Ontology Mapping by Concept Similarity

Robert Villa, Ruth Wilson and Fabio Crestani
Department of Computer and Information Sciences, University of Strathclyde
Livingstone Tower, 26 Richmond Street, Glasgow, UK
Email: robert.villa@cis.strath.ac.uk, ruth.wilson@cis.strath.ac.uk, fabio.crestani@cis.strath.ac.uk

Key Words
Ontologies, information retrieval

Abstract
This paper presents an approach to the problem of mapping ontologies. The motivation for the research stems from the Diogene Project which is developing a web training environment for ICT professionals. The system includes high quality training material from registered content providers, and free web material will also be made available through the project's "Web Discovery" component. This involves using web search engines to locate relevant material, and mapping the ontology at the core of the Diogene system to other ontologies that exist on the Semantic Web. The project's approach to ontology mapping is presented, and an evaluation of this method is described.

1. Introduction

Ontologies have been defined as "explicit conceptualisation[s] of a domain" (Gruber 1993), in which objects, concepts and the relationships between them are defined as a set of representational terms, enabling knowledge to be shared and reused. Given the decentralised nature of Semantic Web development, it is likely that the number of ontologies will greatly increase over the next few years (Doan et. al. 2002), and that many will describe similar or overlapping domains, providing a rich source of material. The ability to automatically and effectively perform "mappings" between these ontologies will become an important requirement. This paper describes an approach to mapping ontologies in the same or similar domains, based on the similarity between their concepts.

We begin by outlining the Diogene Project which, through its web training environment, provides the motivation for this work. In particular, the project's "Web Discovery" component is described, the responsibility of which is to populate a core ontology of concepts with training material freely available on the web. Common problems encountered when mapping ontologies are outlined and discussed, before we put forward our own approach. Finally, we briefly describe some initial experimentation into the effectiveness of our method.

2. The Diogene Project

Diogene is an EC funded project under the 5th Framework Programme - Information Society Technologies (contract IST-2001-33358). Its main objective is to design, implement and evaluate

an innovative web training environment for ICT professionals. This environment will be able to support learners during the whole training cycle, from the definition of objectives to the assessment of results, through the construction of custom self-adaptive courses.

Innovative features of the project include dynamic learning strategies, Semantic Web openness, web services for learning object handling and property rights management, curriculum vitae generation and searching facilities, freelance teacher support, and assisted definition of learning objectives. The system will use several state-of-the-art technologies, including fuzzy learner modelling, intelligent course tailoring, and cooperative online training support.

The system will be accessible through the web exploiting a distributed architecture. Once logged onto a Diogene network, a learner can select a particular set of topics from the Diogene ontology and let the system arrange a personalised self-adaptive course about these topics (personalisation will be based on learner profiling).

In addition, the system will be able to:

- Enable freelance teachers to subscribe to Diogene and to describe (in a formal way) their professional abilities. In this way, teachers can be considered as "learning resources" which can be exploited by students requesting guidance during the learning process.
- Individuate learners with similar needs and/or profiles and provide them with a cooperative environment to support social interactions, mentoring and the exchange of information. The same environment will be used to interface freelance teachers with their students, synchronously and asynchronously.
- Apply a learner model based on an emerging standard format to represent learner-assessed achievements and obtain, for each learner, an electronic curriculum vitae, which will be published with respect to privacy requirements. A CV search engine will be developed which allows third parties to search for qualified professionals.
- Apply sound learning strategies with respect to the web, based on individual needs and learning styles, and provide the ability to automatically improve such strategies by exploiting information about knowledge assessments before and after training experiences.

An important feature of Diogene is the possibility to use, in the dynamic course composition process, high quality contents from registered content providers. Content will cover a set of ICT related topics, and will be indexed according to a knowledge representation methodology able to describe learning objects in a machine-understandable way. Indexed contents will remain on content provider servers but will be available to Diogene as web services.

At the core of the system's knowledge representation framework is an ontology covering the ICT domain. Ontologies provide a common language for sharing knowledge between members of a community of interest. Diogene's ontology will be based on the vocabulary and structure of the ACM Computing Classification, with the following relations linking its concepts:

- Has Part: *HP (x, y_1…y_n)* means that concept *x* is composed of the concepts *y_1 to y_n*; that is to say, to learn *x* it is necessary to learn *y_1* to *y_n*.
- Requires: *R (x, y)* means that, to learn *x*, it is first necessary to learn *y*.
- Suggested Order: *SO (x, y)* means that it is preferable to learn *x* and *y* in this order.

