Medical Image Retrieval: ISSR at CLEF 2009

Waleed Arafa , Ragia Ibrahim
Department of Computer & Information Sciences
Institute of Statistical Studies and Research (ISSR), Cairo University, Egypt
{waleed_arafa,ribrahim}@issr.cu.edu.eg

Abstract

This paper represents the first participation of the Institute of Statistical Studies and Research at Cairo University group in CLEF 2009-Medical image retrieval track. Our system uses Lemur toolkit for text retrieval. The main objective is to carry out retrieving medical image depending on associated image text. We experimented with different text features such as article title, image caption and the article paragraph(s) denoting to the image. We propose a simple and effective extraction method to find relevant paragraphs based on the structure of HTML files. Automatic translation of queries in different languages other than collection language is also experimented. In this paper the results of 9 runs are presented in order to compare retrieval based on different text features and the effect of stop word lists and the use of relevance feedback.

Categories and subject descriptors
H. Information Systems; H.3 Information Storage And Retrieval; H.3.1 Content Analysis and Indexing; H.3.4 Systems and Software;

Keywords
Information retrieval, textual retrieval, medical retrieval, text extraction, relevance feedback, linguistic processing.

Introduction

Medical image retrieval is a challenge for Cross Language Information Retrieval (CLIR) as well as for information retrieval. Since medical image annotations, associated text can be written in more than a single language (multilingual), and the language used to express the associated texts or textual queries should not affect retrieval, then it should be searchable in other languages. [1]
Other limitation in medical image retrieval is the lake of standardization for metadata associated with medical image [2]

Müller, et al. (2004) reported that medical image retrieval information can be used in research, diagnosis and teaching. Accordingly, researchers can benefit from such a system through the use of visual image query.
In diagnosis, the need to search for specific patient by name still required, but to reach the medical information system goal is highly recommended to retrieve images that support specific diagnosis.
In teaching, it can aid lecturers and students to visually inspect (study) results found by browsing educational images repositories [3].
This paper presents system description, experiment results, analysis, and finally future work.

In our experiments, we focus on different selected textual features .We examine and compare the use of title, image caption, and extracted paragraph. In addition, since stop word list have significant effect on information retrieval system, removing too many will hurt effectiveness and not removing may cause problems in ranking [4]. On CLIR retrieval, crossing issue can be done by using one of two ways rather translating queries (source language) into target language (document language) or translate documents to query language [5]. Query translation advantages are: less expensive than document translation from computations point of view and effective with short quires. Used approaches for query translation include Machine translation (MT), Corpus based methods and Dictionary based Methods [6, 7].

Google translation is based on statistical machine translation. It feeds its translation system with billions of words of text, parallel text containing examples of human translations among the languages and monolingual text in the target language [8].

System Description

The ISSR IR's Image Retrieval system was built with ready-made components. For text retrieval Lemur toolkit has been used. The Lemur Toolkit is an open-source toolkit designed to facilitate research in language modeling and information retrieval. Lemur supports a wide range of industrial and research language applications such as ad-hoc retrieval, site-search, and text mining [9]. Indri search engine for language modeling and indexing has been used. Indri search engine is integrated with Lemur toolkit.

We used our own text extraction algorithm to add extra annotation to images before indexing, so that more relevant terms are added for the image to increase its probability to be retrieved. Following is the outline of the algorithm:

- ImageCLEFMed track organizer distributed an XML file, contains image unique name. Image caption, article title, image URL and article URL.
- Extract URL addresses of all images from the XML file to create list of URLs. Download only HTML files referred to by the URLs since HTML files have explicit tags for paragraphs and other elements of the document.
- For each image, the HTML downloaded file is used to extract the paragraph(s) relevant to the image. Extraction process concentrates on HTML structure e.g. <P> and syntax to get each image related paragraph(s) from the article using the following steps:
    - Images are mentioned in paragraph by many ways such as: Fig/Figure n, where n if the figure number, or Figs/Figures n_1, n_2 … n_n to refer to many figures. Image name may contain letter after the number e.g. Fig 1b. So regular expression is used to search for figure reference in all paragraphs.
    - If the above failed, check if the image name contains letter after figure number, search using figure number only without the letter.
    - If the above failed, search in HTML tags in addition to normal text; since we noticed some files contain links to the figure without explicitly mentioning image name.
    - If the above failed, search in the text about the word Figure or Fig only without any number, since some articles have only one figure and refer to it without any number.
    - HTML tags don't distinguish between the text under the image, i.e. the image caption, and the normal text in the article, so check whether found paragraph is the image caption, ignore it in this case.
    - If the found paragraph(s) has other terms that are not found in the image caption add them to the image file. This is done after normalizing the caption and the paragraph.

