Chapter 18
The Use of Ontology for Data Mining with Incomplete Data

Hai Wang
Saint Mary’s University, Canada

Shouhong Wang
University of Massachusetts Dartmouth, USA

ABSTRACT
Ontology has recently received considerable attention. Based on a domain analysis of knowledge representations in data mining, this chapter presents a structure of ontology for data mining as well as the unique resources for data mining with incomplete data. This chapter demonstrates the effectiveness of ontology for data mining with incomplete data through an experiment.

INTRODUCTION
Knowledge management has become one of the important topics in the database management field (Zhang and Zhao 2006). There have been two important research themes in knowledge management: data mining (Cunningham et al. 2006; Fayyad et al. 1996) and ontology (Green and Rosemann 2004; Kim 2002). Data mining is the process of trawling through data to find interesting patterns (Hand 1998). As such a process reveals previously unknown relationships among the data, data mining has become a widely used knowledge discovery technique (Brachman et al. 1996). On the other hand, ontology is a science that studies explicit formal specifications of the resources and relations among them in the domain (Gruber 1993). An ontology is a specification of a conceptualization (Gruber 1995), and intended for knowledge sharing among applications (Welty 2003). In the past few years, the two themes have become well-recognized substrate for research into knowledge management (Nilakanta et al. 2006; Li and Zhong 2006). Yet, potential benefits of joining the two themes have not been explored.

Intuitively, the use of ontology for data mining can be beneficial for knowledge management in the following aspects.

1. To share common understanding of the context of data mining among data miners. For example, given a set of marketing survey...
The Use of Ontology for Data Mining with Incomplete Data

Figure 1. Primitive Ontology

Ontology for Data Mining

Ontology

According to Resource Description Framework (RDF) (W3C 2007), a primitive ontology consists of a pair of resource objects and a relational linkage between them. It is formalized as shown in Figure 1. A large ontology for an entire domain is a composition of a set of primitive ontology.

Resources in ontology are knowledge representations, including data, procedures, rules, ideas that guide actions and decisions (Beckman 1999; Alter 1996; Tobin 1996; van der Spek and Spijkervet 1997). In this study, an ontology is a network of all these resources that shows the paths of data mining actions for the data miner to achieve a certain goal.

Categories of Resources of Data Mining

An ontology for the domain is usually large. To make a large ontology manageable to the developer and user, the entire ontology must be partitioned into parts. The partition is done through categorizing resources and identifying their relations pertinent to the domain. Taxonomy of formalized generic resource categories can help people to better understand and share the ontology. Based on the available limited literature on ontology associated with data mining (e.g., (Bernstein et al. 2005; Kim 2002; Li and Zhong 2006; Welty 2003), and their references) we propose generic resource categories for the domain of data mining as follows.

Clearly, applications of ontology to data mining can be promising for effective knowledge management. However, little research on this issue has been reported in the literature. In this paper, we first discuss the key knowledge elements of data mining, and propose a generic structure of ontology for the domain. We then place the emphasis on ontology development for novel data mining with incomplete data. Through a project of ontology-based data mining system, we demonstrate the effectiveness of ontology in data mining.