Chapter IV
Interactive Quality–Oriented Data Warehouse Development

Maurizio Pighin
IS&SE-Lab, University of Udine, Italy

Lucio Ieronutti
IS&SE-Lab, University of Udine, Italy

ABSTRACT

Data Warehouses are increasingly used by commercial organizations to extract, from a huge amount of transactional data, concise information useful for supporting decision processes. However, the task of designing a data warehouse and evaluating its effectiveness is not trivial, especially in the case of large databases and in presence of redundant information. The meaning and the quality of selected attributes heavily influence the data warehouse’s effectiveness and the quality of derived decisions. Our research is focused on interactive methodologies and techniques targeted at supporting the data warehouse design and evaluation by taking into account the quality of initial data. In this chapter we propose an approach for supporting the data warehouses development and refinement, providing practical examples and demonstrating the effectiveness of our solution. Our approach is mainly based on two phases: the first one is targeted at interactively guiding the attributes selection by providing quantitative information measuring different statistical and syntactical aspects of data, while the second phase, based on a set of 3D visualizations, gives the opportunity of run-time refining taken design choices according to data examination and analysis. For experimenting proposed solutions on real data, we have developed a tool, called ELDA (Evalutation DAta warehouse quality), that has been used for supporting the data warehouse design and evaluation.
INTRODUCTION

Data Warehouses are widely used by commercial organizations to extract from a huge amount of transactional data concise information useful for supporting decision processes. For example, organization managers greatly benefit from the availability of tools and techniques targeted at deriving information on sale trends and discovering unusual accounting movements. With respect to the entire amount of data stored into the initial database (or databases, hereinafter DBs), such analysis is centered on a limited subset of attributes (i.e., datawarehouse measures and dimensions). As a result, the datawarehouse (hereinafter DW) effectiveness and the quality of related decisions is strongly influenced by the semantics of selected attributes and the quality of initial data. For example, information on customers and suppliers as well as products ordered and sold are very meaningful from data analysis point of view due to their semantics. However, the availability of information measuring and representing different aspects of data can make easier the task of selecting DW attributes, especially in presence of multiple choices (i.e., redundant information) and in the case of DBs characterized by an high number of attributes, tables and relations. Quantitative measurements allow DW engineers to better focus their attention towards the attributes characterized by the most desirable features, while qualitative data representations enable one to interactively and intuitively examine the considered data subset, allowing one to reduce the time required for the DW design and evaluation.

Our research is focused on interactive methodologies and techniques aimed at supporting the DW design and evaluation by taking into account the quality of initial data. In this chapter we propose an approach supporting the DW development and refinement, providing practical examples demonstrating the effectiveness of our solution. Proposed methodology can be effectively used (i) during the DW construction phase for driving and interactively refining the attributes selection, and (ii) at the end of the design process, to evaluate the quality of taken DW design choices.

While most solutions that have been proposed in the literature for assessing data quality are related with semantics, our goal is to propose an interactive approach focused on statistical aspects of data. The approach is mainly composed by two phases: an analytical phase based on a set of metrics measuring different data features (quantitative information), and an exploration phase based on an innovative graphical representation of DW ipercubes that allows one to navigate intuitively through the information space to better examine the quality and distribution of data (qualitative information). The interaction is one of the most important feature of our approach: the designer can incrementally define the DW measures and dimensions and both quality measurements and data representations change according to such modifications. This solution allows one to evaluate rapidly and intuitively the effects of alternative design choices. For example, the designer can immediately discover that the inclusion of an attribute negatively influences the global DW quality. If the quantitative evaluation does not convince the designer, he can explore the DW ipercubes to better understand relations among data, data distributions and behaviors.

In a real world scenario, DW engineers greatly benefit from the possibility of obtaining concise and easy-to-understand information describing the data actually stored into the DB, since they typically have a partial knowledge and vision of a specific operational DB (e.g., how an organization really uses the commercial system). Indeed, different organizations can use the same system, but each DB instantiation stores data that can be different from the point of view of distribution, correctness and reliability (e.g., an organization never fills a particular field of the form). As a result, the same DW design choices can produce different informative effects depending on the data actually stored into the DB. Then, although
Related Content

Graph Transformations and Neural Networks
www.igi-global.com/chapter/graph-transformations-neural-networks/10655?camid=4v1a

On Clustering Techniques
www.igi-global.com/chapter/clustering-techniques/10831?camid=4v1a

Development of Data Warehouse Conceptual Models: Method Engineering Approach
www.igi-global.com/chapter/development-data-warehouse-conceptual-models/28159?camid=4v1a

Semantic Data Mining
www.igi-global.com/chapter/semantic-data-mining/10744?camid=4v1a