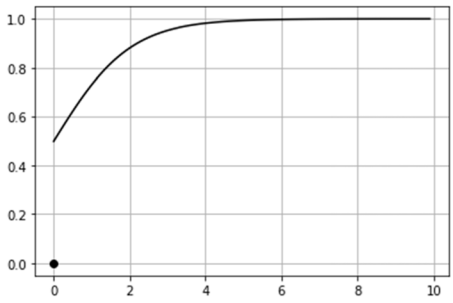

FIGURE 2. The graph of $S(x)$. The function can limit arbitrary input values without the change of relative sizes.

V. EXPERIMENT AND RESULTS

In this section, our algorithm is compared with other popular algorithms of QE. Then we report our evaluation and discuss our findings.

A. DATASETS

First, the stroke patients' EMR documents from CSDC are used for evaluation. Second, a Chinese medical knowledge graph in OpenKG is downloaded for the extraction of expansion terms. The knowledge graph is composed of 135485 entities and 617499 triples, from which various types of expansion terms can be extracted for QE [36]. Finally, the CBOW model trained with the medical corpus is used for semantic similarity calculation. Based on experience, the vector dimension is set as 400, the window size is set as 5, and the word number of negative sampling is set as 10.

B. BASELINES

Our algorithm is compared with four benchmark algorithms. The details are described as follows:

1) CO-OCCURRENCE

The exclusive use of co-occurrence frequencies is widely used in QE. Here we adopt the method of Jain *et al.* [22], by which 20 expansion terms with top co-occurrence frequencies are selected from the medical corpus, and the weights are calculated based on the co-occurrence frequencies, which are shown as follows.

$$co(e,q) = \sum_{q,e\in C} \frac{f(q\cap e)}{f(q\cup e)} \quad (12)$$

$$w(e) = \frac{co(e,q)}{\max(co(e,q))} \quad (13)$$

where $f(q\cap e)$ is the number of documents that contain both the expansion term and the query term, and $f(q\cup e)$ is the

number of documents that contain the expansion term or the query term.

2) SEMANTIC CORRELATION

Here we consider the semantic correlation in medical knowledge graph and adopt the method of ALMasri *et al.* [37] to extract 20 expansion terms with strong semantic correlations. The basis of term selection and weight assignment is shown as follow:

$$SIM\ (q, e) = \frac{|I\ (q) \cap I\ (e)| + |O\ (q) \cap O\ (e)|}{|I\ (q) \cup O\ (q)| + |I\ (e) \cup O\ (e)|} \tag{14}$$

where R^q is the document set with the feedback of the original query, $df\ (t, R^q)$ is the number of documents where e occurs, $idf = \frac{N}{df\ (e,C)}$, where N is the document number in the collection and $df\ (e, C)$ is the document frequency of the expansion term in the collection.

3) SCORING FUNCTION

Here the scores of expansion terms are calculated based on the term frequencies of the original retrieval results. The method of Paik *et al.* [38] is used to select 20 expansion terms and assign the corresponding weights, and the formulas are shown as follows:

$$SC\left(e, R^q\right) = \log_2\left(df\left(e, R^q\right)\right) \times idf\ (e, C) \tag{15}$$

$$w\ (e) = \frac{SC\ (e, R^q)}{\max\left(SC\ (e, R^q)\right)} \tag{16}$$

where R^q is the document set with the feedback of the original query, $df\ (t, R^q)$ is the number of documents where e occurs, $idf = \frac{N}{df\ (e,C)}$, where N is the document number in the collection and $df\ (e, C)$ is the document frequency of the expansion term in the collection.

4) KULLBACK-LEIBLER DIVERGENCE

Kullback-Leibler Divergence (KLD) is a definition to measure the relative entropies and has been used in natural language and speech processing applications. Here the idea of Carpineto *et al.* [39] is used to calculate the weights of expansion terms and select the top 20 expansion terms based on the deformation formula of KLD, which is described as follow:

$$Score_{KLD}\ (t) = \sum_{t \in V} p\ (t \mid D_R) \cdot log \frac{p\ (t \mid D_R)}{p\ (t \mid D_C)} \tag{17}$$

where V represents the sets of expansion terms. $p\ (t \mid D_R)$ is the probability of occurrence of t in the PRF documents. $p\ (t \mid D_C)$ is the probability of occurrence of t in the document collection.

