Pseudo-Relevance Feedback for Information Retrieval in Medicine Using Genetic Algorithms

Lanh Nguyen[1(✉)] and Tru Cao[1,2]

[1] Faculty of Computer Science and Engineering,
Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam
nguyenlanh2580@gmail.com, tru@cse.hcmut.edu.vn
[2] John von Neumann Institute, Vietnam National University at Ho Chi Minh City,
Ho Chi Minh City, Vietnam

Abstract. Pseudo-Relevance Feedback is one of the methods for improving search engine results. By automatically extracting information from a previous search result, a new query is posed as an expansion of the original query, and then it is searched again. In this paper, we apply a genetic algorithm to improve the Pseudo-Relevance Feedback method in searching medical texts. First, a set of candidate terms is constructed by extracting keywords from the documents returned from the initial search using the original query. Then, the seed terms are selected from the candidate term set using our proposed genetic algorithm, to be merged with the original query to create a new query. The new query is searched again, returning a final ranked list of documents. Experimental results on the TREC 2014 CDS dataset show that the proposed method outperforms the baseline method that does not use a genetic algorithm for Pseudo-Relevance Feedback.

Keywords: Medical case report · Clinical question type
Query expansion · Candidate terms · Jaccard similarity coefficient

1 Introduction

Information retrieval focuses on organization of a collection, and storage and communication of different kinds of data structures (e.g. texts, images and sounds) [1]. The purpose of an information retrieval system is to provide users with those documents that satisfy their information need without taking much time. There are search engines that may be available to the public, such as Educational Resources Information Center[1] system for education research and information, or FinAstronomy[2] and Infotopia[3] systems in the domains of science and biology, etc.

[1] http://eric.ed.gov.
[2] http://www.findastronomy.com.
[3] http://www.infotopia.info.

N. T. Nguyen et al. (Eds.): ACIIDS 2018, LNAI 10752, pp. 395–404, 2018.
https://doi.org/10.1007/978-3-319-75420-8_38

In the medical domain, relevant biomedical documents to a patient case report are searched by search engines [2]. For example, a case report may describe information such as a patient's current symptoms, the patient's medical history, the patient's diagnosis, or the steps taken by a physician to treat the patient. The challenges of medical information retrieval include the followings [3]:

- Highly specialized information.
- Different kinds of medical information.
- Medical documents in multiple languages.

Today, users can access online search engines to search for medical information. For instance, PubMed was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. PubMed consists of more than 26 million citations for biomedical literature. PubMed is a free resource and also provides access via websites.

Entrez[4] is the online search engine developed by the NCBI to search on PubMed. This search engine requires a user to enter search terms that are used to search for relevant documents. Therefore, physicians can easily seek out information about how to best care for their patients to make a clinical decision.

Usually, a user's query is not expressive enough to represent the actual user's information need. Therefore, the initial list of search results may not be satisfactory. There are some methods to improve the search effectiveness due to incomplete representation of a query.

The relevance feedback method performs the interaction between the search system and users, in which a user gets improved retrieval performance thanks to their feedback to the system. The idea behind relevance feedback is using the user intervention and feedback to the initial search results to pose a revised query. When this automatic technique works without the user interaction, it is called Pseudo-Relevance Feedback (PRF), also known as Blind Relevance Feedback [4,5].

Query expansion is often combined with PRF where the original query is expanded with some terms automatically extracted from selected documents in the initial search results [5–7]. Further, genetic algorithms (GA) have also been applied to extract those added terms, which are crucial to improve the search performance [8]. However, to the best of our knowledge, GA-based PRF has not been employed for medical information retrieval.

Therefore, in this paper, we propose a genetic algorithm used with PRF for searching medical texts. The main role of GA is to select the seed terms in the top-ranked documents resulted from the initial search, based on their similarity with all of their context terms. A new query is then generated by adding the seed terms to the original query. Finally, the new query is searched again to return the final ranked list of documents.

The remainder of the paper is organized as follows. Section 2 defines the problem of discourse and describes the system architecture, baseline method, and proposed method. Section 3 presents details of the dataset, evaluation methods, and the main experiments. Finally, the conclusion and future work are given in Sect. 4.

[4] https://www.ncbi.nlm.nih.gov/gquery/.

