Chapter 10
An Extraction, Transformation, and Loading Tool Applied to a Fuzzy Data Mining System

Ramón A. Carrasco
University of Granada, Spain

Pedro Villar
University of Granada, Spain

Miguel J. Hornos
University of Granada, Spain

María A. Aguilar
University of Granada, Spain

ABSTRACT

In this chapter, we address the problem of integrating semantically heterogeneous data (including data expressed in natural language), which are collected from various questionnaires published in different websites, into a Data Warehouse. We present an extension of the sentences and architecture of data mining Fuzzy Structured Query Language as an extraction, transformation, and loading tool to integrate semantically heterogeneous data from these websites. Moreover, we show a case study using the questionnaires (carried out during several years) about the courses on Information and Communication Technologies which are taught in the Business Studies implanted at the University of Granada (Spain). With this integrated information, the Data Warehouse user can make several analyses with the benefit of an easy linguistic interpretability. The solution proposed here can be used to similar integration problems.

INTRODUCTION

A *Data Warehouse* (DW) is defined as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process" (Inmon, 2005). Data is extracted from the sources and then loaded into the DW using various data loaders and *Extraction, Transformation* and *Loading* (ETL) tools. We can define *Data Mining* (DM) as the process of extracting interesting information from the data stored in databases. According to (Frawley et al, 1991), a discovered knowledge is interesting when it is novel, potentially useful and non-trivial to compute. A series of new functionalities there exists in DM, which reaffirms that it is an independent area (Frawley et al, 1991): high-level language on the

DOI: 10.4018/978-1-4666-2455-9.ch010

discovered knowledge and for showing the results of the user's information requests (e.g. queries); efficiency on large amounts of data; handling of different types of data; etc. There is a symbiotic relationship between the activity of DM and the DW. The DW sets the stage for effective DM. DM can be done where there is no DW, but the DW greatly improves the chances of success in DM (Wang, 2009; Inmon, 1996). The *World Wide Web* (WWW) has become an important resource of information for the DM process. Consequently, the integration of the WWW information into a DW is important in order to get a more effective DM.

One of the most complex issues about the integration and transformation interface is the case where there are multiple sources for a single element of data in the DW. Our proposal is to integrate semantically heterogeneous data from various websites with opinions about educational issues in order to obtain a more effective DM on this information. Similar integration problems have already been solved in various platforms of the so-called Web 2.0, where people are encouraged to post reviews or express their opinions on several subjects, such as: education (PlanetRate, 2010), tourism (Booking.com, 2010; eDreams.com, 2010; TripAdvisor.com, 2010), etc., using numerical values and/or natural language (forums, news groups, etc.). The general approach of these websites is to compute only the accurate numerical information given by users in order to provide a ranking value (e.g. see Figure 1). However, the opinions expressed by the users in natural language are an important source of information. Therefore, the overall problem is the integration of information collected in these questionnaires which are available on various websites and formats, including also linguistic information.

Many aspects of different activities in the real world cannot be assessed in a quantitative form, but rather in a qualitative one (i.e., with vague or imprecise knowledge). In these cases, a better approach may be to use linguistic assessments instead of numerical values. The fuzzy linguistic approach, which was introduced by (Zadeh, 1975), is a theory that facilitates the coding of human knowledge in the form of linguistic concepts, and proposes a tool for modelling qualitative information in a problem. Consequently, the fuzzy linguistic approach seems to be an appropriate framework for solving our problem.

There are some DM tools based on this concept (Galindo, 2008). One of these tools is dmFSQL (data mining Fuzzy Structured Query Language) (Carrasco et al, 2006), which integrates flexible queries, clustering, fuzzy classification techniques (Carrasco et al, 2002) and fuzzy global dependencies (GDs) (Carrasco et al, 2000). There is a dmFSQL server, programmed in PL/SQL, which allows us to use the language dmFSQL on DW implemented on Oracle© Databases. This enables us to evaluate the DM process at both theoretical and practical levels. This dmFSQL architecture has been satisfactorily used in many problems, such as: finances, marketing, tourism, information retrieval, decision-making, RFID systems, etc. The WWW has become an important resource of information in such applications. However, the ETL process to convert the original information stored in websites to the dmFSQL architecture has been developed ad-hoc, depending on the particular problem. In fact, the standardization

Figure 1. Example of rating on education in http://www.planetrate.com/category/education