Chapter V
Ontology–Based Interpretation and Validation of Mined Knowledge:
Normative and Cognitive Factors in Data Mining

Ana Isabel Canhoto
Henley Management College, UK

ABSTRACT

The use of automated systems to collect, process, and analyze vast amounts of data is now integral to the operations of many corporations and government agencies, in particular it has gained recognition as a strategic tool in the war on crime. Data mining, the technology behind such analysis, has its origins in quantitative sciences. Yet, analysts face important issues of a cognitive nature both in terms of the input for the data mining effort, and in terms of the analysis of the output. Domain knowledge and bias information influence, which patterns in the data are deemed as useful and, ultimately, valid. This chapter addresses the role of cognition and context in the interpretation and validation of mined knowledge. We propose the use of ontology charts and norm specifications to map how varying levels of access to information and exposure to specific social norms lead to divergent views of mined knowledge.

INTRODUCTION

Data mining has been described as a process where '*the interrogation of the data is done by the data mining algorithm rather than by the user. Data mining is a self organizing, data influenced (...) approach to data analysis. Simply put, what data mining does is sort through masses of data*

to uncover patterns and relationships, then build models to predict behaviour' (Chan & Lewis, 2002). This description suggests that the use of data mining techniques tend to minimize the influence that analysts have in the process. Yet, as many practitioners would point out, the reality is very different: even though data mining is a largely quantitative and automated process, the analyst, and hence subjectivity, still plays a crucial role in several steps. Far from being a straightforward and objective process, data mining requires the use of '*intuition and creativity as well as statistical know-how, and you have to hope you have identified the right things to test'* (Humby et al., 2003).

This chapter looks at the role of the analyst in interpreting and validating the results of a data mining exercise. It suggests that, while the data mining community has long identified ways in which the analyst influences the data mining exercise, it has dedicated little attention, so far, to the understanding of the reasons why, and the mechanism how, this happens. As a result, the same literature has limited prescriptive and corrective value for subjectivity in data mining. We respond to the call for further research into the cognitive aspects of data mining (e.g., Chung & Gray, 1999; Kohavi et al., 2002; Pazzani, 2000), by proposing a framework to capture the cognitive and contextual elements shaping the process.

Nowadays, numerous organizations routinely capture and mine customer records to develop profiles of who their users are and what they do in order to inform future decision making. The areas of application range from improving service or performance, to analyzing and detecting terrorist activities (Hosein, 2005). Technologies, such as RFID and mobile technology, are likely to augment the mass of information that must be coped with, and accelerate the extension of profiling to ever more areas of social life. Therefore, the technical process of development of a profile—data mining—is a problem domain of central significance bound up in the growing

debate on the role of profiling in the information society (Hildebrandt & Backhouse, 2005). This chapter contributes to an understanding of how the process, despite being highly automated, is still likely to be affected by cognitive and contextual elements, which may fundamentally affect the effectiveness of the exercise and its outcomes. In the next section, the author presents, in broad terms, the ongoing discussion regarding the nature of data mining, as well as to what extent the analyst may interfere in the process. Then, two theories that specifically deal with subjectivity in the interpretation and classification of various stimuli—classification theory and semiotics—are presented, compared, and contrasted. It is noted that these two theories complement each other in the sense that classification theory analyzes the cognitive process, whereas semiotics analyzes the contextual factors affecting that same process. Following from this, the author proposes a framework to capture the effect of affordances and social norms in shaping the cognitive process of the data mining analyst. The framework is applied to a short case study, and suggestions are given regarding areas for further research into the role of subjectivity in data mining.

BACKGROUND

The processing of data in a data mining exercise includes several steps ranging from data selection and preparation to the interpretation of the emerging results. The input to the data mining process is a collection of data objects organized in a database, and the actual data interrogation process will usually start with the specification of the problem domain and an understanding of the goals of the project. The following stage comprises an assessment of the existing knowledge, as well as of the data that needs to be collected. The target dataset resulting from this stage is treated and, later, interrogated in order to *dig pieces of knowledge from the database* (Bruha, 2000). The final

Related Content

Implementation and Testing Details of Document Classification
 (2021). *Developing a Keyword Extractor and Document Classifier: Emerging Research and Opportunities* *(pp. 159-169).*
www.irma-international.org/chapter/implementation-and-testing-details-of-document-classification/268469

Microblog Sentiment Analysis Using User Similarity and Interaction-Based Social Relations
Chuanmin Mi, Xiaoyan Ruanand Lin Xiao (2022). *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines (pp. 1887-1904).*
www.irma-international.org/chapter/microblog-sentiment-analysis-using-user-similarity-and-interaction-based-social-relations/308581

User-Centric Similarity and Proximity Measures for Spatial Personalization
Yanwu Yang, Christophe Claramunt, Marie-Aude Aufaureand Wensheng Zhang (2010). *International Journal of Data Warehousing and Mining (pp. 59-78).*
www.irma-international.org/article/user-centric-similarity-proximity-measures/42152

Inventory Control and Big Data
Meghna Sharmaand Niharika Garg (2016). *Big Data: Concepts, Methodologies, Tools, and Applications* *(pp. 1543-1554).*
www.irma-international.org/chapter/inventory-control-and-big-data/150229

Case Study: Efficient Faculty Recruitment Using Genetic Algorithm
Amit Verma, Iqbaldeep Kaur, Dolly Sharmaand Inderjeet Singh (2019). *Extracting Knowledge From Opinion Mining (pp. 299-311).*
www.irma-international.org/chapter/case-study/211566