Using External Knowledge to Solve Multi-Dimensional Queries

Saïd RADHOUANI and Gilles FALQUET
Centre Universitaire d'Informatique. Université de Genève
24, rue Général-Dufour, CH-1211 Genève 4, Switzerland
{Said.Radhouani, Gilles.Falquet}@cui.unige.ch

Abstract. To improve the precision of an information retrieval system in a specific domain we propose a new indexing scheme based on external knowledge resources such as thesauri or ontologies. We introduce the notion of domain dimension, which is a substructure of a knowledge resource, to formally represent the different aspects of a domain that appear in a document. Then, we identify dimensions in documents and queries using a conceptual indexing. The result of this indexing is a representation of each document along its semantic dimensions. We also propose a query processing based on multi-dimensional indexing. It is comprised of a dimensional filtering followed by a dimensional ranking.
Experimental results on medical imaging documents (ImageCLEFmed-2005 collection) show that the dimensional filtering, using three dimensions, can improves the mean average precision by about 25%.

Keywords. Conceptual indexing, multi-dimensional analysis, external knowledge, UMLS

Introduction

In specific domains such as medicine, the information retrieval task differs from general case the vocabulary is precise and less ambiguous than in the general language. When a user seeks information, he or she can express his information need using the specific technical vocabulary containing terms whose meaning is precise. To obtain precise answer during the IR process, documents and queries should be interpreted by minimizing the risk of error. In this paper, we investigate which effects can be achieved for precision of information retrieval by integrating external knowledge although a large meta-thesaurus to process medical queries.

In order to evaluate our approach, we carry out an empirical evaluation on the ImageCLEFmed-2005 collection. This collection contains images annotated in three languages. The queries are composed of image-examples and a textual description explaining the goal of research. The example shown in Figure is one of the 25 queries of the CLEF Medical Image Retrieval Task [4]. In this query, it is clear for a human reader that we are looking for images that contain two elements: one part of the **anatomy**, namely *femur*, and one **pathology**, namely *fracture*. These two elements are semantically related. The fracture is a pathology of a bone such as the femur. These two elements should be described in images whose **modality** is *x-ray*. Thus, x-ray images that contain "*a frac-*

ture of a cranium" or "*a femur without fracture*" are supposed not relevant to this query. Also, we suppose that images containing "*a fracture of a femur*" and whose *modality* is not *x-ray* are not relevant to this query.

Figure 1. A query example extracted from the ImageCLEFmed-2005 base

Observing the set of queries in ImageCLEFmed-2005 collection, we have noticed a regularity of this phenomenon. Indeed, almost all queries contain these three elements (anatomy, pathology, and modality). Hence we call them the **dimensions** and we define them as follow: "*a dimension of a domain corresponds to a point of view according to which one can see this domain. It is comprised of a part of the domain vocabulary and has an internal semantic consistency*".

We suppose that an author uses dimensions of his domain of interest to represent the theme of his document[1]. We also suppose that a user uses dimensions of his domain interest to describe his information need. Hence, we make the assumption that a relevant document to one query with dimensions is the one that fulfils correctly to these dimensions.

In order to solve such multi-dimensional queries, we propose to take into account this concept of dimensions during the information retrieval (IR) process. Existing IR approaches are based on statistical methods that use distributions of key-words to compute the similarity between the query and the documents of the collection. These approaches can not solve multi-dimensional queries because they do not take into account their dimensions. Indeed, these approaches consider documents and queries as bags of words.

To solve multi-dimensional queries, we propose to represent the semantic documents (queries) content using their dimensions. For this reason, the dimensions should initially be defined and then, identified in the documents and queries. The dimensions depend on the organization of the studied domain. We suppose that external knowledge, described through an external resource (linguistic: thesaurus or semantic: ontology), can define the domain dimensions. The external resources contain concepts and semantic relations that constitute the dimensions. Hence, the identification of the dimensions from documents (queries) requires extracting concepts that describe them in documents (queries). This can be done through a conceptual indexing using the selected external resource. Thus, these are the main problems that we face:

- Choosing the external resource and defining the dimensions;
- Extracting the concepts from text to be indexed;
- Identifying the dimensions from documents and queries;
- Taking into account the dimensions during the indexing and querying processes.

[1] For instance, a doctor uses the dimensions anatomy, pathology, and modality to write a report corresponding to a patient.

