

## Exploiting disambiguated thesauri for information retrieval in metadata catalogs <sup>?</sup>

J. Nogueras-Iso, J. Lacasta, J.A. Banares, P.R. Muro-Medrano, and  
F.J. Zarazaga-Soria

Computer Science and Systems Engineering Department, University of Zaragoza  
María de Luna, 1. 50018-Zaragoza (Spain)  
fjnog, jlacasta, banares, prmuro, javyg@unizar.es

**Abstract.** Information in Digital Libraries is explicitly organized, described, and managed. The content of their data resources is summarized into small descriptions, usually called metadata, which can be either introduced manually or automatically generated. In this context, specialized thesauri are frequently used to provide accurate content for subject or keyword metadata elements. However, if a Digital Library aims at providing access for the general public, it is not reasonable to assume that casual users will use the same terms as the keywords used in metadata records. As an initial step to fill the semantic gap between user queries and metadata records, the authors of this paper already created a method for the semantic disambiguation of thesauri with respect to an upper-level ontology (WordNet). This paper presents now the integration of this disambiguation within an information retrieval system, in this case adapting the vector-space retrieval model. Thanks to the disambiguation, both metadata records and queries can be homogeneously represented as a collection of WordNet synsets, thus enabling the computing of a similarity value, which ranks the results.

### 1 Introduction

As opposite to the largely unstructured information available on the Web, information in Digital Libraries (DLs) is explicitly organized, described, and managed. In order to facilitate discovery and access, DL systems summarize the content of their data resources into small descriptions, usually called metadata, which can be either introduced manually or automatically generated (index terms automatically extracted from a collection of documents). The focus of this paper is DLs working with metadata records using an agreed metadata schema.

---

<sup>?</sup> The basic technology of this work has been partially supported by the Spanish Ministry of Science and Technology through the projects TIC2000-1568-C03-01 from the National Plan for Scientific Research, Development and Technology Innovation and FIT-150500-2003-519 from the National Plan for Information Society; and by the Aragón Government through the project P089/2001. The work of J. Lacasta has been partially supported by a grant from the Aragón Government and the European Social Fund (ref. B139/2003)

Indeed, most DLs use structured metadata in accordance with recognized standards such as MARC21 (<http://lcweb.loc.gov/marc/marc.html>) or Dublin Core (<http://www.dublincore.org>). Moreover, in order to provide accurate metadata, metadata creators use specialized thesauri to fill the content of typical keyword sections. According to ISO-2788 (norm for monolingual thesauri), a thesaurus is a set of terms that describe the vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (e.g. synonyms, broader terms, narrower terms and related terms) are made explicit. Thesauri provide a specialized vocabulary for the homogeneous classification of resources and for supplying users with a suitable vocabulary for the retrieval.

Works like [1,2] present systems that use thesauri as the basis for discovery services. For instance, first work aims at identifying human experts in different subjects of an application domain. There, a concept index was built manually and experts were associated with these concepts. After the user specifies a set of concepts, the system searches for experts who either know about one of those concepts or know about concepts "closely" related to "the user's concepts of interest". That is to say, the system evaluates the semantic relatedness using the network representation of the thesaurus. The hits returned are ranked according to the distance between query concepts and the concepts assigned to each expert.

However, if a DL aims at providing access to the general public (not only constrained to the community of experts that created the resources in the DL), it is not reasonable to assume that casual users will use the same query terms as the keywords used in metadata records. This discordance between query terms and metadata keywords is even worse in the case of DLs handling resources from different application domains, where metadata creators have probably used different thesauri (increasing the heterogeneity of keywords). This situation implies that discovery in DLs cannot be implemented as a simple word matching between the user queries and metadata records. On the contrary, a DL should be able to understand the sense of the user's vocabulary and to link these meanings to the underlying concepts expressed by metadata records.

In order to fill the semantic gap between user queries and metadata records, we proposed in [3] a method for the semantic disambiguation of thesauri with respect to an upper-level ontology, which is closer to the user expressions. Concepts contained in user queries are usually extracted by means of natural language processing techniques (beyond the scope of this paper) that also make use of similar upper-level ontologies. Therefore, it seems reasonable to use the semantic disambiguation of thesauri as a mechanism that harmonizes concepts in metadata records and user queries. In particular, our method provides the disambiguation against WordNet [4], a large-scale lexical database developed from a global point of view that can provide a good kernel to unify, at least, the broader concepts included in distinct thesauri. Our method can be classified as an unsupervised disambiguation method and applies a heuristic voting algorithm that makes profit of the hierarchical structure of both WordNet and the thesauri. Whereas thesaurus hierarchical structure provides the disambiguation

context for terms, the hierarchical structure of WordNet enables the comparison of senses from two related thesaurus terms.