For linking the documents to concepts in the ontology, a further relation is defined:

- Explains: *Explains (x, z)* means that concept *x* is explained by the material in document *z*.

Diogene extends the 'explains' relation by allowing an optional weight to be added, representing the classifier's confidence in the "explains" link. Documents from registered content providers will be classified within this ontology. For these manually attached documents, the confidence weight will typically be '1', representing a correct link. Automatic techniques, which are unlikely to find perfect matches to concepts, may vary this value between zero and one, to reflect the individual algorithm's confidence in the link.

These relations were previously employed in the Learning Intelligent Advisor computer-based learning system (Capuano, et. al. 2003).

3. "Web Discovery" Component

In addition to the high quality content from registered providers, Diogene will also provide users with the opportunity to use free web content in their domain of interest. Such material will have limited pedagogical value but can be used as additional material during training sessions.

The Web Discovery component has the responsibility of discovering this material. It performs searches on the web and Semantic Web in order to discover and acquire didactical web pages, which it stores, catalogues and makes available to learners. Through a keyword-based text categorisation algorithm it is able, where absent, to automatically link textual documents to ontology concepts. Through a mixed approach based on keywords and ontologies, moreover, it is able to bypass compatibility problems between different ontological representations of the same domain. The Web Discovery component also offers a research service, matching documents to users' requests.

It has the following responsibilities in the Diogene system:

- To discover and index new content and metadata on the web.
- To discover and index new content and metadata on the Semantic Web.
- To maintain a repository of metadata describing discovered content.
- To provide a personalised metadata list to the user, matching the characteristics specified in a metadata query.
- Finally, to provide a link to the web documents themselves.

There are several main use scenarios in which the Web Discovery component plays a key role in Diogene. In a typical scenario, the learner requests training material from the system, specifying his or her own learning objectives. A list of concepts and learner preferences is passed to the Web Discovery component, which it matches against its repository of metadata for web documents. The Web Discovery component then provides a list of links to suitable resources to the learner.

Free web content is discovered in two ways:

- Searching the conventional web using automatically generated queries representing individual concepts in the ontology.
- Mapping external ontologies on the Semantic Web to Diogene's ontology, and incorporating their associated material into Diogene's training environment.

3.1. Content Discovery on the Conventional Web

Content discovery on the conventional web involves three steps:

- Constructing a query for each concept in Diogene's ontology
- Executing that query using a web search engine
- Downloading and filtering the results against the concept definition

These steps are shown in Figure 1.

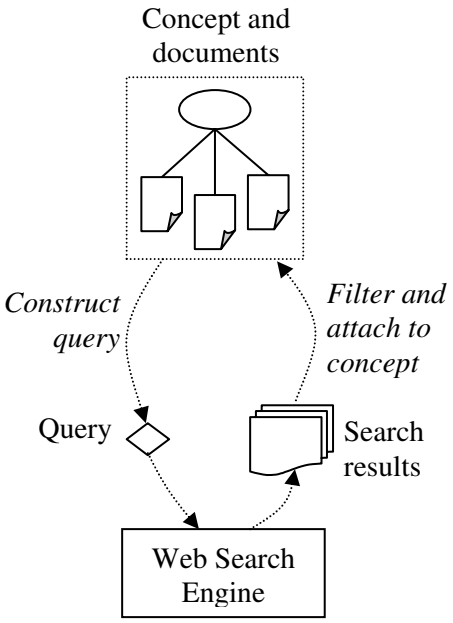

Figure 1. Searching for a concept on the web

Queries are constructed using textual information about concepts, such as their names, descriptions and the documents already attached to them. A search can then be carried out using a web search engine such as Google, producing a list of hypertext links to potentially relevant material.

Next, these results are filtered in order to remove broken links and out-of-date material, and to measure more accurately the similarity between the retrieved content and Diogene's concept. This stage involves categorising the new material in Diogene's ontology according to the results of a "training" phase, in which documents assigned to each category are analysed to extract the elements that characterise them; thereafter, when new documents are input to the system, they can be assigned to categories based on their similarity to these features. In short, when an input document is fed to the system, it is tokenised and stemmed, then:

1. The weight of each word in the document is calculated
2. The document is projected into the space of the category *c*
3. A similarity value is computed between the document and the category (concept)
4. The document is assigned to the concept with the highest similarity value

Finally, markup is extracted from the web resource, and a link is recorded in the Web Discovery component's local store.

