Automated Integration of Heterogeneous Data Warehouse Schemas

Marko Banek, University of Zagreb, Croatia
Boris Vrdoljak, University of Zagreb, Croatia
Min Tjoa, Vienna University of Technology, Austria
Zoran Skočir, University of Zagreb, Croatia

ABSTRACT

A federated data warehouse is a logical integration of data warehouses applicable when physical integration is impossible due to privacy policy or legal restrictions. In healthcare systems federated data warehouses are a most feasible source of data for deducing guidelines for evidence-based medicine based on data material from different participating institutions. In order to enable the translation of queries in a federated approach, schemas of the federated warehouse and the local warehouses must be matched. In this paper we present a procedure that enables the matching process for schema structures specific to the multidimensional model of data warehouses: facts, measures, dimensions, aggregation levels and dimensional attributes. Similarities between warehouse-specific structures are computed by using linguistic and structural comparison. The calculated values are used to create necessary mappings.

Keywords: bipartite graph; data warehouse; data warehouse federation; data warehouse integration; mapping; multidimensional model; schema matching; semantic similarity

INTRODUCTION

Increasing competitiveness in business and permanent demands for greater efficiency (either in business or government and non-profit organizations) enforce independent organizations to integrate their data warehouses. Data warehouse integration provides a broader base for decision-support systems, knowledge discovery and data mining than each of the separate independent warehouses could offer. Large corporations integrate their separately developed regional warehouses, newly merged companies integrate their warehouses to enable the business to be run centrally, while independent organizations join together their warehouses leading to a significant benefit to all participants and/or their customers.
In the field of healthcare the emergence of evidence-based medicine has made data integration a sine-qua-non topic. Evidence-based medicine (EBM) is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients (Sackett, Rosenberg, Muir Gray, Haynes & Richardson, 1996). It complements an existing clinical decision making process with the most accurate and most efficient research evidence. Application of its concepts speeds up the transfer of clinical research findings into practice, leading to a higher percentage of healed patients, to cost reduction, both for patients and health insurance organizations, as well as to the improvement of the healthcare process as a whole.

Successful application of evidence-based medicine is strongly related to the data it relies on. The central part of an evidence-based medical information system is a large data warehouse that unites all relevant internal healthcare data of an institution with the evidence-based guidelines coming mostly from outside scientific sources. In order to enhance the productivity of their administration, healthcare organizations practicing evidence-based medicine are striving to a better cooperation with other related organizations. The more data is joined together and the larger and thus more reliable data patterns are created, the more knowledge can be gathered from them. Thus, healthcare organizations practicing evidence-based medicine need to join their data into a single data warehouse, which serves as the foundation of the knowledge discovery system.

The traditional approach to integrating two or more data warehouses (often called component warehouses or local warehouses) is to create a new warehouse, to which the data from all component warehouses are physically copied and thus joined into a single system. However, when independent organizations share their data for mutual purposes, their privacy policies or legal limitations may reduce the access to some data and restrict or completely forbid any physical copy of their data to be created in any system that is out of their full control. The highly confidential and legally protected healthcare records, whose integration has motivated this article, are a typical example of such data.

An alternative approach to a “fully” physical integration that is able to cope with the imposed restrictions could be based on the bus architecture (Inmon, 1996; Hackney, 1997). The local warehouses would contain both detailed and summarized data and a central, physical data warehouse containing exclusively summarized data would be used for OLAP. However, discovering rules in evidence-based medicine requires processing of detailed data that corresponds to the basic (i.e. most detailed) granularity level in the component warehouses. Thus, all data from the components should be transferred into the joint warehouse and not only the summarized data. Moreover, the bus architecture requires the component warehouses to be developed simultaneously, while in our case the integration of independently developed, heterogeneous warehouses is needed.

In our opinion, the proper solution to integrating healthcare data warehouses is a data warehouse federation (Sheth & Larson, 1990; Jindal & Acharya, 2004). The integration is performed from a logical point of view, using a common conceptual model (Jindal & Acharya, 2004), while only heterogeneous local source warehouses exist physically. The user of such a federated solution observes the whole federation as a single unit i.e. she must not notice that several heterogeneous parts of the warehouse actually exist (Sheth & Larson, 1990). The logical existence of the federated warehouse does not have any impact on the users of local component warehouses. In our previous work (Banek, Tjoa & Stolba, 2006; Stolba, Banek & Tjoa, 2006), we developed the conceptual model of a federated data warehouse that unifies data warehouses of different health insurance organizations.

Queries on the federated data warehouse must be translated into sub-queries that correspond to the conceptual models of local warehouses. Furthermore, results of the queries need to be translated to match the conceptual model of the federation and then merged into a single
Related Content

Ethical Data Mining and Social Science Data Exploration and Description: Scope and Limitations in Social Science Research
www.irma-international.org/chapter/ethical-data-mining-social-science/76255/

Referential Horizontal Partitioning Selection Problem in Data Warehouses: Hardness Study and Selection Algorithms
www.irma-international.org/article/referential-horizontal-partitioning-selection-problem/37402/

Semi-Automatic Design of Spatial Data Cubes from Simulation Model Results
Hadj Mahboubi, Sandro Bimonte, Guillaume Deffuant, Jean-Pierre Chanet and François Pinet (2013). International Journal of Data Warehousing and Mining (pp. 70-95).
www.irma-international.org/article/semi-automatic-design-spatial-data/75616/

A Query Beehive Algorithm for Data Warehouse Buffer Management and Query Scheduling

Untangling BioOntologies for Mining Biomedical Information
www.irma-international.org/chapter/untangling-bioontologies-mining-biomedical-information/21732/