C. EVALUATION PROCESS

First, 10 queries with multiple terms are constructed for experiments based on the needs of clinical diagnosis and clinical scientific research for EMR retrieval. Due to the cost of the manual evaluation, we haven't constructed more queries. As the EMR documents for evaluation are selected from China Stroke Data Center, the 10 queries are carefully designed based on the typical characteristics of stroke patients and can be considered representative for retrieval evaluation. The queries are composed of diseases and symptoms, which clinicians use for assisted diagnosis and clinical research frequently. Besides, some non-standard but frequently-used terms are designed into queries to verify the validity of QE. We confirm that the opinions of clinicians have been fully considered in the design of queries, and all queries meet actual clinical needs. The details are shown in Table 4.

TABLE 4. Details of queries.

	Contents	Length	Number of Non-standard Terms
1	感冒, 高血压, 糖尿病 (cold, hypertension, diabetes)	3	1
2	胃癌, 高血压, 心力衰竭 (gastric cancer, hypertension, cardiac failure)	3	1
3	支气管炎, 鼻炎 (bronchitis, rhinitis)	2	0
4	风心病, 脑血栓 (rheumatic heart disease, cerebral thrombosis)	2	1
5	头昏, 上呼吸道感染, 心肌炎 (dizziness, infection of the upper respiratory tract, myocarditis)	3	1
6	脑梗死, 高血压 (cerebral infarction, hypertension)	2	0
7	慢性肾脏病, 水肿, 乏力 (chronic renal failure, edema, weakness)	3	0
8	糖尿病, 四肢麻木, 多饮, 多尿 (diabetes, numbness of limbs, polydipsia, urorrhagia)	4	0
9	视物模糊, 心虚 (blurred vision, palpitation)	2	1
10	-房颤, 糖尿病 (atrial fibrillation, diabetes)	2	0
Avg		2.6	0.5

Second, the corresponding expansion terms of every query are extracted, and the expansion term weights are calculated by every algorithm. Then the EMR documents are retrieved, and the retrieval results of all the algorithms mentioned above are acquired.

Third, we make questionnaires and invite three experts in medical informatics to evaluate whether the document in the results is relevant or not. In the questionnaires, we set options for every retrieval result, and experts give opinions for results based on clinical needs. Then the golden standards of EMR retrieval are proposed based on the summarization of the opinions of medical experts,

Finally, the precision at top 10 (P@10), recall (R), and mean average precision (MAP) are selected for evaluation. These metrics are all widely to measure the performance of IR algorithms[40]. P@10 measures the accuracy of algorithms, R measures the comprehensiveness, and MAP evaluates the accuracy based on the rank of retrieval results.

D. OVERALL RESULTS

The overall results are shown in Table 5.

TABLE 5. Experimental results on our algorithm and baselines.

Metrics	P@10	R	MAP
Co-occurrence	0.150	0.140	0.159
Semantic Correlation	0.150	0.132	0.150
Scoring Function	0.150	0.142	0.166
KLD	0.130	0.147	0.237
Our Algorithm	**0.200**	**0.280**	**0.311**

TABLE 6. Experimental results in performance analysis.

Metrics	P@10	R	MAP
No strategy	0.100	0.178	0.098
Improved Weight Assignment	0.180	0.211	0.139
Retrieval score Adjustment	0.120	0.202	0.127
Full Algorithm	**0.200**	**0.280**	**0.311**

Our algorithm achieves the best performance in all metrics, which confirms that the algorithm can significantly improve the performance of EMR retrieval. Compared to the best metrics among baselines, our algorithm gets a 33.3% increase in the precision, a 90.4% increase in the recall, and a 13.2% increase in MAP. The dramatic increases of metrics show that the improved weights and adjusted retrieval scores can satisfy the requirement of QE and promote the accuracy and comprehensiveness of EMR retrieval, which is helpful in promoting the effectiveness of clinical applications.

E. PERFORMANCE ANALYSIS OF WEIGHT ADJUSTMENT AND RETRIEVAL SCORE ADJUSTMENT

Here the performances of improved weight assignment and retrieval score adjustment are analyzed further. The basic framework of our algorithm is reserved, and the strategies are gradually added to the framework. Three algorithms are designed for performance analysis, and the details of which are described as follows:

1) THE ALGORITHM WITH NO STRATEGY

The algorithm is designed with no strategy. That is, the expansion terms are selected from the same medical KG, and the corresponding weights are calculated by cosine similarities based on the trained CBOW model with the medical corpus.