3.2. Content Discovery on the Semantic Web

Discovering content on the Semantic Web involves mapping Diogene's ontology to other ontologies in the same domain, in order that relevant material from external ontologies can automatically be included in the Diogene system. In other words, rather than searching the whole web for free resources, as described above, this method will focus on ontologies that are known to exist and map them, concept-to-concept, to the project's ontology. The remainder of the paper is concerned with the problem of how best to map one ontology to another, within the context of the project.

4. The Mapping Problem

Noy and Musen have explored the difficulties involved in this process (Noy and Musen 2001). Ontologies, even those in the same domain, may be quite different because, during the conceptualisation phase, the semantics of a domain will be encoded in different ways. Different choices will be made about which classes, instances and relations to use to represent the domain, according to the individual requirements of the system. Therefore, achieving an effective mapping can be problematic, as correspondences in the meanings of the concepts in each ontology have to be discovered and understood.

Potential problems include:

- Different ontologies may use different names to represent the same concept, or the same name to represent different concepts (Ceri and Widom 1993).
- Different ontologies may use different values to represent the same concepts. For example, one ontology may use pounds sterling to refer to the price of a product, while another uses US dollars.
- The same concept may be represented at the "class" level in one ontology and at the "instance" level in another ontology.
- Different ontologies may use different structural representations of concepts, e.g. "de la Mare" and "Mare, de la" may be used in two ontologies to refer to the same person.

Indeed, (Visser, et. al. 1997) identifies eight possible ways in which elements in an ontology can be mismatched, including class mismatches, relation mismatches, concept name mismatches, and concept description mismatches.

The problems involved in mapping ontologies have been tackled in three ways:

- In the manual approach, similar concepts from the ontologies are identified by a human expert and mapped by hand. This is labour-intensive, error-prone and impossible on the scale of the web.
- Semi-automatic approaches try to assist in the mapping process by automating certain stages.
- Finally, fully automatic approaches attempt to complete the whole process of ontology mapping with no human involvement at all.

Semi-automatic systems, in particular, are prevalent. Chimaera (McGuiness, et. al. 2000), for example, addresses a portion of ontology mapping tasks by providing an environment for uploading and viewing the ontologies to be mapped, and suggesting potential matches by analysing the similarity of names and definitions, expanding acronyms, and so on. The user can then choose whether or not to map the concepts. Anchor-PROMPT (Noy and Musen 2001), on the other hand, takes as input pairs of related terms ("anchors") defined by the user, and then performs the remainder of the mapping process automatically: from the set of previously defined anchors, new pairs of semantically close terms are produced by analysing the paths between the anchors in the corresponding ontologies.

Similar problems have also arisen and been addressed in the context of thesauri and schema integration. Sintichackis amd Constantopoulos (1997) outline a method for monolingual thesauri merging, based on detecting equivalent terms in the thesauri being merged. This method focuses on the properties and semantics of thesaural relationships. In a similar vein, Lee and Dubin (1999) use a spreading activation network for mapping two controlled vocabularies, both of which index documents from the NASA Astrophysics Data System. The differences that hinder unambiguous mapping of thesauri are discussed by Doerr (2001), and a notion of optimal mapping is developed, based on limiting the fuzziness introduced when one vocabulary

transforms into another. Yet another approach by Bergamaschi et. al. (2001) creates a common thesaurus between structured and semi-structured data sources using ODL (an object-oriented language based on Description Logics) descriptions. The sources are then mapped semi-automatically, providing an integrated view.

Also relevant to the problem of mapping ontologies, Tower et. al. (2001) look at mapping topical hierarchies such as Yahoo! and UseNet news, in particular tackling the problem of "docking" narrow, focused and refined hierarchies into broader ones. Two algorithms are used: the first matches textual features (short descriptions or longer exemplary passages) of concepts, and the second is based on a tree matching algorithm which extends the first to also uses hierarchical structure to match concepts.

