Diao et al. EURASIP Journal on Wireless Communications and Networking (2018) 2018:105
https://doi.org/10.1186/s13638-018-1124-3

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

The research of query expansion based on medical terms reweighting in medical information retrieval

Lijuan Diao[1,2*], Hong Yan[3], Fuxue Li[3], Shoujun Song[4], Guohua Lei[1,2*] and Feng Wang[5]

Abstract

In recent years, information retrieval technology is widely used in the medical industry. How to effectively retrieve electronic medical record has become a hot issue in the field of information retrieval. Medical terms occupy an important position in the electronic medical record (EMR) retrieval, and they are usually used to limit the retrieval conditions, so they suggest the user's search intention. Aiming at the importance of medical terms, a method of query reformulation based on medical term reweighting is proposed to improve the performance of EMR retrieval. Firstly, the method filtrates medical terms from the origin query. Then, each medical term is weighted by its own self-information which can be estimated from the document set. Finally, the new query is constructed by combining the weighted medical terms and origin query proportionally. Experimental results on the TREC dataset show that our method performed above the baseline in three performance metrics: MAP (+ 14.2%), bpref (+ 8.0%), and P@10 (+ 9.6%).

Keywords: Medical information retrieval, Electronic medical record, Query reformulation, Medical termination

1 Introduction

With the information of the medical system, the electronic medical record (EMR) system is widely used by medical institutions. An electronic medical record contains the patient's clinical information, such as the history of the disease, the examination of the report, and the treatment of drugs. The rich medical information can assist doctors in diagnosing the patients' diseases and providing personalized health care for patients and is more conducive to clinical research. Although the structured text in medical records can be effectively accessed in the electronic medical record system, structured text is difficult to describe the difference between patients. So, the main content of electronic medical records is composed of a large number of free texts [1]. Unstructured text provides convenience for doctors to record medical records but brings huge difficulties to retrieval of electronic medical records. Therefore, how to effectively retrieve electronic medical records has become a hot issue in the field of information retrieval.

Electronic medical record retrieval is a search task for selecting the set of medical records that conforms to the restrictive conditions. The restrictive conditions usually contain a variety of medical terms, for example, glaucoma, amoxicillin, and endoscopy. These medical terms describe the patient's disease, the drug used, the examination, and other important medical information. Intuitively, medical terms express the user's search intentions and should increase their weight. Based on this hypothesis, this paper presents a query reconfiguration method based on the weight adjustment of medical terms (UMLS-W). In the framework of probability model, we extract the medical terms in query statements, measure the weight of these terms with self-information, and reconstruct the query sentences combined with non-medical terms in query statements.

Query reconstruction is a common technology in the field of information retrieval. There are generally two ways of query reconfiguration, namely, the query keyword expansion [2–4] and the keyword weight adjustment [5]. Many researchers present the methods of medical information retrieval by semantic similarity [6], the approach using concept-based medical information retrieval [7], and the method of leverage medical thesauri and physician

* Correspondence: lijuandiaolw@126.com
[1]Binzhou Medical University, Yantai 264003, China
Full list of author information is available at the end of the article

feedback for improving medical literature retrieval [8]. Zhu et al. tries to extract new keywords from different external medical resources and join the original query to construct new queries to improve the quality of medical record retrieval [9]. They also study the size and quality of the resources and the impact of mixed use of different resources on the effect of query refactoring. The experimental results show that the reconstruction methods of these extended keywords have been improved to different degrees, relative to the original query. The University of DELaware the MiXture relevance model (UDELMX) [9] method has the best effect on all evaluation. However, He et al. use the external resources related to medicine, and the retrieval results have only a little promotion than their respective benchmark lines, or even no promotion [10]. The main reason is that the method of extending the keyword is easy to have the problem of query drift. Query drift refers to the subject of extended query deviates from the original search intention of the user and leads to the decrease of precision. In order to reduce the effect of theme drift, Dihn et al. reconstruct the original query by using the method of keyword weight adjustment [11]. The experimental results show that the Log-logistic model (LGD) [11] method has the best retrieval effect among all the evaluation. From the above analysis, we can see that both the method of udelxm and LGD do not make use of medical information. Therefore, we propose the algorithm to add the medical-related information to the weight adjustment and more emphasis on the weight adjustment of the medical terms of the query sentence in this paper.

