Chapter 42
Heterogeneous Text and Numerical Data Mining with Possible Applications in Business and Financial Sectors

Farid Bourennani
University of Ontario Institute of Technology, Canada

Shahryar Rahnamayan
University of Ontario Institute of Technology, Canada

ABSTRACT

Nowadays, many world-wide universities, research centers, and companies share their own data electronically. Naturally, these data are from heterogeneous types such as text, numerical data, multimedia, and others. From user side, this data should be accessed in a uniform manner, which implies a unified approach for representing and processing data. Furthermore, unified processing of the heterogeneous data types can lead to richer semantic results. In this chapter, we present a unified pre-processing approach that leads to generation of richer semantics of qualitative and quantitative data.

INTRODUCTION

There is much interest from the industry in Heterogeneous Data Mining (HDM). This interest seems to be proportional to the heterogeneity level of the data, i.e., the more heterogeneous the data types are, the greater interest is in automatic data processing. This interest is likely due to the availability and abundance of data which potentially offers richer information when they are combined. However, especially due to Internet expansion, information access is virtually unlimited. Therefore, the amount of available data can exceed human data processing capacity in a reasonable time. That is why there is much interest in *automatic* extracting of richer patterns and coincident

DOI: 10.4018/978-1-4666-2455-9.ch042

classification and clustering, of the combined heterogeneous data types. For example, business intelligence sectors and financial institutions are extremely eager to extract coincident clustering results from textual (e.g., business reports) and numerical (especially financial) data based on the content (Bourennani et al., 2009). In a similar manner, many other sectors are attracted by the HDM with the intent of extracting richer patterns from the processing of two or even more combined data types. In brief, the HDM is of importance in many industrial fields. However, it is complex to extract coincident classification results based on the content of heterogeneous data types, likely because of the difficulty of combining the results (Bourennani et al., 2009b). The problem is that, most of the times, non-overlapping research communities work on mining *homogeneous* data types. Therefore, these dissimilar data types are usually pre-processed or represented using completely different techniques. Nevertheless, from a user point of view, these heterogeneous data types should behave and reflect information in a similar way. To respond to this need, several research groups have been working, in the last couple of years, to solve this challenging HDM problem.

In this chapter, different perspectives for mining heterogeneous data types are reviewed. Particularly, by focusing on the pre-processing phase, it is shown how similar representations of heterogeneous data types generate more convergent clustering results (Bourennani et al., 2009c). Self-Organizing Maps are used as clustering methods in our experiments.

LITTERATURE REVIEW ON HETEROGENEOUS DATA TYPES MINING

Based on our best knowledge, Back et al.. (2001) were the first researchers to start working on the HDM in 2001. Their project focused on the HDM of texts and numerical data for benchmarking activities. The same researchers worked on the HDM for financial and business report data (Ecklund et al., 2001), (Kloptchenko et al., 2004), (Magnussona et al., 2005). The reason for using data mining in those kinds of projects is that the tremendous amount of available financial data simply exceeds the interest of the managers and investors to analyze the data (Adriaans and Zantinge, 1996). Furthermore, "the purpose of benchmarking is to compare the activities of one company to those of another, using quantitative or qualitative measures, in order to discover ways in which effectiveness could be increased" (Ecklund et al., 2001). In these works, the qualitative data are actually from the companies' respective CEOs reports, Business reports depending on the project. The quantitative data are nine financial ratios, namely, Return on Total Assets (ROTA), Return on Equity (ROE), and others. The number of companies was from 3 to 76 depending on the projects. The Self-Organizing Map (SOM) which is an unsupervised clustering algorithm was used for processing the heterogeneous data types. The selection of SOM is judicious because it permits the clustering of the data without knowing the expected number of clusters prior to the data mining operations. In addition, the SOM's trained map facilitates the visual exploration of the clusters and the data relationships. The SOM was successfully applied to purely quantitative data (homogenous data) (Back et al., 2001); however, the clustering results were *divergent* when SOM was applied to heterogeneous textual and numerical data. A couple of inappropriate configurations contributed to divergent clustering results.

Firstly in these works, the heterogeneous textual and numerical data types were mined separately using SOM, then they were combined (their respective heterogeneous trained maps) to form a unified map which was divergent. Actually, matching of the SOM maps is very complex because even with the same input data, the results can be different after every training. It would be more appropriated, in our opinion, to process the