IRIT at ImageCLEF 2010: medical retrieval track

Duy Dinh, Lynda Tamine

University of Toulouse,
118 route de Narbonne, 31062 Toulouse, France
{dinh,lechan1i}@irit.fr

Abstract. We reported some experiments conducted by our members in the SIG team at the IRIT laboratory in the CLEF medical retrieval task, namely ImageCLEFmed. In 2010, we are particularly interested in the *case-based retrieval* task. Our information retrieval approach integrates a hybrid method of concept extraction for enhancing the semantics of the document as well as of the query. More precisely, we applied a knowledge-based concept extraction method combined with statistical information obtained by scoring identified terms denoting concepts both in the document and query. The experiments carried out on the ImageCLEF 2010 show that our information retrieval approach based on the proposed method of concept extraction show an improvement of 7.07% in terms of MAP (for the best configuration trained on ImageCLEF 2010) over the baseline.

Key words: Concept extraction, Document Expansion, Query expansion, Biomedical Information Retrieval

1 Introduction

This paper describes the contribution of the SIG team[1] (Generalized Information Systems) at the IRIT[2] (Institute for Research in Informatics of Toulouse) laboratory in its first year participation at the medical retrieval track.

Started from 2004, the ImageCLEFmed (medical retrieval task) aims at evaluating the performance of medical information systems, which retrieve medical information from a mono or multilingual image collection, using *visual features* and/or *textual features*. The ImageCLEFmed 2010 task consists of three subtasks: *modality classification, ad-hoc retrieval* and *case-based retrieval* [1]. Participating the first time in ImageCLEFmed 2010, we are particularly interested in the case-based retrieval task, which was firstly introduced in 2009. This is a more complex task than the two other tasks, but one that is designed to be a step closer to the clinical workflow [2]. Clinicians often seek information about

[1] http://www.irit.fr/-SIG-team
[2] http://www.irit.fr

patient cases with incomplete information consisting of patient demographics, symptoms, findings, test results and a set of images. The goal of this sub-task is to retrieve relevant cases that might best suit the provided case description. Motivated by the challenging characteristic of this sub-task, we are particularly interested in providing clinicians relevant information related to their requests.

The rest of this paper is organized as follows: Section 2 describes our conceptual indexing and retrieval framework, which integrates a hybrid approach of knowledge-based and statistical methods of concept extraction from medical documents as well as of the query. Identified terms denoting concepts extracted from the Medical Subject Headings[3] thesaurus will be used to normalize the semantics of the documents (cases) or the queries (case topics). Submitted results will be presented and discussed in section 3. We conclude the paper in section 4 by outlining some perspectives for future work.

2 Conceptual indexing and retrieval framework

The conceptual indexing and retrieval framework consists of three main components: (1) *concept extractor*, (2) *conceptual indexer* and (3) *conceptual retriever*.

2.1 Concept extractor

Our concept extractor relies on a knowledge-based and statistical concept extraction method. Given a patient case, which is typically a textual document including title and image captions, medical terms denoting MeSH concepts are firstly recognized using MeSH lexicon[4]. The concept extraction is processed through three steps: (1) pre-processing, (2) term recognizer and (3) term weighting.

In the pre-processing step, original documents are aggregated by two parts, namely *title* and *image captions*, from each unique article. Documents are then converted into the TREC-like format. During the main processing step, each document is splitted into sentences using TreeTagger [3]. Medical terms in each sentence are automatically recognized using the Medical Subject Headings (MeSH) thesaurus as the only lexical knowledge source. The longest string in each sentence is used to match with concept entries (both preferred and non-preferred terms) in MeSH. We used the Left Right Maximum Matching [4] algorithm to find the longest string that matches an entry in the MeSH lexicon. Finally, the outcome of the term recognition is a list of candidate terms denoting concepts. Since medical terms may be multi-word or single-word based, in order to distinguish a multi-word term (e.g., "breast cancer", "blood test", ...) to a single-word term (e.g., "brain", "pain", ...), we used the '_' symbol to delimit constituents of a given multi-word term (e.g., "breast_cancer", "blood_test").

[3] `http://www.nlm.nih.gov/mesh`

[4] MeSH lexicon contains all meaningful terms (preferred or non-preferred terms) in the thesaurus

Many researchers think that IR techniques could be used to extract technical terms denoting concepts for conceptual indexing purposes. However, most of works dealing with IR techniques for concept extraction are based on word-based representations. For instance, recent works such as [5, 6] have proposed methods of MeSH categorization by ranking a list of MeSH descriptors (concepts) returned by an IR system based on single words. The shortcoming of such approaches is related to the fact that many concepts sharing the same words may be returned. For example, concept names such as "Receptor Parathyroid Hormone Type 2; Receptor Parathyroid Hormone Type 1; Parathyroid Hormone-Related Protein;" are various ones that share common words with the concept "Parathyroid Hormone" and therefore may add some kind of negative noise to the document loosing the semantics of the document. In our system, in order to cope with the shortcoming of the word-based representations, our approach typically relies on (1) recognizing in the first stage medical terms denoting concepts and then (2) weighting the recognized medical terms using IR models based on concept-based representations (full term indexing).

