S

Semantic Integration and Interoperability among Portals

Konstantinos Kotis
University of the Aegean, Greece

George Vouros
University of the Aegean, Greece

INTRODUCTION

In distributed settings, such as that of the World Wide Web, where a large number of information sources and services reside, portals provide a single point of global access via a single and unified view. This view is circumscribed by a specific conceptualization and a specific vocabulary whose entries provide lexicalizations of the concepts used for shaping information, data, and services provided. Ontologies play a key role to shaping information, as they provide conceptualizations of domains. Different portals may use different or partially overlapping ontologies for shaping information, or even different schemata for storing data. This affects the integration of information from different portals, and the interoperability between the services that portals provide. Consequently, this situation affects recall and precision of information retrieval, and sets limitations to the composition (and decomposition) of services among portals for serving clients' (users or software agents) requests.

Semantic integration refers to the set of problems that appear between disparate information sources and concern matching ontologies or schemas, detecting duplicate tuples, reconciling inconsistent data values, and reasoning with semantic mappings. The goal is to integrate information and data under a single view, preserving the semantics of the sources.

Service invocation in a distributed and open setting involves discovering the appropriate services, selecting among a set of candidates that match the requirements of the client, interacting with the selected service, and interpreting service replies. Much of the work to be done toward services' interoperability concerns publishing semantic service descriptors which clients will readily exploit. The goal is for software agents to discover, interact with, and fetch the results of services automatically.

Both problems concern the mapping, aligning, translating, and merging of ontologies. This article aims to provide a review to the techniques for semantic integration and interoperability of portals by exploiting ontologies. It does not aim to provide an in-depth and exhaustive presentation of the existing approaches.[1] There exist some excellent surveys on the methods and techniques proposed, for instance, in Shvaiko and Euzenat (2005) or in Noy (2004). Instead it provides definitions and a roadmap to the existing research efforts toward this exciting research topic which is of much importance for any Web user, community, enterprise, organization, and government.

BACKGROUND

Although the terms semantic integration and semantic interoperability are used interchangeably in many contexts, we consider them to be distinct, although tightly intertwined: integration concerns information, while interoperability concerns functionality. The common denominator to both problems, as it will be discussed in subsequent subsections, is *sharing the semantics*.

The ISO/IEC 2382 Information Technology Vocabulary[2] defines interoperability as *the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units*.

Dealing with semantic interoperability, we require software units (let us call them agents) to be able to find, use, execute, and interpret outcomes of services provided by other agents. Toward this aim, agents need to publish machine-exploitable descriptions of their capabilities and interaction/communication models. Service capabilities have to be matched against agents' goals and requirements. Matchmaking services can be offered by dedicated agents (translators / mediators) and be distributed to various places, or by the client and service provider agents. The client agents will invoke services by choosing among those matching their requirements and deduce from their descriptions the content of the messages required for interaction. Finally, exploiting the semantics of the service descriptions, clients can interpret the service responses. In more advanced settings, agents may compose multiple services toward achieving a unique goal by reasoning about the effects of services (e.g., for comparing the prices of products offered from different retailers). This is extremely valuable for portals offering a single-point of access to information: they may discover and invoke remote services based on their semantic descriptions and the goals of the (human or software) agents using the portal.

Considering the architecture implied from the above description, this comprises agents that offer and request services, as well as a number of middle-agents that help clients achieve their aims. Of major interest are semantic matchmakers that act like search engines or yellow pages, and ontology mapping registries that help agents bridge the gap between agents' conceptualizations, ensuring a complete and consistent mapping between concepts, relations, individuals, and rules for service related reasoning. Burstein and McDermott (2005, p. 72) have argued that "it may at times be difficult for mediators to relieve functional agents (clients and services) of this responsibility," pointing that "we expect particular agents to be responsible for translating the content of messages produced at different stages of their interaction."

Semantic service descriptions are developed using general-purpose standard ontologies (e.g., those specified by OWL-S[3] or WSMO[4]) and domain specific ontologies. Therefore, the problem of semantic interoperability largely depends on the ability of agents to *align* the ontologies involved, solving the semantic integration problem.

Concerning information integration, two agents are integrated if they can successfully communicate with each other, meaning that they can adequately interpret information communicated between them. Being semantically integrated, after information has been sent to the receiver, the receiver will associate this information to specific concepts (i.e., it will interpret it by means of a specific conceptualization) and will draw all these implications that the sender would exactly have drawn with the same information. In other words, for meaningful information exchange or integration, providers and consumers need compatible semantics.

A traditional example for information integration is the Catalog Integration example (Figure 1) (Shvaiko & Euzenat, 2005). B2B applications represent and store their products in electronic catalog-type models. Catalogs are very simple ontologies, tree-like structures that organize concepts' descriptions hierarchically. A typical example of such a model is the product directory of http://www.amazon.com. In order for a company to participate in a specific marketplace in which amazon.com participates, it must identify correspondences between entries of its catalogs and entries of the catalogs of www.amazon.com. Having identified the correspondences between the entries of the catalogs, it can be assumed that the catalogs are aligned.

Achieving this semantic integration manually (by means of specifying semantic matches) is extremely laborious and error prone and thus very costly. For instance, Doan and Halevy (2005) report that an integration project at the GTE telecommunications company involving 40 databases with a total of 27,000 attributes of relational tables estimated to take more that 12 person years. This was a typical case because the original developers of the databases were not involved. In another example reported by Doan and Halevy (2005), the U.S. Department of Defense standardization effort aimed to produce a single standard data model exceeding 10^5 entities and 10^6 attributes. By the year 2000, they recognized the need for a new approach to this scale of information integration. As one can imagine, things become worse in a distributed and open setting such as the (semantic) Web. New information sources may appear here and there, with numerous data and information being structured using different schemata or ontologies, even for the same domain.

To manage such cases, Uschold and Gruninger (2002, 2005) point out that semantics can be managed effectively

Figure 1. Two catalog schemata from two different companies in a common marketplace