Proceedings of the 5th AGILE Conference on Geographic Information Science. 2002, p. 61-66

5th AGILE Conference on Geographic Information Science- Palma (Mallorca, Spain) 25-27 April 2002

Semantic Disambiguation of Thesaurus as a Mechanism to Facilitate Multilingual and Thematic Interoperability of Geographical Information Catalogues

E. Mata[1], J.A. Bañares[2], J. Gutierrez[2], P.R. Muro-Medrano[2], J. Rubio[1]

[1] Dpto. De Matemáticas y Computación. Universidad de la Rioja,
{eloy.mata, julio.rubio}@dmc.unirioja.es

[2] Dpto. De Informática e Ingeniería de Sistemas. Universidad de Zaragoza
{banares, adsogu, prmuro}@posta.unizar.es

Abstract: Nowadays, there is a growing interest within the geographic information system community to find the location of geographic data through the Internet and to know the features and the possibilities of these data. In order to make this information easily accessible both to trained users and to the general public it is necessary to implement a specific infrastructure of geographical information providing advanced classification and searching services. The ISIGIS project is aimed at creating such an infrastructure in Spain. In this context, when a geographic information Catalog collects the description of geospatial information from different contexts, it is not reasonable to assume that a particular user outside the community where the data were created will describe the sought information by means of the same terms or keywords in the Catalog. Then, the search service of the Catalog cannot be a mere keyword recognition process. It should be able to understand the sense of the user's vocabulary and to link these meanings to the terms in the thesaurus (or thesauri) used in the catalog. In this paper we present an heuristic method based on a voting system to tackle this disambiguation problem in geographic Catalogs related to the ISIGIS project. Our proposal can be considered as an unsupervised disambiguation method based on the hierarchical structure of both WordNet and the thesauri.

Keywords: Information Retrieval, Semantic Disambiguation, Thesauri, WordNet, Catalog, Geographic Information Systems

1 Introduction

The OpenGIS Consortium (http://www.opengis.org) uses the term Catalog to describe the set of service interfaces supporting the organisation, discovery and access to geospatial information (see [1]). Catalog services help the users or the application software to find information in any point of a distributed computing environment. A Catalog can be considered as a specialized database of information on the geospatial resources available to a group or community of users.

The Catalog contains metadata that describe the capacities and the contents of the geospatial data. The metadata management should be assisted by tools in order to get data adjusted to the metadata standards. The use of standard representations of the metadata, such as the American Federal Geographic Data Committee (FGDC) standard or the ISO TC 211 standard (currently in draft version), facilitates the interoperability tasks. Since the different metadata standards share a good number of fundamental elements, it is sensible to produce programming tools which communicate several systems with different metadata dialects by means of convenient pipelines.

A key point in the metadata organisation is the keywords section. Usually, these keywords are extracted from thesauri. This allows the analyst to classify the geodata through a family of terms which represent concepts or categories in one or more knowledge domains (that it to say, a thesaurus in this context plays an ontological role). Such a classification facilitates particularly the retrieval of information from keywords. Nevertheless, a main difficulty arises due to the semantic differences existing among the thesauri used in different communities.

The idea of our work is to take advantage of both the available lexical resources (thesauri, dictionaries, lexical knowledge bases, ontologies, …) and the natural language processing techniques with two aims; firstly, to improve the organization of the information and its retrieval, and secondly, to obtain a friendly and multilingual access to geographic information. The main difficulty in the implementation of a concept-based search is the problem of the semantic disambiguation. In our approach, this problem is tackled by means of a heuristic method supported by the hierarchical structure of WordNet. Given an ambiguous term, a voting system is used to determine the "closest" meaning to the meaning of some words obtained from a given context, namely a complete branch in a thesaurus. The votes carry some weight that has been experimentally determined in order to avoid experiencing limitations in our first algorithm (without heuristic weights). This method has been integrated into a thesauri management application which is part of a larger project aimed at developing the technology necessary to create a Spanish geographic information infrastructure; i.e., the ISIGIS project [2].