2) THE ALGORITHM WITH IMPROVED WEIGHT ASSIGNMENT

The algorithm is designed with improved expansion weights, and the retrieval scores are not adjusted.

3) THE ALGORITHM WITH RETRIEVAL SCORE ADJUSTMENT

The algorithm only adjusts the retrieval scores, and the weights are directly calculated by cosine similarities.

The results are shown in TABLE 6.

The algorithm with no strategy gets the highest recall and the lowest P@10 and MAP compared with baselines, which shows that the extracted expansion terms from medical knowledge graphs by cosine similarities can cover more relevant terms, but the corresponding weights are not proper. The results also show that the single improved weight assignment and the retrieval score adjustment can both promote the performances of QE, and the improved weight assignment gets more superior metrics than retrieval score adjustment, which confirms that the proper weights are important to the quality of QE. Finally, the full algorithm gets better metrics than the other algorithms, which indicates that the combination of improved weight assignment and retrieval score adjustment can improve the performance of QE more significantly without any negative effect on each other.

F. RESULT DISCUSSION

It is first noticed that all baselines get similar values of P@10 and R, and it shows that the selected expansion terms have similar contributions to QE, although the expansion terms are extracted from different sources. Besides, the values of R and MAP of KLD are highest among the four baselines. It shows that the expansion terms from the top retrieved documents are more effective, and the weights calculated by KLD are more proper compared with another three baselines. It also reflects that the method of PRF is helpful for QE of EMR retrieval, and the idea can be integrated into our algorithm in future work.

Our algorithm gets the highest promotion in the recall, which shows that our reservation strategy with the improved expansion term weights is effective for selecting expansion terms with high qualities. Also, the promotions of P@10 and MAP show that the improved weight assignment and retrieval score adjustment ensure the accuracy of EMR retrieval based on the comprehensiveness of retrieval results. Thus, our algorithm decreases the influences of complexity and ambiguity in EMR retrieval, which can help clinical staff find accurate and comprehensive information from EMRs in a short time and has important significance in clinical applications.

By further analysis, the reason for the improved performance can be concluded as follows:

First, the algorithm combines semantic similarities, topic similarities and category weights into weight assignment so that the weights of expansion terms are more comprehensive. Specifically, semantic similarities can ensure that the selected expansion terms are semantically close to the query terms so that the expanded query and the original query have the same meaning. Topic similarities enhance the correlation between the expansion terms and the queries with multiple terms, and the experiment results also show that the consideration of topic similarities can significantly improve the performance of multiple-term retrieval of EMRs. The category weights by expert consultation introduce the medical knowledge into the weight assignment, which increases the clinical significance of weights and adds the rationalities of QE. Therefore, our strategy of weight assignment takes full consideration of the characteristics of EMR retrieval and achieves satisfactory performance.

Second, the algorithm adjusts the retrieval scores based on the characteristics of expansion terms, which makes up for the limitations of a single adjustment of queries. Experiment results show that the single adjustment of retrieval scores can slightly promote the performance of QE, and the combination of improved weights and adjusted retrieval scores can further promote the performance significantly, especially R and MAP, which can prove that our adjustment scheme can help users get useful information in the top retrieval results.

In conclusion, the improvements for QE promote the accuracy and comprehensiveness of EMR retrieval, which has important clinical significance. However, it can't be ignored that there are still weaknesses in our research. On the one hand, some removed expansion terms have strong relevance to the top documents and are helpful for QE. On the other hand, there are plenty of negated, hypothetical, and historical terms in EMRs, and the strategies of term match can't ensure high precision of retrieval. Therefore, the strategy of expansion term classification should be optimized further, and the different semantics in EMRs needs to be considered in the future.

VI. CONCLUSION

In this paper, we propose a QE algorithm based on improved expansion term weights and adjusted retrieval scores. Our algorithm considers the characteristics of Chinese EMR documents and the clinical needs of EMR retrieval, so it has significant improvement in accuracy and comprehensiveness. In the future, our research will further focus on the structures of Chinese EMRs and improve the QE algorithm to help clinical staff to conduct clinical applications better.

