Query Expansion Methods for Text-based Retrieval of Medical Articles

Ivan Kitanovski, Katarina Trojacanec

Faculty of Computer Science and Engineering, University "Ss. Cyril and Methodius", Skopje, Macedonia
{ivan.kitanovski,katarina.trojacanec}@finki.ukim.mk
http://www.finki.ukim.mk

Abstract. This paper presents strategies for text-based retrieval of medical articles by applying query expansion. The goal of the paper is to investigate whether the propose query expansion techniques can improve the retrieval performance. The query expansion strategies include expansion of the queries by adding medical concepts or synonyms and traditional query expansion techniques based on the retrieved document collection. The articles for the experiment are PubMed medical articles. The text from the medical articles is used to create a field-based textual representation. We used Terrier IR search engine due to its flexibility and scalability. The results showed that query expansion can increase the performance of the medical articles retrieval.

Keywords: Medical articles retrieval, Query expansion, PubMed articles, Terrier IR

1 Introduction

Advances in technology are constantly making modern life easier and in the same time more reliant on the ever-growing data collections that power the systems and tools that we use everyday. From simple decision making, to solving complex medical cases. Thorough research of such data collections is needed. The data is digital and can appear in multiple forms like texts, images, web pages etc. The value that the data provides depends on the context where it is provided.

Acquiring the right information related to a medical question is a very complex, but important task as it can help the doctor with the current case he is working on or educate him on a certain topic. Having in mind, that there are over 12400 different categories of medical conditions [1], the existence of an appropriate information retrieval system is vital.

Our interest in this paper lays in medical information retrieval systems, which is related to analysis, organization and retrieval of medical information [2]. More precisely, we focus on retrieval of medical articles (cases), since they contain a lot of biomedical information. Already, there are systems which provide such services: Pubmed [3], eTBLAST [4], Pubget [5] etc. The systems accept keywords

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

as input, whereas a more practical scenario would be where the queries are of narrative nature, where the user can explain the situation in more details [6].

In this paper we try to improve the retrieval by using methods for query expansion in text-based retrieval of medical articles. The proposed approach processes the articles in the standard way and uses indexing and retrieval techniques that are implemented in many search engines. The initial retrieval phase is further tuned by applying query expansion techniques.

The rest of the paper is organized in the following manner: The related work is presented in Section 2. Section 3 contains the query expansion methods used in the experiments. The experimental setup and evaluation measures are described in Section 4. The experimental results and discussion are presented in Section 5. Finally, the concluding remarks and planned future work is provided in Section 6.

2 Related word

Already there are many approaches at solving the medical articles retrieval problem. There are many systems that offer this kind of services, which are usually based on existing search engines.

Vanegas et al [7] proposed a more custom approach where they implement their own version of the Okapi BM25 weighting model using the NLP toolkit for Python. This technique was proposed for the text-based ad-hoc image retrieval task. It achieved good results on the CLEF 2012 workshop. The proposed approach is generic and which means that it can be reused for medical articles retrieval.

The approach presented by Simpson et al [8] uses the Essie search engine for indexing and retrieval. The proposed approach is multimodal,, i.e... the retrieval is consisted of a text-based and content-based approach. For the text-based part, Essie implements a query expansion feature by including UMLS concepts. For the content-base approach they use low level features to describe the images and retrieval is done by computing Euclidean distance between them. The results from both retrieval types are fused in the end. This approach did not provide good results due to the poor content-based part, but is promising as it allows to investigate on the expansion methods it uses.

Similarly, Mourao et al [9] presented two distinct strategies for retrieval,, i.e... text-based and content-based retrieval of medical articles. The text-based retrieval utilized the Lucene search engine as a retrieval platform. As for data, they used the entire text content of the articles even the image captions. The text-based result is further boosted by applying pseudo-relevance feedback. On the other side, the content-based approach was using a combination of low level features: color histograms, FCTH, and LBP histograms. The content-based retrieval of the medical articles was performed on the basis of the images which the articles contained. Unlike Simpson et al [8], they did not provide a multimodal approach to compare the individual and merged results.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

The approach reported by Herrera [10] also uses the Lucene search engine as a retrieval platform. Their approach did not apply any additional processing of the articles, but included the entire text content in the indexing: title, abstract, paper text (also referred as full-text) and image captions. The interesting note is that it provided best results on the CLEF 2012 workshop for the specified subtask.

