LABERINTO at ImageCLEF 2012 Medical Image Retrieval Task

Mariano Crespo, Jacinto Mata, Manuel J. Maña

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

Abstract. This paper shows the experimentation and the results obtained for LABERINTO research group at the ImageCLEF 2012 medical task. We focus our work on image retrieval based on textual information related to the image. Last year we demonstrated that query expansion exploiting the hierarchical structure of the MeSH descriptors achieved a significant improvement in image retrieval systems. This year our goal is to improve the results obtained last year adding a relevance factor to the query terms. In addition, we have developed a new strategy combining the expansion strategy based on the hierarchical MeSH structure with another expansion strategy very popular among researchers in this field, where the query terms are expanded using MMTx program. The experiments carried out have shown that a relevance factor for the query terms achieves a significant improvement for the results of the different expansion strategies.

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

1 Introduction

This paper describes the contribution of the LABERINTO research group in its second 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 2005-2011. Participants will be given a set of 22 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.

Due to the good results obtained in last year's edition, the aim of this year is to improve the effectiveness of the expansion strategy used. We use the MeSH [2] ontology for the expansion of queries, and our new proposal for this year is the inclusion of relevance factors in the query terms and the development of a new

[1] http://www.ncbi.nlm.nih.gov/pmc/

strategy, which is a combination between the expansion strategy based on the hierarchical MeSH structure and a strategy that expands the query terms using MetaMap Transfer program (MMTx) [3].

Ontologies represent a particular knowledge domain in the form of a set of concepts and relations between them. There are many terminological and ontological resources available in the biomedical domain, along with a wide range of applications in NLP: information retrieval, question answering, automatic summarization and classification amongst others. The two resources used in this work were the MeSH ontology and the MMTx program using the source vocabulary SNOMED-CT [4].

MeSH is a controlled vocabulary used for indexing Medline papers. It is comprised of term sets or descriptors, organized into a hierarchical structure to allow searches at various levels of specificity. At present, MeSH encompasses 26,142 descriptors or Main Headings. This is the vocabulary used to index Medline citations. Alternative forms, synonyms and terms related to the descriptors are known as *Entry Terms*. There are over 177,000 *Entry Terms* in MeSH.

MetaMap is a highly configurable program developed by Dr. Alan Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus [5] or, equivalently, to discover Metathesaurus concepts referred to in text.

Finally, we used Lucene [6] to assign the relevance level of matching images based on the terms found.

The rest of the paper is organised as follows. Section 2 describes the expansion strategies used in the experiments and the technique that adds a relevance factor to the query terms. 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. Various works report on studies into the effect of using the MeSH ontology for query expansion. For example, in [7] the authors base the expansion on the hierarchical structure of MeSH. When this technique locates a MeSH descriptor in the query, it ascends the tree to higher levels to search for more general descriptors and adds those it finds to the user query.

In [8] the authors explore a strategy for query expansion using a process of advanced queries known as Automatic Term Mapping (ATM) in PubMed. The study employed a collection of 64 queries and around 160,000 MEDLINE citations which were used in the 2006 and 2007 TREC Genomics Track. One of the main results was an increase in the F measure [9] of 21.5% and 23.3% in the 2006 and 2007 collections respectively through the use of query expansion. The researchers conclude that query expansion through MeSH in PubMed can improve the effectiveness of retrieval, but that in real situations the improvement may not prove to be significant for PubMed users.

In [10, 11] the authors also employ MeSH ontology to expand both the collection and queries. The approach uses two strategies for expanding the document collection.

The first of these extracts the set of MeSH descriptors from the image captions and article titles by means of a classifier, which returns a list of descriptors in order of relevance. The image captions, article titles and the first five descriptors returned by the classifier are then indexed. The second strategy indexes not only the above information, but also the MeSH descriptors associated with the corresponding article in MEDLINE. The query is expanded by extracting the set of MeSH descriptors using the classifier. Only those descriptors belonging to the A and C branches (diseases and anatomical concepts) of the MeSH tree are selected, from which only the three most relevant are added to the query.

As these studies show, various authors have taken advantage of the MeSH ontology in order to expand queries and improve information retrieval systems. In doing so, they have utilized the different cross-reference systems provided by the ontology (synonyms, entry terms, associative relationships among descriptors, …). In this study we present a query expansion strategy that uses the MeSH Tree Structure. Our proposal focuses on the choice of terms to be expanded and demonstrates that the expansion is most efficient when the UMLS Metathesaurus is used, in controlled fashion, for determining which terms are expanded.

2.1 Techniques based on *MeSH Tree-structure*

This strategy is based on the tree structure whereby MeSH organises its descriptors. Figure 1 shows a short extract from the MeSH tree diagram in which it can be seen that the descriptor *Neoplasms by Site* includes six more specific descriptors (children) while the descriptor *Breast Neoplasms* includes only four.

