Managing Multiple Information Sources through Ontologies: Relationship between Vocabulary Heterogeneity and Loss of Information

Eduardo Mena* — Facultad de Informática, UPV/EHU, San Sebastián. Spain, jibmenie@si.ehu.es

Vipul Kashyap — LSDIS Lab, UGA & CS dept., Rutgers Univ., New Brunswick, NJ 08903, kashyap@cs.uga.edu

Arantza Illarramendi — Facultad de Informática, UPV/EHU, San Sebastián. Spain, jipileca@si.ehu.es

Amit Sheth — LSDIS Lab, http://lsdis.cs.uga.edu/, UGA, Athens, GA 30602, amit@cs.uga.edu

Abstract. The ability to deal with a huge number of independent and heterogeneous repositories is the most critical problem in Global Information Systems. One approach to enable efficient query processing is by utilizing semantic descriptions (organized as ontologies) of such repositories whenever available.

In this context semantic relationships among ontologies can be used by Query Processors. Three kind of relationships are considered: synonyms, hyponyms and hypernyms. Using synonyms the semantics of the query is preserved; however, when synonyms are not available and hypernyms or hyponyms are used there exists some loss of information that must be measured.

1 INTRODUCTION

The ability to deal with a huge number of independent repositories is the most critical problem in Global Information Systems. One approach to enable efficient query processing is by utilizing semantic descriptions of such repositories whenever available. We view *domain specific ontologies* as tools to capture the semantics of the underlying repositories. In the OBSERVER[1] architecture (see Figure 1) the content of each data repository (which may be composed of several data sources) of the Global Information System is described by an ontology created using a system based on Description Logics (DLs) [4].

A critical problem due to the autonomy of the component systems in this framework is *vocabulary heterogeneity* which arises due to the use of different vocabularies to characterize similar information across domains (e.g., "dictionary" in ontology A and "thesaurus" in ontology B). In this paper, we discuss the issues involved in enabling *vocabulary sharing*.

In order to answer a user query (which is a DL expression) formulated using terms in one component (user) ontology we have to translate the query using terms of other (target) ontologies of the system and then access the data underlying those ontologies. This translation should preserve the semantics of the user query. This requires the availability of semantic relationships among ontologies to the Query Processors, e.g., *synonyms* in different ontologies. In OBSERVER [8] this is supported by storing the synonyms in the Interontology Relationships Manager (IRM) module. When the user query is

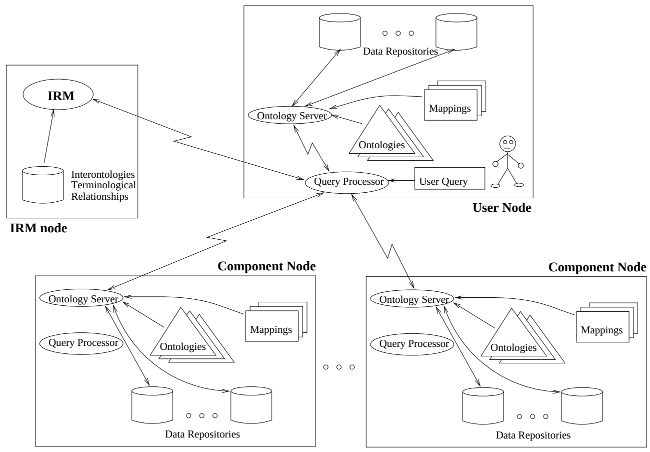

Figure 1: OBSERVER Global Architecture

translated into the "language" of some concrete ontology, the underlying data (that can be distributed among several heterogeneous repositories) is accessed, correlated and presented to the user. Mappings that link each term in an ontology with structures in underlying data repositories are combined in order to access and retrieve data from such repositories. A complete description of this process can be found in [8].

It can so happen that the user query may not be *fully translated* into terms of some ontology due to lack of synonym terms in the target ontology for some terms in the query. Therefore, mechanisms to deal with *partial translations* and incremental enrichment of the answer presented to the user must be implemented [8].

Furthermore, synonym relationships between terms in independent developed ontologies are very infrequent. On the contrary, and real examples confirm it, hierarchical relationships like *hyponym* and *hypernym* are very much frequent. The substitution of a term by its hypernym or hyponym, though providing answers which were not available earlier, however changes the semantics of the query. This leads to *loss of information*, and techniques to estimate such a loss must be developed.