2 Proposed Method

2.1 Problem Definition

To approach and extract the information from biomedical literatures, some modern search engines have been developed. In some cases, the needs of physicians may be found in a document or collection of documents. However, when the medical information become large and overlap, finding the relevant information for patient care becomes a significant challenge.

The focus of the TREC 2014 Clinical Decision Support Track is to retrieve relevant biomedical articles for answering generic clinical questions about medical records [9]. A detailed description of its input and output is presented below.

Input. The input data are divided into two parts: the first part is a document collection and the second is a query, also called as a topic. The document collection for the track is an open access subset of PubMed Central, an online repository of free available full-text biomedical literature. A database is created by indexing the text of the articles in the collection.

Each topic consists of a medical case report and one of the three generic clinical question types as follows:

- Medical case report: A case report typically describes a challenging medical case and is represented in the free-text format.
- Clinical question type: The three most common generic clinical question types are *diagnosis*, *test*, and *treatment*, which account for a majority (52.72%) of the clinical questions posed by primary care physicians [10]. Each case report has one associated clinical question type. Table 1 describes the meanings of these clinical question types.

Table 1. Description of the clinical question types.

Question type	Description
Diagnosis	Question about determining the diagnosis of the patient
Test	Question about suggesting relevant interventions for diagnosing the patient
Treatment	Question about suggesting the best treatment plan for the condition exhibited by the patient

For each of the clinical question types, the resulting documents should be relevant to it. A topic with the *diagnosis* label, for example, requires the system to retrieve those documents that a physician would find useful for determining the diagnosis for the patient described in its case report. Meanwhile, for the *test* label, the search results should suggest required medical tests to be conducted for the patient. Finally, for the *treatment* label, the retrieved documents should suggest to a physician appropriate treatment plans for the condition exhibited by the described patient.

Output. The expected search result is a ranked list of retrieved documents based on their relevance to the query. As in [9], our method evaluates results in 1,000 top-ranked documents for each topic.

2.2 System Architecture

To solve the problem defined above, we first present a general architecture for full-text medical information retrieval systems. Moreover, we briefly describe the baseline method used for each module [11]. This architecture has six main modules, as illustrated in Fig. 1 and presented below.

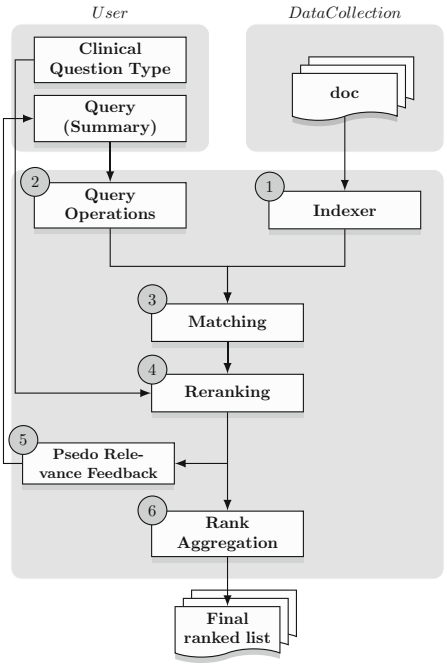

Fig. 1. System architecture for medical information retrieval.

Module 1. Indexing is the process of tagging search terms or phrases to each document to facilitate faster search and retrieval. As in [11], the raw documents is indexed by using Lucene[5].

Module 2. The query operations include two stages: query preprocessing and query expansion. The preprocessing stage may include the steps such as spelling checking, identification of words, phrases, sentences, stop words elimination, stemming, etc. Next, query expansion is the process of reformulating the original query such as finding synonyms or various morphological forms of words, in

[5] https://lucene.apache.org/.

order to better represent the meaning of the query. As in [11], the second stage expands queries by using the Medical Subject Headings (MeSH)[6] thesaurus.

Module 3. This module matches documents and query terms. In this module, [11] employed the Vector Space Model (VSM) [12], Language Model (LM) [13, 14], Best Matching (BM) with the two variants BM25L and BM25$^+$ in [15,16].

Module 4. Based on the clinical question type of a query, this module reranks the search results. The method in [11] counted the number of occurrences of candidate terms by using the MeSH thesaurus for each question type. If a document contains more candidate terms corresponding to the query's question type, it is more relevant.

Module 5. The PRF module is for simulating the user behaviour to select seed terms for query expansion from initial search results. In this work, we apply GA with PRF to improve the system's performance.