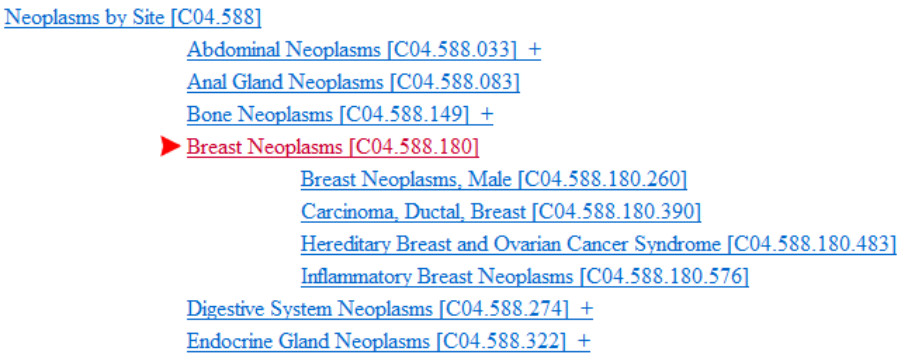

Fig. 1. Excerpt from MeSH Tree

The expansion strategy developed in this section is governed by the following criteria for expanding search terms:

- If the search term is a MeSH descriptor and contains more specific descriptors, it is expanded using these.
- If the search term is a MeSH descriptor but does not contain more specific descriptors, no expansion is performed.

– If the search term is not a MeSH descriptor, no expansion is performed.

In many cases a descriptor comprises more than one term, and performing the expansion at the level of the term is not so efficient. For example, if the search for *Mitral Valve* treats each term independently, neither the term *Mitral* nor the term *Valve* corresponds to a descriptor. Nevertheless, the two terms in combination correspond to the descriptor "*Mitral Valve*", a biomedical concept.

In order to discover the medical concepts within the queries, in this study we have used the National Library of Medicine's MetaMap Transfer program (MMTx) using the source vocabulary SNOMED-CT of the UMLS metathesaurus version 2011AA.

The concepts labelled in this phase were mapped to the MeSH hierarchy in order to perform an expansion of each one. Each labelled concept is sought within the MeSH tree. If the concept is a descriptor, its children are retrieved and added to the query according to the general scheme described above. In this approach, in addition to the terms expanded via the MeSH tree, the UMLS concepts identified are added, as illustrated in figure 2, which provides a schematic representation of the expansion process for the query *lymphoma MRI images*.

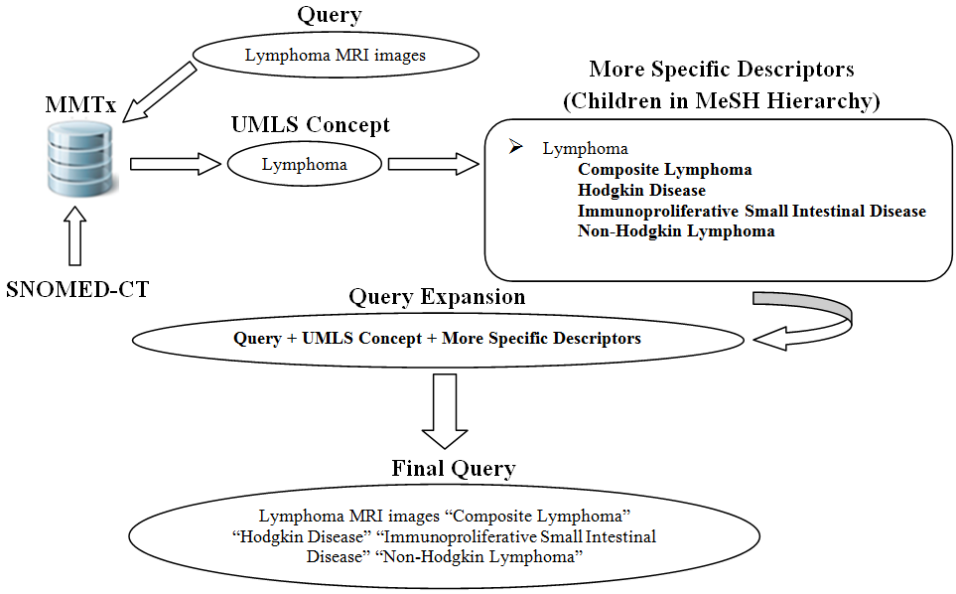

Fig. 2. Example of query expansion process

2.2 Techniques based on *MMTx*

This strategy is based on MMTx. As in the previous section, we used the National Library of Medicine's MMTx to discover Metathesaurus concepts in figure captions. MMTx employs a series of language-processing modules to map text to concepts in

