Template Guided Association Rule Mining from XML Documents

Rahman AliMohammadzadeh
Database Research Group
Faculty of ECE,
School of Engineering
University of Tehran, Iran,
+98-912-5068928
r.mohammadzadeh@ece.ut.ac.ir

Sadegh Soltan
Database Research Group
Faculty of ECE,
School of Engineering
University of Tehran, Iran,
+98-912-3092377
s.soltan@ece.ut.ac.ir

Masoud Rahgozar
Control and Intelligent Processing
Center of Excellence,
Faculty of ECE, School of Engineering,
University of Tehran, Iran,
+98-21-82084304
rahgozar@ut.ac.ir

ABSTRACT

Compared with traditional association rule mining in the structured world (e.g. Relational Databases), mining from XML data is confronted with more challenges due to the inherent flexibilities of XML in both structure and semantics. The major challenges include 1) a more complicated hierarchical data structure; 2) an ordered data context; and 3) a much bigger size for each data element. In order to make XML-enabled association rule mining truly practical and computationally tractable, we propose a practical model for mining association rules from XML documents and demonstrate the usability and effectiveness of model through a set of experiments on real-life data.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Design, Experimentation, Performance

Keywords

XML, Data Mining, Association Rule Mining.

1. INTRODUCTION

Data mining is usually used to extract interesting knowledge from large amounts of data stored in databases or data warehouses. This knowledge can be represented in many different ways such as clusters, decision trees, decision rules, etc. Among them, association rules have been proved effective in discovering interesting relations in massive amounts of data.

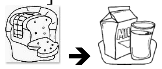
Currently, XML is penetrating virtually all areas of Internet application programming and is bringing about huge amount of data encoded in XML. With the continuous growth in XML data sources, the ability to extract knowledge from them for decision support becomes increasingly important and desirable [3]. Due to the inherent flexibilities of XML, in both structure and semantics, mining knowledge in the XML Era is faced with more challenges than in the traditional structured world.

In this paper, we propose a practical model for mining association rules from XML documents. Our model is based on XML-enabled association rule framework that was introduced by Feng [2]. XML-AR framework extends the notion of associated items to XML fragments to present associations among trees rather than simple-structured items of atomic values. Although this framework is flexible and powerful enough to represent simple and complex structured association rules in XML documents [4] but to our best knowledge no implementation model has been proposed yet.

Copyright is held by the author/owner(s).
WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

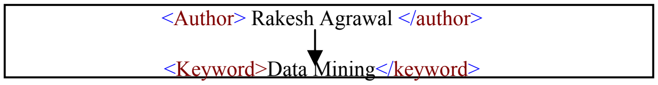
2. XML Association Rules

Association rules were first introduced by Agrawal et al. to analyze customer habits in retail databases. Association rule is an implication of the form $X \Rightarrow Y$, where the rule *body* X and *head* Y are subsets of the set I of *items* ($I = \{I1, I2, \ldots, In\}$) within a set of *transactions* D and $X \cap Y = \varphi$. A rule $X \Rightarrow Y$ states that the transactions T that contain the items in X are *likely* to contain also the items in Y. Association rules are characterized by two measures: the *support*, which measures the percentage of transactions in D that contain both items X and Y; the *confidence*, which measures the percentage of transactions in D containing the items X that also contain the items Y [Figure 1]. In XML context, both D and I are collections of trees [1], in the same way X and Y are XML fragments [Figure 2].



[Support = 2%, Confidence = 95%]

Figure 1. Association rule between bread and milk



<Author> Rakesh Agrawal </author>

<Keyword>Data Mining</keyword>

Figure 2. XML Association rule

3. XML Association Rule Mining (Practical Model)

We consider the problem of mining XML association rules from content [3] of XML documents based on user provided rule template. We suggest an implementation model for the XML-AR framework that was introduced by Feng [2]. Our practical model consists of 5 steps (see Figure 6): Filtering, Generating Virtual Transactions, Finding Association Rules, Converting extracted rules to XML AR rules and Visualizing.

Filtering and Generating virtual transactions are most important steps in this model so we describe these two steps in more details. Filtering step uses the XML-AR template and extracts only those parts of XML that are interesting for the user. In the next step, we define a transaction context, based on tag nesting in XML document and use it to generate virtual transactions that can be used as input format by association rule mining algorithms (e.g. Apriori). As an example, consider the problem of mining frequent associations among people who appear as coauthors, with our

XML-AR template we formulate this task by the following statement:

```
<dblp><*><author>? </author></*></dblp>➔
<dblp><*><author>? </author></*></dblp>
```

Figure 3. XML AR template for finding coauthors

Above statement has two parts (*body* and *head*) and each part has 3-level XML fragment and we are going to find patterns in 3rd level or author tag, so <author>? </author> is equal to items $i \in I$, in addition <author>? </author> occur in <*></*> so <*></*> is our transactions $t \in T$ and <dblp></dblp> is equal to database *D*. [Table 1] displays generated virtual transactions based on XML AR template in [Figure 3] and following XML fragment of DBLP collection:

```
<dblp>…
<incollection key="books/mit/fayyadPSU96/AgrawalMSTV96">
  <author>R. Agrawal</author>
  <author>H. Mannila</author>
  <author>R. Srikant</author>…
  <title>Fast Discovery of Association Rules.</title>
  <publisher>AAAI/MIT Press</publisher>
</incollection>…
<article key="journals/tkde/AgrawalS96">
  <author>R. Agrawal</author>
  <author>J. C. Shafer</author>
  <title>Parallel Mining of Association Rules.</title>
  <journal>TKDE</journal>
</article>….</dblp>
```

Figure 4. Sample XML document

Table 1. Virtual Transactions for coauthoring XML AR

TID	F1	F2	F3
1	R.Agrawal	H.Mannila	R.Srikant
2	R.Agrawal	J. C. Shafer	

4. Experimental Results

Two sets of experiments are performed on DBLP[1] collection. We implement prototype of our model by VS .Net 2005 (C#).

