XML-Enabled Association Analysis

Ling Feng
Tsinghua University, China

INTRODUCTION

The discovery of association rules from large amounts of structured or semi-structured data is an important data mining problem [Agrawal et al. 1993, Agrawal and Srikant 1994, Miyahara et al. 2001, Termier et al. 2002, Braga et al. 2002, Cong et al. 2002, Braga et al. 2003, Xiao et al. 2003, Maruyama and Uehara 2000, Wang and Liu 2000]. It has crucial applications in decision support and marketing strategy. The most prototypical application of association rules is market basket analysis using transaction databases from supermarkets. These databases contain sales transaction records, each of which details items bought by a customer in the transaction. Mining association rules is the process of discovering knowledge such as "*80% of customers who bought diapers also bought beer, and 35% of customers bought both diapers and beer"*, which can be expressed as "*diaper* $\Rightarrow$ *beer"* (35%, 80%), where 80% is the *confidence* level of the rule, and 35% is the *support* level of the rule indicating how frequently the customers bought both diapers and beer. In general, an association rule takes the form $X \Rightarrow Y$ (s, c), where X and Y are sets of items, and s and c are support and confidence, respectively.

In the XML Era, mining association rules is confronted with more challenges than in the traditional well-structured world due to the inherent flexibilities of XML in both structure and semantics [Feng and Dillon 2005]. First, XML data has a more complex hierarchical structure than a database record. Second, elements in XML data have contextual positions, which thus carry the order notion. Third, XML data appears to be much bigger than traditional data. To address these challenges, the classic association rule mining framework originating with transactional databases needs to be re-examined.

BACKGROUND

In the literature, there exist techniques proposed to mine frequent patterns from complex tress and graphs databases. One of the most popular approaches is to use graph matching, which employs data structures like adjacency matrix [Inokuchi et al. 2000] or adjacency list [Kuramochi and Karypis 2001]. Another approach represents semi-structured tree-like structures using a string representation, which is more space efficient and relatively easy for manipulation [Zaki 2002]. This work concentrated on mining frequent tree structures within a forest, which can be extended to for mining frequent tree structures in XML documents. [Zhang et al. 2004, Zhang et al. 2005] proposed a framework, called XAR-Miner, which is directly applicable to mining association rules from XML data. Raw data in the XML document are preprocessed to transform to either an Indexed Content Tree (IX-tree) or Multi-relational databases (Multi-DB), depending on the size of XML document and memory constraint of the system, for efficient data selection and association rule mining. Task-relevant concepts are generalized to produce generalized meta-patterns, based on which the large association rules that meet the support and confidence levels are generated. Recently, Confronted with huge volume of XML data, [Tan, et al. 2005] proposed to generate candidates by model-validating, so that there is no time wasted in deriving invalid candidates which will be discarded at later stages. The algorithm processes an XML document directly taking into account the values of the nodes present in the XML tree, so the frequent item-sets generated contain both node names and values in comparison to the TreeMiner approach, which only generates frequent tree structures. The experiments with both synthetic and real life data sets demonstrate the efficiency of this approach.

MAIN FOCUS

The Framework

Under the traditional association rule framework, the basic unit of data to look at is database *record*, and the construct unit of a discovered association rule is *item* which has an *atomic value*. These lead us to the following two questions: 1) *what is the counterpart of record* and 2) *what is the counterpart of item* in mining association relationships from XML data? [Feng et al. 2003, Feng and Dillon 2004]. This investigation focuses on rule detection from a collection of XML documents, which describe the same type of information (e.g., customer order, etc.). Hence, each of XML documents corresponds to a database record, and possesses a tree-like structure. Accordingly, we extend the notion of associated item to an XML fragment (i.e., tree), and build up associations among trees rather than simple-structured items of atomic values. For consistency, we call each such kind of trees a **tree-structured item** to distinguish it from the traditional counterpart item. With the above extended notions, we propose an **XML-enabled association rule framework**. From both structural and semantic aspects, XML-enabled association rules are more powerful and flexible than the traditional ones.

Definition 1 *Let T denote a set of trees (tree-structured items). An* **XML-enabled association rule** *is an implication of the form $X \Rightarrow Y$, which satisfies the following two conditions:*

1. $X \subset T$, $Y \subset T$, *and* $X \cap Y = \phi$;
2. *for* $\forall T', T' \in (X \cup Y)$, *there exists no such tree* T'' *where* T' *is a subtree of* T *and* T'.

Different from classical association rules where associated items are usually denoted using simple structured data from the domains of basic data types, the items in XML-enabled association rules can have a hierarchical tree structure, as indicated by the first clause of the definition. Here, it is worth pointing out that when each of the tree-structured items contains only one basic root node, the XML-enabled association rules will degrade to the traditional association rules. The second clause of the definition requires that in an XML-enabled association rule, no common sub-trees exist within any two item trees in order to avoid redundant expression.

Figure 1 illustrates some XML-enabled association rule examples. Thanks to XML, XML-enabled association rules are more powerful than traditional association rules in capturing and describing association relationships. Such enhanced capabilities can be reflected from both a structural as well as a semantic point of view:

- Association items have hierarchical tree structures, which are more natural, informative and understandable (e.g., Rule 1 & 2 in Figure 1).
- Associated items inherently carry the *order* notion, enabling a uniform description of association and sequence patterns within one mining framework (e.g., Rule 1 states the sequence of books to be ordered, i.e., "*Star War I*" proceeding "*Star War II* " on a customer's order).
- Associated items can further be constrained by their context positions, hierarchical levels, and weak/strong adhesion in the corresponding XML data to be mined. (e.g., Rule 1 indicates the contextual appearances of BOOKs on the order).
- Association relationships among structures and structured-values can also be captured and described (e.g., Rule 2 states that a student orders some flowers from a shop, and leaves detailed content of FLOWER element such as the kind of flowers and quantity, etc. aside).
- Auxiliary information which states the occurrence context of association relationships can be uniformly self-described in the mining framework (e.g., Rule 1 indicates that only male people have such as order pattern).

Similar to traditional association rules, we use support and confidence as two major measurements for XML-enabled association rules.

Definition 2 *Let D be a set of XML documents. The* **support** *and* **confidence** *of an XML-enabled association rule $X \Rightarrow Y$ are defined as follows:*

$$\text{support}(X \Rightarrow Y) = \frac{|D_{xy}|}{|\mathrm{D}|}, \quad \text{confidence}(X \Rightarrow Y) = \frac{|D_{xy}|}{|D_x|}$$

where $D_{xy} = \{doc | \forall T \in (X \cup Y)(T \in_{tree} doc)\}$, and $D_x = \{doc | \forall T \in X(T \in_{tree} doc)\}$.