2 The use of thesauri to classify metadata and do advanced searches

The description of geospatial information is made in an objective way, completing the information required by the corresponding metadata standard. However, the classification by keywords is a subjective task. With the aim of reducing the subjectivity of these classifications, we used controlled lists of predefined keywords, where possible. These controlled lists define a set of terms that represent concepts and categories of thematic domains and let us classify data from different sources by means of related terms. The use of predefined keywords facilitates the connection between a selected vocabulary and a collection of geographic data.

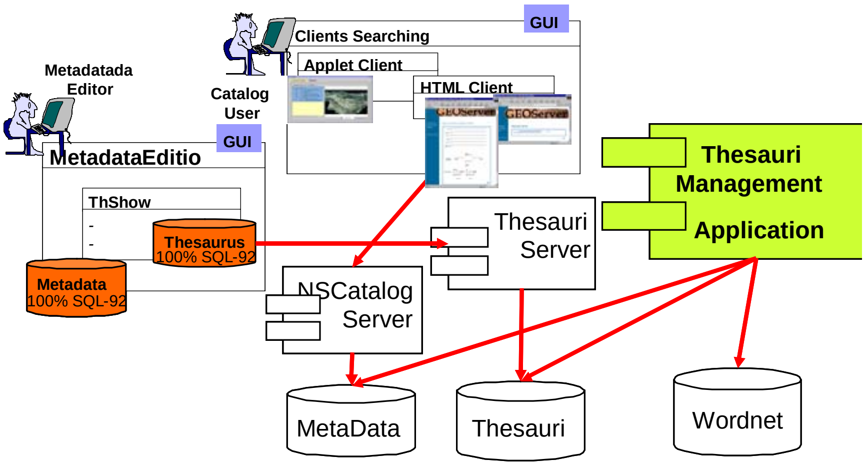


Fig. 1. Metadata capture, thesauri management and searching components.

However, due to the subjectivity of the classification task, even into the same organization or domain, it is possible to find different classifications for the same data. Thus, the controlled lists can be implemented by means of a thesaurus, because in addition to the terms, they provide information on the relationships among them. When the user classifies the information using the terms in a thesaurus, he does it from his own point of view, but the Catalog should allow to share the information among several organisations using different terms, or even languages, to name to the same concepts. Then, not only the relations among the terms in a thesaurus are needed, but also links between different thesauri for the classification of metadata. Moreover, this imprecision or subjectivity in the classification are obvious because although GIS are powerful quantitative tools for geospatial reasoning, man thinks of geographic space in a qualitative way (see [3]). Other additional problem is the use in many thesauri of English terms which are hardly translatable by a user with different a native language or culture.

Ontologies or lexical knowledge bases may be used as a mechanism for thesauri interconnection. There is extensive literature proposing practical ontologies and exploring their properties. In [4] it is shown how ontologies can be used on GIS and how they can contribute to building better information systems. We focus on such as *WordNet* [5] and *EuroWordNet* [6]. They are large-scale lexical databases developed from a global point of view. The point is to enable the user to enter a query so that the system retrieves information on the basis of the concepts in the request. These knowledge bases, unlike dictionaries, organise the words into groups of synonyms and contain relations like *is-a* and *part-whole*. Conceptual queries are inherently hierarchical in nature. We may expand a concept or keyword in the query into a large number of synonyms and narrower or general concepts related by *part-whole* and *is-a* relationships. For example, in order to find spatial references for a political unit, such as a *village*, the query may be expanded with the knowledge on the political subdivisions such as the regions in a country, the villages in a region, etc. The system might lack data at the *village* level, but higher level data including *village* may be enough to answer the query. The main difficulty the system encounters to solve the conceptual query is “to understand” the meaning of the terms in the request, namely the semantic disambiguation problem.

3 Description and evaluation of our semantic disambiguation method.

Word sense disambiguation is perhaps the greatest existing problem at the lexical level in natural language processing [7], and this skill is also applicable to tasks such as machine translation, speech synthesis and information retrieval. A word is polysemic if its sense changes depending on the context. The problem of disambiguation consists in determining which one of the senses of an ambiguous word is invoked in a particular context composed of a set of words related to the ambiguous word.

The different statistical approaches to solve the problem depend on the training material available. Supervised disambiguation methods are based on a labelled training set. Unsupervised disambiguation methods only use unlabeled text corpora. There are also methods that use lexical resources such as machine-readable dictionaries, lexical knowledge bases or thesauri.