5. Mapping Ontologies by Concept Similarity

Diogene has several specific requirements from the mapping process:

- Where possible, techniques for automatic mapping should be adopted. However, as precision is key to the success of the mapping process, human judgement will be introduced where appropriate.
- Diogene's own ontology should not be altered during the mapping process. The motivation to map stems from the desire to make available resources from other systems, and not to transform the ontology according to new knowledge from these systems. Therefore, we assume that the Diogene ontology contains a complete description of the ICT domain, and that any new knowledge in the domain will be added manually.
- Finally, in order that the system is open to incorporating material from as many sources as possible, no assumptions will be made about the external ontologies which will be mapped to Diogene, in terms of the languages in which they are represented, their domain coverage or their structures.

Our methods will use keyword-based IR techniques for similarity matching, in which simple statistics have proved themselves over more complex natural language processing techniques used in AI.

They will comprise two main stages:

- First of all, measuring the similarity between any two concepts in the two ontologies; and
- Secondly, based on the similarity measure, deciding which concepts to map.

5.1. Measuring the Similarity of Concepts

The first stage is a question of calculating how similar any two concepts are. For this, a similarity function can be defined, which calculates a value between 0 and 1:

$$similarity\_function(d, c) \in [0.1]$$

where d is a concept from Diogene's ontology, and c is a concept from some external ontology.

We propose to implement this function by comparing concept descriptions. Descriptions will typically take the form of unstructured text, to which standard IR indexing techniques can be applied in order to build a statistical representation of the concept. This process involves the removal of stop words, stemming, and the calculation of weights for the individual terms (using term frequency – inverse document frequency). The similarity between concepts was computed

using the cosine similarity measure between the concepts' description vectors (Salton and McGill 1983). An n by m similarity matrix was then generated, where n = size of Diogene's ontology, m = size of an external ontology.

Concept descriptions are sometimes manually created as part of the ontology development process; these will be in the style of dictionary definitions or thesaurus entries, and are designed for human consumption. However, where an ontology exists without concept descriptions, these can be automatically generated from an analysis of the documents in the ontology. In this case, the textual content of all the documents attached to a single concept is extracted and indexed using the IR techniques described above. The terms with the highest weight are taken to characterise the concept.

5.2. Performing the Mapping

Once the similarity of the concepts in two ontologies has been calculated, the next stage is to decide how to use this information to perform a mapping. Our approach has to take into account two main problems:

5.2.1. Differences in Scope

Diogene's ontology and the external ontology to which it is being mapped may not cover exactly the same domain, and the external ontology may contain concepts which are outside Diogene's scope. In line with our second requirement, that Diogene's ontology should not be altered during the mapping process, such concepts should be excluded in a mapping.

The quantity of relevant concepts in any external ontology will vary. Therefore, methods based on relative measures, such as always mapping a certain percentage of the concepts in an external ontology, are unsuitable, as the cost of including irrelevant documents in the Diogene system is high. Rather, an absolute threshold will be set which specifies that all concepts with similarity values above a certain level should be mapped or put forward for human judgement, while those with low similarity values should not be mapped. This threshold must be set carefully: too high and too few external concepts will be mapped; too low, and too many erroneous mappings are likely to occur.

5.2.2. Differences in Levels of Granularity

Moreover, the two ontologies being mapped may be defined at different levels of granularity, and therefore our techniques require a degree of flexibility capable of accommodating a variety of scenarios, for example:

- One-to-one mappings, in which a single external concept is mapped to a single Diogene concept. This simplifies the mapping algorithm but may be too restrictive.
- Many-to-one mappings, in which many external concepts are mapped to a single Diogene concept. This is appropriate where the external ontology is specified at a greater level of granularity, e.g. Diogene contains the concept "functional programming" while the external ontology contains concepts for specific types of functional programming, such as "programming LISP" and "programming ML".
- One-to-many mappings, in which a single external concept is mapped to more than one Diogene concept. This is appropriate where Diogene's ontology has greater granularity than the external ontology.

5.2.3. The Mapping Algorithm

Our algorithm generates mappings from the similarity matrix. The point in the matrix with the highest similarity value corresponds to the most similar two concepts from the input ontologies, and the algorithm finds this point (outputting the corresponding mapping) before returning to find the next most similar two concepts, and so on, in a loop, until a threshold value is reached. The threshold specifies the minimum similarity value which must exist between two concepts before they can be mapped. For this to work, the mapping generated in each iteration of the loop must be removed from the matrix, the easiest way of doing this being to set the appropriate similarity score to zero. Different kinds of mapping can be generated by changing which elements are zeroed, at this step. A one-to-many or many-to-one mapping can be created by zeroing either a whole column or whole row of the matrix, which has the effect of removing a single concept from the mapping process. A one-to-one mapping can be generated by removing both mapped concepts from the matrix (zeroing both the column and row of the relevant concepts), and a one-to-many mapping by only zeroing the single cell corresponding to the previously generated mapping.