In the rest of this paper, we present some related works (cf. section 1). We define the external resource and the dimensions in section 2. We introduce the document module and the query module in sections 3 and 4 respectively. For the evaluation (cf. section 5), we investigate the ImageCLEFmed-2005 collection. Finally, we conclude and present our future works (cf. section 6).

1. Related works: external knowledge-based information retrieval

The idea of using external knowledge for IR has been largely explored but with relatively little success. The principal proposals relate to the query expansion. For example, Voorhees [19] who extends queries by using semantic lexical relations (WordNet synsets). The major problem which is faced is related to the ambiguity (i.e. chose the synsets that contain the correct meaning of the terms). Experimental evaluations did not give good results even with manual disambiguation. Qiu and Frei [18] obtain better result by choosing the concepts which are semantically related to the entire query rather than to the individual terms of the query. In the same way, Baziz [6] studied the use of different type of semantic relations for the query expansion. He proposed a technique that makes it possible to choose "the best" concept to add during the expansion process. Based on this technique, he proposed two expansion methods: (1) the "moderated expansion" which consists to add, for each term of the query, only one concept by type of relation, (2) the "careful expansion" that consists to add, for each query, only one concept by type of relation. The experiments show that the careful expansion gives the best results. This conclusion confirms that drawn by Qiu and Frei [18].

The external resources were also used for terms disambiguation. For example, Gonzalo et al. [13] studied the impact of the terms ambiguity based on manual disambiguation and the introduction, voluntarily, of errors of disambiguation. They show as well as the system functions better with conceptual indexing with less than 30% of disambiguation errors. Baziz et al. [5] propose a disambiguation technique and show that an indexing based on a combination of concepts and words, improves the quality, contrary to an indexing based only on concepts. The authors think that the failure is due to the weak covering by the used resource (WordNet) on the vocabulary of the corpus.

In the medical domain, several works used the UMLS meta-thesaurus for the indexing of medical documents [14]. The benefit of such indexing is not very clear, and it is sometimes, while combining once again, the conceptual indexing with the word-based indexing that a small improvement can be obtained [2,3].

Concerning the concept of dimensions, Hyvönen et al. [15] have proposed an interface for navigation in a base of images. The navigation is based on manually built ontologies. Each ontology describes one of the dimensions present in the base (ex. person, event, etc). The interface thus makes it possible to carry out an access to the base according to different dimensions. Each dimension corresponds to a point of view according to which one can explore the base. In the same way, Aussenac-Gilles and Mothe [4] proposed an ontology-based interface for navigation in a base of textual documents. The authors assume that ontology is complex and makes the interface not easily usable. They thus propose to divide ontology into different hierarchies; each one corresponds to one of the dimensions present in the base (example of the astronomy domain: astronomical objects, measuring instruments, observatories, etc). Hence, the multi-dimensional access to the base is made through the defined hierarchies.

From the existing works, we can note that the use of external resources is a good solution to have a precise representation of documents and queries. Indeed, the conceptual indexing combined with a word-based indexing allows the improvement of the answers precision. Good results have also been obtained with the query expansion, which represents a means for enhancing the recall. However, the conceptual indexing and the query expansion are not sufficient to solve multi-dimensional queries. Indeed, in these two cases, documents (queries) are considered as bags of concepts and dimensions are thus ignored. We think, that in addition to the conceptual indexing, a transverse organization on the concepts is necessary to define dimensions and to organize research according to these dimensions. Possible solutions of dimensions definition have been proposed in [15,4]. However, in these works, there is no solution to solve multi-dimensional queries. In the following sections, we present the use of an external resource for solving multi-dimensional queries.

2. External knowledge-based dimensions definition

To be able to set up a conceptual indexing and handle dimensions, we need external resources that must at least have a lexical structure (association between terms to concepts), and a semantic structure (relationships between concepts, e.g. an *is-a* hierarchy). Thesaurus or ontology, for example, can have these characteristics.

