Enhancing Data Warehouse management through semi-automatic data integration and complex graph generation

Abstract

Strategic information is one of the main assets for many organizations and, in the next future, it will become increasingly more important to enable the decisionmakers answer questions about their business, such as how to increase their profitability. A proper decision-making process is benefited by information that is frequently scattered among several heterogeneous databases. Such databases may come from several organization systems and even from external sources. As a result, organization managers have to deal with the issue of integrating several databases from independent data sources containing semantic differences and no specific or canonical concept description.

Data Warehouse Systems were born to integrate such kind of heterogeneous data in order to be successively extracted and analyzed according to the manager's needs and business plans.

Besides being difficult and onerous to design, integrate and build, Data Warehouse Systems present another issue related to the difficulty to represent multidimensional information typical of the result of OLAP operations, such as aggregations on data cubes, extraction of sub-cubes or rotations of the data axis, through easy to understand views. In this thesis, I present a visual language based on a logic paradigm, named Complexity Design (CoDe) language, and propose a semi-automatic approach to support the manager in the automatic generation of a Data Warehouse answering to his/her specific requirements. In particular, the manager expresses his questions in natural language and selects the needed information among different data sources. The selected data are imported from the databases, a data tuning process is enacted, and an association rule mining algorithm is run to extract the conceptual model underneath the data. Successively, the CoDe models are provided and the Data Mart is generated from the models. Finally, the CoDe model and the DW data are used to generate a graphical report of the required information. To shows the information the manager needs to make his decision according to the strategic questions, the report adopts Complex Graphs. In this thesis, I also evaluate the effectiveness of the proposed approach, in terms of comprehensibility of the produced visual representation of the data extracted from the data warehouse. In particular, an empirical evaluation has been designed and conducted

to assess the comprehension of graphical representations to enable instantaneous and informed decisions. The study was conducted at the University of Salerno, Italy, and involved 47 participants from the Computer Science Master degree having management, information systems and advanced database systems and data warehouse) knowledge. Participants were asked to comprehend the semantic of data using traditional dashboard and complex graph diagrams.

The effort required to answer an evaluation questionnaire has been assessed together with the comprehension of the data representations and the outcome has been presented and discussed. The achieved results suggest that people reached a higher comprehension when using Complex Graphs like the ones produced by our approach as compared to standard Dashboard Graphs. Furthermore, the analysis of the time needed to comprehend the data semantic shows that the participants spent significantly less time to understand the representation adopting a Complex Graphs visualization as compared to the standard Dashboard Graphs. Both skilled and inexperienced users took advantages from the complex graph representation in term of comprehension and effort, with most skilled participants taking a greater advantage in comprehension time. This finding could be very relevant for the decision maker. Indeed, complex graphs represent an effective visualization approach that enables a quicker and higher comprehension of data, improving the appropriateness of the decisions and reducing the decision making effort.

Based on this result, an automatic generation of data warehouse has been presented that uses complex graphs to visualize the main facts that a decision maker should use for strategic decisions. In particular, the CoDe language has been adopter to represent such information.

When assessing the automatic generation of data warehouse, the most critical phase of the process has been the data integration, due to the need of an important contribution from the manager. So, to encourage the implementation of the process in a real setting, a semi-automatic data integration process has been proposed and tested on 2 case studies to assess the goodness of the approach. The results, indicate that, despite some limitations, in one of the two case studies we obtained encouraging results.