Towards A Better Performance for Medical Image Retrieval Using An Integrated Approach

Zheng Ye[1,2], Xiangji Huang[1], Hongfei Lin[2]
[1]Information Retrieval and Knowledge Managment Lab, York University, Toronto, Canada
[2]Information Retrieval Lab, Dalian University of Technology, Dalian, China
{yezheng,jhuang}@yorku.ca, hflin@dlut.edu.cn

Abstract

In this paper, we propose an integrated approach for medical image retrieval. In particular, we present a series of experiments in medical image retrieval task. There are three main goals for our participation of this task. First, we will test traditional well-known weighting models used in text retrieval domain, such as BM25, TFIDF and Language Model (LM), for context-based image retrieval. Second, we will evaluate statistical-based feedback models and ontology-based feedback models. Third, we will investigate how content-based image retrieval can be integrated with these two basic technologies of traditional text retrieval. The experimental results have shown that 1) traditional weighting models can work well in context-based medical image retrieval task especially when the parameters are tuned properly; 2) statistical-based feedback models can improve the retrieval performance when a small number of documents are used; however, the medical image retrieval can not benefit from ontology-based query expansion; 3) the retrieval performance can be slightly boosted by integrating content features.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software;

General Terms

Measurement, Performance, Experimentation

Keywords

CBIR, Visual and Textual Retrieval, Weighting Model, Pseudo Relevance Feedback, Ontologies, MeSH

1 Introduction

This is the first year that we participate in ImageCLEF campaign. Specifically, we participate in Medical Image Retrieval for the retrieval of similar images task and only focus on English language.

The data set used contains all images from articles published in Radiology and Radiographics including the text of the captions and a link to the html of the full text articles. Over 70,000 images are included in the dataset of Medical Image Retrieval 2009. More information for the dataset and topics can be found in [4].

For the first year of the participation in ImageCLEFmed task, we first test traditional well-known weighting models used in text retrieval domain, such as BM25, TFIDF and LM, for context-based image retrieval. Since the text context information (captions) is very short, which is different from traditional adhoc text collections, it is necessary to test and adapt traditional weighting models for this particular task. Second, on the basis of the baseline results, we use statistical-based pseudo relevance feedback and ontology-based (MeSH [1]) query expansion approaches to enhance the retrieval performance. Finally, we note that many images share the same context text for comparison reason in an article, but the circumstance is always that only one of these images is what we are looking for. So it is impossible for us to filter other images with the same context text using only the context-based image retrieval technologies. We explore different image content features to enhance context-based image retrieval technologies.

The remainder of this paper is organized as follows. In section 2, we describe the comparisons of basic retrieval models. In section 3, we present the experimental results for statistical-based and ontology-based query expansion. In section 4, we propose an integrated approach for medical image retrieval. In section 5, we conclude the paper with a discussion of our findings and a look at future work.

2 Weighting Models

In the previous medical image retrieval tasks, a number of different information retrieval (IR) toolkits, such as Lemur, Jirs and Lucene, are used as the basic retrieval systems for context-based medical image retrieval. However, there is no systematic comparison of different weighting models for ImageCLEFmed task. In addition, it is not clear that whether the default parameters in these models empirically tuned for traditional adhoc datasets are optimal.

In this paper, we have made comparisons for four well-known weighting models: BM25 [3], JM - LM [6], TFIDF and DFR_In_expB2 [2]. The corresponding weighting functions are as follows.

- BM25

$$\omega = \frac{(k_1+1)*tf}{k_1*((1-b)+b*dl/avdl)+tf} * \log\frac{N-n+0.5}{n+0.5} * \frac{(k_3+1)*qtf}{k_3+qtf} \tag{1}$$

- JM - LM

$$\omega = (1+\frac{\mu}{1-\mu}*\frac{tf*FreqTotColl}{l*F_t}) \tag{2}$$

- TFIDF

$$\omega = qtf*\frac{k_1*tf}{tf+k_1*(1-b+b*\frac{dl}{avdl})}*\log(1+\frac{N}{n}) \tag{3}$$

- DFR

$$\begin{aligned} \omega &= TF*qtf*NORM*log_e(\frac{N+1}{n\_exp}) \\ TF &= tf*log_2(1+avdl/dl) \\ NORM &= (tf+1)/(df*(TF+1)) \\ n\_exp &= idf*(1-e^{-f}) \\ f &= qtf/df \end{aligned} \tag{4}$$

[1] http://www.nlm.nih.gov/mesh/

where w is the weight of a query term, N is the number of indexed documents in the collection, n is the number of documents containing the term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined), K equals to $k_1 * ((1-b) + b * dl/avdl)$.

In our experiments, the values of k_1, k_2, k_3 and b in the BM25 function are set to be 1.2, 0, 8 and 0.75 respectively; the value of μ is set to 0.15. In addition, the image context texts are preprocessed in the same way for all experiments in order to make reasonable comparison. We use blank delimiter to separate words for indexing and searching and stopwords are removed. Beside these two simple steps, no further technologies have been used.

