LABERINTO at ImageCLEF 2011 Medical Image Retrieval Task

Jacinto Mata, Mariano Crespo, Manuel J. Maña

Dpto. de Tecnologías de la Información. Universidad de Huelva
Ctra. Huelva - Palos de la Frontera s/n. 21819 La Rábida (Huelva)
{jacinto.mata, mariano.crespo, manuel.mana}@dti.uhu.es

Abstract. This paper shows the experimentation and the results obtained for LABERINTO research group at the ImageCLEF 2011 medical task. We focus our work on image retrieval based on textual information related to the image. The initial hypothesis is that query expansion could improve the effectiveness of image retrieval systems. In this proposal, three different types of indexes were built and several information elements contained in MeSH ontology were used to expand the queries. The experiments carried out show that the expansion strategies using the MeSH ontology obtain good results for this task.

Keywords: Text-based image retrieval, medical domain, query expansion, ontologies, MeSH.

1 Introduction

This paper describes the contribution of the LABERINTO research group in its first participation at the Medical Image Retrieval task [1].

This task of ImageCLEF 2011 uses a subset of PubMed Central[1]. This year, the organization proposed three types of subtasks: *Modality Classification*, *Ad-hoc Image-based Retrieval* and *Case-based Retrieval*. We are particularly interested in the *Ad-hoc Image-based Retrieval*. This is the classic medical retrieval task, similar to those organized in 2005-2010. Participants will be given a set of 30 textual queries with 2-3 sample images for each query. The queries will be classified into textual, mixed and semantic, based on the methods that are expected to yield the best results.

In this work, we have used the MeSH[2] [2] ontology for query expansion in order to improve our medical image retrieval system. Query expansion is used in a search engine when new terms are added to the user's query in order to increase the efficiency in retrieval. Recently, systems based on query expansion are significantly improving their results, making use of external resources such as ontologies and lexical hierarchies.

MeSH is an initiative from the U.S. National Library of Medicine. It is a controlled vocabulary used for indexing articles from Medline. It consists of sets of terms called

[1] http://www.ncbi.nlm.nih.gov/pmc/
[2] http://www.nlm.nih.gov/mesh/meshhome.html

descriptors, arranged in a hierarchical structure that enables the search at different levels of specificity. There are currently 26,142 MeSH descriptors or Main Headings. There are also over 177,000 alternative expressions, synonyms and terms related to these descriptors, named entry terms.

The rest of the paper is organised as follows. Section 2 describes the expansion strategies used in the experiments. In Section 3 the results obtained are shown and discussed. Finally, conclusions and future works are outlined in Section 4.

2 Query Expansion using MeSH

MeSH ontology offers many possibilities for expanding the query terms. There are several works where studies on the effect of the use of the MeSH ontology for query expansion are presented. In [3], the authors investigate a query expansion strategy process using an advanced PubMed search called Automatic Term Mapping (ATM). For this task, we have used several strategies for expansion based on the *entry terms* similar to those used in [4] and other strategy based on the tree structure whereby MeSH organises its descriptors [5].

Many times a descriptor or *entry term* is made up of more than one term. For example, if the query *Mitral Valve* was made for each term independently, neither *Mitral* or *Valve* correspond to a descriptor or *entry term*. However, the union of the two terms corresponds to a descriptor itself as "*Mitral Valve*", which is a biomedical concept.

That is the reason why each query was pre-processed by dividing it into n-grams, with the aim of exploring all the possibilities offered by the query to obtain sequences that are MeSH descriptors or *entry terms*. Below is an example of processing a query with n-grams.

Query: Breast cancer mammogram

N - Grams

(1): Breast
(2): Breast cancer
(3): Breast cancer mammogram
(4): cancer
(5): cancer mammogram
(6): mammogram

Where the n-gram 2 and 4 are *entry terms* and 1 is a descriptor.

The following sections describe the strategies used to expand the queries.

2.1 Techniques based on *MeSH Tree-structure*

This strategy is based on the tree structure whereby MeSH organises its descriptors. In this case, if the descriptor is a parent node, it is expanded with its child descriptors. If the descriptor does not have any children there is no expansion. Figure 1 shows a brief MeSH tree excerpt which indicates that the Brain descriptor has seven children while the Central Nervous System descriptor has three.

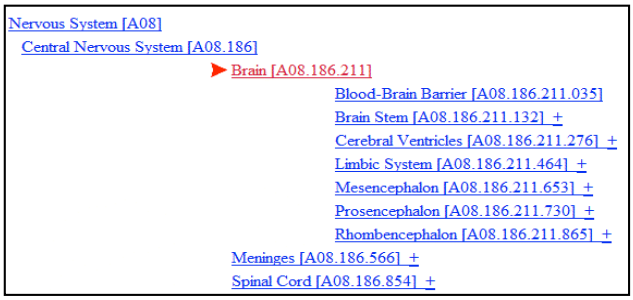

Fig. 1. Excerpt from MeSH Tree

2.2 Techniques based on *Entry Terms*

The first expansion strategy consists in exploring the MeSH tree by checking if the query n-gram is a descriptor. If the n-gram is a descriptor, the query is expanded using all the *entry terms* of the descriptor. If the n-gram is not a descriptor, we check if it is an *entry term*. If so, its descriptor and all the entry terms of that descriptor are added to the expansion.

The second strategy has only a small variation from the first. When a n-gram in the query is a descriptor, the query is expanded with the *entry terms* of the *preferred concept*, instead of all the *entry terms* of that descriptor.