Among the related works we can find in the literature the following are the more remarkable. We briefly comment the differences with our approach:

- in TSIMMIS [5] no ontology based on Description Logics is used to describe data sources and the problem of using different terms (different "languages") to make the same

* This work was supported by a grant of the Basque Country Government.

[1] The OBSERVER system is our approach of using multiple pre-existing ontologies to access heterogeneous, distributed and independently developed data repositories [8].

query is not considered.

- In Information Manifold [7] they do not consider the vocabulary sharing problem, i.e., the world view and external site descriptions must be in the same "language".
- In SIMS [1] the vocabulary sharing problem is solved but only when there is no loss of information. They substitute a term by its parents/children only when some properties guarantee that there is no loss of information.

2 GENERATING TRANSLATIONS WITH LOSS OF INFORMATION

In the process of refining[2] the answer presented to the user, s/he can choose between translating the query into new ontologies using synonyms or trying to fully translate the unused partial translations already found by substituting the non-translated terms using hyponym or hypernym relationships.

We substitute a non-translated term by the intersection of its immediate parents or the union of its immediate children. The loss of information is measured in both cases and the translation with less loss of information is chosen. This method is applied recursively until a full translation of the conflicting term is obtained. Using hyponym and hypernym relationships as described above can result in several possible translations of a non-translated term into a target ontology. Very simple intuitive measures depending on the extensions of the terms in the underlying ontologies may help in choosing the translations and minimizing the loss of information.

To obtain the immediate parents and children of a term in the target ontology two different kind of relationships related to the conflicting term are involved:

1. Synonyms, hyponyms and hypernyms between terms in the user and target ontologies.
2. Synonyms, hyponyms and hypernyms in the user ontology.

The first three types of relationships are those stored in the IRM repository. The second three types are relationships between terms in the same ontology; synonyms are equivalent terms, hyponyms are those terms subsumed by the non-translated term and hypernyms those terms that subsume the conflicting term.

The task of getting the immediate parents/children is not easy to perform. To obtain the parents/children within the user ontology, the corresponding functions (e.g., subsumption) of the DL systems can be used. But we must combine that answer with the immediate parents/children in the target ontology. Taking into account that some relationships stored in the IRM can be redundant (they were independently defined by different ontology administrators) such a task can be quite difficult. We would need a DL system dealing with "distributed" ontologies.

In Figure 2, we show two ontologies with some relationships between them (arrows are hyponyms relationships, double arrows are synonyms, and dashed lines are interontology relationships) and on the right the integrated ontology (synonyms are grouped into one term). We can see that obtaining the immediate parents is not evident; for instance, to get the immediate parents of B4 we must deduce that A1 is a child of B1. There are also redundant relationships like the one between A2 and B2.

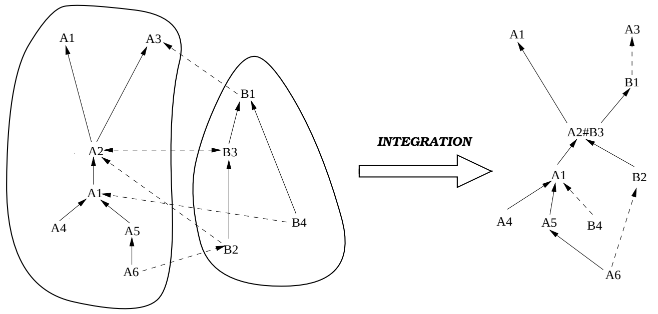

Figure 2: Integrating two ontologies

In order to work with the two previously presented kinds of relationships in an homogeneous way the solution seems to be the integration of the user and the target ontology and using the deductive power of the DL system to obtain the immediate parents/children [2]. The properties between terms in the different ontologies are exactly the interontology relationships stored in the IRM, so no intervention of the user is needed. Although some of the previous relationships can be redundant the DL system will classify the terms in the right place in the ontology. To know if the resulting terms of the integrated ontology are *primitive* or *defined* (it depends on A and B) we apply the rules described in [3].

3 EXTENSIONAL AND INTENSIONAL LOSS

The change in semantics caused by the use of hyponym and hypernym relationships must be measured not only in order to decide which substitution minimizes the loss of information but also to present to the user some kind of "level of confidence" in the new answer. The loss of information can be measured in an extensional manner (based on the number of instances of each term) and in an intensional manner (based on the terminological difference between the user query and the translated expression used to access the data).

