Computers in Biology and Medicine 39 (2009) 396–403

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Query expansion with a medical ontology to improve a multimodal information retrieval system

M.C. Díaz-Galiano*, M.T Martín-Valdivia, L.A. Ureña-López

Departamento de Informática, Campus Las Lagunillas, s/n. University of Jaén, Jaén E-23071 Spain

ARTICLE INFO

Article history:
Received 27 November 2007
Accepted 29 January 2009

Keywords:
Information retrieval system
Natural language processing
MeSH ontology
Query expansion
Knowledge integration

ABSTRACT

Searching biomedical information in a large collection of medical data is a complex task. The use of tools and biomedical resources could ease the retrieval of the information desired. In this paper, we use the medical ontology MeSH to improve a Multimodal Information Retrieval System by expanding the user's query with medical terms. In order to accomplish our experiments, we have used the dataset provided by ImageCLEFmed task organizers for years 2005 and 2006. This dataset is composed of a multimodal collection (images and text) of clinical cases, a list of queries for each year, and a list of relevance judgments for each query to evaluate the results. The results from the experiments show that the use of a medical ontology to expand the queries greatly improves the results.

1. Introduction

Biomedical information available in electronic format is increasing rapidly. In fact, there are large collections of medical data that contain visual and textual information available for researchers, health care providers and all types of customers interested in this kind of information [1–3]. However, it is not always easy to access this large volume of data and, consequently, consumers need to avail themselves with tools to enhance data accessibility and management of retrieval systems.

In this research line, an effort is made to develop resources in order to assist the end user of these large volumes of biomedical information [4,5]. The integration of these resources improves the results of information systems [6–8]. One of the resources most widely used is an ontology. According to Gruber [9], an ontology is a specification of a conceptualization that defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain. The specification takes the form of the definitions of representational vocabulary (classes, relations, and so on), which provide meanings to the vocabulary and formal constraints on its coherent use.

Ontologies have been used for several natural language processing tasks [10], such as automatic summarization [11], word sense disambiguation [12], knowledge acquisition [13], etc. There are many thesauri and ontologies in the biomedical domain to store and classify the medical knowledge, i.e: GO[1] [14], UniProt[2] [15], SwissProt[3] [16], MeSH[4] [17], UMLS[5] [4], which have been applied to a wide range of applications in biomedicine [6,18,19].

This paper describes the use of the medical ontology MeSH to improve a multimodal retrieval system by expanding the user's query with MeSH terms. The evaluation of this system is carried out using the collection, queries and relevance judgments provided by the ImageCLEF medical task organization. Moreover, we compare the use of a traditional Information Retrieval (IR) system, an IR system with medical knowledge, a Content-Based Information Retrieval (CBIR) system and a mixed system with information from these systems.

The structure of the paper is as follows. Section 2 introduces several related studies which describe approaches to query expansion. Section 3 describes the ImageCLEFmed task and the corpus used in the experiments. Section 4 briefly introduces the MeSH ontology. Section 5 explains the approach used for the expansion of the queries. Section 6 shows the experiments carried out and their results. And finally, the conclusions of the experiments are presented in Section 7.

* Corresponding author. Tel.: +34 953 212882; fax: +34 953 212472.
E-mail addresses: mcdiaz@ujaen.es (M.C. Díaz-Galiano), maite@ujaen.es (M.T Martín-Valdivia), laurena@ujaen.es (L.A. Ureña-López).

[1] http://www.geneontology.org/
[2] http://www.ebi.uniprot.org/index.shtml
[3] http://www.expasy.org/sprot/
[4] http://www.nlm.nih.gov/mesh/meshhome.html
[5] http://www.nlm.nih.gov/research/umls/

0010-4825/$ - see front matter Crown Copyright © 2009 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2009.01.012

2. Background

Ontologies have been used for several natural language processing tasks such as automatic summarization [20], text annotation [21], and word sense disambiguation [12], among others.

One of the main applications of ontologies is the term expansion found in free textual documents/queries. Several studies indicate the advantage of using ontologies for query expansion in order to improve IR systems. For example, Bhogal [7] presents a review of various query expansion approaches including relevance feedback, corpus-dependent knowledge models and corpus-independent knowledge models. Her work also analyzes query expansion using domain-specific and domain-independent ontologies.