the UMLS Metathesaurus. Then we configured MMTx with the option MMI (-N) that displays, in a separate section, a ranked list of all the mappings assigned to the text. Additional data such as the PMID of the citation, CUIs, abbreviated Semantic Types are also included. Finally, we selected those candidates which semantic types are: *Diagnostic Procedure (diap), Disease or Symdrome (dsyn), Body Part, Organ, or Organ Component (bpoc), Neoplastic Process (neop), Injury or Poisoning (inpo), Body Location or Region (blor), Pathologic Function (patf) or Cell (cell).* In order to elaborate this list, we did a study of the most repeated semantic types in the queries for the last four years of ImageCLEF. We also asked an expert to make the final selection. Figure 3 shows a schematic representation of the expansion process for the query *lymphoma MRI images*.

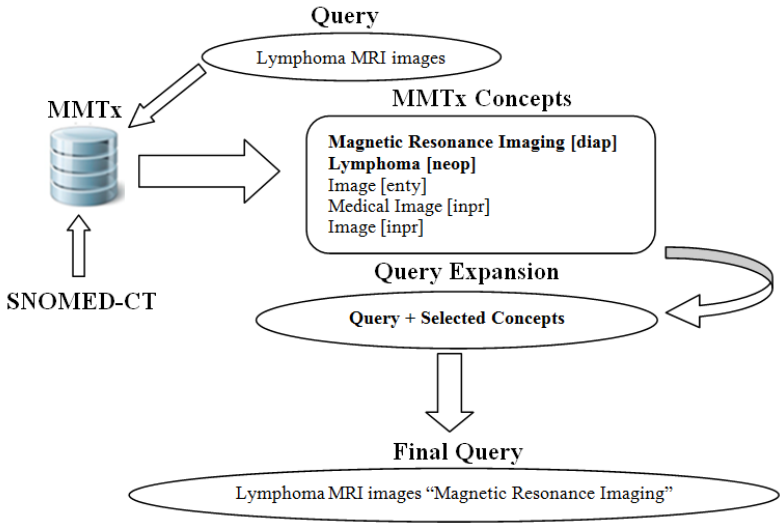

Fig. 3. Example of query expansion process

2.3 Adding relevance levels to the query terms

The technique for adding a weight to each query term has been developed with the *boost factor* given by Lucene. It provides the relevance level of matching documents based on the terms found. To boost a term, the caret ("^") symbol with a boost factor (a number) at the end of the term is used. The higher the boost factor, the more relevant the term will be. By default, the boost factor is 1. For the experiments, different boost factors values based on the Inverse Document Frequency (*idf*) were tested. The *idf* is used to weigh the term information value in general, based on frequency of use or appearance. It is averaged as shown in (1).

$$IDF(t,doc) = \log_2\left(\frac{N}{DF}\right) \quad \textbf{(1)}$$

Where:

t: Term.
doc: document.
N: Total number of documents in the collection.
DF: Frequency of occurrence of the term (*t*) in the document (*doc*).

To carry out the experiments, the *idf* for each query term was calculated independently and then, it was multiplied by a factor α or β (between 0.1 and 1) depending on original or expanded query terms. Finally, the relevance factor is calculated as shown in (2).

$$RF(t,doc) = (\alpha \mid \beta)\frac{-\log_2\left(\frac{N}{DF}\right)}{\left|\log_2\left(\frac{N}{DF}\right)\right|} \quad \textbf{(2)}$$

Where:

t: Term.
doc: document.
α: Original terms factor.
β: Expanded terms factor.
N: Total number of documents in the collection.
DF: Frequency of occurrence of the term (*n*) in the document (*doc*).

The best results were obtained setting $\alpha = 1$ and $\beta = 0.1$, i.e. when exists a greater difference of the relevance factor between the original terms and the expanded terms, but always adding greater weight to the original query terms.

3 Experiments and Results

This section details the experiments that were conducted to evaluate various expansion strategies. For this aim, seven different runs were sent:

- **Laberinto_BL:** Original queries.
- **Laberinto_BL_MSH:** Queries expanded with techniques based on MeSH Tree-Structure.
- **Laberinto_MSH_PESO_1:** Relevance factor of original terms α = 1, relevance factor of expanded terms β = 0.1 and queries expanded with techniques based on MeSH Tree-Structure.