We hypothesize that a MeSH concept can be thought of as a document containing biomedical terms describing itself. Each concept in the MeSH thesaurus, which can be distinguished from others by its concept unique identifier (CUI), contains many textual fields such as: MAIN HEADING (concept name or preferred term), ENTRY (synonyms or lexical variants or non-preferred terms), QUALIFIERS, SCOPE NOTE etc. Different synonyms and lexical variants of this concept could be found in the *ENTRY* field.

Here, we are mainly interested by concept entries (MAIN HEADING, ENTRY) since they constitute the most common indexing and retrieval features used in the domain. Let's denote $Entries(C)$ the set of preferred and not preferred terms of concept C. According to our approach, MeSH thesaurus is viewed as a collection of textual concepts. Formally, each concept C_i of the MeSH thesaurus is represented as a vector of linearly basis vectors namely basic terms in the MeSH lexicon: $C = (c_1, c_2, \ldots c_{Nc})$ where Nc is MeSH lexicon size, c_j is a weight measuring the aboutness of term c_j in a document D, computed according to the BM25 weighting schema [7]:

$$c_j = \frac{tf_{c_j}^C * (k_3 + 1) * tf_{c_j}^D}{(k_3 + tf_{c_j}^D) * \frac{k_1 * (1-b) + b * cl}{avgcl + tf_{c_j}^C}} * \log \frac{N_c - n_j + 0.5}{n_j + 0.5} \quad (1)$$

where $tf_{c_j}^C$ is the number of occurrences of term c_j in concept C, N_c is the total number of concepts in MeSH, n_j is the number of concepts containing term c_j, cl is the concept length of C (i.e. total number of distinct terms occurring in its textual features), and $avcl$ is the average concept length in MeSH, $k_1 = 1.2, k_3 = 8, b = 0.75$ are the constants used in the experiments reported here.

We applied the BM25 weighting model to measure the degree of expressiveness of each recognized terms (both multi-word and single-word terms) denoting

concepts. In such as a way, our concept extraction approach is typically based on the combination of both the knowledge-based and statistical based methods, allowing to recognize a list of candidate terms denoting concepts in the document that are ranked in an decreasing order of their ability to describe the document. Given a list of recognized terms denoting concepts in the document, each of them is assigned by a score based on the the state-of-the-art term scoring function BM25 [7]. Finally, the top-ranked terms are translated into their preferred form[5], i.e. main heading, for a conceptual representation of the document.

Inspired by recent works dealing with medical concept extraction for document and query expansion [6, 8], we also used MeSH terms identified by our concept extraction method to expand the document/query using their preferred form, i.e. main headings, in an attempt to normalize and standardize the vocabulary used by different authors/search users. Figure 1 illustrates the overview processing of the concept extraction from a given document. The outcome of the concept extraction is then used to expand the document or the query.

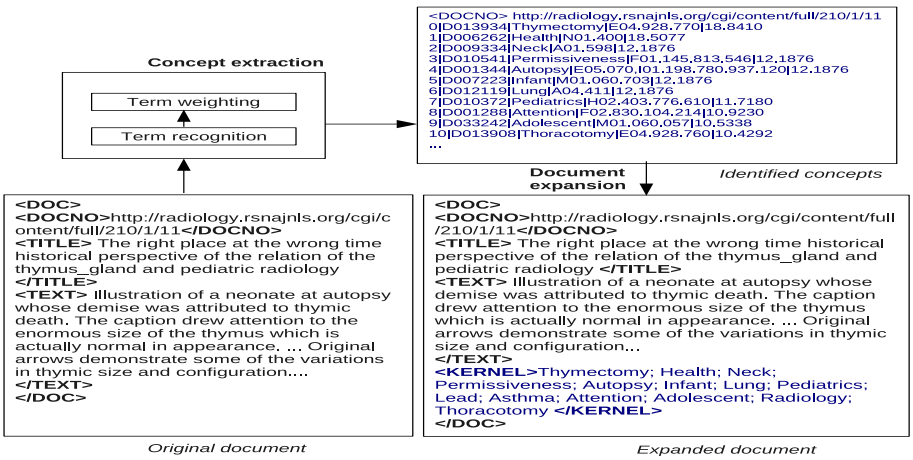

Fig. 1: Concept extraction for document expansion

2.2 Conceptual indexer

The conceptual indexer component aims at gathering statistical information (e.g., word/term frequency, document frequency, positions, etc.) about words in the original document and terms denoting concepts that have been identified for each document into the appropriate index structures. For such as task, we used Terrier [9] with some modification so that multi-word terms are also taken