WordNet [10] is a large electronic lexical database of the English language and was conceived as a full-scale model of human semantic organization, where words and their meanings are related to one another via semantic and lexical relations. WordNet has been used for query expansion in several estudies. Voorhees [22] expands query nouns using WordNet synsets and the "is-a" relation. An early and relevant study by De-Buenaga-Rodríguez et al. [23] showed that the expansion of document text by adding terms from synsets where the categories are included enhances system performance. Gonzalo et al. [24] use WordNet synset for manually indexing a disambiguated test collection of documents and queries derived from the SemCor semantic concordance. Navigli and Velardi [25] suggest that query expansion is suitable for short queries. They use WordNet 1.6 and Google for their experiments. Martín-Valdivia et al. [12] integrate the WordNet concepts (synsets) of several semantic relationships in order to improve two different NLP tasks: a word sense disambiguation system and a text categorization system. The results obtained show that this expansion is a very promising approach. Vallet et al. [26] build an ontology and associate it to a corpus of news. They annotate the documents with the ontology concepts and then analyze the query in order to generate a semantic query. Their results achieve measurable improvements with respect to keyword-based search systems. The problem with domain-independent ontologies such as WordNet is that they have a broad coverage. Thus, ambiguous terms within the ontology can be problematic and researchers prefer to use domain-specific ontologies because the terminology is less ambiguous. For narrower search tasks, domain-specific ontologies are the preferred choice. A domain-specific ontology models terms and concepts which are proper to a given domain. For example, Aronson and Rindflesch [6] use the MetaMap program for associating UMLS Metathesaurus concepts with the original query. They conclude that the optimal strategy would be to combine query expansion with retrieval feedback. Hersh et al. [27] observe that the UMLS Metathesaurus can provide benefits for IR tasks such as automated indexing and query expansion in a substantial minority of queries and highlight the importance of studying which queries are better suited to expansion. Nilsson et al. [28] use synonyms and hyponyms from domain-specific ontology based on the Stockholm University Information System (SUiS) to carry out query expansion. Their experiments have shown a precision increase.

As for the biomedical domain, some recent studies such as Yu [29] review current methods in the construction, maintenance, alignment, and evaluation of medical ontologies.

3. Corpus description

For our experiments, we have used a corpus with multimodal medical information supplied by the Cross Language Evaluation Forum (CLEF)[6] [30] organization for the ImageCLEFmed[7] task [31]. CLEF is a European forum which promotes investigation in multilingual information access, testing and evaluating IR systems operating on European languages in both monolingual and cross-language contexts. CLEF performs several evaluation tasks: multilingual retrieval (Ad-Hoc), interactive retrieval (iCLEF), question-answering (QA@CLEF), image retrieval (ImageCLEF), geographical information retrieval (GeoCLEF), etc. Many of these tasks are divided into subtasks. For example, in 2006 the ImageCLEF task comprised two general subtasks: general photographs and medical images (ImageCLEFmed). In these tasks the organizers gave the same data to the participants in order to obtain comparable results from all participant systems.

The ImageCLEFmed goal is to enable researchers to assess and compare system performance and for this reason the organizers develop realistic test collections which simulate real-world medical retrieval tasks.

The ImageCLEFmed task includes a multimodal and multilingual collection based in medical cases. Every year the organizers develop several topics and the task consist in finding images of the collection relevant for these topics.

The ImageCLEFmed collection contains about 50,000 images arranged in four data sets: CASImage, MIR, PEIR, and PathoPIC. These data sets are very heterogeneous. PEIR and PathoPIC collections contain medical images (scans, X-rays, CT, MRI, …), with one textual annotation per image. The CASImage and MIR data sets are organized by clinical cases. Each clinical case contains a group of medical images and textual notes in XML format containing information about the illness displayed in the image.

In a first step, we have pre-processed the collection in order to extract the textual information associated with each visual image. We have used English for the document collection as well as for queries. Annotations in other languages have been translated into English. Thus, French annotations in the CASImage collection were translated into English, using the Reverso on-line Machine Translator[8] integrated into the SINTRAM [32] system, and then were incorporated into the collection. The Pathopic collection has annotations in both English and German (actually, Pathopic is a parallel corpus). We used only English annotations in order to generate the Pathopic documents, discarding German annotations.

For each image, one textual document has been generated in order to create the whole textual collection. Note that each case can include more than one image (Fig. 1). In this case, we have generated more than one textual document per case (one per image) by duplicating the same text.

Finally, the collections have been pre-processed as usual, applying a stop-word list and stemmer methods before indexing [33].

Together with this collection we have used two sets of queries. These sets are supplied for the ImageCLEFmed task for years 2005 and 2006. In 2005, the organizers developed 25 queries and 30 in 2006. The queries were classified based on topic categories reflecting whether they were more suitable for retrieval using visual, textual, or mixed approaches. Each set is composed of textual and visual queries (Fig. 2). The visual queries consist of one or more medical images. The textual queries are in three languages: English, German, and French, but only English queries have been used here because the goal of our experiments is only to improve a multimodal IR system without using multilingual information.

4. MeSH ontology

The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary compiled by the National Library of Medicine (NLM)

[6] http://www.clef-campaign.org/

[7] http://ir.ohsu.edu/image/

[8] http://www.reverso.net