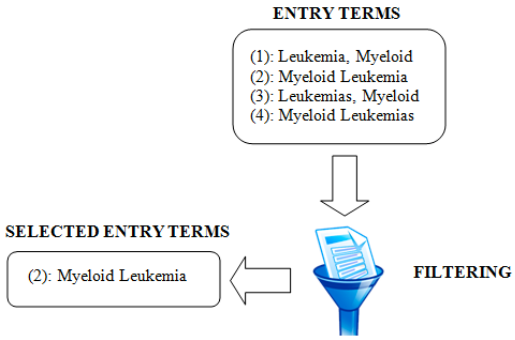

Fig. 2. Example of a filtering process.

When the results of these expansion strategies were calculated, it was found that they introduced too much noise into the queries and the results were not as good as expected. To this end, a filtering of the query was carried out to reduce redundant *entry terms*. Figure 2 shows an example of a filtering process.

3 Experiments and Results

This section details the experiments that were conducted to evaluate various expansion strategies. For this aim, three different indexes were created:

- **Captions (C):** This index contains the text of the captions of each image.
- **Image Reference (IR):** In this index, the sections of the paper that reference each image were indexed. For this indexing, the text of the papers was split into sentences using OpenNLP[3] software and we have only indexed the sentences which refer to an image.
- **Full Text (FT):** This index contains the full text of each paper.

For this edition, three different runs for each indexing were sent:

- **Baseline (B):** Original queries.
- **Concept Tree (CT):** Queries expanded with techniques based on *MeSH Tree-Structure*.
- **Entry Terms Peferred Concept (ETPC):** Queries expanded with techniques based on *Entry Terms*.

Moreover, an additional run based on **Entry Terms (ET)** was sent.

[3] http://incubator.apache.org/opennlp/

In order to perform text indexing and run the different queries, Lucene[4] search engine was used with the default settings. Table 1 shows the results obtained with each run.

Table 1. Results from LABERINTO research group in ImageCLEF 2011.

Ranking	Run	MAP	P10	P20	Rprec	Bpref	Num_Rel_Ret
1	laberinto_CTC	**0.2172**	**0.3467**	0.3017	**0.2369**	**0.2402**	1471
4	laberinto_BC	0.2133	0.3400	**0.3067**	0.2363	0.2384	1469
16	laberinto_ETPCC	0.1939	0.2933	0.2617	0.2089	0.2198	**1526**
44	laberinto_BIR	0.1496	0.3400	0.3000	0.1908	0.1992	1292
48	laberinto_CTIR	0.1466	0.3433	0.2950	0.1868	0.1953	1293
50	laberinto_ETPCIR	0.1411	0.3000	0.2850	0.1766	0.1887	1325
57	laberinto_BFT	0.1146	0.2533	0.2267	0.1621	0.1786	1355
58	laberinto_CTFT	0.1101	0.2500	0.2333	0.1512	0.1691	1348
59	laberinto_ETFT	0.1050	0.2567	0.2250	0.1302	0.1640	1292
60	laberinto_ETPCFT	0.1014	0.2400	0.2200	0.1253	0.1571	1310

Looking at specific runs comparisons, we can further draw the following conclusions:

The best results were obtained using the index of the image captions. On the other hand, the most effective expansion strategy was the expansion based on the MeSH Tree Structure for all the indexes. The best result among all our runs was *laberinto_CTC* (Concept Tree with Captions), which reached a MAP value of 0.2172. This value was the highest among all the runs for textual retrieval type. With respect to the strategy based on *Entry Terms*, we can observe that it retrieve more relevant images than the other strategies. We think that it is also an effective strategy and we will work to improve the MAP and to keep high values for relevant images retrieved.

4 Conclusions and Future Work

In this paper we have presented different query expansion strategies using one of the most widely used ontologies in the medical domain, with the aim of enhancing the efficacy of a textual content-based image retrieval system. Different MeSH ontology elements were chosen for expansion.

The results of our experiments showed that the expansion strategies using the hierarchical structure whereby MeSH organises its descriptors, obtain good results for this task. This work verified the difficulty of finding an appropriate strategy for query expansion. We think that there are information elements or element combinations in MeSH that might be used to expand the queries and could substantially improve an image retrieval system.

In future work, we will continue researching into other query expansion strategies and the use of other ontologies, such as UMLS[5] [6]. Moreover, we plan to build

[4] http://lucene.apache.org/

[5] http://www.nlm.nih.gov/research/umls/

indexes using only medical concepts extracted from the image captions. Finally, we want to experiment expanding as the queries as the indexed text.

5 Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation, the Spanish Government Plan E and the European Union through ERDF (TIN2009-14057-C03-03).

References

1. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., Garcia Seco de Herrera, A. and Tsikrika, T. 2011. The CLEF 2011 medical image retrieval and classification tasks. CLEF 2011 working notes, Amsterdam, The Netherlands.
2. Nelson, S.J., Schopen, M., Savage, A.G., Schulman, J.L. and Arluk, N. 2004. The MeSH translation maintenance system: structure, interface, design and implementation. M. Fieschi, et al. (Ed.). Proceedings of the 11th World Congress on Medical Informatics, pp.67–69.
3. Lu, Z., Kim W. and Wilbur, W. 2009. Evaluation of query expansion using MeSH in PubMed. Information Retrieval, Vol. 12, No. 1, pp. 69-80.
4. Díaz, M.C., Martín, M.T. and Ureña, L.A. 2009. Query expansion with a medical ontology to improve a multimodal information retrieval. Computers in Biology and Medicine, 4, 396-403.
5. Mata, J., Crespo, M. and Maña, M. 2011. Estudio del uso de ontologías para la expansión de consultas en recuperación de imágenes en el dominio biomédico. Procesamiento del Lenguaje Natural, nº 47.
6. Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(2004) 267–270.