[5] Each concept has its preferred (main heading) and non-preferred form (lexical variants)

into consideration. During the indexing, each word/term in the document is processed through a highly configurable "Term pipeline", which transforms them in various ways, using plugins such as n-gram indexing, stemming, removing stopwords, and so on. We have added in the Term pipeline the "Synonym finder" in order to transform any terms denoting the same concept to its preferred form. After the conceptual indexing stage, an index of four main data structures is written out: *lexicon*, *document index, direct index*, and *inverted index*. We refer details about each data structure to the article [9].

2.3 Conceptual retriever

The retriever component aims at finding the most relevant documents (search results) in response to a user query. At this stage, documents are retrieved and ranked on the basis of a relevance estimation, which is usually incorporated into a term weighting model (e.g., TF-IDF, PL2, BM25 ...). We used Terrier with appropriate settings (described later in section 3) to perform the retrieval. In such settings, documents and eventually queries are expanded with concept names (or preferred terms) identified by our hybrid concept extraction method. The relevance score of the document D_i with respect to the query Q is given by:

$$RSV(Q, D_i) = RSV(Q^w, D_i^w) + RSV(Q^c, D_i^c) \tag{2}$$

where $RSV(Q^w, D_i^w)$ is the *TF-IDF* word-based relevance score and $RSV(Q, D_i^s)$ is the concept-based relevance score of the document w.r.t the query, computed as follows:

$$\begin{aligned} RSV(Q^w, D_i^w) &= \sum_{q_k^w \in Q^w} (1 + \alpha_w) * TF_i(q_k^w) * IDF(q_k^w) \\ RSV(Q^c, D_i^c) &= \sum_{q_k^c \in Q^c} (1 + \alpha_c) * TF_i(q_k^c) * IDF(q_k^c) \end{aligned} \tag{3}$$

where TF_i: the normalized term frequency of the word q_k^w or preferred term q_k^c in document D_i, IDF: the normalized inverse document frequency of q_k^w or q_k^c in the collection, α_w: the word score modifier, α_c: the preferred term score modifier. The values of the parameter α are obtained by training the retrieval on an IR benchmark.

3 Results and discussion

The goal of our experiments is to evaluate the retrieval effectiveness based on our concept extraction method as well as the impact of the document expansion (DE) and query expansion (QE) using an appropriate number of preferred terms. Terms appearing in a specific field may have a different relevance score to others. *Title*[6], *image captions*[7] and *kernel*[8] are the three main fields of the document.

[6] article title of the patient case
[7] aggregated text obtained by combining all image captions in a patient case
[8] the expanded preferred terms to the document

We carried out two sets of experiments: the first one is based on the classical index of titles and image captions of patient cases using Terrier standard configuration based on the state of the art weighting scheme OKAPI BM25 [7], used as the baseline, denoted *BM25* (run 1). The second set of experiments concerns our conceptual indexing method and consists of four scenarios:

1. the first one is only based on document expansion using identified preferred terms denoting concepts, denoted *DE* (run 4),
2. the second one is based on document expansion (DE) and field indexing, denoted *DE+field* (run 2),
3. the third one is based on document expansion (DE) and query expansion (QE), denoted *DE+QE* (run 5 & 6),
4. the fourth one is based on both document expansion (DE), query expansion (QE) and field indexing, denoted *DE+QE+field* (run 3).

We use both terms representing MeSH concepts (main headings or preferred terms) and single words that do not match any entry in the thesaurus. In the classical approach, documents, i.e. patient cases, were first indexed using the Terrier IR platform (`http://ir.dcs.gla.ac.uk/terrier/`). It consists in processing single words occurring in the documents through a pipeline: removing stop words, and stemming[9] of English words.

In our conceptual IR approach for case-based retrieval, documents and/or queries are firstly analyzed to extract an appropriate number of concepts, namely N and indexed with an appropriate term weighting schema. The parameter N is an experimental variable that must be learned from an IR benchmark by regarding the MAP value or a MEDLINE sub-collection by regarding the F-measure. It very depends on the IR/concept extraction performance of the underlying system. Through some experiments on the ImageCLEFmed 2009 [2] and OHSUMED [10] collections, we obtained two possible values of N, which are 28 and 34 respectively. In addition, we also take into account the position of each word/term in the document. For this reason, we modified by adding the score of word/term in *title* and *kernel* field with a percentage of $\alpha_w^{title} = 5\%$, $\alpha_w^{caption} = 0\%$ and $\alpha_c^{kernel} = 85\%$, which have been trained on the OHSUMED collection (see formula 3).