ACKNOWLEDGMENT

The authors extend our most sincere thanks to the China Stroke Data Center for sharing the EMRs used in Word2Vec model training and algorithm evaluation. The EMR data are under license in the using process and can't be provided by our research group for the authorization requirement. However, the data can be available under the permission of the China Stroke Data Center with a reasonable request on the website.[4]

The codes, the trained Word2Vec model and the medical corpus without EMRs in the paper are open for researchers and can be downloaded from our repository in GitHub.[5]

REFERENCES

[1] W. Hersh, *Information Retrieval: A Health and Biomedical Perspective*. New York, NY, USA: Springer, 2008.

[2] S. Robertson, C. Van Rijsbergen, P. A. Williams, and R. N. Oddy, *Information Retrieval Research*. London, U.K.: Butterworths, 1981.

[3] D. C. Blair and M. E. Maron, "Full-text information retrieval: Further analysis and clarification," *Inf. Process. Manage.*, vol. 26, no. 3, pp. 437–447, Jan. 1990.

[4] H. Wang, Q. Zhang, and J. Yuan, "Semantically enhanced medical information retrieval system: A tensor factorization based approach," *IEEE Access*, vol. 5, pp. 7584–7593, 2017.

[5] E. N. Efthimiadis, "Query expansion," *Annu. Rev. Inf. Sci. Technol.*, vol. 31, no. 8, pp. 87–121, 1996.

[6] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits, "Predicting query performance by query-drift estimation," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 1–35, May 2012.

[7] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," *Int. J. Semantic Web Inf. Syst.*, vol. 2, no. 3, pp. 55–73, Jul. 2006.

[8] L. Zighelnic and O. Kurland, "Query-drift prevention for robust query expansion," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2008, pp. 825–826.

[9] J. Lei, P. Sockolow, P. Guan, Q. Meng, and J. Zhang, "A comparison of electronic health records at two major Peking University Hospitals in China to United States meaningful use objectives," *BMC Med. Informat. Decis. Making*, vol. 13, no. 1, p. 96, Dec. 2013.

[10] T. Shu, H. Liu, F. R. Goss, W. Yang, L. Zhou, D. W. Bates, and M. Liang, "EHR adoption across China's tertiary hospitals: A cross-sectional observational study," *Int. J. Med. Informat.*, vol. 83, no. 2, pp. 113–121, Feb. 2014.

[11] Z. Liu and W. W. Chu, "Knowledge-based query expansion to support scenario-specific retrieval of medical free text," *Inf. Retr.*, vol. 10, no. 2, pp. 173–202, Feb. 2007.

[12] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[13] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1698–1735, Sep. 2019.

[14] L. Nguyen and T. Cao, "Pseudo-relevance feedback for information retrieval in medicine using genetic algorithms," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Cham, Switzerland: Springer, 2018, pp. 395–404.

[15] A. R. Aronson and T. C. Rindflesch, "Query expansion using the UMLS metathesaurus," in *Proc. AMIA Annu. Fall Symp.*, 1997, p. 485.

[16] D. Zhu and B. Carterette, "Using multiple external collections for query expansion," in *Proc. 20th Text Retr. Conf.*, 2011, pp. 1–6.

[17] C.-J. Lee, R.-C. Chen, S.-H. Kao, and P.-J. Cheng, "A term dependency-based approach for query terms ranking," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2009, pp. 1267–1276.

[18] J. H. Park and W. B. Croft, "Query term ranking based on dependency parsing of verbose queries," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2010, pp. 829–830.

[19] B. Xu, H. Lin, L. Yang, K. Xu, Y. Zhang, D. Zhang, Z. Yang, J. Wang, Y. Lin, and F. Yin, "A supervised term ranking model for diversity enhanced biomedical information retrieval," *BMC Bioinf.*, vol. 20, no. S16, pp. 1–11, Dec. 2019.

[20] D. Zhu and B. Carterette, "Improving health records search using multiple query expansion collections," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Oct. 2012, pp. 1–7.

[21] A. Boubacar and Z. Niu, "Concept based query expansion," in *Proc. 9th Int. Conf. Semantics, Knowl. Grids*, Oct. 2013, pp. 160–169.

[22] H. Jain, C. Thao, and H. Zhao, "Enhancing electronic medical record retrieval through semantic query expansion," *Inf. Syst. e-Bus. Manage.*, vol. 10, no. 2, pp. 165–181, Jun. 2012.

[23] M. Mitra, A. Singhal, and C. Buckley, "Improving automatic query expansion," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1998, pp. 206–214.

