SINAI at ImageCLEF 2010 medical task

M.C. Díaz-Galiano, M.T. Martín-Valdivia, Arturo Montejo-Raez, M.A. García-Cumbreras
University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,maite,amontejo,mgarcia}@ujaen.es

Abstract

Recent researches demonstrate that the use and integration of several knowledge sources improves the quality and efficiency of information systems. In this paper, we present the system developed for the ImageCLEF 2010 medical task. We show the effect of using the medical ontology MeSH to expand terms found in textual queries. This year we have applied a strategy for deciding when include the information extracted from MeSH. The experiments carried out show that our machine learning approach to determine when to perform expansion did not resulted in any improvement over our base line, that where no expansion was performed at all.

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

Algorithms, Experimentation, Languages, Performance

Keywords

Query expansion, Document expansion, MeSH ontology, Information Retrieval

1 Introduction

This paper presents the system developed by the SINAI research group at the ImageCLEF medical retrieval task 2010.

The goal of the medical task is to retrieve relevant images based on an image query[6].

The efficient access to multimodal information is becoming more and more difficult every day. For this reason, it is necessary to develop search strategies for easier retrieval of useful information. One of these strategies includes the use of linguistic resources in order to improve the access and management of information by expanding queries in information retrieval systems, enriching the databases semantically or extracting unknown data from collections.

In previous years we have experimented with the expansion of the queries with medical ontologies[5, 3]. Last year we continued our usage of the MeSH ontology for query expansion, but comparing it to term expansion within the documents in the collection[4].

This year, our main goal is to study the expansion of the collection and queries with the same ontology MeSH, applying a strategy for deciding when include the information extracted.

This year, the same collection as 2009 and 2008 is used but with a larger number of images. 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, more than 77,000 images.

Previous years we developed a system that test different aspects, such as the application of Information Gain in order to improve the results[2], that obtained poor results, the expansion of the topics with the MeSH[1] ontology[3], the expansion of the topics again with the UMLS[2] metathesaurus and minor textual information but more specific[5], and the term expansion within the documents in the collection[4].

The following section describes the strategy for deciding when include the information extracted. In Section 3, we explain the experiments and obtained results. Finally, conclusions are presented in Section 4.

2 Expansion with medical ontology. Machine learning for decision making

We have used the MeSH ontology to expand the queries included in the GoldMiner collection. First, we have extracted two types of terms from the MeSH descriptors:

- MeSH Heading (MH), which is a term composed by one or more words. Each record contains only one MH term.
- Entry, which is composed of one or more words too. Each record contains several Entry terms. These terms are a different way of writing the MH term, that is, they are synonyms.

This set of terms of a same descriptor constitutes a bag of terms. We have used the bags of terms to expand the queries. If all the words of a term are in the query, we generate a new expanded query by adding all the terms in this bag. To compare the words of a particular term and those of the query, we first put all the words in lowercase without removing stopwords. It does not matter the order in which these component words occur in the query.

In order to reduce the number of terms available to expand the query, we have only used those that are in the MeSH categories [A] Anatomy, [C] Diseases, and [E] Analytical, Diagnostic and Therapeutic Techniques and Equipment[7].

Given the hardness of the "when to expand" problem, a Machine Learning perspective was followed. Using the gain or loss in MAP as a label, with a true value as label when the MAP increased with the expansion and a false value as label when the MAP decreased, the direction was to generated models in order to determine, automatically, when to expand or not.

The well known learning algorithms Support Vector Machine (SVM)[1] was applied in these experiments. The learning approach consists, therefore, in training a binary classifier with part of the data (the queries of previous years) and test its performance with the queries tested in 2010.

3 Experiment Description and Results

By using the data in 2008 and 2009 ImageCLEFmed campaigns, we have carried out several experiments. We found that there were no clear correlation (slightly more than 0,20 in the Pearson's coefficient) between the success of the expansion for a query and the number of relevant documents found in associated qrels lists. Therefore, the expansion seemed not related to the expected number of relevant documents.

Despite the hardness of the "when to expand" problem, a Machine Learning perspective was followed. Using the gain or loss in MAP as a label, with a true value as label when the MAP

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

increased with the expansion and a false value as label when the MAP decreased, the direction was to generate a model in order to determine, automatically, when to expand or not.

Based on the Ranking Status Value (RSV) returned by the retrieval engine for both queries (expand and original): the RSV values of the top 100 documents in each of the two retrieved lists are used as features for computing the model. Therefore, each query is a sample with 200 features. Feature weights (frequencies in the case of word-based features, RSV in the second case) were normalized using Z-transformation and the dimensionality were reduced by applying Singular Value Decomposition (SVD).

Run	**MAP**
No expansion	0.2764
Always expand	0.2616
Expand sometimes	0.2672

Table 1: MAP values on Ad Hoc runs

Results are shown in Table 1. As we can see, there is a very small variation in MAP among different strategies. Nevertheless, this year our expansion approach did not improved over the base line.

4 Conclusions

The machine learning approach to determine when to perform expansion did not resulted in any improvement over our base line, that where no expansion was performed at all. It is clear that our expansion strategy should be studied in depth, as the behaviour of the expansion over queries is not consistent and seems not related to the gain in MAP.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project TEXT-COOL 2.0 (TIN2009-13391-C04-02), a grant from the Andalusian Government, project GeOasis (P08-TIC-41999), and a grant from the University of Jaen, project RFC/PP2008/UJA-08-16-14 and project UJA2009/12/14.

References

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[2] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña López. Using Information Gain to Improve the ImageCLEF 2006 Collection. In *CLEF*, volume 4730 of *Lecture Notes in Computer Science*, pages 711–714. Springer, 2006.

[3] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña López. Integrating MeSH Ontology to Improve Medical Information Retrieval. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *Lecture Notes in Computer Science*, pages 601–606. Springer, 2008.

[4] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, and L.A. Ureña López. SINAI at ImageCLEF 2009 medical task. In *On-line Working Notes, CLEF 2009*, 2009.

[5] M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, L.A. Urea-López, and A. Montejo-Ráez. Query Expansion on Medical Image Retrieval: MeSH vs. UMLS. In Carol Peters et al., editor, *CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 732–735. Springer, 2009.

[6] Henning Mller, Jayashree Kalpathy-Cramer, Ivan Eggel, Steven Bedrick, Charles E. Kahn Jr., and William Hersh. Overview of the CLEF 2010 medical image retrieval track. In *In the Working Notes of CLEF 2010*, 2010.

[7] S. Radhouani, J. Lim, J.-P. Chevallet, and G. Falquet. Combining textual and visual ontologies to solve medical multimodal queries. In *in Proc. IEEE ICME*, 2006.