The remainder of this paper is organized as follows: we give methodology that explains how to automatically extract keywords from a given electronic medical record in Section 2 and Section 3. We present the term weighting model in query reformation in Section 3. Section 4 illustrates the experimental evaluation, and its result is described in Section 5. Section 6 summarizes the conclusions.

2 Methodology

2.1 Query reconstruction UMLS-W

We present UMLS-W that is the method of query reformation based on medical term weighting model by analyzing the query statements and the retrieval targets of the medical record retrieval. The specific process of the query method is shown in Fig. 1.

Our method consists of three main steps as follows:

(1) The mainly medical key phrases are extracted, and the stop words are removed from a given medical topic.
(2) The weighting of medical terms that is from the up steps is computed.
(3) The new query terms are reconstructed using of original query words and the weighting of medical terms.

2.2 Medical term extraction

The Unified Medical Language System (UMLS) is a biomedical concept thesaurus for coordinating health and medical vocabularies. UMLS contains three major components: (1) the “Metathesaurus” which includes data from MeSH, SNOMED, RxNorm, and other collection; (2) the “SPECIALIST lexicon and lexical Tools”; and (3) the “Semantic Network” which provides general categories and relationships. Meanwhile, the Metathesaurus is the core database in the UMLS and MetaMap is a medical terminology recognition tool based on the Metathesaurus of UMLS, which is developed by Aronson and Lang [12]. By analyzing the text to be recognized, MetaMap [12] extracts some phrases and each phase has a

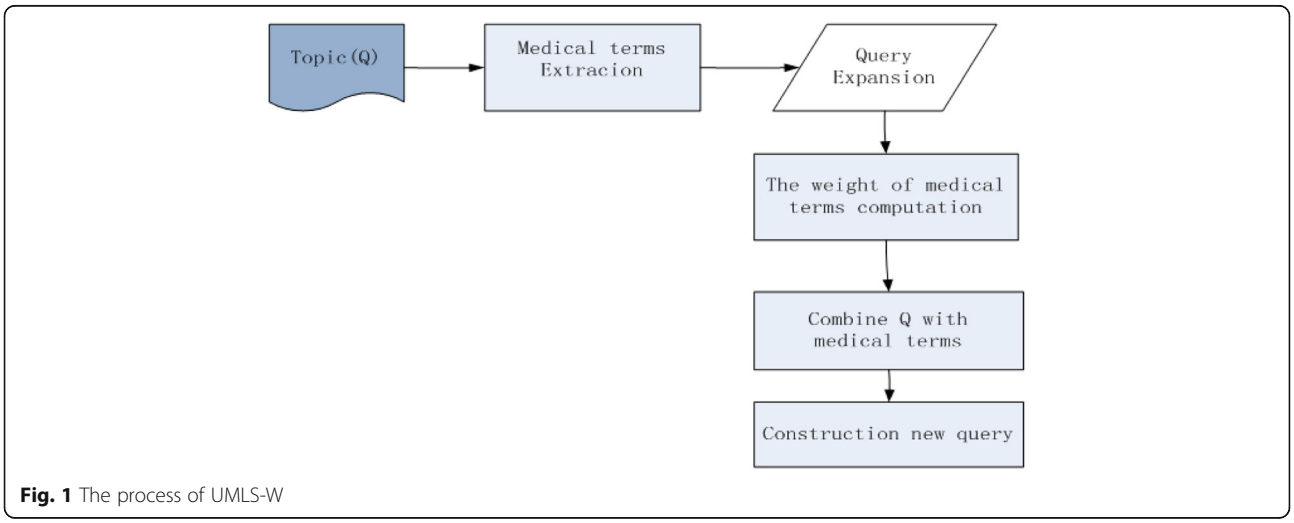

Fig. 1 The process of UMLS-W

mapping list that is mapping from phrase to medical term. MetaMap scores each mapping, and the higher the score, the greater the possibility that the phrase can be mapped to the medical term.