- **Laberinto_MSH_PESO_2:** Relevance factor of original terms $\alpha = 2$, relevance factor of expanded terms $\beta = 0.1$ and queries expanded with techniques based on MeSH Tree-Structure.
- **Laberinto_MMTx_MSH:** Queries expanded with mixed expansion strategies based on MeSH Tree-Structure and MMTx.
- **Laberinto_MMTx_MSH_PESO_1:** Relevance factor of original terms $\alpha = 1$, relevance factor of expanded terms $\beta = 0.1$ and queries expanded with mixed expansion strategies based on MeSH Tree-Structure and MMTx.
- **Laberinto_MMTx_MSH_PESO_2:** Relevance factor of original terms $\alpha = 2$, relevance factor of expanded terms $\beta = 0.1$ and queries expanded with mixed expansion strategies based on MeSH Tree-Structure and MMTx.

In order to perform text indexing and run the different queries, Lucene 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 2012.

Ranking	Run	MAP	GM-MAP	Bpref	P10	P30
10	Laberinto_MSH_PESO_2	**0.1859**	0.0537	**0.1939**	**0.3318**	0.1894
18	Laberinto_MSH_PESO_1	0.1707	0.0512	0.1712	**0.3318**	0.1894
20	Laberinto_MMTx_MSH_PESO_2	0.1680	**0.0555**	0.1711	0.3227	0.1909
22	Laberinto_MMTx_MSH_PESO_1	0.1677	0.0554	0.1701	0.3182	0.1879
24	Laberinto_BL	0.1658	0.0477	0.1667	0.3000	**0.1939**
30	Laberinto_BL_MSH	0.1613	0.0462	0.1812	0.2682	0.1864
41	Laberinto_MMTx_MSH	0.1361	0.0438	0.1570	0.2091	0.1758

Looking at specific runs comparisons, we can further draw the following conclusions. Adding a relevance factor to the query terms considerably improved the results, especially when greater weights to the original terms than the expanded terms were given. The best result among all our runs was *Laberinto_MSH_PESO_2*, which reached a MAP value of 0.1859. With respect to the strategy that combines MMTx with MeSH hierarchy, we can observe that the results are somewhat inferior, but both techniques improve the results if adding a relevance factor to the query terms.

4 Conclusions and Future Work

The principal objective was to improve the effectiveness of image retrieval system through textual content.

In the course of our experimentation we gained an understanding of the difficulties of finding an appropriate strategy for performing query expansion.

The results of our experiments showed that the expansion strategy employing medical concepts alongside the MeSH hierarchy successfully improved the effectiveness of the system. The results achieved with the mix MMTx and MeSH hierarchy strategy are somewhat inferior. On the other hand, the results show that add

a relevance factor to the query terms considerably improved the results, especially if we give a greater weight to the original terms than the expanded terms.

In future studies we also intend to perform expansion on the medical concepts occurring in the text used for constructing the index. We will also explore the new expansion strategies both MeSH as UMLS and new techniques to assign a relevant factor to the query terms. Finally, we also intend to dedicate future studies to analyzing queries in detail so as to extract information from abbreviations, type of image to search for [12] (eg, radiographs, tomographs) and so on. After all, the most essential thing for an image retrieval system to work well is to know exactly what one is searching for.

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. 2012. The CLEF 2012 medical image retrieval and classification tasks. CLEF 2012 working notes, Rome, Italy.
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. Aronson AR. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001, pp.17–21.
4. SNOMED Clinical Terms. International Health Terminology Standards Development Organisation (IHTSDO). Available at: http://www.ihtsdo.org/snomed-ct/. Accessed: Aug 16, 2012.
5. Bodenreider O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, vol. 32, pp. 267–270.
6. Cutting D, Busch M, Cohen D, et al. Apache Lucene. 2008. Available at: http://lucene.apache.org/. Accessed: Aug 16, 2012.
7. Gobel G, Andreatta S, Masser J, et al. 2001. A MeSH based intelligent search intermediary for Consumer Health Information Systems. Int J Med Inform. vol. 64, pp. 241–51.
8. Lu Z, Kim W, Wilbur W. 2009. Evaluation of query expansion using MeSH in PubMed. Information Retrieval, vol. 12, pp. 69-80.
9. Van Rijsbergen CJ. 1979. Information Retrieval. 2nd ed. Butterworths, London, UK.
10. Gobeill J, Theodoro D, Patsche E, et al. 2009. Taking benefit of query and document expansion using mesh descriptors in medical imageclef 2009. Working Notes of CLEF.

11. Gobeill J, Ruch P, Zhou X. 2009. Query and Document Expansion with Medical Subject Headings Terms at Medical ImageCLEF 2008, CLEF 2008. LNCS, Springer. vol. 5706, pp.736–743.
12. Rahman M, Antani S, Fushman D, et al. 2012. Biomedical Image Retrieval Using Multimodal Context and Concept Feature Spaces, in: Medical Content-based Retrieval for Clinical Decision Support. LNCS. vol. 7075, pp. 24-35.