This disambiguation facilitates a unifying system to express user queries and metadata records but it does not constitute itself the final objective. The final purpose is to integrate this disambiguation within an Information Retrieval System (IRS). One example of this integration could be the case of Oracle InterMedia Text package [5], which offers theme-based retrieval for Document Object Like data. This package, by means of the ABOUT operator, enables the querying for documents that are about certain theme or concepts. Themes are extracted from documents and queries by parsing them using an extensive lexicon together with a knowledge base of concepts and relations. High precision is achieved by a disambiguation and ranking technique called "theme proving" whereby a knowledge base relation is verified in the lexical and semantic context of the text in a document. Two themes prove each other if they are closely connected in the knowledge base either hierarchically or through cross-references. This eliminates many bad hits arising from word sense ambiguities.

As a logical continuation of [3], this work aims at verifying the applicability of our disambiguation method within an information retrieval system. In particular, this paper presents the adaptation of the vector-space retrieval model [6] to the context of metadata catalogs. Other classical models, like the probabilistic or neural-net based models, would probably perform better in more heterogeneous contexts. However, the initial hypothesis was that in this context, where metadata records are the summary of the desired resource, a simple model may provide satisfactory results.

The indexing with WordNet synsets is not new in the context of general text retrieval, [7] shows some experiments and revises some related works. In general, the conclusion of these works is that WordNet indexing can improve performance whenever the disambiguation accuracy rate is high (in some cases not less than 90%). These conclusions are probably not extensible to the IRS proposed in this paper because they were indexing free text and this IRS is constrained to the keywords section of metadata. However, it is expected that the disambiguation accuracy in our IRS will be very high. The first reason is that we are disambiguating the own keywords. As opposed to free text retrieval, we are not going to extract concepts from words that are not essential to the document meaning. Additionally the thesaurus hierarchy provides an accurate and limited context for disambiguation.

The rest of the paper is organized as follows. Section 2 presents the information retrieval system with the adaptation of the indexing technique to the specific features of metadata schemas. The indexing technique makes profit of the metadata keywords section, whose content has been strategically filled in by selecting terms from disambiguated thesauri. Thanks to the disambiguation, both metadata records and user queries can be homogeneously represented as a collection of WordNet synsets (concepts in WordNet), thus enabling the computing of a similarity value, which ranks the results returned by the digital library. Section 3 presents some of the results from the initial experiments of the retrieval

system. It has been tested against a geographic catalog, i.e. a catalog containing metadata records that describe data with some kind of location reference. And finally, this work ends with some conclusions and future lines.

## 2 The retrieval model

An information retrieval model can be defined as the specification for the representation of documents, queries, and the comparison algorithm to retrieve the relevant documents. The vector-space retrieval model [6] proposes a framework in which partial matching is possible and it is characterized by the use of a weight vector representing the importance of each index term with regard to a metadata record (document). Hence, the framework  $F$ , which represents the collection of records and the user queries, consists of a  $M$ -dimensional vector space, where each dimension corresponds with each distinct index term in the glossary (denoted as  $T$  and being  $M$  the size of the glossary). Following expressions show vector representations of a document  $d_j \in D$  (documents in the collection) and a query  $q \in Q$  (set of user queries):

$$d_j = ((t_1; w_{1,j}); (t_2; w_{2,j}); \dots; (t_M; w_{M,j})); q = ((t_1; w_{1,q}); (t_2; w_{2,q}); \dots; (t_M; w_{M,q})) \quad (1)$$

where  $t_1; t_2; \dots; t_M \in T$  are the  $M$  synsets belonging to the glossary;  $w_{i,j}$  represents the weight given to an index term with respect to  $d_j$ ; and  $w_{i,q}$  is the weight given to an index term with respect to  $q$ . Finally, this model provides a function to compute the degree of similarity between each metadata record and a user query  $q$ , enabling the ranking of records with respect to  $q$ . Following equation shows the exact formula to compute the similarity value (denoted as  $Sim(d_j; q)$ ) which is based on the cosine of the angle formed by the vector representing the metadata record and the vector of the user query [8].

$$Sim(d_j; q) = \frac{b_j \cdot l_q}{\|d_j\| \cdot \|q\|} = \frac{\sum_{k=1}^M w_{k,j} \cdot w_{k,q}}{\sqrt{\sum_{k=1}^M w_{k,j}^2} \cdot \sqrt{\sum_{k=1}^M w_{k,q}^2}} \quad (2)$$

Next subsections explain the process to obtain the index terms of metadata records and queries and their weights.