Method	**RunID**	**MAP**	**bpref**	**P10**	**P20**
baseline (BM25)	1	0.2103	0.1885	0.2786	0.2571
DE	4	0.2085	0.20.83	0.3143	0.2857
DE+field	2	0.2182	0.2267	**0.3571**	**0.3107**
DE+QE	5	0.2085	0.20.83	0.3143	0.2857
	6	0.2193	0.2139	0.3286	0.2857
DE+QE+field	3	**0.2265**	**0.2351**	0.3429	0.3071

Table 1: Results of our submitted runs for the Case-based retrieval task

[9] http://snowball.tartarus.org/

Method	RunID	MAP
DE	4	-0.86%
DE+field	2	+3.76%
DE+QE	5	-0.86%
	6	+4.28%
DE+QE+field	3	**+7.07%**

Table 2: Improvement rates over the baseline

Table 1 depicts the IR performance of the baseline and our various runs based on the document and/or query expansion with/without field indexing. Table 2 shows the improvement rates in terms of MAP-value of our methods over the baseline. We obtained the following results: both run 4 and 5 are observed with a decrease of -0.86% of MAP. Most of the remainder runs are observed with an improvement rate from $+3.76$ % to $+7.07\%$. Run 4 has been designed for only document expansion with $N = 34$ preferred terms added. As mentioned, this number is selected based on learning from a corpus and depends on the test queries and also the document length. For example, if the document is short but the number of selected concepts is high, this could be the reason of the decrease of the IR performance. Run 5 has been designed for document and query expansion with $N = 34$ preferred terms added. The decrease of the IR performance may be explained by the same reason. Indeed, in run 6, which has been designed for the same purpose as run 5 but the parameter N has been set to 28, we observed an improvement rate of $+4.28\%$. Run 2 has been designed for document expansion with field indexing with the following configuration: $N = 34, \alpha_w^{title} = 5\%$, $\alpha_w^{caption} = 0\%$ and $\alpha_c^{kernel} = 85\%$. We observed an improvement rate of $+3.76\%$ in terms of MAP over the baseline. The combination of those runs is revealed in run 3, which is document and query expansion and field indexing, with the following configuration: $N = 28, \alpha_w^{title} = 5\%$, $\alpha_w^{caption} = 0\%$ and $\alpha_c^{kernel} = 85\%$. We dramatically observed the best improvement rate of $+7.07\%$ in terms of MAP. We conclude from those experiments that the concept extraction must generate an appropriate number of concepts so that we can use them to expand the document and the query to normalize and standardize the vocabulary used by different authors/users.

4 Conclusion

This article describes the conceptual retrieval approach of the SIG team for the ImageCLEF 2010 medical retrieval track, especially the case-based retrieval task. The results obtained by our submitted runs prove that our method of concept extraction is useful to enhance the semantics of the document, which could be an interesting evidence to improve the retrieval effectiveness of medical retrieval systems. However, the retrieval performance can be better improved by state-of-the-art query expansion techniques.

References

1. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Jr., C.E.K., Hersh, W.: Overview of the clef 2010 medical image retrieval track. In: Working Notes of CLEF 2010
2. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Jr., C.E.K., Hersh, W.: Overview of the clef 2009 medical image retrieval track. In: Working Notes of CLEF 2009
3. Schmid, H.: Part-of-speech tagging with neural networks. In: Proceedings of the 15th conference on Computational linguistics. (1994) 172–176
4. Dinh, D., Tamine, L.: Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients (short paper). In: Conférence francophone en Recherche d'Information et Applications (CORIA), Sousse, Tunisie, 18/03/2010-21/03/2010, Hermès (mars 2010) 325–336
5. Ruch, P.: Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics **22**(6) (March 2006) 658–664
6. Gobeill, J., Theodoro, D., Patsche, E., Ruch, P.: Taking benefit of query and document expansion using mesh descriptors in medical imageclef 2009. In: Working Notes of CLEF 2009
7. Robertson, S.E., Walker, S., Hancock-Beaulieu, M.: Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive. In: TREC. (1998) 199–210
8. Le, D.T.H., Chevallet, J.P., Dong, B.T.T.: Thesaurus-based query and document expansion in conceptual indexing with umls: Application in medical information retrieval. In: Research, Innovation and Vision for the Future, 2007 IEEE International Conference on. (2007) 242–246
9. Ounis, I.;Lioma, C.C.V.: Research directions in terrier. Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper (2007)
10. Hersh, W., Buckley, C., Leone, T.J., Hickam, D.: Ohsumed: an interactive retrieval evaluation and new large test collection for research. In: SIGIR'94. (1994) 192–201