Knowledge Acquisition from Semantically Heterogeneous Data

Doina Caragea
Kansas State University, USA

Vasant Honavar
Iowa State University, USA

INTRODUCTION

Recent advances in sensors, digital storage, computing and communications technologies have led to a proliferation of autonomously operated, geographically distributed data repositories in virtually every area of human endeavor, including e-business and e-commerce, e-science, e-government, security informatics, etc. Effective use of such data in practice (e.g., building useful predictive models of consumer behavior, discovery of factors that contribute to large climatic changes, analysis of demographic factors that contribute to global poverty, analysis of social networks, or even finding out what makes a book a bestseller) requires accessing and analyzing data from multiple heterogeneous sources.

The Semantic Web enterprise (Berners-Lee et al., 2001) is aimed at making the contents of the Web machine interpretable, so that heterogeneous data sources can be used together. Thus, data and resources on the Web are annotated and linked by associating meta data that make explicit the ontological commitments of the data source providers or, in some cases, the shared ontological commitments of a small community of users.

Given the autonomous nature of the data sources on the Web and the diverse purposes for which the data are gathered, in the absence of a universal ontology it is inevitable that there is no unique global interpretation of the data, that serves the needs of all users under all scenarios. Many groups have attempted to develop, with varying degrees of success, tools for flexible integration and querying of data from semantically disparate sources (Levy, 2000; Noy, 2004; Doan, & Halevy, 2005), as well as techniques for discovering semantic correspondences between ontologies to assist in this process (Kalfoglou, & Schorlemmer, 2005; Noy and Stuckenschmidt, 2005). These and related advances in Semantic Web technologies present unprecedented opportunities for exploiting multiple related data sources, each annotated with its own meta data, in discovering useful knowledge in many application domains.

While there has been significant work on applying machine learning to ontology construction, information extraction from text, and discovery of mappings between ontologies (Kushmerick, et al., 2005), there has been relatively little work on machine learning approaches to knowledge acquisition from data sources annotated with meta data that expose the structure (schema) and semantics (in reference to a particular ontology).

However, there is a large body of literature on distributed learning (see (Kargupta, & Chan, 1999) for a survey). Furthermore, recent work (Zhang et al., 2005; Hotho et al., 2003) has shown that in addition to data, the use of meta data in the form of ontologies (class hierarchies, attribute value hierarchies) can improve the quality (accuracy, interpretability) of the learned predictive models.

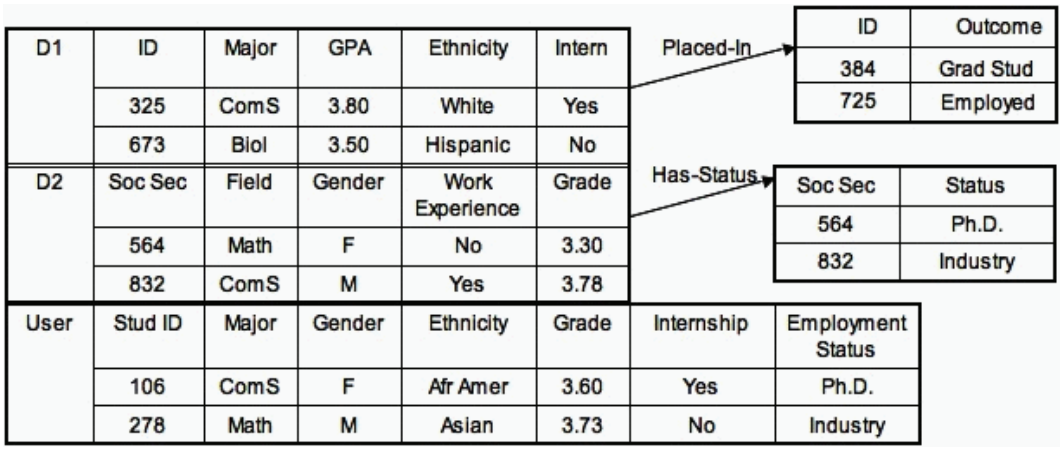
The purpose of this chapter is to precisely define the problem of knowledge acquisition from semantically heterogeneous data and summarize recent advances that have led to a solution to this problem (Caragea et al., 2005).

BACKGROUND

Motivating Example

The problem addressed is best illustrated by an example. Consider two academic departments that independently collect information about their students (Figure 1). Suppose a data set D_1 collected by the first department is organized in two tables, *Student* and *Outcome*, linked by a *Placed-In* relation using

Figure 1. Student data collected by two departments from a statistician's perspective



ID as the common key. Students are described by *ID*, *Major*, *GPA, Ethnicity* and *Intern.* Suppose a data set D_2 collected by the second department has a *Student* table and a *Status* table, linked by *Has-Status* relation using *Soc Sec* as the common key. Suppose *Student* in D_2 is described by the attributes *Student ID*, *Field*, *Gender, Work-Experience* and *Grade.*

Consider a user, e.g., a university statistician, interested in constructing a predictive model based on data from two departments of interest from his or her own perspective, where the representative attributes are *Student ID*, *Major, Gender*, *Ethnicity*, *Grade, Internship* and *Employment Status.* For example, the statistician may want to construct a model that can be used to infer whether a typical student (represented as in the entry corresponding to D_U in Figure 1) is likely go on to get a *Ph.D.* This requires the ability to perform queries over the two data sources associated with the departments of interest from the user's perspective (e.g., *fraction of students with internship experience that go onto Ph.D*). However, because the structure (schema) and data semantics of the data sources differ from the statistician's perspective, he must establish the correspondences between the user attributes and the data source attributes.

Ontology-Extended Data Sources and User Views

In our framework each data source has associated with it a data source description (i.e., the schema and ontology of the data source). We call the resulting data sources, *ontology extended data sources* (OEDS). An OEDS is a tuple $\mathscr{D}=\{D,S,O\}$, where D is the actual data set in the data source, S the data source schema and O the data source ontology (Caragea et al., 2005). The formal semantics of OEDS are based on ontology-extended relational algebra (Bonatti et al., 2003).

A *data set D* is an instantiation $\mathscr{I}(S)$ of a schema. The *ontology O* of an OEDS $\mathscr{D}$ consists of two parts: *structure ontology*, O_S, that defines the semantics of the data source schema (entities and attributes of entities that appear in data source schema *S*); and *content ontology*, O_I, that defines the semantics of the data instances (values and relationships between values that the attributes can take in instantiations of schema *S*). Of particular interest are ontologies that take the form of *is-a* hierarchies and *has-part* hierarchies. For example, the values of the *Status* attribute in data source D_2 are organized into an *is-a* hierarchy.

Because it is unrealistic to assume the existence of a single global ontology that corresponds to a universally agreed upon set of ontological commitments for all users, our framework allows each user or a community of users to select the ontological commitments that they deem useful in a specific context. A *user's view of data sources* $\mathscr{D}_1, \mathscr{D}_2 \ldots \mathscr{D}_n$ *is specified* by user schema S_U, user ontology O_U, together with a set of semantic *correspondence constraints IC*, and the associated set of *mappings* from the user schema S_U to the data source schemas $S_1,\ldots,S_n$ and from user ontology O_U to the data source ontologies $O_1,\ldots,O_n$ (Caragea et al, 2005).