Chapter 4.8
An XML-Based Database for Knowledge Discovery:
Definition and Implementation

Rosa Meo
Università di Torino, Italy

Giuseppe Psaila
Università di Bergamo, Italy

ABSTRACT

Inductive databases have been proposed as general purpose databases to support the KDD process. Unfortunately, the heterogeneity of the discovered patterns and of the different conceptual tools used to extract them from source data make the integration in a unique framework difficult. In this chapter, we explore the feasibility of using XML as the unifying framework for inductive databases, and propose a new model, XML for data mining (XDM). We show the basic features of the model, based on the concepts of data item (source data and patterns) and statement (used to manage data and derive patterns). We make use of XML namespaces (to allow the effective coexistence and extensibility of data mining operators) and of XML-schema, by means of which we can define the schema, the state and the integrity constraints of an inductive database.

INTRODUCTION

Data mining applications are called to extract descriptive patterns, typically used for decision making, from the data contained in traditional databases and recently also from other unconventional information systems such as the web.

Examples of these applications are the *market basket analysis*, that extracts patterns such as association rules between purchased items, sequential patterns (that extract temporal descriptions between observed events), classification, clustering and link analysis as in Quinlan (1993) and Agrawal, Imielinski, and Swami (1993)

(that provide, in other words, user profiles, text mining, graph mining, and so on). Furthermore, these patterns can be used to give an explanation of the patterns themselves. In this case the data patterns are considered as data to be analysed (and not necessarily with the same analysis tool that was used to obtain them).

Inductive databases have been launched in Imielinski and Mannila (1996) as general-purpose databases in which both the data and the patterns can be represented, retrieved, and manipulated with the goal to assist the deployment of the *knowledge discovery process* (KDD). Thus, KDD becomes a querying sequence in a query language designed for a specific data mining problem (Boulicaut, Klemettinen, & Mannila, 1998). Consequently, an inductive database should integrate several heterogeneous data mining tools that deal with very different heterogeneous and complex data models. For example, source raw data may be represented as flat tables, or, nowadays, by loosely structured documents containing data coming from the Web as well. Also, the conceptual models are different: classification tools usually adopt a data model that is a classification tree, while basket analysis usually represents patterns by means of set enumeration models.

In this chapter, we propose a semi-structured data model specifically designed for inductive databases and, more generally, for *knowledge discovery systems*. This model is called XDM (XML for data mining). It is based on XML and is devised to cope with several distinctive features at the same time (Bray, Paoli, & Sperberg-McQueen, 1997).

- At first, it is semi-structured, in order to be able to represent an a-priori infinite set of data models.
- Second, it is based on two simple and clear concepts, named *Data Item* and *Statement*: a data item is a container of data and/or patterns; a statement is a description of an operator application.
- Third, with XDM the inductive database state is defined as the collection of data items and statements, and the knowledge discovery process is represented as a set of relationships between data items and statements.
- Fourth, it provides a definition of the database schema by means of the set of integrity constraints over inputs and outputs for operators. Moreover, it constitutes the meta-data of the KDD process (i.e., in terms of the kind of data produced by the operators). The database schema was obtained with the aid of XML-schema, which makes possible to define constraints that must hold on some specific data items or operators, thus ensuring a certain level of correctness of data and patterns. XML-schema specifications constrain the structure of XML documents and overcome the limitations of classical XML DTDs, by adding the concept of data type for attributes. Refer to Thompson, Beech, Maloney, and Mendelson (2001) and Biron and Malhotra (2001) for detailed descriptions on XML-schema.

The above discussed features of the model set the foundations to achieve operator interoperability within a unique framework (provided that the various operators' API are XML compliant). Finally, the adoption of XML as syntactic format provides several benefits; in particular, the concept of *namespace* opens the way to the integration of several data formats and operators inside the same framework (Bray, Hollander, & Layman, 1999).

XDM provides several interesting features for inductive databases:

- At first, source raw data and patterns are represented at the same time in the model.
- Second, the pattern derivation process is stored in the database: this is determinant for the phase of pattern interpretation and