The formal model of an external resource S is a 5-tuple $[C, \leq c, R, T, F]$ where:

- C is a set of concepts $\{c_i, \ldots, c_s\}$;
- $\leq c$ is a partial order on C, called the concept hierarchy;
- R is a set of binary relations $\{R_1, \ldots, R_k\}$ on C, where each R_i corresponds to a semantic relation type (typical types are *part-of, instance-of*, *consequence*, etc.);
- T is the lexicon of the external resource. It consists of a set of terms $\{t_i, \ldots, t_r\}$;
- $F \subseteq T \times 2^C$ is a function that associates each term to the set of concepts it designates (if t_i is polysemous $F(t_i)$ has more that one element).

Then, we propose to define the *dimensions* relatively to one or several external resources. A dimension Dim_i is a substructure of the external resource S. For instance, the pathology dimension of the medical domain is a substructure of the UMLS metathesaurus. In some cases a dimension can be the whole external resource. Finally, we define a *domain* as a set of dimensions.

It should be noted that it is illusory to think that a single hierarchical classification can satisfies all the experts of a domain, because any classification presents the reality in an always-debatable point of view. This is important because the use of a classification for indexing means imposing a point of view on any user of an IRS. This is why we use an external knowledge model that can be made of several, non-related, knowledge resources. Thus the dimensions need not originate from the same knowledge resource.

In a practical perspective, our experiments have shown that it is relatively easy to manually extract a dimension from a vast knowledge resource such as UMLS.

3. Document module

We propose to take into account the dimensions to represent the semantic content of the document. Our idea is that a theme developed in a document D is described through a set

of dimensions of the domain to which D belongs. Each dimension is represented in the document by a set of concepts, denoted by terms, and contributes to expose the theme present in this document. Thus, we analyze documents at two levels:

- *Dimensions*: to identify the dimensions that describe the theme present in the document;
- *Concepts*: to identify the concepts that describe each dimension.

Hence, the questions that we face are: How to identify concepts and dimensions from the document? How to use concepts and dimensions to represent the semantic content of the document?

In the following we assume that each document of the considered corpus belongs to one domain and hence contains concepts from at least one dimension of its domain.

3.1. Conceptual indexing

Let us consider a document $D = \{t_1, \ldots, t_n\}$, where the t_i's are the terms occurring in D. The conceptual indexing consists in selecting, relatively to the external resources, a set of concepts that will represent this document. This step substitutes to each term, one or more concept identifiers. The terms that are not associated to any concept in the external knowledge resource are ignored. The resulting conceptual document D_c is a set of concepts $D_c = \{c_1, \ldots, c_m\}$, where each c_i belongs to $F_k(t_i)$ for some term t_i and some external resource S_k. At this stage, there is no attempt at disambiguating terms. A term will be replaced by two or more concepts if it belongs to more than one resource or if it is ambiguous.

3.2. Identification of the document dimensions

The identification of document dimensions consists in distributing each concept c_j of D_c in one sub document d_i depending on its belonging to the dimension Dim_i. Finally, the document is represented as follow: $D_{dim} = \{d_1, \ldots, d_m\}$, where m is the number of dimensions occurring in D_c, $d_i = \{c_1^i, \ldots, c_{k(i)}^i\}$ is the sub document corresponding to the dimensions Dim_i, $c_{k(i)}^i$ is a concept belonging to the dimension Dim_i, and $k(i)$ is the number of concepts in the document dimension d_i. The set of document dimensions pertaining to the dimension Dim_i is denoted by Dim_i^d.

3.3. Multi-dimensional indexing

We suppose that a document dimension d_i can be an answer unit to a query that asks only the dimension Dim_i. Thus, we consider d_i as an independent document. In order to query each document dimension, we index it using the Vector Space Model (VSM) [20]. The document dimension d_i is hence represented by a vector of concepts $\overrightarrow{d_i} = (w_{c1}, \ldots, w_{ck})$, where each w_{cj} is the weight of the concept c_j in d_i. It corresponds to the importance of the concept c_j in the document dimension d_i. The importance of a concept depends on its frequency in the document dimension, and on its relations with the other concepts of the same document dimension. We suppose that the more frequent the concept in the document dimension, the more important it is. We also suppose that the more semantic relations the concept has with other concepts of the doc-

ument dimension, the more importance it has. Thus, w_{cj} is calculated by taking into account the normalized frequency $F(c_j)$ in the document dimension d_i (see formula 1), and the cumulative semantic similarity of the concept with the other concepts in the same document dimension.

$$F(c_j) = \frac{\text{Freq}(c_j)}{max_{x=1..n}(\text{Freq}(c_x))} \tag{1}$$

- Freq(c_j): the absolute frequency of c_j in a document dimension;
- n: the number of all different concepts occurring in Dim_i^d;
- $max_{x=1..n}(.)$: the maximum value of a concept frequency in a document dimension.

