An Experiment with Ontology Mapping using Concept Similarity

Robert Villa, Ruth Wilson & Fabio Crestani
Dept. Computer & Information Sciences
University of Strathclyde
Glasgow, UK
{robert.villa, ruth.wilson, fabio.crestani}@cis.strath.ac.uk

Abstract

This paper describes a system for automatically mapping between concepts in different ontologies. The motivation for the research stems from the Diogene project, in which the project's own ontology covering the ICT domain is mapped to external ontologies, in order that their associated content can automatically be included in the Diogene system. An approach involving measuring the similarity of concepts is introduced, in which standard Information Retrieval indexing techniques are applied to concept descriptions. A matrix representing the similarity of concepts in two ontologies is generated, and a mapping is performed based on two parameters: the domain coverage of the ontologies, and their levels of granularity. Finally, some initial experimentation is presented which suggests that our approach meets the project's unique set of requirements.

1. Introduction

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.

At the core of the system's knowledge representation framework is an ontology covering the ICT domain, within which content from registered providers will be classified. In addition, Diogene will provide users with the opportunity to use free Web content in their domain of interest; this material may have a limited pedagogical value but can be used as additional material during training sessions. One method for drawing useful freeware resources from the Web and making them available to the system is to "map" external ontologies in the same domain to Diogene's ontology. 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.

This paper is concerned with the problem of how best to map one ontology to another, within the context of the project. The following section describes Diogene's ontology, and section 3 describes our motivations for this research and our specific requirements. Section 4 briefly summarises the ways in which mappings have been tackled in other research, while section 5 details our approach, which involves measuring the similarity of concept descriptions. Then, in sections 6 and 7, some initial experimentation into the effectiveness of this approach is presented. Finally, conclusions are drawn and some ideas for future research are outlined.

2. Diogene's Ontology

Diogene's ontology is based on the ACM Computing Classification (CCS), used for classifying books, journal articles and conference proceedings in the field of computing in a four-level subject hierarchy. However, it differs from the CCS in several ways. First, some concepts such as "general" and "miscellaneous" have been removed. Second, new concepts have been be added where the CCS's level of detail or coverage is insufficient to represent content providers' specialist areas of interest. Finally, and most importantly for our mapping technique, Diogene's ontology is not hierarchical; rather, it can be seen as a direct acyclic graph, with the following relations linking its concepts:

- Has Part: *HP* $(x, y_1 \ldots y_n)$ 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.

These relations were previously employed in the Learning Intelligent Advisor computer-based learning system (Capuano et al, 2002). Figure 1 provides an example of how they might be put to use in the project.

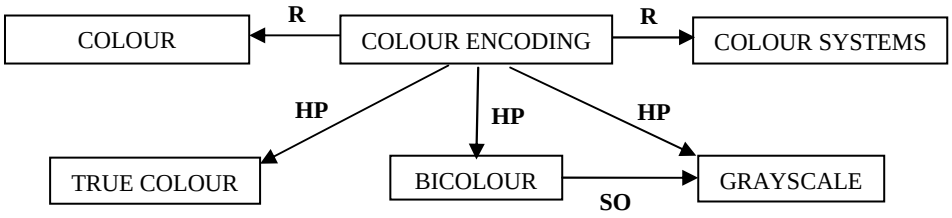

Figure 1: Part of Diogene's ontology

The relations form the basis of a dynamic course-generation process, suggesting a sequence of topics based on learners' objectives. In the example provided in Figure 1, they specify that a user wishing to study colour encoding will first be *required* to know about colour and colour systems, then he will learn about true colour, bicolour and grayscale (which are *part of* colour encoding), and it will be *suggested* that he study grayscale before bicolour.

Training material from registered providers and from the Web will be categorised within the ontology, and delivered to learners to explain each concept in the learning path.

3. Motivations

The particular motivation for Diogene's ontology mapping algorithms is to incorporate ICT learning material from other ontologies on the Web, while preserving the quality of the system. Specifically, this means that:

- For our purposes, the mapping process will involve identifying similar concepts in external ontologies and adding their associated content to the appropriate concept(s) in Diogene. The aim is not to transform the ontology according to new knowledge from these systems by incorporating new concepts or structures, but just to incorporate their learning material. We assume that the Diogene ontology contains a complete description of the ICT domain, and that any new concepts in the domain will be added manually.
- Ontologies have been defined as "explicit conceptualisation[s] of a domain", in which objects, concepts and relationships between them are defined as a set of representational terms, enabling knowledge to be shared and reused (Gruber, 1993). McGuinness discusses the spectrum of specifications which people have termed ontologies, including controlled vocabularies, glossaries, thesauri, web hierarchies such as Yahoo!, subclass hierarchies, formal instance relationships, frames, value restrictions, and logical constraints (McGuinness, 2002). As noted by Berners-Lee et. al. (2001), the availability of heterogeneous data sources such as these is the essential property of

the World Wide Web and underlying our mapping algorithm is a desire to take full advantage of all such quality sources. This has two main consequences. Firstly, where possible, automatic techniques will be adopted, so that mapping from multiple ontologies is not an expensive or time-consuming process. Secondly, in order that the system is versatile enough to incorporate material from as broad a selection of these sources as possible, our definition of the "ontologies" to which we might map is deliberately loose. No assumptions are made about the ontologies which will be mapped to Diogene, in terms of:

  - Their structures. In the above spectrum of ontologies, some have simple hierarchical structures while others use various sophisticated relations between concepts. Our system should be capable of mapping to any of them.
  - Their domain coverage. The purpose of the system will be to map other ontologies in the ICT domain to Diogene. These may be broader or narrower in scope; either way, they may include relevant material, and the mapping algorithms should handle them appropriately.
  - Their levels of abstraction/detail. Other ontologies in the ICT domain may be defined to higher or lower levels of granularity than Diogene's ontology, and the system should be able to correctly perform mappings in either situation.
  - The languages in which they are represented. These may be languages especially formulated for expressing ontologies such as DAML+OIL or OWL, or structures represented in RDF, XML or HTML.

- Finally, although the quality of material from external ontologies cannot be guaranteed, the quality of the mappings themselves is a key consideration. Precision is given greater importance than recall, ensuring that any automatic mappings to our ontology are "correct" or highly likely to be correct, even if there are very few, rather than allowing many incorrect mappings to occur. Human judgement may be introduced where appropriate.

Furthermore, these requirements must be met within the project's practical constraints of time and cost.

4. Other Approaches to the Problem

Noy and Musen have explored the difficulties involved in mapping ontologies:

> A domain expert who wants to determine a correlation between two ontologies must find all the concepts in two source ontologies that are similar to one another, determine what the similarities are, and either change the source ontologies to remove overlaps or record a mapping between the sources for future reference (Noy & 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 a correct mapping can be problematic, as correspondences in the meanings of the concepts in each ontology have to be discovered and understood.

The problems involved in mapping ontologies have been tacked 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. Chimaera (McGuinness et al, 2000) and Anchor-PROMPT (Doan et al, 2002) are two examples, and (Bergamaschi et al, 2001) describes an approach which uses ODL (an object-oriented language, with an underlying Description Logics) descriptions of the source ontologies to set a shared vocabulary, in the form of a common thesaurus.
- Finally, fully automatic approaches attempt to complete the whole process of ontology mapping with no human involvement. These tend to operate on simpler structures such as hierarchies (Tower, Chaisson & Belew, 2001).

5. Ontology Mapping by Concept Similarity

In response to one of our primary requirements, to be open to mapping to any of a spectrum of ontologies and not just those conforming to the traditional definition, our algorithms focus on measuring the similarity of the concepts. Unlike some other recent approaches to the mapping problem, we do not use the relations between concepts to inform the mapping process; rather, our method is to "flatten" the concerned ontologies and apply traditional keyword-based IR techniques for similarity matching. Two main stages are involved:

- 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.

Although the techniques themselves are well known, we believe that, in applying them to the new problem of mapping ontologies, we provide a solution that can be applied to any set of concepts, no matter how structured, thereby enabling maximum versatility.

5.1. Concept Descriptions and Similarity

The first stage is a question of calculating how similar any two concepts are. This function is implemented by comparing concept descriptions, which typically take the form of unstructured text. Standard Information Retrieval (IR) indexing techniques are applied to these descriptions (Salton & McGill, 1983) in order to build a statistical representation of each concept. This process involves the removal of stop words, stemming, and the calculation of weights of the individual terms (using term frequency – inverse document frequency). For example, given this concept name and description:

> Distributed database: A database that is physically decentralized and handled by a database management system in a way that provides a logically centralized view of the database to the user.

A vector such as the following will be generated:

central	1.91	Decentr	2.46	manag	1.40	system	0.81	logic	1.58
database	2.48	Handl	1.68	physic	1.63	user	1.03	view	1.54

Figure 2. Giving weights to description terms

Concept descriptions are sometimes manually created as part of the ontology development process; these are often 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 objects in the ontology. In this case, the textual content of all the material attached to a single concept is extracted and indexed using the IR techniques just described, and the terms with the highest weights are taken to characterise the concept. Figure 3 shows the weights given to terms in five documents attached to the same concept in an ontology. Of these, the highest ranking overall are taken as a description of the concept, in this case "retrieval", "text", "statistical", "language", "metadata", and so on. Underlying this technique is the notion that a concept is defined by the objects attached to it.