In this paper, we use MetaMap to identify medical terms in query statements. Firstly, queries are submitted to MetaMap's online tools. Then, the highest score mapping of each phrase is chosen, and the corresponding medical terms are what we need. However, medical terms identified by MetaMap are not necessarily important, such as patient, doctor, treat, diagnosis, and other words. These words that appear frequently in the electronic medical records are insignificant and are removed as stop words in this paper. There is no standard medical stop word list at present, and we give the medical stop words that are the original form of words in the Table 1. When the stop words are removed, the remaining medical keywords are recorded as $M_{\mathrm{UMLS}} = \{M_1, M_2, .., M_x, ..., M_n\}$.

Medical terms are usually made up of several keywords, so each medical term can be expressed as $M_x = \{t_1, t_2, ..., t_y, ..., t_{m_x}\}$. For example, the topic that is in the test set is that patients diagnosed with localized prostate cancer and treated with robotic surgery. After extracting by MetaMap, we get patients, diagnosed, localized prostate cancer, treated, and robotic surgery, where patients, diagnosed, and surgery are medical stop words. Finally, the medical term set is represented as follows:

$$M_{\mathrm{UMLS}} = \{\text{localized prostate cancer}, \text{robotic surgery}\}$$
$$M_1 = \{\text{localized}, \text{prostate}, \text{cancer}\}$$
$$M_2 = \{\text{robotic}, \text{surgery}\}$$

3 Query expansion

3.1 The weighting of medical term computation

Self-information is used to measure the amount of information contained in the occurrence of a single event. Assuming that the probability of the occurrence of random event ω_n is $p(\omega_n)$, the definition of the self information $I(\omega_n)$ is a formula (1).

$$I(\omega_n) = -\log(p(\omega_n)) \tag{1}$$

As we can find from the definition, the lower the probability of an event happens, the greater the self-information it contains when the event really happens.

Table 1 Medical stop words

ADMISSION	DIAGNOSIS	GIVE	PATIENT	BLOOD
ADMIT	DISCHARGE	HOSPITAL	RECEIVE	GRADE
CARE	DISEASE	MEDICINE	TAKE	POSITION
DIAGNOSE	DOCTOR	NURSE	TREAT	ARRHYTHMIA

According to the definition of self-information, the weighting of medical terms w_x can be represented as

$$w_x = -\ln(p(M_x|\theta_C)) \tag{2}$$

where M_x is the medical term, θ_C is the document set model, and $p(M_x|\theta_C)$ is the probability of medical terms M_x generated by a document collection model θ_C. The natural logarithm is adopted in this paper. How to compute the probability of medical terms $(M_x|\theta_C)$? The document is modeled using unigram language model, and they are independent between words. The concrete model is as follows:

$$p(M_x|\theta_C) = \prod_{t \in M_x} p(t|\theta_C)^{tf(t,M_x)} \tag{3}$$

where $tf(t, M_x)$ is the number of times t that appears in M_x and $p(t|\theta_C)$ is the probability of medical term t generated by a document collection model θ_C .

As long as a term t that does not appear in the document set, $p(t|\theta_C)$ is zero according to formula (3) therefore, $p(M_x|\theta_C)$ is also zero. In order to prevent the occurrence of zero probability, we need to remove the words that have only appeared in M_x but have not appeared in the document collection and it is represented as M'_x. So, formula (4) is as follows:

$$p(M'_x|\theta_C) = \prod_{t \in M'_x} p(t|\theta_C)^{tf(t,M'_x)} \tag{4}$$

Finally, the weighting of medical term can be expressed as the following formula (5).

