Query-Initiated Discovery of Interesting Association Rules

Jongpil Yoon\textsuperscript{1} and Larry Kerschberg\textsuperscript{2}

\textsuperscript{1} Sookmyung W. University, Computer Science Department
Seoul, 140-742, Korea
jyoon@sookmyung.ac.kr
\textsuperscript{2} George Mason University
Information and Software Systems Engineering Dept.
Fairfax, VA 22030-4444, USA
kersch@gmu.edu

Abstract. The approach presented in this paper is to discover association rules based on a user’s query. Of the many issues in rule discovery, relevancy, interestingness, and supportiveness of association rules are considered in this paper. For a given user query, a database can be partitioned into three views: a positively-related-query view, a negatively-related-query view, and an unrelated-query view.

We present a methodology for data mining and rule discovery that incorporates pattern extraction from the three types of query views with pattern spanning to enlarge the scope of a pattern and its derived association rule. The rule discovery process involves several interrelated steps: 1) pattern extraction from both positively- and/or negatively-related query views, 2) pattern association across attributes to enhance the semantics of patterns, while performing 3) pattern spanning within an attribute domain to enhance the supportiveness of the resulting pattern.

The contributions of the paper includes the specification and development of the data mining method and associated tool that combines the operations of association and spanning on query views to derive semantically interesting patterns. These patterns can then be used in decision making because the patterns were mined from the user’s original hypothesis as expressed by the query.

1 Introduction

A number of studies [1,4,8,9,10,13] have concentrated on developing methods of rule discovery from databases. Although association rule discovery is becoming increasingly important, the discovered association patterns may not be semantically rich enough to assist users in understanding the application domain or in making decisions based on the discovered knowledge. It is partly because association patterns are not interesting or relevant to user’s intentions with respect to the application domain, and partly because they only take into account the item sets or attribute values, but not their frequency (or occurrence). In this paper, we propose a viable and efficient strategy for discovering interesting patterns in
large databases, and thus extend the work reported in previous research on this subject.

The interesting, relevancy, and supportiveness of patterns discovered from the databases are discussed in this paper.

- **Interestingness.** Interestingness as unexpectedness can be defined in general [5,11]. Discovered patterns are interesting if they are unexpected.
- **Relevancy.** So many patterns are discovered, but not all of them are related to a user’s concerns or needs. Some patterns, if discovered from a point of view corresponding to a users’ query, are related to the user’s concerns. A user query is a way of specifying a user’s intention to a database.
- **Supportiveness and Confidence.** Discovered patterns can be used for applications only when they exceed the given confidence thresholds. The higher the supportiveness of a discovered pattern, the more applicable it may be. The higher the confidence of a discovered pattern is, the more it can be asserted for that database.

With these three pattern features in mind, we describe correlations among them. Knowledge discovery and data mining in databases (KDD) aims to extract patterns that are not previously known. Patterns are of interest if they are unexpected and are also useful to users. They should be relevant to a user’s query as well. See the following example.

**EXAMPLE 1.1:** For example in a soil database, suppose that a farmer expects (or queries for) information about soil composition ratio amongst calcium, iron, and sodium as they relate to higher yield. A pattern as to what soil composition ratios produce higher yield is relevant to the user’s expectation (or an answer in typical database processing), but it may not be interesting because it is pre-expected (asked) or known. On the other hand, a pattern as to which customers want bigger fruits may be interesting because it is not expected. However, it is not at all relevant to the user’s query. Increasing the interestingness of a pattern may result in decreasing its relevancy, and vice versa. Our goal in this paper is then to extract patterns which are both related to and interesting to the user, based on the initial query. For the example, a pattern which characterizes those composition ratios that do not lead to higher yield is both related and interesting.

To achieve our goal, we classify a database into the three views according to a user’s query. Consider a query posed to a table in a database. The first view is the answer (query) from the table. The second view is the remaining of the tuples which are not in the answer. Finally the third view consists of the other tables in the database, not specifies in the query. These views are explained later in detail in Section 3. The key idea is that patterns extracted from the (first and) second view(s) are both interesting and relevant to the query issued by a user.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes background knowledge, and defines the relationships amongst the properties of relevancy, interestingness, and supportiveness of patterns. Section 4 describes pattern spanning for a higher support. Section 5 describes our method for mining patterns by extracting, associating, and spanning