Interesting Knowledge Patterns in Databases

Rajesh Natarajan
Indian Institute of Management Lucknow (IIML), India

B. Shekar
Indian Institute of Management Bangalore (IIMB), India

INTRODUCTION

Knowledge management (KM) transforms a firm's knowledge-based resources into a source of competitive advantage. Knowledge creation, a KM process, deals with the conversion of tacit knowledge to explicit knowledge and moving knowledge from the individual level to the group, organizational, and interorganizational levels (Alavi & Leidner, 2001). Four modes—namely, socialization, externalization, combination, and internalization—create knowledge through the interaction and interplay between tacit and explicit knowledge. The "combination" mode consists of combining or reconfiguring disparate bodies of existing explicit knowledge (like documents) that lead to the production of new explicit knowledge (Choo, 1998). Transactional databases are a source of rich information about a firm's processes and its business environment. Knowledge Discovery in Databases (KDD), or data mining, aims at uncovering trends and patterns that would otherwise remain buried in a firm's operational databases. KDD is "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). KDD is a typical example of IT-enabled combination mode of knowledge creation (Alavi & Leidner, 2001).

An important issue in KDD concerns the glut of patterns generated by any knowledge discovery system. The sheer number of these patterns makes manual inspection infeasible. In addition, one cannot obtain a good overview of the domain. Most of the discovered patterns are uninteresting since they represent well-known domain facts. The two problems—namely, rule quality and rule quantity—are interdependent. Knowledge of a rule's quality can help in reducing the number of rules. End-users of data mining outputs are typically managers, hard pressed for time. Hence, the need for automated methods to identify interesting, relevant, and significant patterns. This article discusses the interestingness of KDD patterns. We use the association rule (AR) (Agrawal, Imielinski, & Swami, 1993) in a market-basket context as an example of a typical KDD pattern. However, the discussions are also applicable to patterns like classification rules.

BACKGROUND

The Rule Quantity Problem: Solution Perspectives

The rule quantity problem may be a result of the automated nature of many KDD methods, such as AR mining methods. In one study, Brin, Motwani, Ullman, and Tsur (1997) discovered 23,712 rules on mining a census database. Approaches to alleviate this problem aim at reducing the number of rules required for examination while preserving relevant information present in the original set. Redundancy reduction, rule templates, incorporation of additional constraints, ranking, grouping, and visualization are some of the techniques that address the rule quantity problem.

In AR mining, additional constraints in conjunction with support and confidence thresholds can reveal specific relationships between items. These constraints reduce the search space and bring out fewer, relevant, and focused rules. Rule templates (Klemettinen, Mannila, Ronkainen, Toivonen, & Verkamo, 1994) help in selecting interesting rules by allowing a user to pre-specify the structure of interesting and uninteresting class of rules in inclusive and restrictive templates, respectively. Rules matching an inclusive template are interesting. Such templates are typical post-processing filters. Constraint-based mining (Bayardo, Agrawal, & Gunopulos, 2000) embeds user-specified rule constraints in the mining process. These constraints eliminate any rule that can be simplified to yield a rule of equal or higher predictive ability. Association patterns like negative ARs (Savasere, Omiecinski, & Navathe, 1998; Subramanian, Ananthanarayana, & Narasimha Murty, 2003), cyclic ARs (Ozden, Sridhar, & Silberschatz, 1998), inter-transactional ARs (Lu, Feng, & Han, 2000), ratio rules (Korn, Labrinidis, Kotidis, &

Faloutos, 1998), and substitution rules (Teng, Hsieh, & Chen, 2002) bring out particular relationships between items. In the market-basket context, negative ARs reveal the set of items a customer is unlikely to purchase with another set. Cyclic association rules reveal purchases that display periodicity over time. Thus, imposition of additional constraints offers insight into the domain by discovering focused and tighter relationships. However, each method discovers a specific kind of behaviour. A large number of mined patterns might necessitate the use of other pruning methods. Except for rule templates, methods that enforce constraints are characterized by low user-involvement.

Redundancy reduction methods remove rules that do not convey new information. If many rules refer to the same feature of the data, then the most general rule may be retained. "Rule covers" (Toivonen, Klemettinen, Ronkainen, Hatonen, & Mannila, 1995) is a method that retains a subset of the original set of rules. This subset refers to all rows (in a relational database) that the original ruleset covered. Another strategy in AR mining (Zaki, 2000) is to determine a subset of frequently occurring closed item sets from their supersets. The magnitude of cardinality of the subset is several orders less than that of the superset. This implies fewer rules. This is done without any loss of information. Sometimes, one rule can be generated from another using a certain inference system. Retaining the basic rules may reduce the cardinality of the original rule set (Cristofor & Simovici, 2002). This process being reversible can generate the original ruleset if required. Care is taken to retain the information content of the basic unpruned set. Redundancy reduction methods may not provide a holistic picture if the size of the pruned ruleset is large. Further, the important issue of identification of interesting patterns is left unaddressed. For example, a method preserving generalizations might remove interesting exceptions.

Visualization techniques take advantage of the intuitive appeal of visual depiction that aids in easy understanding (Hilderman, Li, & Hamilton, 2002). Various features like use of graphs, colour, and charts help in improved visualization. Rules depicted in a visual form can be easily navigated to various levels of detail by iteratively and interactively changing the thresholds of rule parameters. The main drawback in visualization approaches is the difficulty of depicting a large rule/attribute space. In addition, understandability of visual depiction decreases drastically with increase in dimensions. Hence, a user might fail to detect an interesting phenomenon if it is inlaid in a crowd of mundane facts. However, for browsing a limited rule space, visualization techniques provide an intuitive overview of the domain.

A user might be able to get a good overview of the domain with a few general rules that describe its essentials. Mining generalized association rules using product/attribute taxonomies is one such approach (Srikant & Agrawal, 1995). If all items at lower levels of a product taxonomy exhibit the same relationship, then rules describing them may be replaced by a general rule that directly relates product categories. General Rules, Summaries, and Exceptions (GSE) patterns introduced by Liu, Hu, and Hsu (2000) is an approach to summarization. The general rules, along with summaries, convey an overview while exceptions point to cases differing from the general case. Another approach is to group rules on the basis of exogenous criteria such as economic assessment, profit margin, period of purchase, and so forth (Baesens, Viaene, & Vanthienen, 2000). Clustering techniques group "similar" rules (Gupta, Strehl, & Ghosh, 1999) by imposing a structure on them. Rules within each group can then be studied and evaluated based on this structure. Most of the techniques stated help in consolidating existing knowledge rather than identifying new/latent knowledge.

The Rule Quality Problem: Solution Perspectives

The "rule-quality" problem is a consequence of most of the discovered patterns referring to obvious and commonplace domain features. For example, Major and Mangano (1995) mined 529 rules from a hurricane database of which only 19 were found to be actually novel, useful, and relevant. The most common and obvious domain facts are easily discovered since they have strong presence in databases. In addition, such facts form a core component of the user's domain knowledge due to repeated observation and application. Examination of these patterns is a waste of time since they do not further a user's knowledge. Ranking rules based on their interestingness is one approach that may address the rule-quality problem.

INTERESTINGNESS MEASURES

Interestingness measures try to capture and quantify the amount of "interest" that any pattern is expected to evoke in a user. Interesting patterns are expected to arouse strong attention from users. "Interestingness," an elusive concept, has many facets that may be difficult to capture and operationalize. Some of them may be domain and user-dependent. In other cases, depending on the context, the same features may be domain and user-independent. Capturing all features of interesting-