The method proposed in this article can be considered as an unsupervised disambiguation method based on the hierarchical structure of WordNet, similar to the methods described in [8, 9].

In this work we used the GEMET thesaurus (*GEneral Multilingual Environmental Thesaurus*) proposed by the *European Topic Centre on Catalogue of Data Source* of the Environmental European Agency. GEMET consists of 5,542 terms organised in 109 branches and translated into 13 languages.

We took the terms of GEMET and, for each word in the thesaurus, we tried to determine the "closest" sense to the senses of the rest of the words in the whole branch.

Instead of working on the idea of the closest sense by means of the probability theory as in [10], we chose a voting system which integrates the words in the context, without assuming the independence hypothesis. We also used the hierarchical structure of WordNet based on the assumption that the more similar two terms are, the more hyperonims they have. We considered three criteria to correct the results obtained in our tests. These criteria are slightly related to the aspects that [8] uses to define conceptual distance (the length of a path of concepts in WordNet, the hierarchy and density depth).

1. Lower level WordNet concepts have longer paths and then, share more subhierarchies. We divided them by the depth of the WordNet concept to solve the problem.
2. All the words in the context must not be valued in the same way. We divided them by the distance between two terms in the thesaurus GEMET. In this way, the closest terms in the structure of GEMET are the most important ones.
3. The most polysemic words in the context vote more times since each one of their senses casts a vote. We divided them by the number of senses of the word.

Our method do not implied a training corpus to estimate probabilities, as in [8] and [9], to calculate the semantic similarity, but we took advantage of the fact that GEMET is a thesaurus with a hierarchical structure which enabled us to value the context words on the basis of their position in the thesaurus.

We took the terms in GEMET for the tests of our disambiguation method and searched them in WordNet. Firstly, we observed that 5,070 out of the 5,542 terms were compound terms; 15% of the words were not found in WordNet and only 24 % of the terms were monosemic.

Secondly, we used morphological techniques to reduce the number of not-found words and to search adjectives associated with a noun, for instance, `administrative` is associated with `administration`. The number of not-found words was reduced to 4.44%, corresponding to verbs and adverbs. Some technical terms were not found either, since WordNet is a global knowledge base. Finally, we concluded that WordNet was a suitable tool for our aim.

We calculated that the prior probability to guess correctly the WordNet sense of a word, randomly and without any disambiguating method, was 0.217 and we tried to check whether our method was able to get better results. We restricted our tests to the `administration` branch of the GEMET thesaurus. Only 2.8% of the words in this branch were not found in WordNet and we got a success rate of 0.77% in the process of disambiguation. Although this was a good result, we observed some failures which could be solved: a) some failures were due to the fact that the largest paths were most voted; others b) to the fact that wrong senses were voted by remote terms in the thesaurus; and others c) to the fact that the most polysemic words cast a vote for each one of their senses, many of them erroneous. Then we decided to improve the voting method taking into account these three factors and we reached a success rate of 81%.

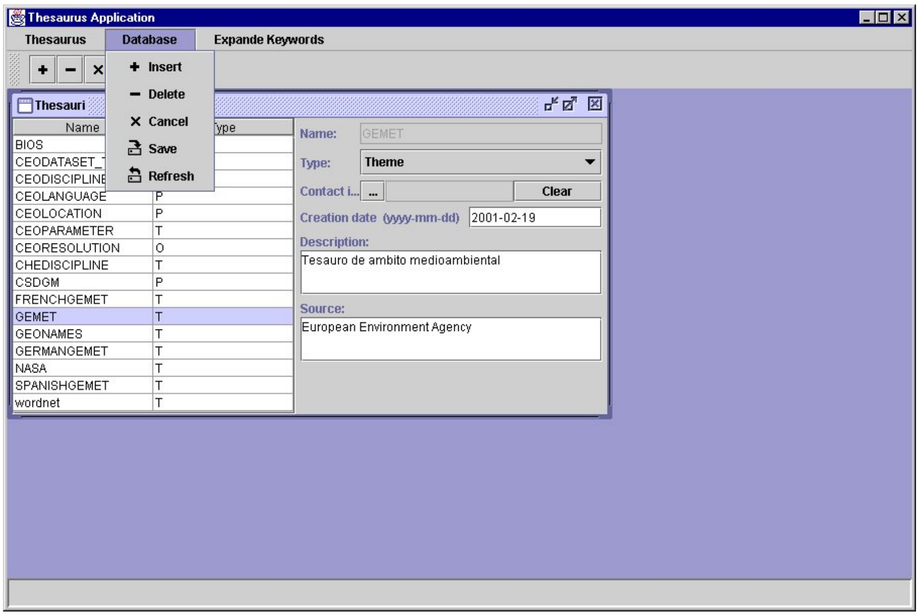
Thirdly, we decided to use other information contents in the thesaurus. We started with synonyms. As for monosemic synonyms, we used the concept they represent; otherwise, in the case of polysemic synonyms, we reduced the polysemy of the original term by joining both terms together. Although GEMET includes few synonyms, the result was a success rate of 91%. The results with other branches were even better with values between 90% and 96%.