The cumulative semantic similarity is based on the semantic similarity between two concepts. There are many ways to define semantic similarities in ontologies or thesauri. In our context, we use the similarity measure defined by [21] which has been tested in [12] and gave good results. It is based on the hierarchical position of the least common subsumer of two concepts.

The cumulative semantic similarity, noted $\widehat{\text{sim}}(c_j)$, is the sum of all the semantic similarities calculated between c_j and all the other concepts included in the studied document dimension d_i. The measure is shown in formula 2, where $\text{sim}(c_j, c_p)$ is the semantic similarity calculated between c_j and c_p.

$$\widehat{\text{sim}}(c_j) = \sum_{c_p \in d_i \setminus \{c_j\}} \text{sim}(c_j, c_p) \tag{2}$$

Finally, the weight w_{cj} is a linear combination of the weighted normalised frequency and the cumulative semantic similarity of c_j (cf. formula 3).

$$w_{cj} = \frac{aF(c_j) + b\widehat{\text{sim}}(c_j)}{a+b} \tag{3}$$

Where a and b are two constants that indicate the relative importance of the frequency and the semantic cumulative similarity.

4. Query module

We propose to take into account the dimensions to interpret the user information need from his query. Our idea is that each user describes his information need through a set of dimensions of his interest domain. Each dimension is represented by a set of concepts and contributes to detail the idea expressed by the user. Thus, these are the questions that we face: How to identify concepts and dimensions from the query? How to use concepts and dimensions to represent the semantic content of the query?

Let us represent a query $Q = \{t_1, \ldots, t_n\}$, where each t_j is a term occurring in Q. For the identification of concepts and dimensions from the query, we use the principle presented in the document module (Section 3). After the conceptual indexing and the dimension identification, a query is represented as follows: $Q_{dim} = \{q_1, \ldots, q_m\}$, where each $q_i = \{c_1^i, \ldots, c_{k(i)}^i\}$ is the sub query corresponding to the dimension Dim_i and m

is the number of dimensions occurring in Q_c. The elements of q_i are the concepts of Q_c that belong to Dim_i. Each q_i is thus a conceptual representation of an aspect of the query.

Our main hypothesis is that a document is relevant for a query if it is relevant for each dimension of this query. Thus, to solve multi-dimensional query, we propose to use its dimensions to filter the documents during the querying process. The relevance of a document D with respect to a query Q is hence given by a combination of two techniques:

- *Filtering* selects documents that contain the query dimensions;
- *Ranking* ranks the filtered documents depending on their relevance to the query.

4.1. Dimensional filtering

One simple way to carry out the filtering is to use the Boolean operators (AND / OR) on the query dimensions. We think that this way can be a constraint for the user, especially when he has doubts, uncertainty, or has some priorities on dimensions of his query. Thus, we propose to use some criterions on the query dimensions in order to have more precision on the user need. Three criterions are thus proposed: "*obligatory*", "*optionally*", and "*priority*".

One dimension marked *obligatory* in a query must appear in the retrieved documents, while an *optional* dimension can appear or not. These criterions can surpass the limits of using the AND/OR operators [16]. It is possible that the user can not use these two criterions but in contrary, has some priorities on dimensions of his query. Thus we propose to use the criterion of *priority* allowing the user to give a priority value between 1 and m, where m is the number of all dimensions present in the query. Hence, a dimension having a priority j must appear in the retrieved documents, else, the dimension having a priority $j + 1$ must appear in the retrieved documents.

Finally, for each query dimension q_i, a criterion is added. The query dimension is thus represented as follow: $q_{i(criterion)}$, where criterion can be "*obligatory*", "*optionally*" or "*priority=value*".

The dimensional filtering consists to conserve only documents that respect the criterions added to the query dimensions. For example, for a query Q containing three dimensions, the sub queries Q_1, Q_2, and Q_3 are constructed. If the user decides that dimensions 1 and 2 are *obligatory*, and dimension 3 is *optionally*, we obtain a query represented as follow: $Q = \{Q_{1(obligatory)}, Q_{2(obligatory)}, Q_{3(optionally)}\}$. This implies that a relevant document must contain the dimensions 1 and 2, and eventually, the dimension 3. Thus, we filter the document collection and we obtain a non-ranked document set D_f that respects the prcised criterions. In order to return, for each query, one ranked document list we use a second technique that we describe in the next section.