Choi et al [11] provided an interesting framework for the medical articles retrieval. Firstly, they focus only on text-based retrieval. Their underlying system is simple in the sens that they index the articles (title, abstract and full-text) with the Indri search engine. In the retrieval phase they use the unigram language model with Dirichlet prior smoothing. The novelty comes into place with the query expansion methods. Here, they use an external bibliographic source, i.e.. MEDLINE. The query is first executed on the MEDLINE database and the MESH terms in the top n documents, which are case reports, are added to the original query. Then, the newly modified query is executed on the medical articles collection. Their approach provided the best results in the CLEF 2013 workshop case-based retrieval subtask.

The presented solutions in the related work, show that the task of medical articles retrieval is a complex one and can be approached from multiple different aspects. Different search engines are used in the presented solutions, but Lucene and Terrier IR, stand out as generic and versatile search engines which can be used in various scenarios. One of the best results were reported with the Terrier IR and it has efficient and effective search methods for large-scale document collections. Hence, we used it in the experiments for this paper. The potential in the query expansion methods also led us to try and investigate whether query expansion will increase the default retrieval performance.

3 Retrieval with query expansion

The related work has provided insight into the state-of-the-art methods for the posed or related of problems. Hence, in this paper we experiments with query expansion methods with the aim to improve the overall medical articles retrieval results.

3.1 MESH term expansion

Medical Subject Headings (MeSH) is a strictly controlled vocabulary with the purpose of indexing journal articles and books in life sciences and frequently used when searching [12]. It is created and maintained by the National Library of Medicine (NLM) in the US. The vocabulary is used by the MEDLINE/PubMed database.

The aim of this approach is to use an external mapping tool, which can analyze the query and extract the MeSH terms associated with it. The existing mapping tools include NLM's Medical Text Indexer (MTI) [13] or MeSH Up [14]. The extracted/detected terms are then added to the original query and it is executed on the medical articles collection.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

3.2 UMLS term expansion

The Unified Medical Language System (UMLS) is a combination of different medical vocabularies [3]. The set provides a mapping structure among the contained vocabularies and in that manner allows for translation among the various terminology systems. It is also perceived as a comprehensive thesaurus and ontology of biomedical concepts. UMLS essentially consists of databases and software tools.

The goal of this approach is similar to the previous one, with the exception that the terms used are UMLS terms. There are also mapping tools that can extract these types of concepts. The most famous being the tool developed by NLM - MetaMap [15]. After the concepts are extracted, they are appended to the query and it is executed over the medical article collection.

3.3 Pseudo-relevance feedback

Query expansion with pseudo-relevance feedback is an industry standard method in information retrieval [16]. The process works by first executing the query on the medical article collections. Once, the results are provided it analyzes the top n documents for m most *informative* terms. The acquired terms are added to the original query and it is again executed on the collection and final results are provided.

4 Experimental setup and evaluation

4.1 Dataset

The data used in the experiments of this paper is provided from the ImageCLEF 2013 [17] collection. The collection contains text and visual data. The subset of the collection meant for case-based retrieval contains 74 654 medical articles (cases), mainly journal articles from PubMed. Each paper is organized in an XML file, which consists of several parts: title, abstract, body text (referred as full-text) and captions from the images associated with the paper. The provided structure is depicted on Figure 1. In the experiments, we also added the MeSH terms associated with article from PubMed.

The collection also provides 35 queries, so that the case-based retrieval can be evaluated appropriately. The queries are short narrative case descriptions. Each text query is accompanied by 2-3 images for multimodal or content-based retrieval experiments. Since the focus of this paper is on text-based medical articles retrieval we use only the text part of the queries. Sample queries:

– **Query 1**. A 50-year-old man with severe right flank pain and hematuria. Renal ultrasound shows a markedly echogenic lesion with a posterior acoustic shadow measuring about 8x10mm in the right kidney.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