### 2.1 The indexing of metadata records

Before applying a retrieval algorithm, documents (metadata records) in the collection must be summarized into a set of representative keywords called index terms. In this context of metadata catalogs, metadata records are precisely a summary of media documents (image, text or whatever). Furthermore, the advantage in this context is that metadata creators introduce explicitly the concepts within the keywords section. Nevertheless, the retrieval model of a metadata catalog cannot be based uniquely on a simple matching between a query word and the words contained in keywords section. On one hand, different metadata creators may not share the same criteria to select a harmonized (homogenous) set of keywords. And on the other hand, this simple matching would

be comparable with a classic Boolean information retrieval model, where query terms are compared with keywords contained in records to decide whether the record is relevant or not without providing any ranking.

As mentioned in the introduction, one way to increment the descriptive potential of the keywords section is to select terms belonging to formalized controlled lists of terms or thesauri. In this way, more sophisticated methods to resolve terminological queries could be applied. However, there is not a universal thesaurus to classify every type of resource and metadata creators make use of different thesauri or controlled lists depending on the application domain. Therefore, the set of keywords, although using thesauri and controlled lists, are still quite heterogeneous. For example, in the context of geographic information, catalogs may include geographic information about topography, cadastre or communications. Hence, we proposed in [3] the semantic disambiguation of thesaurus terms to avoid this heterogeneity. The main objective of this semantic disambiguation method is to relate the different thesauri to an upper-level ontology like WordNet [4].

**Fig. 1.** Example of thesaurus branches

WordNet is structured in a hierarchy of synsets which represent a set of synonyms or equivalent terms. When thesauri are imported into our catalog system database, the initial step of the disambiguation process is to divide the thesaurus into branches (a branch corresponds to a tree whose root is a term with no broader terms and that is constituted by all the descendants of this term in the "broader term/narrower term" hierarchy). The branch provides the disambiguation context for each term in the branch. Secondly, the disambiguation method finds all the possible synsets that may be associated with the terms in a thesaurus branch. And finally, a voting algorithm is applied where each synset related to a thesaurus term votes for the synsets related to the rest of terms in the branch. The main factor of this score is the number of subsumers in synset paths (the synset and its ancestors in WordNet). The synset with the highest score for each term is elected as the disambiguated synset. Table 1 shows the final score of synsets for the branch accident in Fig. 1. For the sake of clarity, some terms and their corresponding synsets have not been shown. A more detailed explanation of the algorithm to obtain the score can be found in [3].

**Table 1.** Disambiguation of a thesaurus branch

Term	Subterm	Synset path	score	lia
accident		event! happening! trouble! misfortune! mishap! accident	3,143	0,551
		event! happening! accident	2,560	0,449
accident! accident source	accident	event! happening! trouble! misfortune! mishap! accident	2,304	0,552
		event! happening! accident	1,873	0,448
	source	entity! object! artifact! creation! product! work! publication ! reference	0,713	0,231
		entity! object! location! point! beginning	0,705	0,228
		entity! object! artifact! facility! source	0,685	0,221
		entity! life_form! person! communicator! informant	0,397	0,128
		entity! life_form! person! creator! maker! generator	0,397	0,128
		psychological_feature! cognition! content! idea! inspiration ! source	0,186	0,060
		abstraction! relation! social_relation! communication ! written_communication! writing! document! source	0,009	0,003
		accident! accident source! oil slick		
		entity! object! ~lm! oil_slick	0,214	1,000
		...		

Therefore, once a new metadata record has been completed, it is possible to obtain the collection of synsets corresponding to the thesaurus terms. Besides, as the metadata creator probably selected terms from different thesauri, there may be repetition of synsets in the obtained collection. Hence, given the keywords section of a metadata record, it is possible to extract a collection of synsets, which are indeed the index terms and may be characterized by a weight proportional to the number of occurrences and the liability of the disambiguated synset.