Module 6. The rank aggregation module is to combine multiple ranked document lists into a single ranked list to improve the rankings produced by individual systems. The Reciprocal Rank Fusion (RRF) algorithm in [17] is used for this module.

2.3 Genetic Algorithm in Pseudo-Relevance Feedback

For PRF, [18] shows that using a search process is more effective than simply using terms from initially retrieved documents. In this paper, we propose to use GA for this searching to select seed terms from a set of candidate terms in medical information retrieval.

Specifically, each document in the initial search results is represented by a string of 0's and 1's as a chromosome, in which each word is a gene. A set of chromosomes together with their associated fitness values is called a population. The population size (N) is the number of chromosomes in each generation.

Specifically, the steps using GA [8] to select the seed terms for PRF are as follows:

- **Step 1:** Generate an initial population of documents from the top of the search result list.
- **Step 2:** Encode retrieved documents into chromosomes in the binary format.
- **Step 3:** Create a new population by carrying out the genetic operations selection, crossover, and mutation on the previous population.
- **Step 4:** Verify the fitness of the new generation of individuals. If converged, stop. Otherwise go to Step 3.
- **Step 5:** Decode the optimized chromosomes to obtain the seed terms for PRF.

[6] https://www.nlm.nih.gov/mesh/.

Details of each step is presented below.

Initial Population. A candidate set is initialized with the top-30 ranked documents.

Chromosome Representation. Each chromosome encodes a binary string. The length of chromosomes depends on the size of the candidate set. When a keyword is present in a document, the corresponding bit is set to 1; otherwise, it is 0.

Fitness Function. As in [8], the fitness of each chromosome in a population is evaluated using the Jaccard similarity measure, which ranges between 0 and 1.

Selection Operation. As in [19], the selection process is to remove some bad (with low fitness) chromosomes. It is based on spinning the roulette wheel in N times.

Crossover Operation. As in [19], let P_c be the crossover probability. This probability gives us the expected number $P_c * N$ of chromosomes

Mutation Operation. Let P_m be the mutation probability. This probability gives us the expected number of $P_m * N$ of chromosomes. Every bit in all chromosomes of the whole population has an equal chance to undergo mutation, that is, change from 0 to 1 or vice versa. According to [8], typically P_c ranges between 0.7 and 0.9 and P_m ranges between 0.01 and 0.03.

3 Experiments

In this section, we evaluate and compare empirical performances of the baseline method and the proposed method on the TREC 2014 Clinical Decision Support (CDS) dataset [20].

3.1 Dataset

The focus of TREC 2014 CDS Track is retrieval of biomedical articles relevant to generic clinical question about medical records. The track dataset is divided into two separate parts: the first part includes medical documents such as full-text biomedical articles and the second part contains case reports as topics. Details are presented below.

Documents. The document dataset is an open access subset from PubMed Central (PMC)[7]. This set contains the abstracts, full texts, and other metadata of 733,138 articles (47.2 GB) in the biomedical domain. The articles are presented in the NXML format using the National Library of Medicine's Journal Archiving and Interchange Tag Set [9].

Each article in the collection is uniquely identified by the PubMed Central Identifier (PMCID) number, which is specified by the <*article-id*> element within its NXML file. The article is named using the same PMCID number.

[7] https://www.ncbi.nlm.nih.gov/pmc/.

Topics. The query set of the dataset includes 30 different topics in total, with 10 topics for each of the query types *diagnosis*, *test*, and *treatment*. Each topic consists of a summary which describes a patient's case report created by expert topic developers at the U.S National Library of Medicine maintaining actual medical records. Figure 2 shows examples of the three topics types used in the task.

```
<topic number="3" type="diagnosis">
    58-year-old female non-smoker with left lung mass on x-ray. Head
   CT shows a solitary right frontal lobe mass.
</topic>
```

```
<topic number="12" type="test">
   25-year-old woman with fatigue, hair loss, weight gain, and cold
   intolerance for 6 months.
</topic>
```

```
<topic number="26" type="treatment">
   Group traveling to the Amazon rainforest, including 3 pregnant
   women. All members' immunizations are up-to-date but they require
   malaria prophylaxis.
</topic>
```

Fig. 2. Three of 30 topics from the TREC 2014 CDS Track.

