WORD-SPACE APPROACH TO CASE-BASED RETRIEVAL

Ivan Kitanovski
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

Katarina Trojacanec
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

Ivica Dimitrovski
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

Suzana Loshkovska
Faculty of Computer Sciences and Engineering
Skopje, Macedonia

ABSTRACT

This paper presents a strategy for medical articles retrieval, a process also known as case-based retrieval. The proposed strategy is a word-space approach and uses Terrier IR search engine for indexing and retrieval. The medical articles were indexed on different parts of their content to determine which part provides best performance in indexing and retrieval. In the retrieval phase, experiments with six different weighting models were performed. The results show that the best performance is achieved by indexing the whole content of the articles (including the image captions) and by using the BM25 weighting model in the retrieval phase.

I. INTRODUCTION

The subject of this paper is the retrieval of medical articles (cases). Medical information retrieval (IR) is concerned with analysis, organization and retrieval of medical information [1]. It is a vital part of everyday medical practice. It provides the physician with relevant medical article(s) or general data, which can help him gain new insight for the case (s)he is currently working on. This information is essential so the physician can provide better and more thorough diagnosis for the current patient, or to establish a more precise treatment based on previous similar cases [2]. The relevance of the retrieved articles is very important and presents an exciting challenge for many researchers in this domain.

The Cross-Language Evaluation Forum (CLEF) is a self-organized body whose main mission is to promote research, innovation and development of information access systems [3]. CLEF defines two types of medical retrieval: image-based retrieval and case-based retrieval. The task of image-based retrieval consists of retrieving medical images which are relevant to a given textual and/or visual query. Case-based retrieval focuses on retrieving the most relevant medical cases to a given textual and/or visual query. This is a more complex task, but one that is considered closer to the everyday clinical workflow. In this task, a case description, with patient demographics, limited symptoms and test results including imaging studies, is provided (but not the final diagnosis). The goal of case-based retrieval is to retrieve cases (including images) that are most similar to the query case description. Thus, the main problem of case-based retrieval is determining whether a document is relevant to a given query.

Already, there are systems which offer medical articles retrieval like Pubmed [4], Pubget [5], eTBLAST [6] etc. Essentially, they treat the articles as a set of terms (word-space approach) and index/retrieve the documents using key words entered as a query. The query in the case-based retrieval is of narrative nature and is not suited for these systems.

In this paper, a strategy for word-space case-based retrieval is proposed and evaluated. The proposed strategy treats the medical articles like a set of terms and uses techniques for indexing and retrieval implemented by many search engines. The existing techniques are configured and combined to implement the proposed strategy. The evaluation task considers evaluation of a strategy for selection of the parts of the medical articles to be indexed and how this strategy influences the retrieval outcome. Several weighting models are compared to find the most appropriate model for this type of data. The results show that this strategy leads to better retrieval performance.

This paper is organized as follows. The section two presents the related work. In section three, the proposed approach is elaborated. The experimental setup and design are presented in section four. Section five shows the results and discussion. The concluding remarks are presented in the section six.

II. RELATED WORK

Medical articles retrieval is an area of active research and there are many attempts to solve this problem, despite the fact that there are commercial systems offering similar services. Approaches from the word-space domain usually integrate an existing search engine and combine or add features on top of that engine.

Simspon et al [7] use the Essie and Lucene search engines for indexing and retrieval. But, as an additional approach they use query expansion with Google Search API. In the retrieval phase the textual query is put on the API and the top five ranking documents are retrieved. Using special toolkits, like Metamap [8], the three most common concepts associated with the documents are extracted and used for query expansion.

The Medgift group [9] uses the Lucene search engine to index the text from the medical cases and afterwards, to perform the retrieval. The group did not perform any additional parameter optimizations and used the default setup to run their experiments. This is used as a baseline in the case-based retrieval task and it provides the best results for this task.

Another interesting approach is presented by Vanegas et al. [10]. They implement their own version of the Okapi BM25 weighting model using the Python NLP toolkit. Although, they participated in the text-based image retrieval, it is important to note their approach was the best for that task. This shows great promise and potential for possible implementation in the case-based retrieval problem, too.

Stathopoulos et al. [11] also use Lucene search engine for text-based image retrieval. They first pre-process the words for indexing by applying stop-words removal and Porter

stemmer. Afterwards, in the retrieval phase they use the default scoring function of Lucene.

