Applying an Ontology on Data Integration

Agustina Buccella and Alejandra Cechich
Departamento de Ciencias de la Computación, Universidad Nacional del Comahue,
Buenos Aires 1400, Neuquén, Argentina
Email: abuccel,acechich@uncoma.edu.ar

Nieves R. Brisaboa
Departamento de Computación, Universidade de A. Coruña,
Campus de Elviña s/n, 15071 – A. Coruña, España
Email:brisaboa@udc.es

Abstract. The term “Federated Databases” refers to the data integration of distributed, autonomous and heterogeneous databases. However, a federation can also include information systems, not only databases. When integrating data, several issues must be addressed. Here, we focus on the problem of heterogeneity, more specifically on semantic heterogeneity – that is problems related to semantically equivalent concepts or semantically related/unrelated concepts. In order to address this problem, we apply the idea of ontologies as a tool for data integration. In this paper, we explain this concept and we briefly describe a method for constructing an ontology by using a hybrid ontology approach.

Keywords: Federated Databases, Ontology, Semantic Heterogeneity.

1. Introduction

Nowadays, large modern enterprise has different portions of the organization using different database management systems to store and search their critical data. Competition, evolving technology, geographic distribution and the inevitable growing decentralization, all contribute to this diversity. All of these databases are very important for the enterprise and they have a different interfaces for their administration. It will be useful for the enterprise to retrieve the information through a common interface to realize, for instance, the full value of the data they contain [15]. The term **Federated Database** emerged to characterize techniques for proving an integrated data access having a set of distributed, heterogeneous and autonomous databases [14]. We briefly explain these concepts:

- ***Autonomy***: The users and the applications can access to the data through a federated system or by your own local system. The autonomy can be classified in three types [17,6]: *design autonomy, communication autonomy* and *execution autonomy.*
- ***Distribution***: Nowadays most computers are connected to some type of network, especially the Internet, and it is natural to think of combining application and data sources that are physically located on different hosts, but that can communicate through the network.
- ***Heterogeneity***: it can be classified into four categories [25]: *structure, syntax, system, and semantic*. The *structure* heterogeneity involves different data models; the *syntax* heterogeneity involves different languages and data representations and the *system* heterogeneity involves hardware and operating systems. The semantic heterogeneity can be classified as follows: *semantically equivalent concepts* (the models use different terms to refer the same concept, e.g. synonymous; the properties are modeled differently by distinct systems, etc.), *semantically unrelated concepts* (the same term may be used by distinct systems to denote completely different concepts) and *semantically related concepts* (generalization/specification, different classifications, etc.). Another similar classification of heterogeneity can be found in [10].

In order to address the problem of semantic heterogeneity previously described, we apply the idea of ontologies as a tool for data integration. Section 2 introduces the concept of ontologies and discusses different approaches for data integration. Then we describe our method to build an ontology. Future work and the conclusion are discussed in section 3.

2. Data Integration based on Ontologies

The term “ontology” across the years has been used in many ways and domains [1,12]. In the computer science world the ontologies are introduced by Gruber [13] as an *“explicit specification of a conceptualization*”. A *conceptualization* refers to an abstract model of how people commonly think about a real thing in the world, e.g. a chair. *Explicit specification* means that the concepts and relations of the abstract model have been given explicit names and definitions [24]. An ontology gives the name and the descriptions of the entities of specific domains using predicates that represent relationship between these entities. It provides a vocabulary to represent and communicate knowledge about the domain and a set of relationship containing the term of the vocabulary at a conceptual level. Therefore, an ontology might be used for data integration tasks because of its potential to describe the semantic of information sources and to solve heterogeneity problems [25,10].

On the other hand, the concepts *data integration*, *application integration* and *application interoperability* are similar but we must differentiate them [7]. ***Data integration*** is concerned with unifying data that share some common semantics but originate from unrelated sources. ***Application interoperability*** attempts to standardize the interfaces between stand-alone applications such that the data generated from one application can flow as the input to another application. ***Application integration*** involves aspects of data integration and of application interoperability. In this paper we focus mainly in the first one.