Finally, we used related terms and the glosses of the definitions, but in this case, the method did not yield better, and furthermore, the complexity both in terms of time and memory was higher.

4 A software tool for thesauri management

In this work we have developed some software tools to help Catalog agents involved in the creation of geographic metadata. This software tools provide different performances depending on the characteristics of the user. At the lowest functional level, the system allows only to create and edit metadata. This tool works with low-cost databases (it only requires that the access can be done through JDBC with some database manager system such as Access, mySQL, Oracle, etc. to preserve the metadata using a model written in SQL-92). This tool includes the necessary thesauri and allows to display and browse into a thesaurus to select the keywords to classify the data.

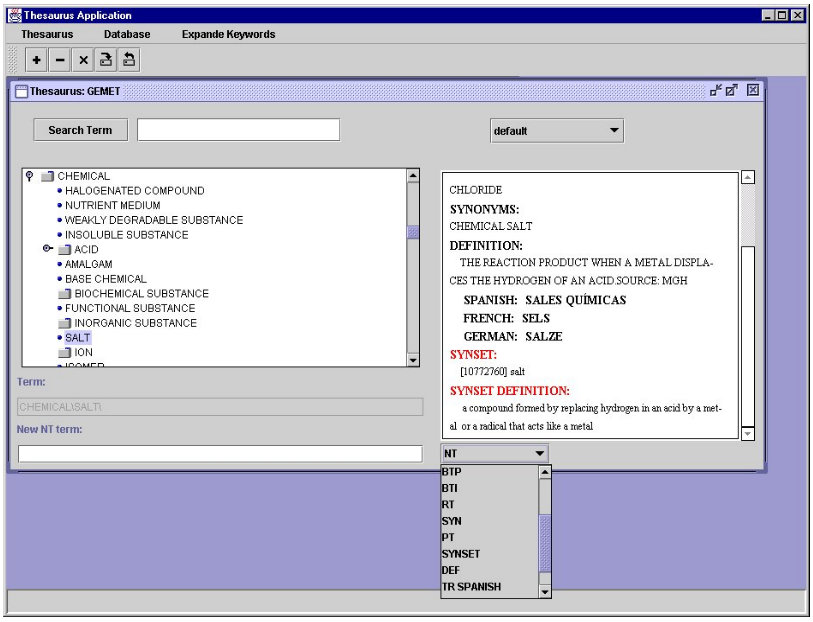
The highest functional tool is aimed at the advanced metadata users, and at Catalog managers too, responsible for the management and improvement of the metadata. The system works in Oracle 8i and takes advantage of its ability to manage spatial objects, text documents and thesauri. This tool presents the required functionality to improve quality of the metadata: thesauri management, derivation and automatic improvement of metadata. The tool also offers the improvement in the Catalog searching utilities supported by the disambiguation method presented in this paper.