As concerns the vector model, one of the best weighting schemes for index terms (the synsets) is the one proposed in [8], which tries to balance the effect of intra-clustering similarity (features that better describe a subset/cluster of documents in the collection) and inter-clustering dissimilarity (features which better distinguish a subset from the remaining documents in the collection) of documents (see equation 3). Assuming this weighting scheme, the first step to calculate the weight of a synset is to obtain the frequency of a synset  $t_i$  in a metadata record  $d_j$ . For a classical information retrieval system, this frequency (denoted as  $freq_{i,j}$ ) would be simply the number of occurrences of an index term. But in this case, we cannot obviate that the disambiguation of thesaurus terms is heuristic and we wanted to consider the score obtained for each synset in the disambiguation process. Therefore, given a thesaurus term  $s$ , we have estimated the liability of the elected synset  $t_i$  with respect to the other non-elected synsets which were initially associated with the term  $s$ . This liability value, denoted as  $lia_{s,t_i}$ , is computed as the division between the score of the elected synset and the sum of the scores of all synsets related to a thesaurus term. Column *lia* in table 1 shows an example of such percentage.  $freq_{i,j}$  is finally computed as the sum of the liability of each synset  $t_i$  that is indirectly referenced by the terms included in a metadata record  $d_j$ . Secondly, it is necessary to obtain the normalized frequency  $f_{i,j}$ , which is computed as the division between  $freq_{i,j}$  and the maximum frequency (computed over all synsets  $t_i$  referenced by  $d_j$ ). Next step is the calculation of the inverse frequency  $idf_i$  of a synset  $t_i$ , i.e. the division

between the size of the collection (denoted as  $N$ ) and the number of records referenced by this synset (denoted as  $n_i$ ). The point here is that if a synset is referenced in many metadata records, it is not very useful to discriminate them. Finally, the total weight  $w_{i,j}$  is computed as the product between  $f_{i,j}$  and  $idf_i$ .

$$freq_{i,j} = \prod_{s \in d_j} lia_{s,i}; f_{i,j} = \frac{freq_{i,j}}{\max_{t_i}(freq_{i,j})}; idf_i = \log N/n_i; w_{i,j} = f_{i,j} \cdot idf_i; \quad (3)$$

Additionally, subsection 3.2 proposes a variant of the indexing to augment the number of index terms for each metadata record.

## 2.2 The indexing of queries

Regarding the queries formulated by users, it is also necessary to find index terms characterizing these queries. Indeed, the query performed by the user specifies, although vaguely, the set of metadata records that he/she wants to discover. As well as metadata records have been summarized into a collection of synsets, queries must be also synthesized into a set of WordNet synsets. That is to say, in parallel to the indexing of metadata records, every word belonging to the query must be searched into WordNet and then, their possible senses, in the form of synsets, should be processed to obtain a representative collection of synsets. The first question here was whether we should also try the disambiguation of queries or not. By disambiguation of queries it is meant the election the synset that better represents each query word among its possible synsets found in WordNet. In the context of our experiments it was assumed that the queries contained only a few words and not necessarily connected (i.e. with no synsets in common). Therefore the final decision was the non-disambiguation of queries. Besides, some works like [9] showed that trying to disambiguate the query in addition to the corpus made the results worse, especially in cases where the query was very short. Additionally, it must be mentioned that the use of synsets provides an implicit expansion of query words because each synset represents a set of synonyms (the word typed by the user and all its possible synonyms). In [10] Voorhees essayed different strategies for query expansion using the different types of associations between WordNet synsets and she concluded that apart from synonyms the use of other types of related terms does not necessarily improves information retrieval.

Finally, regarding query weights, a variant from the weighting scheme in [8] is applied to compute the weight of every synset with respect to the query  $q$  (see eq. 4). This variant, suggested in [11], gives a minimum weight of 0.5 to the normalized frequency. In this case,  $freq_{i,q}$  is computed as the number of indirect references to the synset  $t_i$ .

$$w_{i,q} = (0.5 + 0.5 \cdot (freq_{i,q} / \max_{t_i}(freq_{i,q}))) \cdot idf_i \quad (4)$$

### 3 Testing the retrieval model

#### 3.1 Metadata corpus and conditions of essays

The formal precision and recall measures used to quantify retrieval effectiveness of IR systems are based on evaluation experiments conducted under controlled conditions. This requires a testbed comprising a fixed number of documents, a standard set of queries, and relevant and irrelevant documents in the testbed for each query. This is the case of TREC (<http://trec.nist.gov/>), an annual conference for academic and industrial text retrieval systems, which provides 2 GB document collection with about half a million documents. However, we could not find such a controlled testbed in the context of metadata catalogs. Therefore, our approach has been based on the comparison of our results with the results obtained from other metadata catalogs storing similar metadata records. That is to say, we have had to compare retrieval effectiveness in terms of qualitative statements and the number of metadata records retrieved.

