Interesting Problems in Semantic Integration and Interoperability*

Extended Abstract

Vasilis Vassalos

Department of Informatics
Athens University of Economics and Business
Patission 76, Athens 10434 Greece
tel: +30 210 8203762, email: vassalos@aueb.gr

Abstract. We summarize the discussions of the breakout session on Problem Sharing, held during the Dagstuhl workshop on Semantic Integration and Interoperability. The breakout session brought together people from different communities (databases, AI planning, formal logic, knowledge representation) to share exciting and challenging problems in each community in a common language.

1 Introduction

The challenges of Semantic Integration and Interoperability (I&I) have attracted attention from a large number of researchers from a variety of fields (databases, automated reasoning, knowledge representation and ontology engineering, formal logic, web engineering, etc), as evidenced also by the participation in recent workshops on this topic, e.g., [KSS+04,DHN03].

It has been our experience that each community contributes significantly to the area of Semantic I&I, but that there are more opportunities for synergy than are currently realized. In particular, the use of different nomenclature sometimes obscures the strong relationships between problems and techniques in Semantic I&I that researchers from different communities work on. The purpose of scientific meetings such as the Dagstuhl workshops is indeed to foster collaborative and interdisciplinary work, but sometimes there is an implicit assumption that participants understand and agree on problem definitions.

In what follows we present briefly four problems in semantic integration and interoperability, ranging from specific to broad challenge problems. Each problem was proposed by one of the participants of the breakout session on Sharing Problem Definitions, held during the Daghstuhl workshop on Semantic Interoperability and Integration. These problems are shared to invite collaboration and interdisciplinary work; they also demonstrate the relationships between the Semantic I&I research problems that motivate researchers from different fields.

* Breakout session held at the Dagstuhl Seminar on Semantic Interoperability and Integration on September 23, 2004. Participants: Naveen Ashish, Michael Grüninger, Marco Schorlemmer, Vasilis Vassalos.

Dagstuhl Seminar Proceedings 04391
Semantic Interoperability and Integration
http://drops.dagstuhl.de/opus/volltexte/2005/54

2 Data Integration using Logical Views

The goal of data integration is to provide uniform access to a variety of autonomous, heteregeneous information sources via a high-level language. These sources may contain information that is structurally and semantically diverse and may support different access methods and interfaces. A data integration system should permit integrated querying and transformation of this diverse information while respecting the autonomy and the differing capabilities of the underlying sources. Data integration has been the focus of significant academic and industrial research activity [Hal01,PV02] in the database, knowledge representation and AI planning communities.

Formally, a data integration system can be described as a triple (G, S, M) [Len02], where

- G is the global *schema*, that is, a set of objects, axioms and constraints defining the "world", i.e., our universe of discourse, in an appropriate language
- S is the local *schema*, i.e., a set of objects, axioms and constraints on the information sources in an appropriate language, and
- M is a set of assertions relating elements from G and S.

The many flavors of data integration systems such as local as view, global as view, etc., can be mapped into the above description [Len02].

3 Automatic Structured Query Generation from Natural Language Search Terms

Domain specific search engines or meta-search engines (such as search engines for a scientific discipline such as earth science or bioinformatics) will likely retrieve structured information from the multiple sources they are searching. However, end users are likely to prefer using natural language keywords or sentences as input search terms or queries (such as "precipitation datasets southern california last 2 years"). The challenge is to automatically generate structured queries from such natural language input terms. Domain and source ontologies (and mappings between them) will help in addressing semantic mismatch problems or using uncontrolled keywords, but they do not directly solve the problem.

In related work, natural language interfaces to database systems were investigated heavily in the 70s and 80s [ART95]. Lately, when the data integration community has developed parsers for converting natural language queries to structured queries in a data integration setting, e.g., [YM03,Coh00]. This line of work is a good beginning, but does not completely address our problem as the search terms are more open-ended, and the available semantic information is not taken into account.

4 The limits of semantic integration

The last two problems probe the meaning of the definition of semantic integration. The first of these problems is based on the following scenario. Assume two

agents A and B that share the same ontology, meaning that they agree on the semantics of their nonlogical symbols. The two agents can still disagree on the interpretation of their logical symbols, or on their inference rules. For example, agent A may be using classical first-order logic while agent B may be using first-order logic with the closed world assumption (e.g., in the case of a Prolog system.) Does making A and B semantically interoperable mean that A can ask B whether $\Gamma \vdash \phi$ and use the results in its own reasoning? Or does it simply mean that A can ask B for facts (and vice-versa)? Clearly, realizing the Semantic Web [BL98] requires that A can use B's reasoning service. It should be also clear that this presents a problem in the above scenario, as inferences are not preserved between A and B, despite their shared ontology.

Therefore, the challenge, which we call *proof-theoretic integration*, is to describe and take into account the differences in inference engines. For example, under what circumstances could A use the answers provided by B in its reasoning? Alternatively, how can the agents map their logical capabilities into compatible "logical layers" [BL98]?

Finally, the larger question is finding the right measure for semantic integration and interoperability. Apparently, agreeing on/mapping the semantics of nonlogical symbols is not enough. Is preservation of inferences enough for semantic integration? What are compelling examples/counterexamples?

References

[ART95] I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. Natural language interfaces to databases - an introduction. *Natural Language Engineering*, 1(1), 1995.

[BL98] Tim Berners-Lee. Semantic web road map, 1998. http://www.w3.org/DesignIssues/Semantic.html.

[Coh00] W. W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Inf. Systems*, 18, 2000.

[DHN03] A. Doan, A. Halevy, and N. Noy, editors. *Online Proceedings of the Semantic Integration Workshop (SI'03)*, 2003. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-82/.

[Hal01] Alon Y. Halevy. Answering queries using views: A survey. *The International Journal on Very Large Data Bases*, 10:270 – 294, December 2001.

[KSS+04] Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors. *Online Proceedings of the Dagstuhl Seminar on Semantic Integration and Interoperability*, 2004. http://www.dagstuhl.de/04391/Proceedings.

[Len02] M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of ACM PODS*, 2002.

[PV02] Y. Papakonstantinou and V. Vassalos. Architecture and implementation of an xquery-based information integration platform. *IEEE Data Engineering Bulletin*, 25(1), 2002.

[YM03] B. Yan and R. MacGregor. Translating nave user queries on the semantic web. In *Online Proceedings of the Semantic Integration Workshop (SI'03)*, 2003. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-82/.