

Solving Data Inconsistencies and Data Integration with a Data Quality Manager

Angeles Maria del Pilar, MacKinnon Lachlan M.

School of Mathematical and Computer Sciences, Heriot-Watt University,
Edinburgh, EH14 4AS
pilar.lachlan@macs.hw.ac.uk

Abstract. We propose the development of a Data Quality Manager to establish communication between the process of integration of information, the user and the application, to deal with data inconsistencies and information integration.

Introduction

Data integration is the process of extracting and merging data from multiple heterogeneous sources to be loaded into an integrated information resource. The use of domain ontology, metadata, transformation rules, user, and system constraints have resolved the majority of the problems of domain mismatch associated with schematic integration and global schematic approaches. However, even when all the mappings, semantic and structure heterogeneity are solved in the global schema, consistency may not have been achieved, because the information provided by the sources may be mutually inconsistent. At the same time, each autonomous component database deals with its own quality properties on information, such as accuracy, reliability, availability, timeliness and cost of information access.

State of the Art

Several approaches for inconsistency between databases have been implemented:

- By reconciliation of data, also known as data fusion: different values become just one using a fusion function, depending on the data semantic [3].
- On the basis of individual data properties: associated with each information source (i.e. accuracy, completeness) [1].

A number of approaches for data integration using quality aspects have been addressed:

- Data integration techniques based on data quality aspects [2] with data quality information stored in metadata.
- Information Quality reasoning [4] is defined as the integration of information quality aspects, to the process of planning and optimizing queries against databases and information systems. Such aspects are related through the establishment of information quality criteria, assessment methods and measure.

Proposal

We intend to develop a Data Quality Manager, which will contain a *Reference Model* a *Measurement Model* and an *Assessment Model* to define the quality criteria, the metrics and the assessment methods. The Data Quality Manager will establish the basis for taking decisions during the identification of information sources in heterogeneous systems, such that:

- Based on the Reference Model, to classify the data sources (see Figure 1).
- The use of quality aspects stored in a quality metadata (QMD) as a whole with the user priorities for the selection of the best sources of information (see Figure 2).

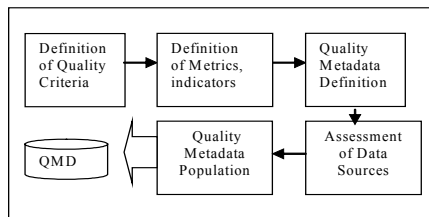


Fig. 1. Data Quality Manager Components.

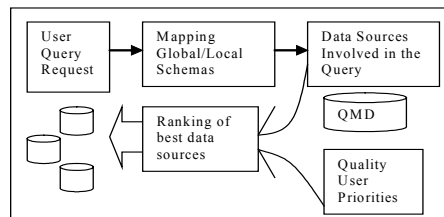


Fig. 2. Selection of best data sources.

- The query planning, considering data quality estimations to find the best combination for the execution plan (see Figure 3).
- After the query execution, and detection of inconsistent data, data quality might be used to perform data fusion to resolve data inconsistencies.
- Integration of the information and ranking the query result with the quality criteria estimated by the user (see Figure 4).

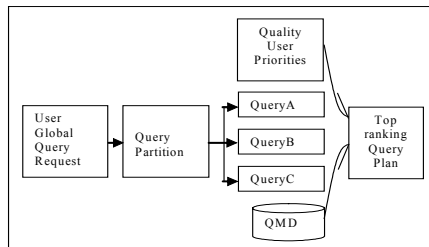


Fig. 3. Query Planning.

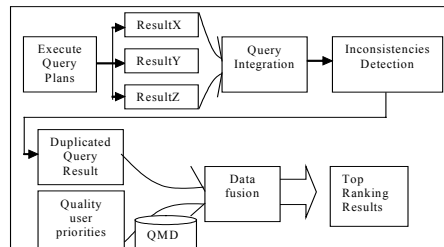


Fig. 4. Ranking Query Results.

References

1. Anokhin P., Motro A.: "Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources" George Mason University, (2003)
2. Gertz M.: "Managing Data Quality and Integrity in Federated Databases" Second Annual Working Conference on Integrity and Internal Control in I. S. Virginia, Kluwer Academic Publishers (1998)
3. Motro A., Rakov I. "Estimating the Quality of Databases" Proceedings of FQAS 98: Third International Conference on Flexible Query Answering Systems (T. Andreasen, H. Christiansen, and H.L. Larsen, Editors), Springer-Verlag, Berlin, Germany, (1998) 298-307.
4. Naumann F. "Quality-Driven Query Answering for Integrated Information Systems" Lecture Notes in Computer Sciences LNCS 2261, Springer-Verlag, Berlin Heidelberg (2002) .