For the purpose of comparison, we selected the Geoscience Data Catalog (<http://geo-nsdi.er.usgs.gov/>) produced by the U.S. Geological Survey. This catalog contains around 1,000 metadata records describing geospatial data (data that is associated directly or indirectly with a location on the Earth's surface). The metadata records are compliant with the american standard CSDGM (Content Standard for Digital Geospatial Metadata, <http://www.fgdc.gov>), which includes a keywords section where the metadata creator can specify different values and the thesauri to which they belong. One of the reasons to select this catalog was our experience in Spatial Data Infrastructures. However, the results of this work are extensible to any type of digital library using metadata schemas that contains a keyword section. Another important reason to select this catalog was that it provides a search engine, based on ISearch software [12], that enables the comparison of results. ISearch is the search component of ISite, an open source package for indexing and searching documents that implements the ISO 23950 information retrieval protocol. And it supports full text and field based searching using the same ranking algorithm as the SMART retrieval system [6], which is precisely the origin of the vector-space retrieval model.

Therefore, ISearch and our IRS should be comparable. To simulate similar conditions for our information retrieval system, the metadata records were downloaded from this catalog and imported in our metadata database. However, only 753 of the imported records contained thematic keywords. The initial problem of the imported records was that we had uniquely disambiguated the thesauri: NGMDB ("National Geologic Map Database Catalog themes, augmented", <http://ngmdb.usgs.gov/>) with 72 terms appearing 1105 times in the collection; and GTE ("Gateway to the Earth", <http://alexandria.sdc.ucsb.edu/> » <http://hill.usgs-terms/usgs/html9/>) with 648 terms appearing only 144 times in the collection. In short, there were only 340 metadata records that included terms of disambiguated thesauri, an average of 3.673 keywords for those records. However, there were 656 records with an average of 7.87 terms belonging to unspecified thesauri, which were entitled in metadata records as "General" or "none". Therefore, we tried to transform



some keywords from none and General thesauri into terms belonging to GEMET, NGMDB and GTE. In particular, we selected GEMET ("General European Multilingual Environmental Thesaurus", <http://www.mu.niedersachsen.de/cds/>) because it is a quite comprehensive thesaurus for geographic information that consists of 5,542 terms organized in 109 branches and translated into 12 languages. In this transformation, we also solved some small morphological differences between the included terms and the terms of the disambiguated thesauri, e.g. difference between singular and plural versions. Thanks to this modification of metadata records, 711 records contain an average of 5.594 theme keywords belonging to the three disambiguated thesauri.

### 3.2 The experiments

After analyzing the synsets that were referenced indirectly by the metadata records of our catalog, we obtained that 201 synsets were referenced by 707 metadata records, each record referencing an average of 4.065 synsets and 20 synsets at maximum. And the minimum, maximum and average values for  $n_i$  were 1, 15.228 and 347.

As a first search example, the query *geology erosion* was performed obtaining the results presented in table 2 (only the 3 first results). The query contains two words that are associated with 5 WordNet synsets, whose weights are displayed in table 3. The weights of synsets referenced by the first hit are presented in table 4 and the similarity for the first metadata record is given by

$$Sim(d_j; q) = \frac{5.46 \cdot 3.27}{3.27^2 + 5.46^2 + 5.86^2} \cdot \frac{0.89^2 + 0.71^2 + 0^2 + 5.46^2 + 0^2}{0.89^2 + 0.71^2 + 0^2 + 5.46^2 + 0^2} = 0.37 \quad (5)$$

Table 2. Returned results for query geology erosion

Order	Title	Sim
1	Beach profile data for Maui, Hawaii	0.375
2	Beach profile data for Oahu, Hawaii	0.375
3	Possible Costs Associated with Investigating and Mitigating Some Geologic Hazards in Rural Parts of San Mateo County, California	0.318
...		