$$w_x = -\sum_{t \in M'_x} (tf(t, M'_x) * \ln(p(t_y|\theta_C))) \tag{5}$$

The problem is converted to how to calculate the probability of the generation word t in a document collection model θ_C . Poisson distribution is used to fit the probability distribution of the number of occurrences T of word t in the document, that is T~Poisson(λ). In the Poisson probability distribution model, we use $P(T \geq 1)$ to estimate $p(t|\theta_C)$. So

$$p(t|\theta_C) = P(T \geq 1) = 1 - P(T = 0) = 1 - e^{-\lambda} \tag{6}$$

In formula (6), we can calculate the parameter λ by using maximum likelihood estimate (MLE) based on the document collection as follows:

$$\lambda = \hat{\lambda}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} k_i \tag{7}$$

where n is the size of the document collection and k_i is the frequency that the word t appears in the document D_i. On the premise of fixed document set, every term corresponding λ can be firstly calculated by

preprocessing, which may avoid repeated computation and reduce computation. According to the estimate of formula (7), the terms localized, prostate, cancer, robotic, and surgery corresponding λ values are 0.0447, 0.0482, 0.1280, 0.0006, and 0.2641 in Topic 104 of test sets. We can obtain the weights of $w_1 = 8.31$ and $w_2 = 8.88$ by the computation of formulas (5) and (6) which the values of the above terms are respectively substituted in.

3.2 Construction new query

Medical terms are only a part of the original query sentence. If the medical terms are used as query which loses some information, the result of the query is not overall. Therefore, we combine the original query sentences Q_{origin} with medical terms to construct new queries. The specific process is as follows:

Firstly, with regard to the original query statement Q_{origin}, the weight distribution among the words that constitute the query Q_{origin} is averaged, and the query clause 1 is obtained.

Secondly, the weight distribution of each word in the medical terms M_x is averaged.

Thirdly, the weight distribution among the medical terms M_x is $w_x/\sum_{x=1}^{n} w_x$, and combined with the second step, the query clause 2 is obtained.

Finally, the weight distribution between the query clause 1 and the query clause 2 is allocated in accordance with the proportion of $\alpha : (1-\alpha)$, and the new query statements are obtained. The range of parameter α is [0,1].

According to the above construction process, the weight of medical term M_x in the whole query statement is obtained from the query clause 1 and the query clause 2. The method of computation is formula (8) as follows:

$$w'_x = \alpha * \frac{|M_x|}{|Q_{\text{origin}}|} + (1-\alpha) * \frac{w_x}{\sum_{x=1}^{n} w_x} \tag{8}$$

where $|M_x|$ is the length of the medical terms, $|Q_{\text{origin}}|$ is the length of the original query statement, and the weight of the words that is not the medical term in the original query Q_{origin} is the same as α/Q_{origin}.

For example, according to formula (8), we can calculate the weights of two medical terms in Topic 104 of test set are respectively

$$w'_1 = \alpha^* \frac{3}{11} + (1-\alpha) * \frac{8.31}{8.31+8.88}$$
$$w'_2 = \alpha^* \frac{2}{11} + (1-\alpha) * \frac{8.88}{8.31+8.88}$$

4 Experimental

For our experiments, we use 35 topics authored for the TREC medical task in 2016. There are 100,866 reports and 17,198 time visits in the dataset. That is, the average one time visit corresponds to 5.86 reports. The description part of each topic consists of 9.79 words on average, with an average of 5.06 words belonging to medical terms.

We use the mean average precision (MAP) that is a common evaluation method for information retrieval, binary preference (bpref), and top ten (P10) which are official evaluation of the TRECMed 2016 in this paper. The three evaluation methods are introduced as follows:

P10: The accuracy of the top ten documents in single retrieval results is measured.

MAP: The mean average precision of a single topic is the average of the accuracy after each related document is retrieved. MAP is the average of the mean average precision of each topic, which is a single value indicator that reflects the performance of the system on all related documents.

bpref: It mainly refers to the number of unrelated documents appearing before the relevant documents, and the specific formula is

$$\text{bpref} = \frac{1}{R}\sum_{r}\left(1 - \frac{|n \text{ ranked higher than } r|}{\min(R, N)}\right) \tag{9}$$

where R is the related result in the decision results for each topic, r represents related documents, n is a subset of the set of unrelated documents in the top R, n ranked higher than r means that there are a number of unrelated documents in the number of the current related documents. It should be noted that the P10 or bpref mentioned in this paper refers to the mean of P10 or bpref of all subjects, and MAP is the most important evaluation criterion. In our experiments, we evaluate the performance achieved on this task according to these relevant judgments.

