

## Chapter 3

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

**Ramón A. Carrasco**

*University of Granada, Spain*

**Miguel J. Hornos**

*University of Granada, Spain*

**Pedro Villar**

*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.*

DOI: 10.4018/978-1-61350-038-5.ch003

## INTRODUCTION

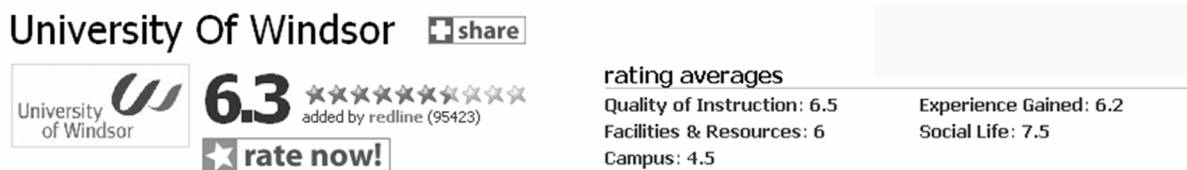
A *Data Warehouse* (DW) is defined as “a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management’s decision-making process” (Inmon, 2005). Data is extracted from the sources and then loaded into the DW using various data loaders and *Extraction, Transformation 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 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.

Figure 1. Example of rating on education in <http://www.planetrade.com/category/education>



22 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

[www.igi-global.com/chapter/extraction-transformation-loading-tool-applied/58411](http://www.igi-global.com/chapter/extraction-transformation-loading-tool-applied/58411)

## Related Content

---

### Effects of Data Envelopment Analysis on Performance Assessment: A Cognitive Approach

Heinz Ahnand Nadia Vazquez Novoa (2015). *International Journal of Business Analytics* (pp. 1-22).

[www.irma-international.org/article/effects-of-data-envelopment-analysis-on-performance-assessment/124179](http://www.irma-international.org/article/effects-of-data-envelopment-analysis-on-performance-assessment/124179)

### Mind the Gap: From Analytics to Action in Student Retention

Qing Huang, Nilupulee Nathawitharana, Kok-Leong Ong, Susan Kellerand Dammina Alahakoon (2019).

*Applying Business Intelligence Initiatives in Healthcare and Organizational Settings* (pp. 218-236).

[www.irma-international.org/chapter/mind-the-gap/208098](http://www.irma-international.org/chapter/mind-the-gap/208098)

### Data Mining for Health Care Professionals: MBA Course Projects Resulting in Hospital Improvements

Alan Olinskyand Phyllis A. Schumacher (2012). *Organizational Applications of Business Intelligence Management: Emerging Trends* (pp. 133-143).

[www.irma-international.org/chapter/data-mining-health-care-professionals/63971](http://www.irma-international.org/chapter/data-mining-health-care-professionals/63971)

### Applying Soft Computing to Clinical Decision Support

José Machado, Lucas Oliveira, Luís Barreiro, Serafim Pintoand Ana Coimbra (2016). *Applying Business Intelligence to Clinical and Healthcare Organizations* (pp. 256-271).

[www.irma-international.org/chapter/applying-soft-computing-to-clinical-decision-support/146072](http://www.irma-international.org/chapter/applying-soft-computing-to-clinical-decision-support/146072)

### Applications of Data Mining in Software Development Life Cycle: A Literature Survey and Classification

Naveen Dahiya, Vishal Bhatnagar, Manjeet Singhand Neeti Sangwan (2016). *Business Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 558-570).

[www.irma-international.org/chapter/applications-of-data-mining-in-software-development-life-cycle/142638](http://www.irma-international.org/chapter/applications-of-data-mining-in-software-development-life-cycle/142638)