Table 3. Computation of synset weights for the query

Word	Synset	$freq_{i,q}$	$f_{i,q}$	$n_i$	$w_{i,q} = (0.5 + 0.5f_{i,q}) \cdot \ln d_i$
Geology	4655198 (a science that deals with the history of the earth as recorded in rocks)	1	1/1	288	$1 \times \ln(707/288)=0.89$
	6691504 (geological features of the earth)	1	1/1	347	$1 \times \ln(707/347)=0.71$
Erosion	9691024 (the mechanical process of wearing or grinding something down)	1	1/1	0	0
	10413485 (condition in which the earth's surface is worn away by the action ...)	1	1/1	3	$1 \times \ln(707/3)=5.46$
	9691547 (erosion by chemical action)	1	1/1	0	0

**Table 4.** Computation of synset weights for the first 2 hits (N = 707)

Id	Thesaurus, Keyword	synset	$ I a_i$	$freq_{i,j}$	$f_{i,j}$	$n_i$	$idf_i$	$w_{i,j}$
1, 2	GEMET, Coastal Erosion	10413485	0.6	0.6	0.6/1	3	5.46	3.27
	"	6801422	1	1	1/1	3	5.46	5.46
	GEMET, Beach	6739108	1	1	1/1	2	5.87	5.86

One effect that can be observed from this first query is the influence of the inverse frequency and the number of keywords used in each metadata record. For instance, that is the reason to explain the relevance of metadata record in 3rd position. Although, it references three synsets (10413485, 4655198 and 6691504) that match with synsets in the query, its similarity to the query is lower than the similarity of 1st record (only one match with query synsets). On one hand, weights for synsets 4655198 and 6691504 have a low weight because of the inverse frequency. The synsets related with *geology* are very frequent in the collection and the vector-space model balance this effect with a low value for the  $idf_i$ . That is to say, this weighting scheme considers that the fewer a term occurs in, the more important it must be. Sometimes this is not satisfactory, but more often it is useful. And on the other hand, metadata record in 3rd position references a total number of 13 synsets. As the number of referenced synsets grows, the norm of the vector representing the record will increase, increasing as well the denominator in the similarity formula. This denominator favours metadata records with fewer keywords. Although some times this means that such metadata records are better focused on a subject, other times is simply due to a worse quality in metadata cataloguing. It was tested the possibility of obviating the denominator (always equals to one). But this variation was rejected because the results were not satisfactory. There was almost no graduation (a great deal of hits shared the same similarity value) for the similarity in simple queries as the previous one. Besides, as the number of terms in the query increases, the norm of the vectors representing the records is not so influential.

After this first example, we wanted to test one of the obvious advantages of our information retrieval system in comparison with ISearch software. This advantage was that the queries can contain words that have not been necessarily included in metadata keywords, e.g. synonyms of these keywords that match with the same WordNet synsets. For instance, we performed two queries with two synonyms, *fuel* (or *fuels*) and *combustible*, which correspond with the same WordNet synset (10669661, *a substance that can be burned to*). Our IRS always returned 138 hits but ISearch only returned records (138 hits with same score) for the query *fuels*, which was the word included in the keyword section. Basically, the hits returned by our IRS were graduated by the number of synsets indirectly referenced: 2 synsets for the first 18 hits, 3 synsets for the following 35 hits and so on.

In a second series of tests, we decided to augment the number of synsets representing the metadata records. For this expansion, we included the disambiguated synsets that were associated to the broader terms of the terms included in keyword section. For instance, the broader term of *coal* in GEMET is *fossil*

*fuel*, and thus metadata records with term *coal* will be indexed with the disambiguated synset of *coal* (10628288, *carbonized vegetable matter deposited in :::*) as well as with the disambiguated synset of *fossil fuel* (10527530, *fuel consisting of the remains of organisms preserved in rocks :::*). The idea was that if a user asks for resources about *fossil fuel*, he might be interested in different types of fossil fuels (e.g. *coal*, *natural gas* or *petroleum*). Of course, the weight of the synset for the broader term must be lower than the weight for the real term included in the metadata record. In particular, the liability of the synsets which are associated with broader terms was divided by 2. With this expansion we obtained that 272 synsets were referenced by 709 metadata records, each record referencing an average of 5.988 synsets and 29 synsets at maximum. And the minimum, maximum and average values for  $n_i$  were 1, 16.577 and 362. Thanks to this modification, our IRS returned 121 hits for the query *fossil fuel*, one hit more than the query *coal*. Meanwhile, ISearch returned no hits for query *fossil fuel*. This is due to the fact that ISearch only performs simple word matching, and only the word *fuels* (the plural version of *fuel*) is included in metadata records.