The steps of the experiment are described as follows: (1) Indri retrieval system is used as our experiment retrieval system, indexing and retrieving with report as a unit, and Porter algorithm [13] is used to extract the stem, when building the index. (2) The retrieval model uses the language model supported by Indri [14], using the Dirichlet smoothing method [15, 16], and the parameter takes the default value of 2500. (3) The retrieval results require the aggregation of report to visit. The method of aggregation is to calculate the score of the visit score (v) based on the report ranking, and visit is sorted in descending order of score (v) . The calculation formula of score (v) is as follows:

$$\text{score}\ (v) = \sum_{r \in v} \frac{1}{\text{rank}(r)} \tag{10}$$

where rank(r) is the ranking of a report. In order to verify the effectiveness of this method, a comparative experiment of three kinds of queries is designed. First, the

description of the topic is retrieved directly as a query statement, and the results are used as the baseline (Baseline). Second, referring to the new query construction process of Section 3.2, excepting for the distribution of the weight among M_x is changed to average, the rest is consistent, and the retrieval results are recorded as UMLS-E. Third, using the method proposed in this paper, the retrieval results are recorded as UMLS-W. In order to evaluate the effect of the smooth parameter α, 11 groups of experiments are designed for UMLS-E and UMLS-W, respectively, from 0 to 1 according to the 0.1 step length and set 11 values of α.

The Indri system is adopted in the experiment of this paper. After reconstructing the query statements by UMLS-E and UMLS-W, it is easy to use Indri query syntax to formalize them, as shown in Figs. 2 and 3. Where *#weight* and *#combine* are all operators of Indri query syntax, *#weight* is the distribution weight which is assigned to a given proportion and *#combine* is the average allocation of weights.

5 Results and discussion

5.1 Results

Firstly, we analyze the impact of values α on queries. As shown in Fig. 4, the MAP value of the UMLS-E and UMLS-W is changing at different values α . Whether UMLS-E or UMLS-W, the MAP value obtained using only medical term ($\alpha = 0$) to query is higher than the MAP value obtained using only the original query sentence ($\alpha = 1$). Therefore, medical terms are very important in the retrieval of electronic medical records. In addition to the point of Baseline ($\alpha = 1$), the overall performance of the UMLS-W method is completely superior to the UMLS-E method because the UMLS-W line is always above the UMLS-E line. When $\alpha = 0.6$, the MAP value of UMLS-E and UMLS-W reaches the maximum, so we set the value to 0.6.

Secondly, we analyze the effect of using self-information to measure the weight of medical terms. As can be seen in Table 2, the comparison of the experimental results of three experiments, Baseline, UMLS-E, and UMLS-W, is very obvious. Compared with Baseline, UMLS-E and UMLS-W increased 5~14% on three indicators, which indicates that more consideration of the weight of medical terms helps to improve the retrieval performance of electronic medical records. The results of comparison between UMLS-W and Baseline are better than the results of comparison between UMLS-E and Baseline. It shows that the distribution of weights among medical terms by using self-information is more helpful to improve retrieval performance. Meanwhile, it validates the effectiveness of using the self-information measure for the weight of medical terms.

$$\begin{aligned}&\#weight(\\&\alpha\ \#combine(Q_{origin})\\&(1-\alpha)\ \#combine(\\&\#combine(M_1)\ \#combine(M_2)\ldots\#combine(M_n)\\&)\end{aligned}$$

Fig. 2 The query syntax of UMLS-E

$$\begin{aligned}&\#weight(\\&\alpha\ \#combine(Q_{origin})\\&(1-\alpha)\ \#combine(\\&w_1\ \#combine(M_1)\ w_2\ \#combine(M_2)\ldots w_n\ \#combine(M_n)\\&)\\&)\end{aligned}$$

Fig. 3 The query syntax of UMLS-W

Thirdly, the accuracy and recall of Baseline, UMLS-E, and UMLS-W is shown in Fig. 5. As we can see from Fig. 5, the overall retrieval performance is UMLS-W > UMLS-E > Baseline.