Song et al. [12] use Lucene search engine for indexing and retrieval. They use TF-IDF as a weighting model in retrieval and also implement query expansion using concept mapping tools, like Meshup.

Vahid et al [13] use the Terrier IR search engine for text-based indexing and retrieval of images. The text is pre-processed in the following order: deletion of special characters, stop words removal, token normalization and stemming using Porter stemmer. In the retrieval phase the TF-IDF weighting model is used as the most common and famous weighting model in IR systems. This was among the best performing approaches in its category.

The related work shows many attempts to tackle this problem with word-space approach. Some of the presented approaches do to not try to solve the case-based retrieval directly, but solve the problem of text-based image retrieval. This is similar to case-based retrieval because both are only word-space approaches to a form of text retrieval. Most of the authors use Lucene search engine, although Terrier IR is a search engine which provides efficient and effective search methods for large-scale document collections, too. [1] Our strategy is to utilize the Terrier IR search engine for case-based retrieval, with the attempt to gain better retrieval performance.

III. Proposed Strategy

We propose a strategy that uses a word-space approach to medical articles retrieval i.e. no knowledge about the medical concepts in the medical articles is extracted or analysed to perform the indexing and retrieval. The diagram of the strategy is depicted on Figure 1.

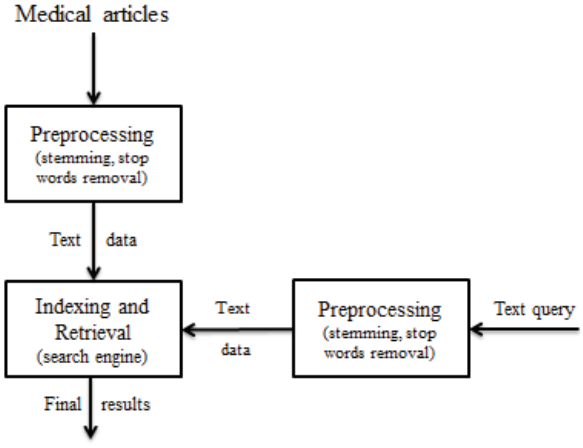

Figure 1: Diagram of the proposed strategy

The medical articles and input queries are first pre-processed. The pre-processing includes several stages. First, the text is cleared from special characters. In this stage, all characters which have no meaning are removed, like punctuation marks or white space. Next, all stop words are removed. These are also words without semantic value, but occur very frequently within the text. The thirds stage is token normalization in which all terms are converted to lower case. The final stage of the pre-processing is stemming. Stemming is process of reducing a word to its base or root. In this stage, Porter stemmer is implemented which is a famous stemmer for removing the commoner morphological endings of words in English [14].

Once the text data is pre-processed the indexing stage begins. During the index stage a classical two-pass indexing is performed [15]. The result of this is stage is an inverted index, which is later used for efficient retrieval.

In the retrieval phase, the system receives queries in the form of textual data as input. The queries are narrative in nature. The queries are subjected to pre-processing in the same way the articles were pre-processed before the indexing was performed. After the queries are pre-processed the retrieval is performed using some weighting model. In this paper, several weighting models were used to compare their performance.

The result of the retrieval stage is a sorted list of medical cases based on their similarity with query.

IV. Experimental Setup and Design

The data for the experiments is provided from the ImageCLEF 2012 [16] collection. The collection contains textual and visual data. This paper focuses on textual data. The case-based collection contains 74 654 medical articles (cases), which are mainly journal articles from PubMed [16]. Each paper consists of several parts: title, abstract, body text (referred as full-text) and captions from the images associated with the paper. The queries for the case-based retrieval are short narrative case descriptions. The collection contains 26 queries in total. The following list presents examples of queries:

Query 1. A 43-year-old man with painless, gross hematuria. Abdominal CT scan revealed a large left renal mass with extension into the left renal pelvis and ureter.

Query 2. A 70-year-old man with a history of alcoholism, now with hemoptysis. Contrast CT shows uniformly enhancing, dilated, tortuous structures surrounding the esophagus. In addition, the liver is small and nodular, and the spleen is enlarged.

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.