4.2. Ranking technique

We rank the documents set D_f in order of relevance by using the VSM. We notice that a document is represented by a set of document dimensions, each one described by a set of concepts. Thus, to evaluate the relevance of a document D to a query Q, we compute the similarity $Sim(D, Q)$ between them by taking into account the similarity between all the dimensions that they share (cf. formula 5).

$$Sim(D, Q) = \frac{1}{m} \sum_{i \in [1,m]} Sim_{dim}(d_i, q_i) \quad (4)$$

- $Sim_{dim}(d_i, q_i) \in [0, 1]$: the similarity between a query dimension q_i and a document dimension d_i. It is computed by the cosine of the angle between the vectors $\overrightarrow{d_i}$ and $\overrightarrow{q_i}$ representing respectively d_i and q_i. This similarity is equal to 1 if q_i and d_i share the same concepts, and 0 if they do not share any concept;
- m: the number of dimensions in Q.

5. Experimental evaluation

The goal of our experiments is to evaluate the impact of taking into account dimensions on the mean average precision (MAP) of the IRS. We also evaluate the impact of using the criterions on the query dimensions. In the current experience, we do not set up the multi-dimensional indexing. We only set up the multi-dimensional querying. The evaluation consists to compare the result obtained by our approach to those obtained by the VSM.

5.1. The corpus and the external resource

As part of the Cross Language Evaluation Forum (CLEF), the ImageCLEF-2005 track [11] that promotes cross language image retrieval has a Medical Image Retrieval (MedIR) task in 2005. The test collection ImageCLEFmed-2005 contains 50,026 images with annotations in XML format. The majority of the annotations are in English but a significant number is also in French and German, with a few cases that do not contain any annotation at all. The 25 queries of the ImageCLEFmed-2005 base have been formulated with example images and short textual descriptions. For the current experience, we used only the English part of he ImageCLEF-2005 collection.

We used the UMLS (Unified Medical Language System) both as an external resource for conceptual indexing, and also as a reference to define dimensions. UMLS, a medical meta-thesaurus, is the result of fusion of many resources (thesaurus). UMLS contains 170 relation types between its concepts. All its concepts are organised, through a hierarchy, in 135 categories called "semantic types" and forming the semantic network. We use this structure to define the dimensions in the medical domain. For the indexing, we used the XIOTA experimental system [9].

5.2. Conceptual indexing

The conceptual indexing is a mean to take into account the dimensions during the IR process. As detailed in our previous work, this process is very difficult to set up [17]. Indeed, we made the hypothesis that only terms present in UMLS and retrieved, with lexical variation in medical text, make it possible to identify one concept[2]. To associate a term to a corresponding concept, we tested some techniques taking into account the size of each term. In order to reduce terms ambiguity and consequently improve the mapping, we carried out some filtering on document text and/or on the meta-thesaurus

[2]This hypothesis is restrictive because the terminology of UMLS does not cover all possible textual forms.

Table 1. Results using dimensions filtering and criterion on dimensions

	LTC		DFR	
	MAP	%	MAP	%
H1	0.222	+4.47%	0.2606	+2.71%
H2	0.2158	+1.55%	0.249	-1.88%
H3	0.2253	+6.02%	0.2606	+2.71%
H4	0.2279	+7.24%	0.27	+6.42%
H5	**0.2655**	**+24.94%**	**0.2897**	**+14.89%**

(e.g. eliminate some specific thesaurus from UMLS (those that are not relevant for our task).

For indexing, we use two weighting schemes: LTC and DFR [1]. Based on conceptual indexing, these two schemes give respectively a Mean Average Precision (MAP) of 0.2125 and 0.2537. In the following sections, these two results are called the *baseline*.

Results show that concept's extraction based on all terms independently of their size give better results than matching based of longer terms. Indeed, the extraction based only on longer terms is very precise, but also gives a lower recall. We also notice that filtering techniques can improve result and surpass those obtained during word-based indexing. Indeed, these filtering reduce ambiguity during the concept-term matching. Finally, despite the incomplete concept extraction, concept-based indexing allows to surpass the word-based indexing [17].