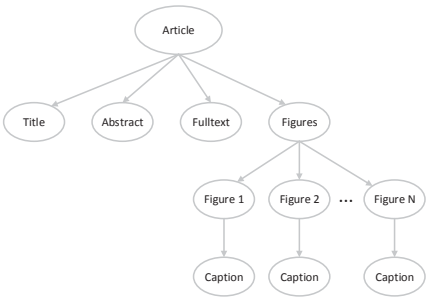

Fig. 1. Diagram of the XML structure of a medical article

- **Query 2**. A 49-year-old woman with a prolapsed mass in the opening of her urethra. Pelvic CT shows a heterogeneously enhanced mass on the female urethra. Pathology shows ramifying papillae, high nuclear/cytoplasmic ratio, and brisk mitotic activity.
- **Query 3**. A 56-year-old woman with Hepatitis C, now with abdominal pain and jaundice. Abdominal MRI shows T1 and T2 hyperintense mass in the left lobe of the liver which is enhanced in the arterial phase.

4.2 Retrieval details

As a retrieval platform, we used the Terrier IR [18] search engine. We used it to index the articles (title, abstract, full-text, image captions and MeSH terms) and to later retrieve them. The articles and text queries are first preprocessed in the indexing and retrieval phase respectively. The preprocessing consists of the following steps: 1. Tokenization using the built-in tokenizer for English; 2. Stopwords removal; 3.Stemming, i.e.. finding the root term using Porter stemmer [19]. We turned to the BM25 [20] weighting model as it is reported as one of the best models used in this context. The model is used to compute a numeric score, which is essentially an estimate of how similar an article is to a give query. The higher the number - the more similar the items. After the score is calculated, the articles are sorted in descending order and returned.

The results are tuned by providing the queries with query expansion. The pseudo-relevance feedback method adds the 4 most informative terms which appear in the top 3 initially returned articles and the terms must appear in a minimum of 2 articles. The MeSH and UMLS expansion methods use mapping tools developed from NLP to find and extract the medical concepts associated with them. For extracting UMLS terms, we used the MetaMap tool. The queries are sent to the tool and it returned UMLS concepts accompanied by a score which determines their relevance. We used all concepts with a score higher thn 10 as the concepts below that are too generic and add no real value. As for the MeSH terms, we used the MTI tool. The MTI tools can accept large chunks of

text (up to 10000 words). We send the queries to that tool and use all MeSH terms that are returned, as they are only related to medical terms.

4.3 Evaluation measures

As evaluation metrics for the experiments, we turned to the standardized ImageCLEF evaluation metrics:

- Mean Average Precision (MAP) - the mean of the average precision scores for each query
- Precision at first 10 (P10) - precision of the first (top) 10 returned articles (cases)
- Precision at first 30 (P30) - precision of the first (top) 30 returned articles (cases)

The total number of documents over which the MAP is processed is 1000, according to the ImageCLEF practices [17]. This means that the approach should return 1000 cases for a given query.

4.4 Experimental questions

The main purpose of this paper was to answer the following questions:

1. *Can the default case-based retrieval be boosted by query expansion?*
2. *Which of our proposed query expansion methods provided the best results?*

To answer the first question, we would compare the results from the default retrieval against all other retrieval runs using query expansion. The second question would be answered, just by comparing the results of the retrieval runs using the query expansion.

5 Results and discussion

We made four types of experiments to try and answer our questions: 1. Baseline (normal) experiment - performs the retrieval with the original query without modification (as is); 2. Retrieval with MeSH concepts - the queries here are extended by adding the extracted MeSH terms; 3. Retrieval with UMLS concepts - in this experiment, the queries are extended with the UMLS terms extracted from them; 4. Retrieval with pseudo-relevance feedback - this experiments expands the queries with the pseudo-relevance feedback method.

The results from the experiments are presented on Table 1. The results show that using query expansion can increase the default performance. The biggest improvement is in the case of the pseudo-relevance feedback method, where a boost is noted in the P10 and P30 metrics. These are very important as they represent the precision for what regular users would experience, since users usually look at the top returned results.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

Table 1. Results from the evaluation of the query modification methods.

	MAP	P10	P30
normal	0.2004	0.2029	0.1381
mesh concepts	0.1556	0.1800	0.1238
umls concepts	0.1625	0.1943	0.1343
pseudo-rel.	0.2005	0.2286	0.1667

It is interesting to note that the baseline retrieval performs second best in all metrics. Implying that the added MeSH and UMLS terms do not add value to the retrieval process, i.e.. they add more false-positives in the top ranked items. This can be due to multiple issues. One issue can be that the queries are not descriptive enough for the mapping tools to extract the appropriate terms. Which means that the extracted terms are not carrying the meaning that is supposed to be attached to the query. That will cause for other not relevant articles to be pushed towards the top ranks.

