Automatic Integration Issues of Tabular Data for On-Line Analysis Processing

Yuzhao Yang*, Jérôme Darmont**, Franck Ravat*, Olivier Teste*

* Institut de Recherche en Informatique de Toulouse -IRIT CNRS (UMR 5505)-
Université de Toulouse, 118, Route de Narbonne, 31069 Toulouse Cedex 9, France
{Yuzhao.Yang, Franck.Ravat, Olivier.Teste}@irit.fr,

**ERIC UR 3083, Université de Lyon, Lyon 2
5 avenue Pierre Mendès France, F69676 Bron Cedex, France
Jerome.Darmont@univ-lyon2.fr

Abstract. Companies and individuals produce numerous tabular data. The objective of this position paper is to draw up the challenges posed by the automatic integration of data in the form of tables so that they can be cross-analyzed. We provide a first automatic solution for the integration of such tabular data to allow On-Line Analysis Processing. To fulfill this task, features of tabular data is analyzed and the challenges of automatic multidimensional schema generation should be addressed. Hence, we propose a typology of tabular data and a automatic process based on different steps to integrate one or more data sources.

1 Introduction

Business Intelligence (BI) plays an important role in numerous companies and administrations to efficiently support decision making processes. With the current digitization trend, even small companies and organizations can exploit a large number of data every day and the rise of open data make various data even more accessible. Nevertheless, the implementation of a BI project needs to be realized by people who have the professional knowledge and deep skills in BI technologies such as data warehousing and data visualization. Such projects are also usually expensive and time-consuming. As a result, it is necessary to find a solution to automate the BI process to allow small enterprises, organizations and even individuals without deep technical expertise to easily analyze data. Up to now, there is no platform that achieves this goal.

In current BI systems, data are extracted and stored in a data warehouse to allow On-Line Analysis Processing (OLAP) and visual data rendering. Thus, automating the data warehousing process is crucial to allow non-specialist to exploit such approaches. There exist various forms of data, but most of the data in small enterprises and organizations, as well as most of the open data are in tabular form from spreadsheet software. Although there are commercial BI tools allowing the exploitation of tabular data such as Excel, Qlikview or Tableau, none of them automates the multidimensional analysis of tabular data.