Table 1: Performance of best offcial runs of each group for textual retrieval

Runs	MAP
LIRIS_maxMPTT_extMPTT	0.4293
sinai_CTM_t	0.3795
york.In_expB2c1.0^o	0.3685
ISSR_text_1	0.3499
ceb-essie2-automatic	0.3484
deu_run1_pivoted	0.3389
clef2009	0.3362
ISSR_Text_2	0.3315
BiTeM_EN	0.3206

Table 2: A comparison of four weighting models: BM25, JM - LM, TF_IDF and DFR

Runs	BM25^o	JM - LM	TF_IDF	DFR
MAP	0.3515	0.3444	0.3608	**0.3730**

Table 3: Performance BM25 model with different parameter (b) settings

b	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MAP	0.3560	0.3554	**0.3586**	0.3577	0.3553	0.3543	0.3515	0.3503	0.3490	0.3450

Table 2 presents the top ten official results. The third run marked by superscript 'o' is our best official textual run. Thereafter, all our official runs are marked in the same way. From table 2, we can see that DFR weighting model has achieved the best performance under default setting.

As for the parameters tuning, we test different settings for b in BM25 and μ in JM-LM. From table 3, BM25 model works steadily. For JM-LM model, as the increase of parameter μ, the performance increases significantly. When μ takes the value of 0.9, JM-LM model outperforms DFR model. Although JM-LM model is not as steady as BM25 model, it is still promising if the parameter can be properly tuned.

3 Query Expansion

3.1 Query expansion with the Bose-Einstein distribution

The pseudo relevance feedback method used in our experiments is DFR-based weighting model described in [2]. The basic idea of these term weighting models for query expansion is to measure the divergence of a term's distribution in a pseudo relevance set from its distribution in the whole

Table 4: Performance JM-LM model with different parameter (μ) setting

μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
MAP	0.3401	0.3486	0.3542	0.3603	0.3677	0.3734	0.3774	0.3796	**0.3817**	0.0390

collection. The higher this divergence is, the more likely the term is related to the query topic. We use Bo1 weighting model in this set of experiments. The Bo1 term weighting model is based on the Bose-Einstein statistics. Using this model, the weight of a term t in the *exp_doc* top-ranked documents is given by:

$$w(t) = tf_x \cdot \log_2 \frac{1 + P_n}{P_n} + \log_2(1 + P_n) \tag{5}$$

where *exp_doc* usually ranges from 3 to 10 [2]. Another parameter involved in the query expansion mechanism is *exp_term*, the number of terms extracted from the *exp_doc* top-ranked documents. *exp_term* is usually larger than *exp_doc* [2]. P_n is given by $\frac{F}{N}$, F is the frequency of the term in the collection, and N is the number of documents in the collection. tf_x is the frequency of the query term in the *exp_doc* top-ranked documents.

The main goal of this set of experiments is to investigate how many top documents and terms should be used for query expansion. For the limitation of space, we only present experimental results on the basis of BM25 model.

Table 5: MAP Performance of Query Expansion – baseline MAP=0.3730

docs/terms	5	10	20	30	50	70	100
5	0.3901	0.3940	0.3947	0.3961	0.3958	0.3954	**0.3963(6.25%)**
10	0.3576	0.3622	0.3643	0.3670	0.3672	0.3688	0.3683
20	0.3443	0.3533	0.3520	0.3526	0.3541	0.3561	0.3613
30	0.3371	0.3415	0.3377	0.3401	0.3434	0.3432	0.3448
50	0.3388	0.3405	0.3345	0.3351	0.3372	0.3379	0.3391

Form table 5 we can see, in general, the performance can be boosted if we can set the parameters properly. When the number of documents for query expansion increases from 5 to 10, the performance drops quickly. The results in table 5 suggests that only a very small number documents are useful for query expansion in context-based medical image retrieval task.

3.2 Query Expansion with MeSH Ontology

In medical domain, terms are highly synonymous and ambiguous. This motivates us to investigate using ontology to expand the original query terms.

The Medical Subject Headings (MeSH) is a thesaurus developed by the National Library of Medicine. MeSH contains two organization files, an alphabetic list with bags of synonymous and related terms, and a hierarchical organization of descriptors associated to the terms. A term is composed by one or more words.

We have used the longest match approach to recognize the MeSH terms in a query. In particular, if all the words of a term are in the query, we add our synonymous terms to the query. To compare the words of a particular term and those of the query, we rst put all the words in lowercase and we do not remove stop words. In order to reduce the number of terms that could expand the query, only three categories of MeSH terms (A: Anatomy, C: Diseases, E: Analytical, Diagnostic and Therapeutic Techniques and Equipment) have been used for query expansion. Table 3.2 presents the MeSH-based query expansion results under four different models.

Unfortunately, MeSH-based query expansion approach does not work well as we expected. Our conjecture is that MeSH-based query expansion may also bring negative terms into query, especially the abbreviation terms. In addition, when the JM-LM is used as the basic retrieval

Table 6: MAP Performance of MeSH-based Query Expansion

Runs	BM25	JM - LM	TF_IDF	DFR
Not-FB	0.3515	0.3444	0.3608	0.3730
MeSH-FB	0.3458^o	0.3056	0.3529	0.3685^o

model, the performance drops remarkably. This is another evidence that the performance of JM-LM is not steady.