5.3. Multi-dimension filtering

In the ImageCLEFmed-2005 base, we have noticed, from the queries, the presence of three dimensions: *Anatomy*, *Pathology*, and *Modality*. We defined these dimensions through the semantic network of UMLS. The following semantic types of UMLS define them respectively:

- *Anatomy*: "Anatomical Structure", "Body System", "Body space or Junction", "Body Location or Region";
- *Pathology*: "Disease or Syndrome", "Finding", "Injury or Poisoning";
- *Modality*: "Diagnostic Procedure", "Manufactured Object".

In order to evaluate the impact of taking into account dimensions, we compare the results obtained here with the *baseline*. To carry out the filtering by dimensions, we have made five implicit hypotheses using different criterions on the query dimensions. Obtained results are presented in Table 1 where rows correspond to the hypotheses, and values correspond to the results and their variation rates compared to the *baseline*. Here we present the hypotheses.

H1: *Relevant documents include all the three query dimensions (if they exist in the query)*. In this case, the request is presented as follows: $Q = \{Anatomy_{(obligatory)}, Pathology_{(obligatory)}, Modality_{(obligatory)}\}$. This hypothesis improves the result both for LTC and DFR. By observing the documents of the collection, we noticed that the *modality* dimension is not clearly stated in the documents. Indeed, the reports generally described a lesion on a part of the body and information on the type of image is often implicit. For this reason, we prefer the following hypothesis:

H2: *Relevant documents include at least one of the three query dimensions (if they exist)*. This hypothesis improves the result for LTC, but causes a slight decrease of the result for the DFR.

H3: *Relevant documents contain the anatomy dimension present in the query*. By forcing only the "anatomy" dimension, we obtain a better result (+6.02%) in LTC and (+2.71%) DFR. We think that the result is better when we force any dimension and it seems that the dimensions are not equivalent. The anatomy is probably important because it is discriminating and non ambiguous, while pathology is more ambiguous (e.g. fracture of a cranium, fracture of a finger, etc.). Thus, we prefer the following hypotheses:

H4: *Relevant documents contain the anatomy, or else the pathology, or else the modality*. In this case, the request is represented as follow: $Q = \{Anatomy_{(priority=1)}, Pathology_{(priority=2)}, Modality_{(priority=3)}\}$. This hypothesis proposes an importance order on dimensions. We still obtain an increase in performance. Finally, and as the modality is not always present in documents, we propose the following hypothesis:

H5: *Relevant documents contain the anatomy and the pathology dimensions*. Thus, we obtain our best result: +24.94% (LTC) and +14.89% (DFR).

The efficiency difference between the dimensions can be explained by the fact that our technique of dimensions identification is not reliable[3]. Results show that taking into account query dimensions can enhance average precision. Actually, our approach allows to structure query, and thus precise it. Result obtained after the filtering by dimensions is complementary to results obtained by conceptual indexing. Indeed, the conceptual representation makes it possible to identify the query dimensions. The multi-dimensional filtering is a way to surpass the limits of the VSM that does not take into account relations between concepts of each vector and thus ignores the semantic content of document/query.

6. Conclusion and future work

In this paper, we proposed a new approach to solve multi-dimensional queries. First, we defined domain dimensions through external resources. Then, using a conceptual indexing and based on the defined domain dimensions, we identified dimensions from documents and queries in order to represent their semantic content. Thus, we proposed a new indexing language that takes into account the dimensions for better interpreting and representing the semantic document content. We also proposed a new query language that takes into account the dimensions for a better interpretation of the user needs and to represent the semantic content of a query.

Through an experimental evaluation on the ImageCLEFmed-2005 collection, we evaluated the filtering part of our approach. We set up the query module and show that our approach leads to an improvement of the mean precision by about 25%.

The results obtained so far encourage us to explore further the multi-dimensional approach. In the near future, we will implement a testing framework to conduct experi-

[3] We should verify manually the extraction and estimate a percentage of reliability.

ments on the entire approach. We will also study how our approach can be generalized to corpuses covering several domains. For this purpose we will introduce a notion of dimension relevance to evaluate how well a dimension describes the content of a document.

References

[1] Amati G. and Van Rijsbergen C.J., Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. ACM Transaction on Information Systems, 20(4) :357-389, October 2002.