Information Storage and Retrieval									
Doc 1		**Doc 2**		**Doc 3**		**Doc 4**		**Doc 5**	
Metadata	2.17	Abstract	1.51	Stemming	1.59	Probabilistic	2.01	Fuzzy	1.94
Library	1.20	Chinese	1.28	Stopword	1.55	Retrieval	2.61	Crisp	1.28
Bibliographic	1.42	Newspaper	0.75	French	1.24	Aboutness	1.17	Set	1.88
Network	1.47	Statistical	2.42	Search	2.09	Vector	1.35	Hierarchy	1.66
Control	1.76	Text	2.48	Language	2.30	Mathematical	1.09	Approximation	0.89

Figure 3. Generating descriptions from documents attached to concepts

The terms selected to describe a concept in our ontology, together with their weights, are compared to the descriptions of concepts in other ontologies. Their similarity is computed using the cosine similarity measure, which is commonly used in information retrieval and document clustering (van Rijsbergen, 1979 ; Sneath and Sokal, 1973).

The output of this process is an n by m similarity matrix, where n = size of Diogene's ontology, m = size of an external ontology, and each element of the matrix contains the calculated similarity value between each pair of concepts in the two ontologies.

5.2 Concept 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 takes into account two main problems.

First of all, as outlined in section 3 above, 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, an absolute threshold is 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.

Secondly, 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.

Our algorithm uses the similarity matrix to generate the mapping. The point in the matrix with the highest similarity value corresponds to the most similar two concepts from the input ontologies. 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. Different kinds of mapping can be generated by excluding concepts once they have been mapped. For example, a one-to-one mapping is created by removing two concepts from the matrix as soon as they have been mapped; one-to-many or many-to-one mappings can be created by removing a single concept once it is mapped.

6. Experimentation

Some initial experiments have been carried out into the success of this approach. In particular, one experiment has investigated the effectiveness of using concept descriptions to measure similarity. It is summarised below.

The two entities we selected for mapping in this experiment were the ACM CCS and INSPEC, from the Institute of Electrical Engineers. Although they are hierarchical classification schemes rather than

ontologies in the traditional sense, since our techniques involve flattening any structure they could be treated in the same way as Diogene's ontology during the mapping process, providing useful experimental results. Indeed, Diogene's ontology was based on the ACM CCS and has a similar coverage and level of granularity. Its domain overlaps with that of INSPEC, so it was possible to find a common document set.

In the CCS, concepts are arranged in a four-level hierarchy (the fourth level comprising subject descriptors), with 11 first-level nodes. It is used to classify all the bibliographic items contained in the ACM Portal, including book, journal and conference listings, and each item is assigned a primary classification and an additional classification.

INSPEC is a database of 5.75 million records, its coverage is broader, comprising four categories (Physics, Electrical and Electronic Engineering, Computers and Control Technology, and Information Technology), each containing two further levels. Articles are assigned to varying numbers of concepts, usually three or four.

Both ACM and INSPEC 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)*. This journal, published monthly, covers issues such as information storage, management and retrieval, social and legal aspects of information, text analysis and communication systems. From a list of all 1996-2000 *JASIS* documents indexed in both ACM and INSPEC, two distinct test collections were generated by randomly removing half those documents from ACM and the other half from INSPEC.

From these test collections, four groups (hereafter referred to as "ontologies" for the purposes of this paper) were generated:

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

The differences in these ontologies are detailed in Table 1.

Ontology	**No. concepts**	**No. documents**	**Average no. documents per concept**	**Average no. words per description**
ACMsml	18	237	5.39	872
INSPECsml	54	106	5.72	935
ACMlrg	18	140	7.78	1,286
INSPEClrg	54	148	7.89	1,278

Table 1: Properties of the four test ontologies

The experiment focused on how the number of documents attached to an ontology affects the success of our approach. Intuition denotes that a more comprehensive understanding of a concept can be gained from ten documents related to that concept than just one or two; therefore, the greater the number of documents attached to a concept, the better the automatically generated descriptions, and therefore the mappings, will be. Stated simply, the experimental hypotheses are:

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

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; only the number of documents varied. INSPECsml was mapped to ACMsml and INSPEClrg to ACMlrg, according to the following procedure:

- The titles and abstracts of the documents in the test ontologies were used to generate descriptions of the concepts to which they were attached, using the techniques described in 4.1 above.
- Based on these descriptions, similarity values between concepts in the pairs of ontologies were calculated.
- 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 the 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. Had the ACM ontology contained more concepts than INSPEC, a one-to-many mapping would have been more appropriate; example scenarios are provided in 4.2 above.