6. Evaluating the Effectiveness of Automatically Generated Concept Descriptions

In order to evaluate the effectiveness of generating descriptions from the textual content of documents and using them to calculate the similarity between concepts, a small experiment was conducted.

The experiment had two main hypotheses:

1. Automatically generated concept descriptions will be effective in calculating the similarity between concepts.
2. The greater the quantity of material from which concept descriptions are derived, the better our algorithms will perform.

The research was conducted in parallel with the creation of Diogene's ontology, and so the ACM Computing Classification System (CCS), on which the project's ontology is based, was selected as the base ontology in the experiment. INSPEC, from the Institute of Electrical Engineers, was selected as an appropriate ontology to map to the ACM, as both systems index articles from the *Journal of the American Society for Information Science (JASIS)*, now called the *Journal of the American Society for Information Science and Technology (JASIST)*, providing a common document set containing article titles and abstracts. From a list of all 1996-2000 *JASIS* articles indexed in both ACM and INSPEC, two distinct test collections were generated by randomly removing half those articles from ACM and the other half from INSPEC. From these test collections, four "ontologies" were generated:

- Two large ontologies (ACMlrg and INSPEClrg), containing article titles and abstracts from years 1996-2000 of *JASIS*.
- Two small ontologies (ACMsml and INSPECsml), comprising titles and abstracts of *JASIS* articles from years 1998-2000.

The number of concepts in ACMsml and ACMlrg and in INSPECsml and INSPEClrg was kept constant in order to isolate the effect of the number of documents contained in the ontologies. INSPECsml was mapped to ACMsml and INSPEClrg to ACMlrg by generating descriptions of their concepts from the titles and abstracts of articles attached to them, and calculating similarity values based on these descriptions in the pairs of ontologies. A many-to-one mapping from INSPEC to ACM ontologies was performed, based on the assumption that, as the INSPEC ontologies have identical coverage to ACM ontologies but contain a greater number of

concepts, it will sometimes be the case that several INSPEC concepts should be mapped to a single ACM concept. Different thresholds were set, in order that the behaviour of the mapping system could be analysed at all levels.

Finally, to calculate the "correctness" of the two sets of mappings, the automatic classification of articles from INSPEC test ontologies in the mapped (output) ontologies was compared to their original classification in the ACM. Where documents were assigned to the same concept in both the automatic and the original classifications, this was considered a correct automatic mapping; where differences occurred, this was considered incorrect.

Recall and precision scores were necessarily limited by the fact that human indexers will classify even identical collections of documents differently. Because the onotlogies are mapped at the concept (class) level, documents will remain grouped according to their original classification in INSPEC. It is therefore inevitable that the newly mapped ontology will not contain the same arrangement of documents as the original ACM classification. To judge the extent of these limitations, and to better understand the success of our mapping procedures, the best possible mappings based on the intersection of articles between INSPEC and ACM concepts were studied. Maximum precision and recall could then be calculated, using the measures described above. Comparing the precision and recall scores for our automatic mapping with these optimum scores enabled the success of our mapping procedures to be gauged more fully.

The results of this experiment have been collated in the following tables, which provide precision and recall scores for many-to-one mappings between the small and large ontologies. Maximum possible scores are also provided, and the percentage actually achieved has been calculated.

Ontologies	Precision	Max precision	% max achieved
INSPECsml to ACMsml	0.288	0.341	84.46
INSPEClrg to ACMlrg	0.347	0.374	92.78

Table 1. Percentage of maximum possible precision scores achieved

Ontologies	Recall	Max recall	% max achieved
INSPECsml to ACMsml	0.396	0.717	55.23
INSPEClrg to ACMlrg	0.412	0.615	66.99

Table 2. Percentage of maximum possible recall scores achieved.

Precision and recall were high, showing that automatically generated concept descriptions are effective in measuring the similarity between concepts in two ontologies. Moreover, precision and recall increased with the size of the ontologies being mapped, suggesting that, the larger the number of textual objects attached to each concept in an ontology, the more effective the automatic mapping procedure will be. Therefore, the two experimental hypotheses were confirmed:

1. Automatically generated concept descriptions are effective in calculating the similarity between concepts.
2. The greater the quantity of material from which concept descriptions are derived, the better our algorithms performed.