4.1 Experiment 1 – Finding Coauthors

In this experiment we extract the Co-Authoring relationship between different authors in DBLP data set. DBLP collection has 328858 bibliography of different publications (article, book, PhD thesis, etc) expressed in XML format. We use rule template introduced in [Figure 3] and input parameters in [Table 2], [Figure 5] and [Table 3] display some of results:

Table 2. Experiment 1 Parameters

Number of Records	Support	Confidence
328858	0.0001	60%

Table 3. Frequent itemsets of experiment 1

1-Itemset	2-Itemset	3-Itemset
2659	76	1

```
<dblp><*><author>Marco Conti</author></*></dblp>➔
<dblp><*><author>Enrico Gregori</author></*></dblp>
```

Figure 5. Sample XML AR Rule (Confidence=0.826)

4.2 Experiment 2 – keyword Relationship

Each bibliography in DBLP has a *key* attribute (see Figure 4) that includes general information like conference name, publication year, publication type, etc. This experiment extracts Co-Occurring relationship between *keys* in different publications. Rule template of this experiment was displayed in [Figure 7], input parameters, was displayed in [Table 4], [Table 5] contains number of extracted frequent itemsets and [Figure 8] visualizes some extracted rules.

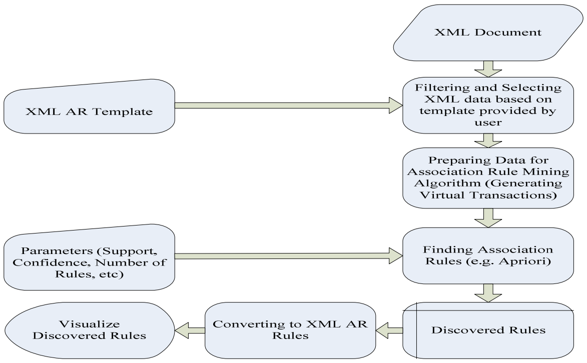

XML Document
Filtering and Selecting XML data based on template provided by user
XML AR Template
Preparing Data for Association Rule Mining Algorithm (Generating Virtual Transactions)
Parameters (Support, Confidence, Number of Rules, etc)
Finding Association Rules (e.g. Apriori)
Visualize Discovered Rules
Converting to XML AR Rules
Discovered Rules

Figure 6. AR Mining from XML documents (Practical Model)

```
<dblp><* key="?"></*></dblp>➔<dblp><* key="?"></*></dblp>
```

Figure 7. XML AR template for finding keyword relationship

Table 4. Experiment 2 Parameters

Number of Records	Support	Confidence
328858	0.0001	60%

Table 5. Frequent itemsets of experiment 2

1-Itemset	2-Itemset	3-Itemset
1250	1170	5

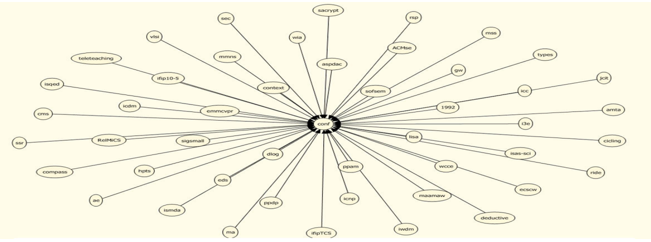


Figure 8- Dependency network between *keys* (center node is *Conf* and the other nodes are major conference names)

5. Conclusions

Main contribution of this paper is extending and implementing the XML-AR framework that was introduced by Feng [2]. In addition we suggested a practical model that can be used for mining association rules from XML documents. Our model uses XML-AR template for filtering data and generating virtual transactions so it can efficiently find the rules in which the user is mostly interested. Applying frequent sub-tree mining techniques and directly mining frequent XML fragments is our future work.

6. References

[1] Braga D., A. Campi, M. Klemettinen, and P. L. Lanzi. Mining association rules from XML data. In Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, September 4-6, Aixen-Provence, France 2002.

[2] Feng L. & T. Dillon. Mining XML-Enabled Association Rule with Templates. In Proceedings of KDID04, 2004.

[3] Nayak, R. Discovering Knowledge from XML Documents , in Wong, John, Eds. Encyclopedia of Data Warehousing and Mining. Idea Group Publications, 2005.

[4]. Tan, H., T.S. Dillon, L. Feng, E. Chang, F. Hadzic, "X3-Miner: Mining Patterns from XML Database," In Proc. Data Mining '05. Skiathos, Greece, 2005.

[1] It is available at http://dblp.uni-trier.de/XML/