Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data

Dimitrios Skoutas, National Technical University of Athens, Greece
Alkis Simitsis, National Technical University of Athens, Greece

ABSTRACT

One of the main tasks in the early stages of a data warehouse project is the identification of the appropriate transformations and the specification of inter-schema mappings from the data sources to the data warehouse. In this article, we propose an ontology-based approach to facilitate the conceptual design of the back stage of a data warehouse. A graph-based representation is used as a conceptual model for the datastores, so that both structured and semi-structured data are supported and handled in a uniform way. The proposed approach is based on the use of Semantic Web technologies to semantically annotate the data sources and the data warehouse, so that mappings between them can be inferred, thereby resolving the issue of heterogeneity. Specifically, a suitable application ontology is created and used to annotate the datastores. The language used for describing the ontology is OWL-DL. Based on the provided annotations, a DL reasoner is employed to infer semantic correspondences and conflicts among the datastores, and to propose a set of conceptual operations for transforming data from the source datastores to the data warehouse.

Keywords: conceptual design; data semantics; data warehousing; ETL; ontology; semantic matching; workflow diagram

INTRODUCTION

Successful planning and decision making in large enterprises requires the ability of efficiently processing and analyzing the organization’s informational assets, such as data regarding products, sales, customers, and so on. Such data are typically distributed in several heterogeneous sources, and stored under different structures and formats. For this purpose, as well as for performance issues, data warehouses are employed to integrate the data and provide an appropriate infrastructure for querying, reporting, mining and for other advanced analysis techniques. One of the main
challenges in the early phases of the design and deployment of a data warehouse is to identify the appropriate data sources, and to specify the operations and transformations needed to overcome the arising structural and semantic conflicts. Specialized tools, commonly known as extraction–transformation–loading (ETL) tools, already have been proposed to facilitate this procedure (Luján-Mora, Vassiliadis, & Trujillo, 2004; Trujillo & Luján-Mora, 2003; Vassiliadis, Simitsis, & Skiadopoulos, 2002).

Moreover, several commercial solutions already exist (IBM, 2005; Informatica, 2005; Microsoft, 2005; Oracle, 2005).

However, the design part of these tools mainly focuses on the representation and modeling of the ETL processes, whereas the identification of the required mappings and transformations needs to be done manually. The lack of precise metadata hinders the automation of this task. The required information regarding the semantics of the data sources, as well as the constraints and requirements of the data warehouse application, tends to be missing. Usually, such information is incomplete, or even inconsistent, often being hard-coded within the schemas or metadata of the sources or even provided in natural language format after oral communication with the involved parties (e.g., business managers and administrators/designers of the data warehouse) (Hüsemann, Lechtenbörger, & Vossen, 2000). As a result, designing ETL processes becomes a very tedious and error-prone task. Given the fact that typical ETL processes are quite complex and that significant operational problems can occur with improperly designed ETL systems, developing a formal, metadata-driven approach to allow a high degree of automation of the ETL design, is critical in employing a data warehouse solution.

The schema of a datastore describes the way that data are structured when stored, but does not provide any information for their intended semantics. Therefore, metadata are required to allow for the understanding, management, and processing of these data. Semantic Web technologies provide a means to formally specify the metadata, so that automated reasoning techniques can be employed to facilitate further processing. The Semantic Web is an extension of the current Web, where information is described by formally defined, machine-processable metadata to further facilitate and automate tasks such as searching, sharing, and combining information. Ontologies are a key enabling technology for the Semantic Web. Borst (1997) defines an ontology as a formal and explicit specification of a shared conceptualization. An ontology provides a way for describing the meaning and relationships of the terms in a domain. In the context of a data warehouse application, ontologies can be used as a conceptual model for describing the semantics of the data sources, allowing reasoning techniques to be applied for inferring correspondences or conflicts among these sources.

Earlier work argues that ontologies constitute a very suitable model for describing the semantics of the data sources in a data warehouse application and for automatically identifying correspondences among these sources (Skoutas & Simitsis, 2006). Specifically, integration from relational sources is considered. The designer uses the application requirements, as well as implicit knowledge of the domain, to create a shared vocabulary, that is a set of terms, and applies this vocabulary to annotate the tables and attributes of each source, such as denoting whether null values are allowed for an attribute or what unit of measurement is used for the values of an attribute. Given the vocabulary and the annotations, an ontology is constructed, based on which reasoning tasks may be performed to facilitate the ETL workflow construction.

On the other hand, even though the relational model has widespread adoption and an RDBMS constitutes the typical solution for storing an organization’s operational data, due to the increasingly important role of the Internet and the World Wide Web in e-commerce and business transactions in general, semi-structured data also play a progressively more important role in this context, significantly increasing the amount of heterogeneity between the data sources, and thus, the complexity of the ETL design process. Semi-structured data refers to
OOPS!: A Pitfall-Based System for Ontology Diagnosis
Maria Poveda-Villalón, Asunción Gómez-Pérez and Mari Carmen Suárez-Figueroa
www.igi-global.com/chapter/oops/196437?camid=4v1a