3.2 Evaluation Methods

In this section, we present some methods to evaluate information retrieval systems. According to [21], Precision and Recall are the basic measures expressed as percentages. Precision and Recall are set-based measures in unordered sets of documents. However, the quality of a search engine is also expressed via ranking of relevant documents retrieved. That is, more relevant documents are expected to be at higher positions in the result list. Therefore, other measures such as $P@k$ and *R-precision* are introduced, as in [22].

3.3 Experimental Results

We conduct 4 experiments on the dataset presented in Sect. 3.1. The first two experiments (Method 1, Method 2) use the baseline method and the others (Method 3, Method 4) use the proposed method. Following the TREC standard, the 1,000 top-ranked documents are retrieved for each query in the evaluation. Also, in the experiments, we set $P_c = 0.783$, $P_m = 0.029$, and $N = 30$ (i.e., top-30 ranked documents).

As shown in Table 2, in the experiments we apply various algorithms such as BM25L, BM25$^+$, VSM, LM in the matching module. Method 1 uses BM25L model while Method 2 combines multiple models by using the RRF algorithm as introduced above. The first two experiments expand queries by using the MeSH thesaurus and PRF. The PRF module in Method 1 and Method 2 only extracts

keywords from the top-3 documents as in [11]. In contrast, we combine GA and PRF in Method 3 and Method 4 that correspond to Method 1 and Method 2, respectively.

Table 2. The overall performance of different methods.

Method ID	Method description	*R-prec*	*P*@10
Method 1	BM25L, MeSH, PRF	0.1850	0.3533
Method 2	RRF: (BM25L, MeSH, PRF) (BM25$^+$, MeSH, PRF) (VSM, MeSH, PRF) (LM, MeSH, PRF)	0.1913	0.3667
Method 3	BM25L, MeSH, PRF, GA	**0.1950**	**0.3733**
Method 4	RRF: (BM25L, MeSH, PRF, GA) (BM25$^+$, MeSH, PRF, GA) (VSM, MeSH, PRF, GA) (LM, MeSH, PRF, GA)	**0.2036**	**0.3933**

As one can see, for *R-prec*, Method 3 outperforms Method 1 by 5.4% (0.1950 vs 0.1850), and Method 4 outperforms Method 2 by 6.4% (0.2036 vs 0.1913). Meanwhile, for *P*@10, the improvement of Method 3 over Method 1 is by 5.6% (0.3733 vs 0.3533), and the improvement of Method 4 over Method 2 is by 7.2% (0.3933 vs 0.3667).

In addition, Fig. 3 shows the *R-prec* measures for each clinical question types. As compared to Method 2, Method 4 outperforms it on all of the three query types. However, Method 3 is only better than Method 1 with the *test* query type. It could be due to different initial search results in different runs and on different question types, where Method 2 and Method 4 use a different matching model from that of Method 1 and Method 3.

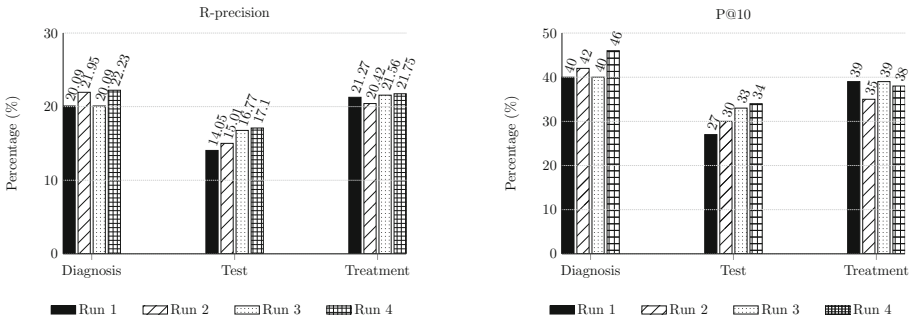

Fig. 3. The overall performance of different methods for each query types.

4 Conclusions

We have first proposed and presented the use of GA in PRF for medical information retrieval. First, relevant documents are retrieved and ranked for the initial search. Second, a set of candidate terms is extracted from the set of top-k retrieved documents. Third, the seed terms are selected from the candidate set by the GA and added to the original query. Finally, the system matches the new query and documents in the database, and returns the final ranked list.

