Ontologies with Semantic Web/Grid in Data Integration for OLAP

Tapio Niemi, Helsinki Institute of Physics, CERN, Switzerland
Santtu Toivonen, VTT Technical Research Centre of Finland, Finland
Marko Niinimäki, University of Tampere, Finland
Jyrki Nummenmaa, University of Tampere, Finland

ABSTRACT

Traditionally, data used in OLAP (online analytical processing) have been limited to the contents of the data warehouse of a company. However, the needs for analysis are often more demanding and data are needed from different sources. In this article, we study how the semantics of data sources can be described to allow combining data from several sources into an OLAP cube. We apply Semantic Web technologies for defining an OWL/RDF ontology for OLAP data sources and OLAP cubes. These definitions are then utilised in OLAP cube formation by posing an OWL/RDF ontology-based query against them. We use Grid technologies to enhance the efficiency of processing and ensuring security. Our primary interest is in the cube construction (i.e., ETL process), and we assume that standard OLAP methods can be used for the actual analysis. Our tests show that the proposed approach can speed up the construction of an OLAP cube for ad hoc queries by supporting a high-level query language and reducing the amount of required data.

Keywords: data warehousing; Grid; online analytical processing(OLAP); Semantic Web; query languages

INTRODUCTION

The corporate environment is increasingly dependent on the speed and accuracy of inter-organizational decision making. Reliable and up-to-date data, together with the necessary knowledge needed to interpret it, probably are the most important resources in any kind of decision making. Business intelligence systems (e.g., OLAP) play a central role in corporate decision support, as they act as an interface to the data and information.

The data sources are often heterogeneous and geographically distributed. They may include operational databases of the company, data files in the intranet, and also data in the public internet. For example Eurostat, the Statistical Office of the European Communities (http://europa.eu.int/comm/eurostat), publishes a lot of statistical information on the World Wide Web. However, retrieving the data must be mainly done manually. Our aim is to develop methods to automate utilising these kinds of datastores.
together with data from other sources. In this way, a researcher in a company could integrate statistical data with the company’s own data to make the analysis more profound.

Online analytical processing (OLAP) is one of the most promising methods used in analysis of business data and it is supported by major database system vendors. However, the deployment of OLAP systems has been slow due to deficiencies in design methods, the need for manual work in data collection and integration, complicated query methods, and excessive computing power requirements. Especially, integrating different databases is seldom straightforward. For example, in relational databases, semantically different attributes still can have the same domain (data type): a person’s age and telephone number can be stored using the same integer data type. This makes it possible to join two tables using the age and telephone number columns. Another problem is that schemas of relational databases are thought to be unique in the sense that attributes having the same name still can have different meanings in different schemas. Use of the resource description framework (RDF) is seen as a useful technical solution to solve these problems (Berners-Lee, 1998), since RDF makes it possible to define the meaning of data in a machine-readable form. In practice, we use common RDF definitions to describe the attributes used in data sources. For example, in the context of OLAP, we can create definitions for commonly used hierarchy types, measurement units, currencies, aggregation types, and so forth. Existing data sources can be transformed to follow the common ontology by writing rules how local attributes are mapped to corresponding concepts of the common ontology.

Here, we present a data source integration solution for OLAP, based on RDF and other semantic web technologies. Some preliminary ideas of this work have been presented as a poster (Toivonen & Niemi, 2004) and a conference paper by Niemi et al. (Niemi, Toivonen, & Niinimäki, 2005). Our implementation that uses distributed computing technologies has been described by Niinimäki (Niinimäki, 2006). The fundamental idea is to describe the data sources using an ontology-based on both global and application area-specific concepts and restrictions related to their instances. Using this ontological information, it is possible to collect and transform information from local data sources and further build an OLAP cube from the collected data.

In our design, a general ontology (or upper ontology) describes OLAP concepts such that there are measures (like value of trade of a certain product between two countries) and dimension hierarchies (say, type of product “High-pressure hydro-electric conduits,” its subgroup—“Tubes, pipes and fittings, of iron or steel” and main group “Iron and steel”). An application-specific ontology is derived from the general one; in our case we use the structure of world trade data as an example. We assume that there is one accessible global domain-specific ontology. Local datastores can be diverted from this ontology, in which case transformations (mappings of the local data formats to the global ontology) are provided. Data itself may be distributed in several sources that can be of different types—relational databases, XML files (The World Wide Web Consortium, 2002), and so forth. In order to integrate the data from different sources, we use an interim RDF data format and an RDF query language. Moreover, we show how distributed computing (Grid) technologies can be utilised for accessing the data and efficiently preparing it for uploading in an OLAP server.

We have constructed a prototype implementation for our method and tested it with a data set of ca. 3.9 million tuples. Our comparison with a commercial and an open source OLAP server shows that our approach can improve the total evaluation time of ad hoc queries, since only the relevant data for the query at hand needs to be loaded into an OLAP server. In the traditional approach, the whole data set is first extracted, transformed in a common format, and finally loaded into the OLAP server. This process generally is known as ETL (Extraction, Transform, Loading). Therefore, we also have tested an open source ETL tool as a means for...
Towards Large-Scale Unsupervised Relation Extraction from the Web
[www.igi-global.com/article/towards-large-scale-unsupervised-relation/74337?camid=4v1a](http://www.igi-global.com/article/towards-large-scale-unsupervised-relation/74337?camid=4v1a)