## 4 Conclusions and future lines

This paper has presented the adaptation of a vector-space information retrieval model to the context of metadata catalogs. The indexing of metadata records assumes that the metadata schema includes a keyword section or subject element, something quite usual in most metadata schemes. Besides, the indexing technique is based on the inclusion in this section of terms selected from disambiguated thesauri. The index terms are precisely the synsets associated with the selected thesaurus term during the disambiguation process of the thesaurus. The viability of the retrieval model has been tested with a catalog containing metadata records, which describe geographic resources. The first experiments showed that the initial proposal of the method provided an acceptable ranking for the expected records. And in a second series of experiments, the indexing of metadata records was modified to augment the number of index terms. Apart from collecting the synsets associated with a thesaurus term, the indexing method also included the synsets associated with the broader terms in the thesaurus hierarchy. These synsets coming from broader terms were assigned a lower weight. This modification was based on the assumption that metadata records represented by these synsets (from broader terms) are still semantically close to queries including the broader concept. This expansion could have been also continued with the synsets associated with other related terms. However, works like [1] suggest not considering concepts at distance two or more from an initial concept. Anyway, it is necessary to test the method with a bigger corpus of metadata records and better classified with additional disambiguated thesauri.

The main disadvantage of the IRS presented in this paper is that the thesauri disambiguation may not be adequate for very specific domain ontologies. WordNet is an upper-level ontology that lacks domain-specific terminology. Nevertheless, the intention of this work is to approximate as much as possible the terms

used in metadata records and the concepts extracted from "general-purpose" queries. And WordNet is a public domain electronic lexical database which may be considered as one of the most important resource available to researchers in computational linguistics, text analysis and many related areas.

On the other hand, an improvement in the computation of the weight of each index term would be to consider the importance of the thesaurus, to which the terms in the keyword section belong. A term selected from a specific thesaurus like GEMET may be more relevant than a term belonging to a thesaurus that compiles only a hundred of categories. Finally, it must be mentioned that this retrieval method could be extended by indexing other metadata fields (or elements) like *title*, or *abstract*. Besides, the value of similarity could be integrated into more complex information retrieval systems as another factor to compute the final value for the degree of similarity.

## References

1. Clark, P., Thompson, J., Holmback, H., Duncan, L.: Exploiting a thesaurus-based semantic net for knowledge-based search. In: Proc 12th Conf on Innovative Application of AI (AAAI/IAAI'00). (2000) 988{995
2. Swoboda, W., Kruse, F., Nikolai, R., Kazakos, W., Nyhuis, D., Rousselle, H.: The UDK Approach: the 4th Generation of an Environmental Data Catalogue Introduced in Austria and Germany. In: Proc. of 3rd IEEE META-DATA Conference, Bethesda, Maryland (1999)
3. Mata, E.J., Ansó, J., Bañares, J.A., Muro-Medrano, P.R., Rubio, J.: Enriquecimiento de tesauros con wordnet: una aproximación heurística. In: Actas IX CAEPIA, Gijón (2001) 593{602
4. Miller, G.A.: Wordnet: An on-line lexical database. Int. J. Lexicography 3 (1990)
5. Mahesh, K., Kud, J., Dixon, P.: Oracle at TREC8: a lexical approach. In: Proc. of 8th Text REtrieval Conference (TREC-8), Maryland (1999) 207{216
6. Salton, G., ed.: The SMART retrieval system - Experiments in Automatic Document Processing. Prentice Hall, Inc., Englewood Cliffs, NJ (1971)
7. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with WordNet synsets can improve Text Retrieval. In: Proc. COLING/ACL'98 Workshop on Usage of WordNet for Natural Language Processing. (1998)
8. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)
9. Voorhees, E.M.: Using WordNet to disambiguate Word Senses for Text Retrieval. In: SIGIR '93, Proc. 16th annual international ACM SIGIR conf. on Research and Development in Information Retrieval. (1993) 171{180
10. Voorhees, E.M.: On Expanding Query Vectors with Lexically Related Words. In: Text REtrieval Conference. (1993) 223{232
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24 (1988) 513{523
12. Nassar, N.: Searching With Isearch, Moving beyond WAIS. Web Techniques magazine, [www.webtechniques.com](http://www.webtechniques.com) (1997)