5.2 Discussion

We compare the methods of UMLS-W and udelmx proposed by Zhu[] and LGD proposed by Dinh[]. Udelmx focuses on expanding keywords, while LGD focuses on the adjustment weight of keywords. We can see the results of the comparison of the three methods of UMLS-W, udelmx, and LGD in Table 3. By comparing the data in Table 3, we can see that the performance of the three method is not much in the bpref, and UMLS-W is just a little better. UMLS-W is 6.6% higher than udelmx on the P10. The reason is that after expanding the keywords of udelmx, too many extended words can not only optimize the original query but also add noise which makes the ambiguity of queries increase. Therefore, it leads to the decrease of precision. However, there is no the above question in the method of UMLS-W and LGD, and even the P10 value of LGD is little better than

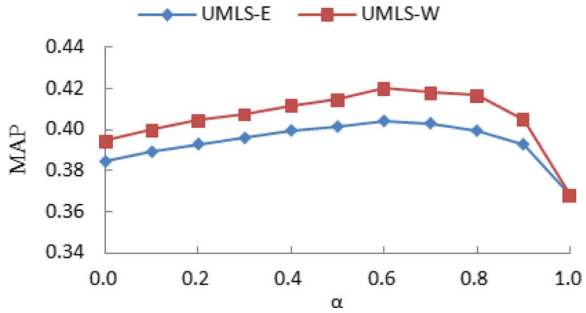

Fig. 4 The effect of α value on UMLS-E and UMLS-W

Table 2 The comparison of Baseline, UMLS-E, and UMLS-W

	MAP	bpref	P10
Baseline	0.3678	0.4835	0.5235
UMLS-E	0.4042(+ 9.9%)	0.5094(+ 5.4%)	0.5559(+ 6.2%)
UMLS-W	0.4200(+ 14.2%)	0.5222(+ 8.0%)	0.5735(+ 9.6%)

UMLS-W. On the MAP index, UMLS-W increased by 4.8 and 6.5% compared with udelmx and LGD, respectively. The reason is that UMLS-W is not subject to the problem of topic drift compared with udelmx and it takes into account the factor of medical term. Therefore, the performance of the method UMLS-W is better than the method LGD.

In general, the UMLS-W method used to measure the weight of medical terms by self-information in this paper is proved to be more reasonable. The reconstructed query can improve the performance of electronic medical record retrieval system.

6 Conclusions

The query sentences in the electronic medical record often contain some medical terms as a limiting condition. In this paper, we studied how to use these medical terms to reconstruct queries. Finally, we proposed a method of using self-information to measure medical terms, combined with original query sentences to reconstruct queries, and improved the performance of electronic medical record retrieval system. Experimental results on the TREC dataset show that our method performed above the baseline in three performance metrics: MAP (+ 14.2%), bpref (+ 8.0%), and P@10 (+ 9.6%). At present, the set values of the parameter α are the same in this paper. In the next work, we plan to study the adaptively set values for different queries and further improve the performance of electronic medical record retrieval.

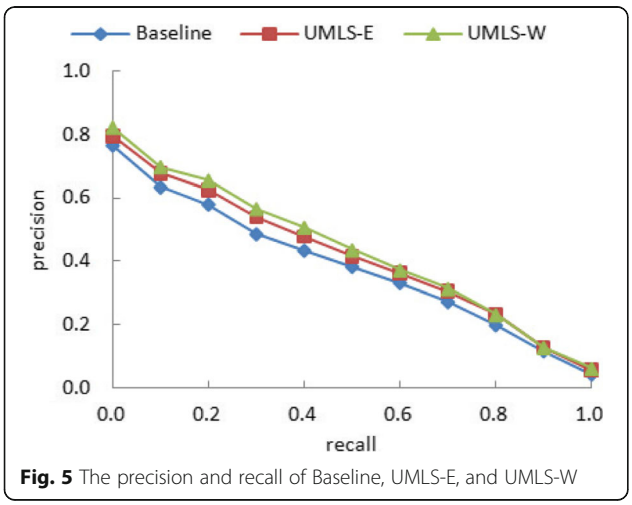

Fig. 5 The precision and recall of Baseline, UMLS-E, and UMLS-W

Table 3 The comparison of LGD, udelmx, and UMLS-W

	MAP	bpref	P10
LGD	0.3944	0.5311	0.5794
udelmx	0.4007	0.5289	0.5382
UMLS-W	0.4200	0.5222	0.5735

Acknowledgements
The research presented in this paper was supported by Binzhou Medical University, Yantai, China; Yingkou Institute of Technology, Yingkou, China; Affiliated Hospital of Binzhou Medical University; Ludong University, Yantai, China.