To determine the best methods for indexing and retrieval available with the Lemur toolkit, 2008 collection with title and caption only has been used since it is proved they give better performance than adding abstract or the whole article to the title and caption in CLEF 2008 [11].

Different stemmers, indexing methods, and language models are tested. Best result is obtained by using Indri indexing and Okapi language model. Then these setting are tested with a free stop word list published by The Information Retrieval Group [10], it gives better performance, we also updated the stop word list to add common terms found in the queries that are not relevant to the medical domain such as 'show me', 'image', and 'photo', this slightly improved the performance. Pseudo relevance feedback included in Lemur toolkit is applied; many feedback settings are checked to determine the best one.

The above setting are applied after adding the relevant paragraph(s) into 2008 collection of title and caption, and there was about 30% increase in mean average precision (MAP).

Experimental Results

Table 1 represents submitted runs. We have submitted nine runs and all runs are automatic. The evaluation measures reported in this paper are standard measures computed with the trec_eval: MAP (Mean Average Precision), recall, and the R-precision.

All runs are for text only, 5 of them are for English queries, 2 for French and 2 for German. Here is the description of the first five runs for English queries.

- Run 1 uses only title and caption as text features.
- Run 2 uses only title, caption and added paragraph(s) as text features.
- Run 3 uses only title and caption as text features with pseudo relevance feedback.
- Run 4 uses only title, caption and added paragraph(s) as text features with pseudo relevance feedback.
- Run 5 uses only title, caption and added paragraph(s) as text features with pseudo relevance feedback with updated stop word list.

Two runs are submitted for French queries after using automatic Google Translation to translate them into English, they are:

- Run 6 uses only title, caption and added paragraph(s) as text features with updated stop word list.
- Run 7 uses only title, caption and added paragraph(s) as text features with updated stop word list with pseudo relevance feedback.

Two runs are submitted for German queries after using automatic Google Translation them into translate into English, they are:

- Run 8 uses only title, caption and added paragraph(s) as text features with updated stop word list.
- Run 9 uses only title, caption and added paragraph(s) as text features with updated stop word list with pseudo relevance feedback.

#	Run Name	Language	<P>	PRF	USWL	MAP	R-Prec	Recall
1	ISSR_Text_1	English	No	No	No	**0.3499**	**0.3827**	0.7269
2	ISSR _Text_2	English	Yes	No	No	0.3315	0.3652	0.7485
3	ISSR_Text_1rfb	English	No	Yes	No	0.277	0.3014	0.6791
4	ISSR _Text_4	English	Yes	Yes	No	0.2672	0.2916	0.7341
5	ISSR _Text_5	English	Yes	Yes	Yes	0.2692	0.2945	0.7358
6	ISSR _Text_FR_1	French	Yes	No	Yes	0.2951	0.3354	0.6956
7	ISSR _Text_FR_2	French	Yes	Yes	Yes	0.3111	0.3338	**0.7667**
8	ISSR _Text_DE_1	German	Yes	No	Yes	0.1997	0.2314	0.6808
9	ISSR _Text_DE_2	German	Yes	Yes	Yes	0.1981	0.2197	0.6088

Table 1. Results of ISSR nine submitted runs. All of them use textual features only for retrieval. They are all automatic, no manual feedback was involved (AUTO). <P>: added paragraphs, PRF: Pseudo Relevance Feedback, USWL: updated stop word list, MAP: mean average precision, R-Prec: R-Precision.

Results and Analysis

Results described in table 1 show that best result obtained when image's caption and title features are only used. Using paragraphs in addition to title and caption decreased the performance as opposed to the results of the same experiment on 2008 collection. Recall is increased after adding paragraph (run 2 more than run 1) and slightly

increased after updating the stop word list (run 5 more than run 4), this results are similar to the results of 2008 collection.