4 An Integrated Approach

Content-based Image Retrieval (CBIR) systems enable users to search a large image database by issuing an image sample, in which the actual contents of the image will be analyzed. The contents of an image refer to its features – colors, shapes, textures, or any other information that can be derived from the image itself. This kind of technology sounds interesting and promising. The key issue in CBIR is to extract representative features to describe an image. However, this is a very very difficult research topic.

According to ImageCLEFmed conference notes [5], CBIR always performs poorly , while context based image retrieval can always achieve good performance in terms of MAP measurement. However, content features are also needed, especially when context information is not easy to obtain or a number of images share the same context. This motivates us to combine these two technologies and give a relatively lower weight to CBIR approaches. In particular, we explore three representative features for medical image retrieval.

1. **Color and Edge Directivity Descriptor (CEDD):** is a low level feature which incorporates color and texture information in a histogram [1].

2. **Tamura Histogram Descriptor:** features coarseness, contrast, directionality, linelikeness, regularity, and roughness. The relative brightness of pairs of pixels is computed such that degree of contrast, regularity, coarseness and directionality may be estimated [7].

3. **Color Histogram Descriptor:** Retrieving images based on color similarity is achieved by computing a color histogram for each image that identifies the proportion of pixels within an image holding specific values (that humans express as colors). Current research is attempting to segment color proportion by region and by spatial relationship among several color regions. Examining images based on the colors they contain is one of the most widely used techniques because it does not depend on image size or orientation.

The final rank list is attained by merging the context-based retrieved score ($S_{context}$) and content-based similarity score ($S_{content}$). In particular, we use linear combination. The formula is described as follows.

$$score = (1 - \lambda) * S_{context} + \lambda * S_{content} \tag{6}$$

Table 7: MAP performance of integrated approach (BM25 basic model)

Runs	BM25	CEDD	Tamura	Color
$\lambda = 0.1$	0.3515	0.3515	0.3514	0.3529
$\lambda = 0.2$	0.3515	0.3552	0.3544	0.3524
$\lambda = 0.3$	0.3515	0.3616	0.3544	0.3516

From table 7, we can see that retrieval performance can be slightly boosted by integrating content features. Among these three features, CEDD can improve the performance most. However, more representative features are needed to be developed.

5 Conclusions

In this study, we first evaluate four well-known weighting models for context based medical image retrieval. The performances of the four weighting models are comparable, but DFR weighting model works best under default settings. JM-LM model is not steady for this task, but if the parameter can be tuned properly, it is still promising. Second, we investigate query expansion technologies for this task. In general, statistical-based QE method outperforms ontology-based methods. The experimental results also suggests that only a small number of top ranked documents are useful for statistical-based QE method. Ontology-based methods sound interesting and useful, however the actual performance is not good. More sophisticated processing for this kind of methods is needed. Finally, we explore three content features for content-based medical image retrieval. The experimental results have shown that retrieval performance can only be slightly improved. The current features extracted from images may not be representative enough to capture the characteristics of images. Better features are required to improve CBIR.

In the future work, we will work on the following two directions. First, we will use data-driven approaches to choose optimal parameters for statistical-based QE. Second, we will explore the correlation of different content features of images. In addition, if more features can be integrated into medical image retrieval properly, we believe the retrieval performance can be further improved.

6 Acknowledgements

This research is jointly supported by NSERC of Canada, the Early Researcher/Premier's Research Excellence Award, Natural Science Foundation of China (No. 60373095 and 60673039) and the National High Tech Research and Development Plan of China (2006AA01Z151).

References

[1] M. Vincze A. Gasteratos and J.K. Tsotsos. Cedd: Color and edge directivity descriptor. a compact descriptor for image indexing and retrieval. *ICVS*, pages 312–322, 2008.

[2] G. Amati. Probabilistic models for information retrieval based on divergence from randomness. *PhD thesis, Department of Computing Science, University of Glasgow*, 2003.

[3] Micheline Hancock-Beaulieu, Mike Gatford, Xiangji Huang, Stephen E. Robertson, Steve Walker, and P. W. Williams. Okapi at trec-5. In *Text REtrieval Conference (TREC) TREC-5 Proceedings*, 1996.

[4] Ivan Eggel-Steven Bedrick Sad Radhouani Brian Bakke Charles Kahn Jr. William Hersh Henning Mller, Jayashree Kalpathy-Cramer. Overview of the clef 2009 medical image retrieval track. In *CLEF working notes 2009, Corfu, Greece*, 2009.

[5] Deselaers T.-Kim E. Kalpathy-Cramer J. Deserno T.M. Clough P. Mller, H. and W. Hersh. Overview of the imageclefmed 2007 medical retrieval and annotation tasks. In *Working Notes of the 2007 CLEF Workshop*, 2008.

[6] Zhaohui Zheng Donald Metzler Ruiqiang Zhang, Chang Yi and Jianyun Nie. Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, 2009.

[7] Shunji Tamura, Hideyuki Mori and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8:460–473, 1978.