6 Conclusion

In this paper we provided novel methods for text-based retrieval of medical articles or case-based retrieval. The methods relied on query expansion techniques. We proposed three techniques for query expansion. More precisely, expansion with MeSH terms, expansion with UMSL terms and expansion with pseudo-relevance feedback. The best results were provided with the pseudo-relevance feedback, which is accepted as an industry standard. The expansion with medical terms from standardized vocabularies and tools, still has potential as there is still room to work with other databases (such as MEDLINE).

In the future we plan to implement new methods for retrieval of medical articles and make them available online. We have already developed a publicly available system for medical image retrieval [1]. We plan to add the case-based retrieval functionality to the system, which would use additional medical knowledge bases to perform the retrieval and preferably to include a content-based component, which would also perofrom the retrieval by visual content of the contained images.

Acknowledgments. This work is partially supported by the Faculty of Computer Science and Engineering, Skopje, Macedonia as a part of the project "Scalable Photo Annotation".

References

1. C. Dye, J. C. Reeder, R. F. Terry, Research for universal health coverage, World Health Organization, 2013.

[1] http://194.149.136.27/images/home

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

2. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, D. Johnson, Terrier information retrieval platform, in: Advances in Information Retrieval, Springer, 2005, pp. 517–519.
3. O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (suppl 1) (2004) D267–D270.
4. M. Errami, J. D. Wren, J. M. Hicks, H. R. Garner, etblast: a web server to identify expert reviewers, appropriate journals and similar publications, Nucleic acids research 35 (suppl 2) (2007) W12–W15.
5. Pubget, `http://pubget.com`, accessed: 2015-05-09.
6. C. Peters, Cross language evaluation forum, D-Lib.
7. J. A. Vanegas, J. C. Caicedo, J. E. Camargo, R. Ramos-Pollán, F. A. González, Bioingenium at imageclef 2012: Textual and visual indexing for medical images, in: CLEF (Online Working Notes/Labs/Workshop), 2012.
8. M. M. R. D. D.-F. S. A. Matthew S. Simpson, Daekeun You, G. Thoma, Iti's participation in the 2013 medical track of imageclef, in: CLEF (Online Working Notes/Labs/Workshop), 2013.
9. F. M. Andre Mouruao, J. ao Magalh aes, Novasearch on medical imageclef 2013, in: CLEF (Online Working Notes/Labs/Workshop), 2013.
10. A. G. S. de Herrera, D. Markonis, I. Eggel, H. Müller, The medgift group in imageclefmed 2012., in: CLEF (Online Working Notes/Labs/Workshop), 2012.
11. S. Choi, J. Lee, J. Choi, Snumedinfo at imageclef 2013: Medical retrieval task, in: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes, 2013.
12. C. E. Lipscomb, Medical subject headings (mesh), Bulletin of the Medical Library Association 88 (3) (2000) 265.
13. A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, W. J. Rogers, The nlm indexing initiative's medical text indexer, Medinfo 11 (Pt 1) (2004) 268–72.
14. D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, D. Rebholz-Schuhmann, Mesh up: effective mesh text classification for improved document retrieval, Bioinformatics 25 (11) (2009) 1412–1418.
15. A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.
16. R. Yan, A. Hauptmann, R. Jin, Multimedia search with pseudo-relevance feedback, in: Image and Video Retrieval, Springer, 2003, pp. 238–247.
17. A. G. S. de Herrera, J. Kalpathy-Cramer, D. Demner-Fushman, S. Antani, H. Müller, Overview of the imageclef 2013 medical tasks, Working notes of CLEF.
18. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, C. Lioma, Terrier: A high performance and scalable information retrieval platform, in: Proceedings of the OSIR Workshop, Citeseer, 2006, pp. 18–25.
19. C. Macdonald, V. Plachouras, B. He, C. Lioma, I. Ounis, University of glasgow at webclef 2005: Experiments in per-field normalisation and language specific stemming, in: Accessing Multilingual Information Repositories, Springer, 2006, pp. 898–907.
20. G. Amati, C. J. Van Rijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Transactions on Information Systems (TOIS) 20 (4) (2002) 357–389.