There are many systems that were designed to address the needs of data integration. The developers of each system have made different choices about the best way to provide the needed services. Some of the most popular systems are: the Garlic System [7], the TSIMMIS System [9], the ObjectGlobe System [21], SIMS System [4], etc. In [19] there are a briefly explication of the first three of them with an analysis of their advantages and disadvantages. All of them have been created to resolve any heterogeneity level. As we have already said, we concentrate only in semantic heterogeneity and for that, there are two different branches: *with ontologies and without ontologies.* On the “without ontologies” branch, there are several research efforts with different level of detail, see [2,3,16,20,11].

2.1. Data Integration using Ontologies

There are a lot of advantages in the use ontologies for data integration. Some of them are [19,22]: the ontology provides a rich, predefined vocabulary that serves as a stable conceptual interface to the databases and is independent of the database schemas; the knowledge represented by the ontology is sufficiently comprehensive to support translation of all the relevant information sources; the ontology supports consistency management and the recognition of inconsistent data; etc. Some researches about the ontology in the integration, can be seen in [23,18,26,5].

Following, we describe our method for building the structure of the ontology. Figure1 shows the algorithm designed to do that.

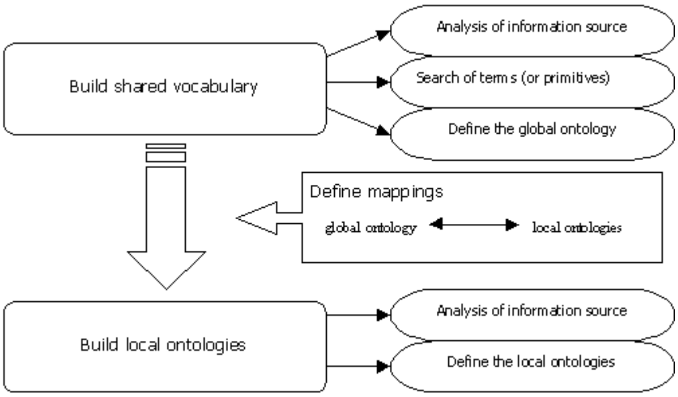
Build shared vocabulary
Analysis of information source
Search of terms (or primitives)
Define the global ontology
Define mappings
global ontology
local ontologies
Build local ontologies
Analysis of information source
Define the local ontologies

Figure 1: Ontology Construction Method

The method has three main stages. Each stage embodies a set of tasks that must be achieved:

1. ***Build shared vocabulary***. It mainly contains three steps:
   - *Analysis of information sources:* It implies a complete analysis of the information sources, e.g., what information is stored, how it is stored, the meaning of this information (the semantics), etc. It must localize the problems about semantic heterogeneity previously explained.
   - *Search of terms (or primitives):* Then, we should produce a list of terms that agree with the shared vocabulary.
   - *Define the global ontology:* With the data of the previous steps we are ready to create the global ontology.
2. ***Build local ontologies***: It contains two main steps: *analysis of information source* and *define the local ontologies*. As the previous stage, we make an analysis of the information source, but in this time we only focus on each database separately, without looking at the others. Then we can begin with the construction of local ontologies.
3. ***Define Mappings***: In this stage we define the mappings (and relations) between the concepts defined in the global ontology and the local ontologies.

Each step described before have a different level of complexity and must be analyzed separately.

3. Conclusions

Our current research on data integration uses the "ontology" concept. We described a method for the construction of a hybrid ontology approach, which will be used to identify all the cases of semantic heterogeneity and to create a guide about how to resolve each of them by using ontologies.

4. References

1. 3rd Millennium, Inc. Practical Data Integration in Biopharmaceutical R&D: Strategies and Technologies. A White Paper. May 2002. http://www.3rdmill.com/
2. Cali, A., Calvanese, D., De Giacomo, G., Lenzerini., M., Naggar, P., Vernacotola, F. IBIS: Data Integration at Work (extended abstract). SEBD 2002.
3. Cali, A., Calvanese, D., De Giacomo, G., Lenzerini., M. On the Role of Integrity Constraints in Data Integration. IEEE Computer Society Technical Committee on Data Engineering. 2002
4. Arens, Y., Hsu, C., Knoblock, C. A. Query processing in the SIMS Information Mediator. Advanced Planning Technology, Austin Tate (Ed.), AAAI Press pp. 61-69, Menlo Park, CA, 1996.
5. Brisaboa, N.R., Penabad, M.R., Places, A.S, Rodríguez, F.J. Ontologías en Federación de Bases de Datos. *Novática (ISSN 0211-2124)*, pp. 45-53. Julio 2002.

