Chapter XII
A Framework for Integrating Ontologies and Pattern-Bases

Evangelos Kotsifakos
University of Piraeus, Greece

Gerasimos Marketos
University of Piraeus, Greece

Yannis Theodoridis
University of Piraeus, Greece

ABSTRACT

Pattern base management systems (PBMS) have been introduced as an effective way to manage the high volume of patterns available nowadays. PBMS provide pattern management functionality in the same way where a database management system provides data management functionality. However, not all the extracted patterns are interesting; some are trivial and insignificant because they do not make sense according to the domain knowledge. Thus, in order to automate the pattern evaluation process, we need to incorporate the domain knowledge in it. We propose the integration of PBMS and ontologies as a solution to the need of many scientific fields for efficient extraction of useful information from large databases and the exploitation of knowledge. In this chapter, we describe the potentiality of this integration and the issues that should be considered introducing an XML-based PBMS. We use a case study of data mining over scientific (seismological) data to illustrate the proposed PBMS and ontology integrated environment.

INTRODUCTION

In the *knowledge discovery from data* (KDD) process, data mining techniques are used to find patterns from a large collection of data (see data mining step in Figure 1). The role of the domain experts in this process is crucial. Their knowledge is used in early stages to prepare data (i.e., to decide for the data cleaning and preparation) and to choose the appropriate parameters for the

data mining algorithms. Their contribution is also necessary for the evaluation and interpretation of the extracted patterns that lead to the generation of knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

In essence, extracted patterns are used from domain experts to explore new relations on data, evaluate theories on the field of interest, and discover unknown and hidden knowledge that will lead to new experiments and theories. However, some of the extracted patterns are considered trivial and some others insignificant, according to the domain knowledge. To evaluate extracted patterns experts have defined a lot of different, either objective or subjective interestingness measures based mostly on statistical properties of the patterns. Nevertheless, analyzing and assessing the usefulness of discovered patterns is a laborious task and is considered a hard problem (Piatetsky-Shapiro, 2000).

Two important issues are raised. The first refers to the *manipulation and management of the patterns in a unified way*, either they have been evaluated or not. Currently, the majority of the available data mining tools support the visualization of patterns, and in the best case storage in relational tables. Combined with the characterization of patterns as complex, compact, and rich in semantics representation of data (Rizzi et al., 2003), this issue raises the challenge for pattern management. In this context, we propose an XML-based pattern base management system (PBMS) for representing, storing, querying, indexing, and updating patterns. This system is data mining engine independent, supporting interoperability and exchange of patterns between different pattern bases.

The second issue is related to the *incorporation of the existing domain knowledge in the data mining process*, and especially in the pattern evaluation phase. Several statistical and interestingness measures have been proposed for the evaluation of patterns (Freitas, 1999; Piatetsky-Shapiro, 1991; Piatetsky-Shapiro, & Matheus, 1994; Silberschatz & Tuzhilin, 1996). These measures are applied either before or during the data mining process. In the first case, they are used to reduce the number of patterns that will be extracted and to speed up the data mining process. While in the evaluation phase, they are used to clean up the patterns considered insignificant.

Nevertheless, no such measure for pattern evaluation is efficient enough as the domain expertise itself. Domain experts can better evaluate the patterns and decide whether they are trivial or not. It is the user who will distinguish interesting rare occurrences of patterns from statistical noise using his/her background knowledge (Pohle, 2003). In order to automate the pattern evaluation process, we need to incorporate the domain knowledge in it. It is generally acceptable that domain knowledge can be represented efficiently using ontologies (Pohle, 2003). An *ontology* is a specification of a conceptualization, a description of the concepts and relationships that can exist for an agent or a community of agents (Gruber, 1993).

We argue that domain knowledge expressed with ontologies could function as a filter in the

Figure 1. The KDD process