Name	ype
BIOS	
CEODATASET_	
CEODISCIPLINE	
CEOLANGUAGE	P
CEOLOCATION	P
CEOPARAMETER	T
CEORESOLUTION	O
CHEDISCIPLINE	T
CSDGM	P
FRENCHGEMET	T
GEMET	T
GEONAMES	T
GERMANGEMET	T
NASA	T
SPANISHGEMET	T
wordnet	T

Name: GEMET
Type: Theme
Contact i...: ...
Creation date (yyyy-mm-dd) 2001-02-19
Description:
Tesauro de ambito medioambiental
Source:
European Environment Agency

Fig. 2. Software application to import thesauri



CHEMICAL\SALT\

CHLORIDE

SYNONYMS:

CHEMICAL SALT

DEFINITION:

THE REACTION PRODUCT WHEN A METAL DISPLACES THE HYDROGEN OF AN ACID.SOURCE: MGH

SPANISH: **SALES QUÍMICAS**

FRENCH: **SELS**

GERMAN: **SALZE**

SYNSET:

[10772760] salt

SYNSET DEFINITION:

a compound formed by replacing hydrogen in an acid by a metal or a radical that acts like a metal

Fig. 3. Software application to display, manage and create relations among thesauri

Figure 2 shows the software application to import, export and update thesaurus. The process of semantic disambiguation is automatically performed during the import process. Figure 3, shows in red, the Wordnet synset that correspond with a selected term from a imported thesaurus.

5 Conclusions and further work

As a result of this work, we have developed a software tool to manage easily of thesauri from different organisations and to create links among them. During the importation process, we have achieved the automatic assignation of concepts and the expansion of keywords that enables the search of geospatial information using several thesauri. The present work has proved the viability and usefulness of the strategy suggested. The automatic assignation of concepts has yielded satisfactory results, with a success rate of 90-96% for the words that could be found in WordNet.

As regards further work, in this paper we have only discussed the links among English thesauri. One of the first remaining tasks is to incorporate the Spanish WordNet. The use of EuroWordNet, which connects WordNets in different languages, will allow the retrieval multilingual information.

Other important line of work will be consider specific thesaurus related with geospatial information, such as the Alexandria Digital Library Feature Type Thesaurus[1] proposed for Gazetteer, and works related with geospatial ontology, in order to obtain a more precise disambiguation of geospatial terms.

Following the motivations of the present project, we are working on an interface based on natural language. This interface is aimed at obtaining the sought concepts and exploits the keywords that catalogue the metadata and the concepts associated with the terms in the

[1] http://alexandria.sdc.ucsb.edu/~lhill/FeatureTypes/index.htm

thesauri to reply a query made in natural language. The tools used to implement the natural language interface use the contributions presented in this paper as for the management of thesauri and the access to WordNet.

Acknowledgements

This work is partially supported by projects MCyT, TIC2000-1568-C03-01 and DGES, PB98-1621-C02-01.

References

1. *The OpenGIS Abstract Specification. Topic13: Catalog Services (version 4)*. OpenGIS Project Document 99-113. OpenGIS Consortium, 1999.
2. J.A. Bañares, M.A. Bernabé, M.Gould, P.R. Muro-Medrano, F.J. Zarazaga. Aspectos tecnológicos de la creación de una Infraestructura Nacional Española de Información Geográfica. *Mapping*, 67:68--77, Ene-2001.
3. J. Donlon and K.D. Forbus. Using a Geographic Information System for qualitative spatial reasoning about trafficability. In *Proceedings of QR99*, Loch Awe, Scotland, 1999.
4. A.U. Frank. Spatial Ontology: A Geographical Point of View. In Spatial and Temporal Reasoning. (O. Stock, ed.), Dordrecht, The Netherlands, Kluwer Academic Publisher, pages 135-153, 1997.
5. G. A. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990. (Special Issue).
6. J. Gonzalo, F. Verdejo, C. Peters and N. Calzolari. Applying EuroWordNet to Cross-Language Text Retrieval. *Computers and the Humanities*, Special Issue on EuroWordNet, 1998. Accesible in http://rayuela.ieec.uned.es/~ircourse/
7. P. Resnik and D. Yarowsky. A perspective on word sense disambiguation methods and their evaluation. In Marc Light, editor, *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?*, pages 79-86, Washington, D.C., 1997.
8. E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, Copenhagen, Denmark, pages 16-22, 1996.
9. P. Resnik. Disambiguating noun groupings with respect to WordNet sense. In *Proceedings of the 3rd Workshop on Very Large Corpora, MIT*, 1995.
10. W. Gale, K.W. Churh, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415-439, 1992.