Chapter 2.11
A Methodology Supporting the Design and Evaluating the Final Quality of Data Warehouses

Maurizio Pighin
University of Udine, Italy

Lucio Ieronutti
University of Udine, Italy

ABSTRACT

The design and configuration of a data warehouse can be difficult tasks especially in the case of very large databases and in the presence of redundant information. In particular, the choice of which attributes have to be considered as dimensions and measures can be not trivial and it can heavily influence the effectiveness of the final system. In this article, we propose a methodology targeted at supporting the design and deriving information on the total quality of the final data warehouse. We tested our proposal on three real-word commercial ERP databases.

INTRODUCTION AND MOTIVATION

Information systems allow companies and organizations to collect a large number of transactional data. Starting from this data, datawarehousing provides architectures and tools to derive information at a level of abstraction suitable for supporting decision processes.

There are different factors influencing the effectiveness of a data warehouse and the quality of related decisions. For example, while the selection of good-quality operational data enable to better target the decision process in the presence of alternative choices (Chengalur-Smith, Ballou, & Pazer, 1999), poor-quality data cause information scrap and rework that wastes people, money, materials and facilities resources (Ballau,

Wang, Pazer, & Tayi, 1998; English, 1999; Wang & Strong, 1996a, 1996b).

We have recently started at facing the problem of data quality in data warehouses (Pighin & Ieronutti, 2007); at the beginning of our research, we have considered the semantics-based solutions that have been proposed in the literature, and then we moved towards statistical methods, since in a real-world scenario data warehouse-engineers typically have a partial knowledge and vision of a specific operational database (e.g., how an organization really uses the operational system) and related semantics and then they need a support for the selection of data required to build a data warehouse. We then propose a context-independent methodology that is able both to support the expert during the data warehouse creation and evaluate the final quality of taken design choices. The proposed solution is mainly focused on statistical and syntactical aspects of data rather on semantics and it is based on a set of metrics, each one designed with the aim of capturing a particular data feature.

However, since most design choices are based on semantic considerations, our goal is to propose a solution that can be coupled with semantics-based techniques (for instance the one proposed by Golfarelli, Maio, and Rizzi (1998)) to effectively drive design choices. In particular, our methodology results effective in the following situations:

- During the construction phase, it is able to drive the selection of an attribute in the case of multiple choices (i.e., redundant information); for example, when an attribute belongs to different tables of a given database or belongs to different databases (that is the typical scenario in these kind of applications). Additionally, it is able to evaluate the quality of each choice (i.e., the informative value added to the final data warehouse choosing a table and its attribute as measure or dimension).
- At the end of the data warehouse design, it measures in quantitative terms the final quality of the data warehouse. Moreover, in the case of data warehouses based on the same design choices (characterized by the same schema), our methodology is also able to evaluate how data really stored into the initial database influences the informative content of the resulting data warehouse.

To evaluate the effectiveness of our methodology in identifying attributes that are more suitable to be used as dimensions and measures, we have experimented proposed metrics on three real ERP (Enterprise Resource Planning) commercial systems. Two systems are based on a DB Informix running on Unix server and one is based on a DB Oracle running on Windows server. In the experiment, they are called respectively *DB01*, *DB02* and *DB03*. More specifically, our metrics have been tested on data collected by the selling subsystems.

In this article, we refer to measures and dimensions related to the data warehouse, and to metrics as the indexes defined in the methodology we propose for evaluating data quality and reliability. Moreover, we use DW and DB to identify respectively a decisional data warehouse and an operational database.

RELATED WORK

In the literature, different researchers have been focused on data quality in operational systems and a number of different definitions and methodologies have been proposed, each one characterized by different quality metrics. Although Wang (1996a) and Redman (1996) proposed a wide number of metrics that have become the reference models for data quality in operational systems, in the literature most works refer only to a limited subset (e.g., *accuracy*, *completeness*, *consistency* and *timeliness*).