Chapter XIV
An Interpretation Process for Clustering Analysis Based on the Ontology of Language

Wagner Francisco Castilho  
Federal University of Rio Grande do Sul (UFRGS), Brazil  
Federal Savings Bank (CEF), Brazil

Gentil José de Lucena Filho  
Catholic University of Brasília, Brazil

Hércules Antonio do Prado  
Catholic University of Brasília, Brazil  
Embrapa Food Technology, Brazil

Edilson Ferneda  
Catholic University of Brasília, Brazil

ABSTRACT

Clustering analysis (CA) techniques consist of, given a set of objects, estimating dense regions of points separated by sparse regions, according to the dimensions that describe these objects. Independently from the data nature, structured or nonstructured, we look for homogenous clouds of points that define clusters from which we want to extract some meaning. In other words, when doing CA, the analyst is searching for underlying structures in a multidimensional space for what one could assign some meaning. To carry a CA application, two main activities are involved: generating clusters configurations by means of an algorithm and interpreting these configurations in order to approximate a solution that could contribute with the CA application objective. Generating a clusters configuration is typically a computational task, while the interpretation task brings a strong burden of subjectivity. Many approaches are presented in the literature for generating clusters configuration. Unfortunately, the interpretation task has not received so much attention, possibly due to the difficulty in modeling something that is subjective in nature. In this chapter a method to guide the interpretation of a clusters configuration is proposed. The inherent subjectivity is approached directly by describing the process with the apparatus of the Ontology of Language. The main contribution of this chapter is to provide a sound conceptual basis to guide the analyst in extracting meaning from the patterns found in a set of data, no matter whether we are talking about data bases, a set of free texts, or a set of Web pages.

Copyright © 2008, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION

According to Fayyad, Piatetsky-Shapiro, and Smyth (1996), knowledge discovery in databases (KDD) is “the non-trivial process of identifying valid, new, potentially useful and ultimately understandable patterns in data.” In other words, KDD refers to the general process of discovering useful knowledge from data. Each stage of the KDD process is complex and involves decisions by analysts and specialists in that domain. The analysts’ previous knowledge heavily influences the KDD process, which involves a series of decisions at each stage, including interpretation and evaluation of results. These decisions are also influenced by the database used, the domain, and the anticipated use of the acquired knowledge. The analysts’ decisions also involve preexisting knowledge structures associated with their experience and knowledge of the domains and database.

All processes of knowledge generation taken from data and information cannot ignore the very phenomenon of interpretation done by human beings that correspond to the result of assessments made by a group of individuals associated with the process. KDD cannot disregard the effect of previous knowledge that permeates the entire process within the perspective of those involved.

Besides the assessments and decisions taken by those involved in the KDD process, it is of special importance to consider the manner these people communicate in the elaboration of these assessments and interpretations and how they coordinate themselves in making decisions and executing actions and procedures. This has a decisive influence on the planning cycle, execution, and evaluation of the results from knowledge creation process.

Regarding the interpretations and communication involved in the knowledge creation process, there are interesting contributions developed based on the philosophy of a language model known as the ontology of language.

This chapter presents the contributions that ontology of language can offer to an interpretation model of KDD results, specifically in clustering analysis (CA). To this end, some fundamentals of the Ontology of Language will be presented in the next section. Later, the CA model will be discussed, taking into consideration the decision making process and the knowledge creation from clusters, its stages and the action coordination cycle, and the methodology of result interpretation in CA. It is also discusses the challenge to aggregate more semantics to the CA process, specifically the influence of previous knowledge and the communication phenomenon that permeates the knowledge creation process.

Usually, KDD refers to databases organized in relational structures. For the sake of simplicity and considering that, for our aim, it does not matter if the data are structured or not, we consider in this chapter a broader definition of KDD that includes the knowledge discovery from nonstructured data, including free texts and Web pages. By doing so, we are emphasizing the generality of our method without losing a clear comprehension of the interpretation task.

ONTOMETRY OF LANGUAGE

Within classical philosophy, ontology is the branch of metaphysics dedicated to the study of being while. Metaphysics, according to Aristotle, means beyond the physical, or rather, beyond what we can perceive with our senses.

Echeverría (1997, pp. 27-28, 37-45) uses a new, nonmetaphysical perspective of the term ontology to support his use based on the tradition established by the German philosopher Martin Heidegger. In this tradition, ontology encompasses the idea he denominates as dasein, which refers to the particular way of existing as human beings. So, ontology refers to the generic comprehension or the interpretation of what it means to be human. This comprehension is the cornerstone of all that
Related Content

Utilization of Data Mining Techniques to Detect and Predict Accounting Fraud: A Comparison of Neural Networks and Discriminant Analysis
[www.igi-global.com/chapter/utilization-data-mining-techniques-detect/25765?camid=4v1a](www.igi-global.com/chapter/utilization-data-mining-techniques-detect/25765?camid=4v1a)

Data Field for Hierarchical Clustering
Shuliang Wang, Wenyan Gan, Deyi Li and Deren Li (2011). International Journal of Data Warehousing and Mining (pp. 43-63).
[www.igi-global.com/article/data-field-hierarchical-clustering/58637?camid=4v1a](www.igi-global.com/article/data-field-hierarchical-clustering/58637?camid=4v1a)

Influence of Domain and Model Properties on the Reliability Estimates' Performance
[www.igi-global.com/article/influence-domain-model-properties-reliability/37405?camid=4v1a](www.igi-global.com/article/influence-domain-model-properties-reliability/37405?camid=4v1a)

A Solution to the Cross-Selling Problem of PAKDD-2007: Ensemble Model of TreeNet and Logistic Regression
Mingjun Wei, Lei Chai, Renying Wei and Wang Huo (2008). International Journal of Data Warehousing and Mining (pp. 9-14).
[www.igi-global.com/article/solution-cross-selling-problem-pakdd/1802?camid=4v1a](www.igi-global.com/article/solution-cross-selling-problem-pakdd/1802?camid=4v1a)