Funding
The authors acknowledge the education of Shandong province "in 12th Five-Year 2015 year" plan of higher education computer teaching project (Grant No. ZBJ15005) and the scientific research fund of Binzhou Medical University (Grant No. BY2015KYQD07).

Authors' contributions
LD was the main writer of this paper. She proposed the main idea, deduced the performance of UMLS-W, completed the simulation, and analyzed the result. HY and FL introduced the Indri system in the experiment. SS verified the medical terms and detected the accuracy of the retrieval results. GL and FW gave some important suggestions for the performance of UMLS-W. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Binzhou Medical University, Yantai 264003, China. [2]Key Laboratory of Language Resource Development and Application of Shandong Province, Yantai, China. [3]Yingkou Institute of Technology, Yingkou, China. [4]Yantai Affiliated Hospital of Binzhou Medical University, Yantai, China. [5]Ludong University, Yantai, China.

Received: 9 February 2018 Accepted: 24 April 2018
Published online: 04 May 2018

References

1. S Choi, J Choi, S Yoo, H Kim, Y Lee, Semantic concept-enriched dependence model for medical information retrieval. J. Biomed. Inform. **47**, 18–27 (2014)
2. X Liu, Y Xia, W Yang, F Yang, Secure and efficient querying over personal health records in cloud computing. Neurocomputing **274**, 99–105 (2018)
3. W Weerkamp, K Balog, M de Rijke, Exploiting external collections for query expansion. ACM Transactions on the Web (TWEB) **6**(4), 18 (2012)
4. J Gao, G Xu, J Xu, Query expansion using path-constrained random walks[C]//proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval. ACM, 563–572 (ACM, Dublin, 2013), https://doi.org/10.1145/2484028.2484058
5. YC Chang, SM Chen, A new query reweighting method for document retrieval based on genetic algorithms. Evolutionary Computation, IEEE Transactions on **10**(5), 617–622 (2006)
6. Hliaoutakis, Angelos, Varelas, Giannis,, Voutsakis, E., Petrakis E, ., and Milios E. Information retrieval by semantic similarity. International Journal on Semantic Web and Information Systems (IJSWIS) 2, 3 (2016).
7. B Koopman, P Bruza, L Sitbon, M Lawley, Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. Australas Med J **5**(9), 482 (2012)
8. P Sondhi, J Sun, C Zhai, R Sorrentino, MS Kohn, Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. J. Am. Med. Inform. Assoc. **19**(5), 851–858 (2012)

9. Zhu D, Carterette B. Using multiple external collections for query expansion[C]// proceedings of the 20th text retrieval conference proceedings TREC. 2011.
10. L Liu, L Liu, et al., A cloud-based framework for large-scale traditional Chinese medical record retrieval. Journal of Biomedical Information. **77**, 21–33 (2018)
11. Dinh D, Tamine L. IRIT at TREC 2011: evaluation of query reformulation techniques for retrieving medical records[C]//Proceedings of the 20th text retrieval conference proceedings TREC. 2011.
12. AR Aronson, FM Lang, An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. **17**(3), 229–236 (2010)
13. MF Porter, An algorithm for suffix stripping. Program: electronic library and information systems **14**(3), 130–137 (1980)
14. V Lavrenko, WB Croft, Relevance based language models[C]//proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM, 120–127 (ACM, New Orleans, 2001)
15. C Zhai, J Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval[C]//proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM, **22**(2), 334-342 (2001)
16. P Clough, M Sanderson, Evaluating the performance of information retrieval systems using test collections. Information Rezearch **18**(2), 655-662 (2013)