We have experimented the proposed GA method on the TREC 2014 CDS dataset. Unlike traditional information retrieval datasets, here each query is associated with a medical question type. The results show that the proposed method improves the system performance, in comparison with the use of PRF without GA.

For the future work, we suggest using semantic relations between medical terms in query expansion and content matching. Besides, we are applying and adapting the proposed method to information retrieval on Vietnamese medical documents.

Acknowledgments. This work is funded by Vietnam National University at Ho Chi Minh City under the grant number B2016-42-01.

References

1. Chou, S., Chang, W., Cheng, C.Y., Jehng, J.C., Chang, C.: An information retrieval system for medical records & documents. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1474–1477 (2008). https://doi.org/10.1109/IEMBS.2008.4649446
2. Goeuriot, L., Jones, G.J.F., Kelly, L., Müller, H., Zobel, J.: Medical information retrieval: introduction to the special issue. Inf. Retr. J. **19**(1–2), 1–5 (2016)
3. Palotti, J., Hanbury, A., Müller, H., Kahn Jr., C.E.: How users search and what they search for in the medical domain - understanding laypeople and experts through query logs. Inf. Retr. J. **19**(1–2), 189–224 (2016)
4. Cao, G., Nie, J., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, 20–24 July 2008, pp. 243–250 (2008). https://doi.org/10.1145/1390334.1390377
5. Lv, Y., Zhai, C., Chen, W.: A boosting approach to improving pseudo-relevance feedback. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, 25–29 July 2011, pp. 165–174 (2011). https://doi.org/10.1145/2009916.2009942
6. Cao, G., Nie, J.-Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 243–250. ACM, New York (2008)
7. Vargas, S., Santos, R.L.T., Macdonald, C., Ounis, I.: Selecting effective expansion terms for diversity. In: Open Research Areas in Information Retrieval, OAIR 2013, Lisbon, Portugal, 15–17 May 2013, pp. 69–76 (2013)

8. Chen, H.: Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. JASIS **46**(3), 194–216 (1995)
9. Simpson, M.S., Voorhees, E., Hersh, W.: Overview of the TREC 2014 clinical decision support track. In: Proceedings of the 23rd Text Retrieval Conference (TREC), Gaithersburg, MD, USA (2014)
10. Del Fiol, G., Workman, T.E., Gorman, P.N.: Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern. Med. **174**(5), 710–718 (2014)
11. Mourão, A., Martins, F., Magalhães, J.: NovaSearch at TREC 2014 clinical decision support track. In: Proceedings of the Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, 19–21 November 2014
12. Singh, J.N., Dwivedi, S.K.: Analysis of vector space model in information retrieval. In: Proceedings Published by International Journal of Computer Applications® (IJCA), vol. 2, pp. 14–18 (2012)
13. Trotman, A., Puurula, A., Burgess, B.: Improvements to BM25 and language models examined. In: Proceedings of the 2014 Australasian Document Computing Symposium, ADCS 2014, Melbourne, VIC, Australia, 27–28 November 2014, p. 58 (2014). https://doi.org/10.1145/2682862.2682863
14. Banerjee, P., Han, H.: Language modeling approaches to information retrieval. JCSE **3**(3), 143–164 (2009)
15. Lv, Y., Zhai, C.: Lower-bounding term frequency normalization. In: Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, 24–28 October 2011, pp. 7–16 (2011)
16. Lv, Y., Zhai, C.: When documents are very long, BM25 fails! In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, 25–29 July 2011, pp. 1103–1104 (2011). https://doi.org/10.1145/2009916.2010070
17. Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, 19–23 July 2009, pp. 758–759 (2009). https://doi.org/10.1145/1571941.1572114
18. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**(1), Article ID 1–1150 (2012). https://doi.org/10.1145/2071389.2071390
19. Gen, M., Liu, B.: A genetic algorithm for optimal capacity expansion. J. Oper. Res. Soc. Jpn. **40**, 1–9 (1997)
20. Roberts, K., Simpson, M.S., Demner-Fushman, D., Voorhees, E.M., Hersh, W.R.: State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. Inf. Retr. J. **19**(1–2), 113–148 (2016)
21. Zuva, K., Zuva, T.: Evaluation of information retrieval systems. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **4**, 35–43 (2012)
22. Mogotsi, I.C., Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008). 482 p. ISBN: 978-0-521-86571-5. Inf. Retr. **13**(2), 192–195 (2010)