Precision was greater than recall and, indeed, it is more important for our system than recall. From the point of view of the Diogene Project, in which the quality of the mappings is key, it is better to perform a small number of correct mappings automatically and allow mappings about which the system is less certain to be considered using human intervention, than to perform a large number of mappings (producing high recall) and sacrifice precision.

7. Conclusions

This paper has investigated the issue of mapping ontologies from the point of view of the Diogene project. An approach involving measuring the similarity of concepts was outlined, and an experiment based on automatically generated concept descriptions was briefly described. It was discovered that such descriptions are effective in calculating similarity, especially where they are derived from large quantities of material.

The findings of this experiment are currently being investigated in more detail. First of all, we will look further into the effect different thresholds on the mappings. We hope to discover an optimum threshold for similarity values, above which all concepts can be mapped automatically; to maintain quality, it may be better to submit concepts with lower similarities for human consideration. We will also examine this notion of importing documents wholesale, according to the concept to which they're attached, as opposed to attaching them individually to Diogene's ontology, comparing the efficiency of these two approaches.

Further, the notion of a "correct" mapping would benefit from more attention. Thus far, this has been judged using automatic techniques, but where experimental conditions are less controlled this becomes difficult to measure. An end-user evaluation of the output of our system is another means by which we will obtain feedback on its success.

Taken together, the results of all experiments will inform the future development of Diogene's ontology mapping algorithms, to produce a system that is operating at its full potential and in line with the project's unique requirements.

Acknowledgements

This research was supported by the Information Society Technologies project Diogene (IST-2001-33358). Further details can be found on the project website: http://www.diogene.org/

References

Bergamaschi, S., Castano, S., Vincini, M. and Beneventano, D. 2001. **Semantic integration of heterogeneous information sources**. *Data and Knowledge Engineering (Special Issue on Intelligent Information Integration)*, **36**.

Capuano, N., Gaeta, M., Micarelli, A. and Sangineto, E. 2003. **An intelligent web teacher system for learning personalisation and Semantic Web compatibility**. *Proceedings of the Eleventh International PEG Conference*, St Petersburg, Russia, 28 June - 1 July.

Ceri, S. and Widon, J. 1993. **Managing semantic heterogeneity with production rules and persistent queues**. *Proceedings of the 19th VLDB Conference*, Dublin, Ireland, 24-27 August.

Diogene. http://www.diogene.org/ (Accessed on 9 December 2003).

Doan, A., Madhavan, J., Domingos, P. and Halevy, A. 2002. **Learning to map between ontologies on the Semantic Web**. *Proceedings of WWW2002*, Hawaii, USA, 7-11 May.

Doerr, M. **Semantic problems of thesaurus mapping**. 2001. *Journal of Digital Information*, **1**(8).

Gruber, T. A. 1993. **A translation approach to portable ontology specifications**. *Knowledge Acquisition*, 5(2).

Lee, J. and Dubin, D. 1999. **Context-sensitive vocabulary mapping with a spreading activation network**. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99),* California, USA, 15-19 August.

McGuiness, D., Fikes, R., Rice, J. and Wilder, S. 2000. **An environment for merging and testing large ontologies**. *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2000),* Colorado USA, 12-15 April.

Noy, N. and Musen, M. 2001. **Anchor-PROMPT: using non-local context for semantic matching**. *Workshop on Ontologies for Information Sharing, Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001),* Washington, USA, 4-10 August.

Salton, G. ad McGill, M. J. 1983. ***Introduction to Modern Information Retrieval***. New York: McGraw-Hill.

Sintichackis, M. and Constantopoulos, P. 1997. **A method for monolingual thesauri merging**. *Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR '97),* Philadelphia, USA, 27-31 July.

Tower, B., Chaisoon, M. and Belew. R. 2001. **Docking topical hierarchies: a comparison of two algorithms for reconciling keyword structures**. *Report no. CS2001-0669*, University of California, San Diego.

Visser, P., Jones, D., Bench-Capon, T. and Shave, M. 1997. **An analysis of ontology mismatches: heterogeneity versus interoperability**. *AAAI Spring Symposium on Ontological Engineering*, Stanford University, California, USA, March 24-26.