For the extensional loss, information about the number of instances of each concept must be available. Based on that information, changes in *precision* and *recall* [6] are calculated at the same time that the translation is performed. Simultaneously, the intensional loss is built based on the difference between the non-translated term and its final translation.

To illustrate how the substitution of a term by a hyponym or hypernym affects the precision and recall parameters, we present the following example:
User Query: 'Get me all the students'

- Answer 1: All the graduate students are returned ⇒ Not all the students may be returned → Loss of recall.
- Answer 2: All the 'persons' are returned ⇒ Not all the answer instances returned belong to 'students' → Loss of precision as irrelevant answers are returned (persons who are not students).

In the previous example, precision and recall can be estimated based on estimates of the size of the extensions of 'students', 'graduate students' and 'persons'.

[2] The user query is translated into each component ontology and the data accessed and correlated in an iterative manner.

A special problem arises when computing intensional loss due to the vocabulary differences. As the intensional loss is expressed using terms of two different ontologies (e.g., "The current answer is about 'medical-books' instead of 'books' (original query)") the explanation could make no sense to the user as it mixes two "vocabularies". Imagine "book" in the user ontology is a hypernym of "book" in some component ontology restricted to medical domain. The explanation "Only 'book' is retrieved instead of 'book' (original query)" does not make any sense because both terms are homonyms. The problem could be even worse if both ontologies were expressed in different natural languages. So, the intensional loss can help to understand the loss only in some cases.

4 OBSERVER: THE PROTOTYPE

We have developed a prototype of OBSERVER, accessible for World Wide Web browsers at http://siul02.si.ehu.es/~jirgbdat/OBSERVER/, that allows accessing different heterogeneous data sources in the domain of bibliographic references. Both data repositories and ontologies describing them have been designed by other working groups and organizations as shown in Table 1. A complete description of ontologies and data repositories can be found in [8].

Ontology	Design source	Terms
WN	WordNet 1.5	73
Stanford-I	Bibliographic-Data (ARPA)	50
Stanford-II	Bibliographic-Data (ARPA)	51
LSDIS	Locally developed	18

Table 1: Details of ontologies

Ont.	Data Source	Data Org.	#Rec.
WN	UGA Main Library (subset)	Files containing MARC records	1.5K
Stanford-I	Library at Monterrey (subset)	Illustra DB storing MARC records	25K
Stanford-II	Library of Congress	Unknown	3.8M
LSDIS	Lab Publications	Text, HTML and Postscript files	70

Table 2: Details of data repositories underlying ontologies

It is important to notice (see Table 2) the heterogeneity among the ontologies (semantic heterogeneity) as well as in the data repositories (structural[3] and operational[4] heterogeneity because they have been developed by different organizations. In this way we want to capture a real case and deal with problems that never arise when ontologies and data repositories are designed under the same point of view.

[3]Different data structures: plain files, databases, WWW documents, etc, and different schemas, if they exist.

[4]Some data repositories are accessed using SQL commands, others by WWW browsers, and some of them they do not even have a defined query language or access method.

References

[1] Y. Arens, C.A. Knoblock, and W. Shen. Query reformulation for dynamic information integration. *Journal of Intelligent Information Systems*, 1996.

[2] J. M. Blanco, A. Illarramendi, A. Goñi, and J. M. Pérez. Using a terminological system to integrate relational databases. In *Information Systems Design and Hypermedia*. Cepadues-Editions, 1994.

[3] J.M. Blanco, A. Illarramendi, and A. Goñi. Building a Federated Database System: An approach using a Knowledge Based System. *International Journal on Intelligent and Cooperative Information Systems*, 3(4):415–455, December 1994.

[4] R. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216, 1985.

[5] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. In *Proc. of the 10th IPSJ, Tokyo, Japan*, 1994.

[6] C. W. Cleverdon. On the inverse relationship of recall and precision. *Journal of Documentation*, 28:195–201, 1972.

[7] A. Levy, D. Srivastava, and T. Kirk. Data Model and Query Evaluation in Global Information Systems. *Intelligent Information Systems*, 5(2), September 1995.

[8] E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Preexisting Ontologies. In *Proc. of the First IFCIS International Conference on Cooperative Information Systems (CoopIS'96), Brussels (Belgium), June*, 1996.