By varying the threshold, the quantity of mappings between the ontologies could be varied. By setting a high threshold, we might expect there to be fewer, higher quality mappings (due to the high similarity value that they all must attain). At the other extreme, a very low threshold may be set, which will output all mappings with even the smallest similarity (generating more, but probably lower quality, mappings).

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

Specifically, two measures were used in considering the results:

$$\text{Precision} = \frac{A \cap B}{A} \text{ and } \text{Recall} = \frac{A \cap B}{B}$$

Recall and precision scores are used extensively in Information Retrieval (van Rijsbergen, 1979), and have been adapted here to suit our specific purposes. They were necessarily limited by the fact that human indexers will classify even identical collections of documents differently. Because the ontologies are mapped at the concept (class) level, documents will remain grouped according to their original classification in INSPEC. For example, when mapping INSPECsml to ACMsml, the best possible match for INSPEC's concept "Publishing and reproduction" is "Document and text processing" in ACM; this is based on the overlap in the documents they classify. However, this is not a perfect match (see Figure 4).

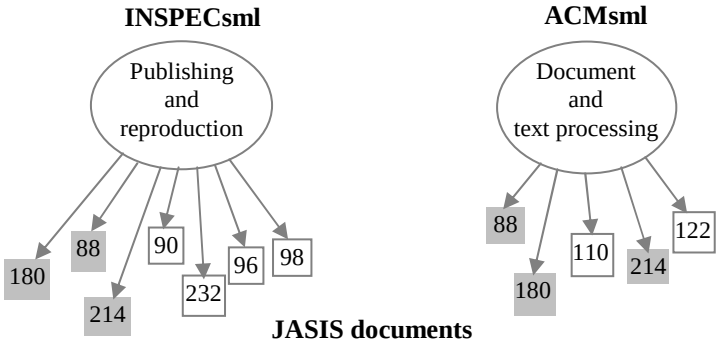

Figure 4: Overlap of documents attached to INSPEC and ACM ontologies

Both concepts contain the documents 88, 180 and 214. However, they both also contain documents not attached to their corresponding concept. Therefore, recall and precision scores are limited, in this case to 3/5 (0.6) and 3/7 (0.429) respectively. Such restrictions will always exist when measuring the success of a mapping by the placement of learning material, for two fundamental reasons:

- Ontologies, even those in the same domain, use different concepts to represent the domain; and
- Human indexers assign documents to concepts according to their own perception and understanding of the domain, and this will vary between individuals (Mai, 2001 ; Saarti, 2002).

7. Results

To test the first hypothesis, and to investigate the operation of the algorithm, many-to-one mappings from INSPEC to ACM were carried out (as described above) on both the small and large ontologies, using a range of different threshold values. Some examples of mapped concepts are shown in Table 2.

ACM Concepts	INSPEC Concepts	Similarity
Information Storage and Retrieval	Information Analysis and Indexing	0.30
Information Search and Retrieval	Information Retrieval Techniques	0.28
Models and Principles	User Interfaces	0.22
Systems and Software	Information Networks	0.20
Content Analysis and Indexing	Information Analysis and Indexing	0.19

Table 2. Some mappings generated in the experiment

For each mapping, precision and recall were calculated; Figure 5 shows the resulting graphs. A threshold of zero indicates that all concepts with any similarity were mapped. Precision and recall both fall to zero towards the right of each graph (at a threshold of between 0.3 and 0.4) indicating that no mappings exist with a larger similarity value (the algorithm produces no output). In both graphs recall decreases and precision increases moderately as the threshold is increased, as was predicted.