For indexing and retrieval Terrier IR [15] was used in this paper. Terrier is a flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents [15]. Terrier implements numerous state-of-the-art indexing and retrieval functionalities, and provides an easy platform for the rapid development and evaluation of large-scale retrieval applications.

Three types of experiments were performed to evaluate the influence of the indexing data on the retrieved outcome. For the first type of experiments, only the title and the abstract of

the medical articles are used for indexing and retrieval. The second types of experiments use the full-text of the medical articles. The third types of experiments use the entire textual data associated with the medical articles i.e. they use title, abstract, full-text and image captions (referred as all-text) for indexing and retrieval.

Regarding the retrieval stage, several weighting models were used to compare their performance on this type of data. The weighting models which were used include: PL2 [17], BM25 [17], BB2 [17], DFR-BM25 [17], TF-IDF [18], DirichletLM [19].

To evaluate the performance of the retrieval different mathematical statistics were calculated: mean average precision (MAP), average of precision at 10 documents retrieved (P10), average of precision at 20 documents retrieved (P20), average precision after R (= number of relevant for topic) documents retrieved (Rprec) [20].

V. Experimental Results and Discussion

The results of the experiments are presented in the following order: Table 1 presents the results from the retrieval when the abstract and title of the medical articles are used for indexing; Table 2 presents results when the full-text is used for indexing; Table 3 depicts the results when the all-text is used for indexing of the medical cases.

Table 1: Results when using title and abstract.

	MAP	Rprec	P10	P20
BM25	0.0755	0.0735	0.0870	0.0630
DFR_BM25	0.0754	0.0735	0.0870	0.0630
TF_IDF	0.0736	0.0709	0.0826	0.0630
BB2	0.0719	0.0656	0.0719	0.0656
PL2	0.0764	0.0697	0.0870	0.0609
DirichletLM	0.0709	0.0741	0.0739	0.0741

Table 2: Results when using full-text.

	MAP	Rprec	P10	P20
BM25	0.0965	0.1181	0.1043	0.0848
DFR_BM25	0.0959	0.1036	0.1000	0.0848
TF_IDF	0.0950	0.1036	0.1000	0.0826
BB2	0.0841	0.1068	0.1087	0.0717
PL2	0.0999	0.1118	0.1130	0.0848
DirichletLM	0.0664	0.0835	0.0696	0.0500

It can be noted that the best overall performance is gained using the BM25 weighting model when the medial articles are indexing using all-text. The best result is 0.1816 MAP, which is better compared to the best performing results on this data set by Medgift of 0.1690 [9]. The difference in the results is mostly due to the used weighting model. The Medgift team uses the TF-IDF model, which is shown to be less effective on this type of data than BM25 [21]. Also, Medgift does not take into account the captions of the images associated with the medical articles. The captions add more information, which is used in this paper in the all-text experiment.

There is a huge difference in MAP between the retrieval using full-text to index the data and when the medical articles are indexed with all-text. The reason for this great difference is due to the nature of the queries. The queries usually contain information about the patient's age, gender, symptoms and information about the modality of the patient's medical images. The captions of images associated with the medical images usually contain most of that information. Example of a caption:

An 84-year-old male with low back pain and right L4 radicular symptoms to the ankle worsened with walking with symptomatic improvement with a right L4-5 transforaminal epidural steroid injection. (a) Right sagittal view of a T2-weighted MRI of the lumbar spine. Note the multilevel degenerative changes and the foraminal stenosis at L4-5 related to disc bulge and facet hypertrophy. (b) Axial view of a T2-weighted MRI through L4-5. Note severe central stenosis on imaging, though symptomatically, he described right L4 radicular symptoms and thus a transforaminal route was chosen

The caption contains a vast amount of data. It contains gender, age, symptoms even image modality. Not all image captions are so descriptive, but usually they contain relatively large quantities of data. Probably the reason for the great advantage of the all-text approach over the full-text approach is due to the inclusion of this additional data.

Table 3: Results when using all-text.

	MAP	Rprec	P10	P20
BM25	0.1818	0.1757	0.1522	0.1391
DFR_BM25	0.1816	0.1767	0.1522	0.1413
TF_IDF	0.1805	0.1662	0.1522	0.1326
BB2	0.1598	0.1604	0.1435	0.1326
PL2	0.1780	0.1861	0.1478	0.1370
DirichletLM	0.1811	0.1744	0.1652	0.1283

VI. Conclusion

A word-space approach to case-based retrieval was presented in this paper. The results show that indexing the medical

articles using all-text provides the most accurate retrieval. The best results present a 0.1818 MAP, which are the best results for this dataset.