6. Busse, S., Kutsche, R.-D., Leser, U., Weber H. Federated Information Systems: Concepts, Terminology and Architectures. Technical Report. Nr. 99-9, TU Berlin. April 1999
7. Carey, M. y colabor.. Towards Heterogeneous Multimedia Information Systems: The Garlic Approach. Fifth International Workshop on (RIDE): Distributed Object Management. 1995.
8. Chandrasekaran, B.; Josephson, R. What are ontologies, and why do we need them? In IEEE Inteligent systems, 1999.
9. Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., Widom, J. The TSIMMIS project: Integration of heterogenous information sources. 16th Meeting of the Information Processing Society of Japan, pp. 7-18, Tokyo, Japan, October 1994.
10. Cheng Hian Goh. Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources. Phd, MIT, 1997. http://ccs.mit.edu/ebb/peo/mad.html - 03/2003
11. Constantinescu, C., Heinkel, U., Rantzau, R., Mistchang, B. SIES: An Approach for a Federated Information System in Manufacturing. www.informatik.uni-stuttgart.de/ipvr/as/personen/constantinescu/ise2001.pdf - 03/2003
12. Event/Process-Based Data Integration for the Gulf of Maine. Campobello Island, New Brunswick. June 12 – 14, 2002. www.spatial.maine.edu/~bdei/bdeippr.pdf - 03/2003
13. Gruber, T. A translation approach to portable ontology specifications. Knowledge Acquisition 1993 - 5(2):199–220. http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html
14. Hasselbring, W. Information System Integration. Comminications of the ACM. June 2000.
15. IBM Federated Database Technology – www7b.boulder.ibm.com/dmdd/library/techarticle/023haas/0203haas.html – 03/2003
16. L. M. Haas R. J. Miller B. Niswonger M. Tork Roth P. M. Schwarz E. L. Wimmers. Transforming Heterogeneous Data with Database Middleware: Beyond Integration. www.almaden.ibm.com/software/km/clio/clio.pdf
17. M.T. Özsu, P. Valduriez, Principles of distributed database systems, 2nd edition, Prentice Hall, 1999.
18. Quddus Chong, Judy Mullins, Rajesh Rajasekharan. An Ontology-based Metadata Management System for Heterogeneous Distributed Databases. CS590L – Winter 2002.
19. Steven Barlow. Data Integration. University of Passau. July 24, 2000.
20. Tejada, S., Knoblock, C.A., Minton, S. Learning object identification rules for Information Integration. *Information Systems* Vol. 26, Nº 8, pp. 607-633, 2001.
21. The ObjectGlobe Homepage. http://www.db.fmi.uni-passau.de:8000/projects/OG/
22. Thomas Adams, James Dullea, Peter Clark, Suryanarayana Sripada, and Thomas Barrett. Semantic Integration of Heterogeneous Information. Sources Using a Knowledge-Based System. In Proc 5th Int Conf on CS and Informatics (CS&I'2000), 2000.
23. Ubbo Visser, Heiner Stuckenschmidt, Christoph Schlieder. Interoperability in GIS – Enabling Technologies. 5th AGILE Conference on Geographic Information Science, Palma (Balearic Islands, Spain) April 25th-27th 2002
24. Visser, U. and Schlieder, C. Modelling with Ontologies. In: The Ontology and Modeling of Real Estate Transactions in European Juristictions Ashgate - 2002, to appear
25. Zhan Cui and Paul O'Brien. Domain Ontology Management Environment. In Proceedings of the 33rd Hawaii International Conference on System Sciences - 2000
26. Zhan Cui, Dean Jones and Paul O'Brien. Issues in Ontology-based Information Integration. IJCAI – Seattle, USA • August 5 2001.