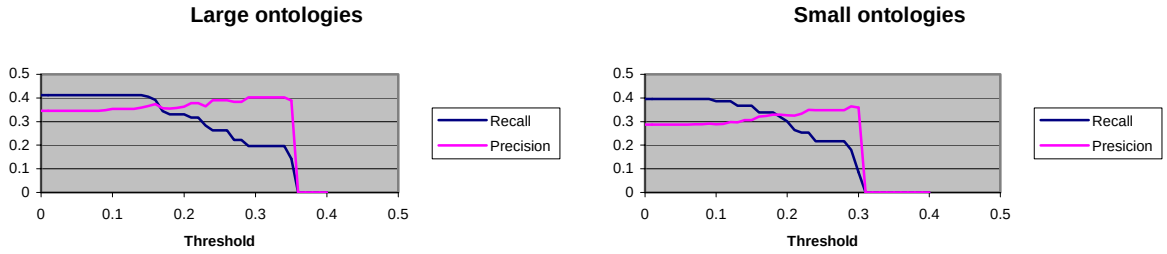

Figure 5. Precision and recall graphs for the large and small ontologies

The recall and precision scores presented in these graphs are absolute values. As outline in section 6, the effectiveness of the algorithm should be judged relative to the best that can be achieved between the two test ontologies, due to the differences in the classifications. To take this into account, and to investigate the second hypothesis (how concept description length alters the operation of the algorithm), scaled recall and precision values were calculated for a threshold of zero, generating all possible mappings. These results are shown in Figure 6. This fixed threshold enabled a “best” mapping to be generated based on the intersection of documents between INSPEC and ACM concepts. Maximum precision and recall could then be calculated, using the measures described above.

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

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

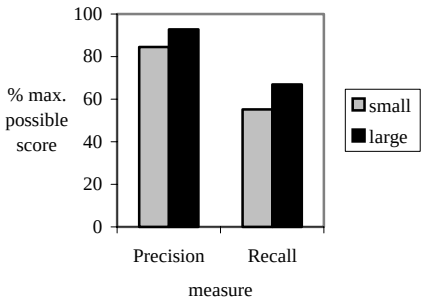

Figure 6. Percentage of the maximum possible precision and recall scores achieved in many-to-one mappings

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 documents 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 (in this case, the number of document abstracts) from which concept descriptions are derived, the better our algorithms performed.

In the scaled results, precision was greater than recall, suggesting that the mappings the algorithm generates are of high precision, but lower recall. This is indicated by the graphs in Figure 5, which show only moderate increases in precision for higher thresholds, in contrast to the relatively rapid decrease in 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, than to perform a large number of mappings (producing high recall) and sacrifice precision.

It is worth noting that this experiment was conducted using academic articles, which are different in nature from the learning objects which will populate Diogene's ontology. However, the project's learning objects will be heavily textual, like the *JASIS* articles, and so our techniques can be applied to them. Moreover, the experiment used only the titles and abstracts of documents in the mapping process, providing a quantity and style of text more comparable to that of learning objects (short, self-contained and concise), and suggesting that our algorithms should also work well with them.

8. Conclusions and Future Work

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 described. It was discovered that such descriptions are effective in calculating similarity, especially where they are derived from large quantities of material.

One of the primary motivations for mapping ontologies by concept similarity is to import documents "en-masse", based on the premise that this is more efficient than considering each one individually. However, a key issue arising from this experiment was the heterogeneity of ontologies, even those in the same domain: we saw how they can differ in terms of terminology, structure, scope and level of granularity. Indeed, Figure 4 showed an example of how documents were classified and grouped differently in our test ontologies. When mapping ontologies that are emerging on the Semantic Web, such differences are certain to exist, and on a larger scale. Although we were able to eliminate the effects of these different classifications for the purposes of the experiment, they would have to be given

consideration in any real world application. For example, in two ontologies, if concept A is found to be similar to concept B with a value of 0.2, this may mean that the two concepts are substantially different (none of their associated documents have much in common) and should not be mapped. On the other hand, it may indicate that just two out of ten documents attached to concept A are highly similar to concept B's documents, while the rest are completely irrelevant. Future experiments will look into harnessing this information about the similarity of concepts to determine situations in which it would be beneficial and efficient to consider importing documents on an individual basis.

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.

9. Acknowledgements

This research was supported by the Information Society Technologies project Diogene (IST-2001-33358). Further details can be found on the project Web site: 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.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The semantic web. *Scientific American*, 17 May.

Capuano, N., Gaeta, M., Micarelli, A. and Sangineto, E. (2002). An integrated architecture for automatic course generation. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*. Kazan, Russia.

Doan, A., Madhavan, J., Domingos, P. and Halevy, A. (2002). Learning to map between ontologies on the Semantic Web. In *Proceedings of WWW2002*. Hawaii, USA.

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

Mai, J. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation*, 57(5).

McGuinness, D. (2002). Ontologies come of age. In, Fensel D., Hendler J., Lieberman H. and Wahlster, W. (eds), *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press.

McGuinness, D., Fikes, R., Rice, J. and Wilder, S. (2000). An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*. Colorado, USA.

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

van Rijsbergen, K. (1979). *Information Retrieval*. London: Butterworths.

Saarti, J. (2002). Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation*, 58(1).

Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy: the Principles and Practice of Numerical Classification*. San Francisco: W.H. Freeman.

Tower, B., Chaisson, 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.