[2] Aronson A. R., Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetamapPprogram. In AMIA 2001 Annual Symposium on biomedical and health informatics, pages 17-27, 2001.

[3] Aronson A. R., Rindflesch T. C. and Browne A. C., Exploiting a Large Thesaurus for Information Retrieval. In Proceedings of RIAO, pages 197-216, New York, October 1994.

[4] Aussenac-Gilles N. and Mothe J., Ontologies as Background Knowledge to Explore Document Collections. In RIAO 2004 , Avignon, pages 129-142, avril 2004.

[5] Baziz M., Aussenac-Gilles N. and Boughanem M., Désambiguisation et Expansion de Requêtes dans un SRI. Revue des Sciences et Technologies de l'Information (RSTI), Hermes, 8(4) :113-136. 2003.

[6] Mustapha Baziz. Indexation Conceptuelle Guidée par Ontologie pour la Recherche d'Information. PhD thesis, Université Paul Sabatier, 2005.

[7] Baziz M., Boughane M. and Aussenac-Gilles N., Conceptual Indexing Based on Document Content Representation. In Ian Ruthven Fabio Crestani, éditeur, Information Context : Nature, Impact, and Role : 5th International Conference on Conceptions of Li-brary and Information Sciences, CoLIS 2005, volume 3507, pages 171-186. Lecture Notes in Computer Science, Jan 2005.

[8] Bodner R. C., Song F., Knowledge-Based Approaches to Query Expansion in Information Retrieval. AI '96, London, 11:146-158. 1996.

[9] Jean-Pierre Chevallet. X-IOTA : An Open XML Framework for IR Experimentation Application on Multiple Weighting Scheme Tests in a Bilingual Corpus. Lecture Notes in Computer Science (LNCS), AIRS'04 Conference Beijing, vol. 3211, 2004, p. 263-280.

[10] Chevallet J-P., Lim J-H. and Radhouani S., Using Ontology Dimensions and Negative Expansion to Solve Precise Queries in CLEF Medical Task. CLEF Workhop, Working Notes Medical Image Track, Vienna, Austria, 21-23 September 2005.

[11] Clough P., Muller H. The CLEF Cross Language Image Retrieval track 2005. *http://ir.shef.ac.uk/imageclef2005*, visited on November 2005.3

[12] Desmontils E. and Jacquin C., "Indexing a Web Site with a Terminology Oriented Ontology". In I.F. Cruz, S. Decker, J. Euzenat et D.L. McGuinness, eds., The Emerging Semantic Web. IOS Press, pages 181- 198, 2002.

[13] Gonzalo J., Verdejo F., Chugur I. and Cigarran J., Indexing with Wordnet Synsets can Improve Text Retrieval. In Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, pages 38-44, Montreal, Canada, 1998.

[14] W. R. Hersh and L. C. Donohoe. Saphire International : A Tool for cCrosslanguage Information Retrieval. In Proceedings of the American Medical Informatics Association Annual Fall Symposium, pages 673-677, 1998.

[15] Hyvönen E., Saarela S., Styrman A. and Viljanen K., Ontology-Based Image Retrieval. In WWW (Posters), 2003.

[16] Kefi L. Berrut C. and Gaussier E.. Un Modèle de RI Basé sur des Critères d'Obligation et de Certitudes. In CORI'06 COnférence en Recherche d'Information, Lyon (France), page 155-160, 15-17 mars 2006.

[17] Radhouani S., Maisonnasse L., Lim J-H., Le T-H-D. and Chevallet J-P., Une Indexation Conceptuelle pour un Filtrage par Dimensions, Expérimentation sur la base médicale ImageCLEFmed avec le méta thésaurus UMLS, in COnférence en Recherche d'Information et Applications CORIA'2006, Lyon France, 15-17 mars, 2006.

[18] Qiu Y. and Frei H-P., Concept-Based Query Expansion. In SIGIR '93 : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 160-169, New York, NY, USA, 1993. ACM Press.

[19] Ellen M. Voorhees. Query Expansion Using Lexical-Semantic Relations. In SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 61-69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[20] Salton, G.: Automatic Text Processing : The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley (1989).

[21] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. In Proceedings of the 32nd annual meeting of the association for computational linguistics, Las Cruses, New Mexico, 1994.