[24] T. Goodwin and S. M. Harabagiu, "UTD at TREC 2014: Query expansion for clinical decision support," Texas Univ. Dallas Richardson, Richardson, TX, USA, Tech. Rep. 1, 2014.

[25] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[26] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[28] X. Rong, "Word2vec parameter learning explained," 2014, *arXiv:1411.2738*. [Online]. Available: http://arxiv.org/abs/1411.2738

[29] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," 2020, *arXiv:2004.13922*. [Online]. Available: http://arxiv.org/abs/2004.13922

[4] http://chinasdc.cn/

[5] https://github.com/chun19920827/QE_Algorithm

[30] M. G. Valdés and M. S. Marín, "Delphi method for the expert consultation in the scientific research," *Revista Cubana de Salud Pública*, vol. 39, no. 2, pp. 253–267, 2013.
[31] C. E. Osgood and E. G. Walker, "Motivation and language behavior: A content analysis of suicide notes," *J. Abnormal Social Psychol.*, vol. 59, no. 1, p. 58, 1959.
[32] G. W. Ryan and H. R. Bernard, "Techniques to identify themes," *Field Methods*, vol. 15, no. 1, pp. 85–109, Feb. 2003.
[33] D. Harman, "Relevance feedback revisited," in *Proc. 15th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1992, pp. 1–10.
[34] Y. Lu, H. Fang, and C. Zhai, "An empirical study of gene synonym query expansion in biomedical information retrieval," *Inf. Retr.*, vol. 12, no. 1, pp. 51–68, Feb. 2009.
[35] D. J. Finney, *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve*. Cambridge, U.K.: Cambridge Univ. Press, 1952.
[36] T. Ruan, M. Wang, J. Sun, T. Wang, L. Zeng, Y. Yin, and J. Gao, "An automatic approach for constructing a knowledge base of symptoms in Chinese," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 1657–1662.
[37] M. Almasri, C. Berrut, and J.-P. Chevallet, "Wikipedia-based semantic query enrichment," in *Proc. 6th Int. Workshop Exploiting Semantic Annotations Inf. Retr. (ESAIR)*, 2013, pp. 5–8.
[38] J. H. Paik, D. Pal, and S. K. Parui, "Incremental blind feedback: An effective approach to automatic query expansion," *ACM Trans. Asian Lang. Inf. Process.*, vol. 13, no. 3, pp. 1–22, Oct. 2014.
[39] C. Carpineto, R. de Mori, G. Romano, and B. Bigi, "An information-theoretic approach to automatic query expansion," *ACM Trans. Inf. Syst.*, vol. 19, no. 1, pp. 1–27, Jan. 2001.
[40] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA: MIT Press, 2005.

SONGCHUN YANG received the B.S. degree in electronic engineering from Tsinghua University. He is currently pursuing the Ph.D. degree in biomedical engineering with the Information Center, Academy of Military Medical Sciences, Beijing, China.

His research interests include information retrieval, knowledge graph, and health informatics.

XIANGWEN ZHENG received the B.S. degree in information and computing science from Peking University. He is currently pursuing the Ph.D. degree in biomedical engineering with the Information Center, Academy of Military Medical Sciences, Beijing, China.

His research interests include healthcare data analytics, semantic networks, openEHR, and knowledge graph.

XIANGFEI YIN received the B.S. degree from Third Military Medical University. He is currently pursuing the M.S. degree in biomedical engineering with the Information Center, Academy of Military Medical Sciences, Beijing, China.

He is also a Medical Associate with the Sansha People's Hospital, China. His research interests include electronic health records and database generation.

HUAJIAN MAO received the B.S. degree from Zhejiang University, in 2007, and the M.S. and Ph.D. degrees from the National University of Defense Technology, China, in 2010 and 2013, respectively, all in computer science.

He is currently an Associate Professor with the Information Center, Academy of Military Medical Sciences, Beijing, China. His current research interest includes health informatics.

DONGSHENG ZHAO received the B.S. degree, the M.S. degree in information systems engineering, and the Ph.D. degree in computer science from the National University of Defense Technology, China, in 1992, 1995, and 1998, respectively.

He is currently the Director and a Professor of the Information Center, Academy of Military Medical Sciences, Beijing, China. He was the past Vice President of China Medical Informatics Association (CMIA), from 2015 to 2017, a Board Member of the Chinese National Population and Health Data Center. His main research interests include health information systems, biomedical big data, and medical AI.