Medical articles retrieval is an important part in everyday medical practice. A better understanding and solving of this problem can aid medical practitioners in their job or it can broaden their knowledge and help them rethink their medical decisions. There are live systems which are currently being used for medical articles retrieval, but there is still room for more improvement of their retrieval performance. One of the possible future steps would be to use prior knowledge about the medical content to expand the indexing space of the medical articles.

ACKNOWLEDGEMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University".

REFERENCES

[1] I. Ounis, G. Amati, and V. Plachouras, “Terrier information retrieval platform,” *Advances in Information Retrieval*, pp. 517-519 2005.

[2] J. Kalpathy-Cramer and H. Müller, “Overview of the CLEF 2011 medical image classification and retrieval tasks”, in *Proceedings of Cross-Language Evaluation Forum*, 2011.

[3] C. Peters, “Cross language evaluation forum,” *D-Lib*, 2000.

[4] A. D. Lindberg, B.L. Humphreys,and A.T. McCray, “The Unified Medical Language System,” *Methods of information in medicine*, vol.32, no.4, pp. 281, 1993.

[5] http://pubget.com/, March 2013.

[6] M. Errami, J.D. Wren, J.M. Hicks, and H.R. Garner, “eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications,” *Nucleic acids research*, vol. 35, no. 2, 2007.

[7] M.S. Simpson, D. You, M.M. Rahman, D. Demner-Fushman, S. Antani and G. Thoma, “ITI’s participation in the ImageCLEF 2012 medical retrieval and classification tasks”, *Working Notes of CLEF*, 2012.

[8] A.R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In *Proceedings of the AMIA Symposium*, p. 17, 2001.

[9] A.G. de Herrara, D. Markonis, I. Eggel, H. Müller, “The medGIFT Group in ImageCLEFmed 2012”, *Working Notes of CLEF*, 2012.

[10] J. Vanegas, J. Caicedo, J. Camargo, R. Ramos-Pollan, F. Gonzalez, “Bioingenium at ImageCLEF 2012: Textual and Visual Indexing for Medical Images”, *Working Notes of CLEF*, 2012.

[11] S. Stathopoulos, N. Sakiotis, T. Kalamboukis, “IPL at CLEF 2012 Medical Retrieval Task”, *Working Notes of CLEF*, 2012.

[12] W.Song, D. Zhang, J. Luo, “BUAA AUDR at ImageCLEF 2012 Medical Retrieval”, *Working Notes of CLEF*, 2012.

[13] A. H. Vahid, A. Alpkocak, R. G. Hamed, N. M. Ceylan, and O. Ozturkmenoglu, “DEMIR at ImageCLEFMed 2012: Inter-modality and Intra-Modality Integrated Combination Retrieval”, *Working Notes of CLEF*, 2012.

[14] C. Macdonald, P. Vassilis, B. He, C. Lioma, and I. Ounis. "University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming", *Accessing Multilingual Information Repositories*, pp. 898-907, 2006.

[15] http://terrier.org/, March 2013.

[16] H. Müller, et al. "Overview of the ImageCLEF 2012 medical image retrieval and classification tasks," *Working Notes of CLEF*, 2012.

[17] G. Amati and C.J. Van Rijsbergen, “Probabilistic models of information retrieval based on measuring the divergence from randomness”, *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357-389, 2002.

[18] D. Hiemstra, “A probabilistic justification for using tf× idf term weighting in information retrieval”, *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 131-139, 2000.

[19] C. Zhai, and J. Lafferty, “A study of smoothing methods for language models applied to information retrieval”, *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 179-214, 2004.

[20] C.D. Manning, P. Raghavan and H. Schütze, *Introduction to information retrieval (Vol. 1)*. Cambridge: Cambridge University Press, 2008.

[21] A. Alpkocak et al., “DEMIR at ImageCLEFMed 2011: Evaluation of Fusion Techniques for Multimodal Content-based Medical Image Retrieval”,*Working Notes of CLEF*, 2011.