The decrease in MAP after adding paragraphs is surprising and not expected, but careful analysis of the results of each query shows an improvement in recall as expected since more relevant terms are added to the image annotation so more relevant images are retrieved. The proposed method significantly increases recall by 60% of the queries, 28% decreased and 12 % unchanged as shown in Table 2.

Measure / Effect	MAP		R-prec		Recall	
	Queries	**Ratio**	**Queries**	**Ratio**	**Queries**	**Ratio**
Increased	9	36%	11	44%	15	**60%**
Same	0	0%	3	12%	3	12%
Decreased	16	64%	11	44%	7	28%

Table 2. Number of queries affected by adding paragraphs to the title and image caption for MAP, R-precision and recall for the 25 English queries.

This unexpected behavior because the additional text has a lot of noise terms, even worth, some paragraphs mention many figures. This caused the similarity between many documents and the query decrease so that they get low rank in the retrieved list; this is the reason of decreased MAP since its value depends on document ranking.

But this noise did not decrease retrieval MAP on 2008 collection for two reasons: firstly, a significant number of images in 2008 dataset don't have HTML file. So, it had no additional text. Secondly, around 2% of the images didn't have captions at all. By applying the proposed technique, annotations to these images are added, so a better retrieval for these images is shown.

Pseudo relevance feedback (PRF) is used in runs ISSR _Text_2 and ISSR_Text_1rfb. PRF is considered a successful simple query expansion technique where most frequent words in top k documents used to expand query terms. In our case, first ranking doesn't include relevant documents, or if it does, it has a lot of noise because of added paragraphs; so many irrelevant terms are added to the modified queries.

For multilingual retrieval task, French and German queries are translated using Google online machine translation [8]. French translated queries increased the MAP than original English queries by about 15% (run 7 more than run 5). On the other hand German quires decreased the MAP by 26% (run 9 less than run 5). However, this result is considered acceptable for statistical automatic translation.

Conclusion and Future Work

A simple syntax-based technique to add relevant text to image annotation is proposed; and this technique is tested on image retrieval using Lemur toolkit. The results show that it is a promising approach. We intend to enhance this approach using semantic extraction methods such as shallow NLP techniques or statistical approaches to extract only relevant sentences from the paragraph denoting the image instead of adding the whole paragraph in order to reduce noise terms.

Acknowledgments

We acknowledge ImageCLEFMed track organizer for providing 2008 data set.

References

[1] http://ir.shef.ac.uk/imageclef/ , last visited September 2007.
[2] Hersh W., Muller H. *Image Retrieval in Medicine: The ImageCLEF Medical Image Retrieval Evaluation*. Bulletin, February/March 2007.

[3] Henning Müller, Nicolas Michoux, David Bandon, Antoine Geissbuhler. *A review of content-based image retrieval systems in medicine - clinical benefits and future directions*, International Journal of Medical Informatics, volume 73, pages 1-23, 2004.
[4] Croft B., Metxler D., Stohman T. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.
[5] T. Tuomas. *Comparable Corpora in Cross-Language Information Retrieval*. University of Tampere, Faculty of Information Sciences, 2008.
[6] A. Pirkola. *The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval*. SIGIR'98 Cross-language Information Retrieval 1998.
[7] Djoerd Hiemstra. *Using language models for information retrieval*. Ph.D. Thesis, Centre for Telematics and Information Technoloy, 2001.
[8] Franz Josef. *Statistical Machine Translation: Foundations and Recent Advances*. Google, Inc. 2005.
[9] *The Lemur Toolkit for Language Modeling and Information Retrieval*. http://www.lemurproject.org, last visited Aug 2009.
[10] *IR Linguistic Utilities*. http://ir.dcs.gla.ac.uk/resources/linguistic_utils, last visited Jul 2009.
[11] M.C. Diaz-Galiano, M.A. Garcia-Cumbreras, M.T. Martin-Valdivia, L.A. Urena-Lopez, A. Montejo-Raez. *SINAI at ImageCLEFmed 2008*.