ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Mining patterns in XML Data using X-Query and Vertical Data Regular Pattern Method

[I]Suneetha Merugula, [II]Jyothi Mandala
[I,II]Asst. Professor, Dept. of IT, GMRIT, Rajam, Andhra Pradesh, India

Abstract

In data mining the regularity of a pattern was treated as an important criterion in several online applications like market basket analysis, network monitoring, web page sequence and stock market. One of the simple methods to mine XML data is probably to transform the data from XML to relations. However, the drawback of this method is that transformation itself is usually complex and time-consuming. Therefore an alternative approach which can mine XML data by XQuery which is query language for XML satisfies the demand for intelligently querying XML data and hence can be used to mine XML data. This paper presents a framework for mining XML data using XQuery and vertical Data regular Pattern method using VDRP-table to generate the complete set of regular patterns in a transactional database for a user given regularity threshold.

Keywords

XQuery, XML, VDRP, Association Rule mining.

I. Introduction

EXtensible Markup Language (XML) has emerged as the main standard for describing data and exchanging data on the web. The ability to extract knowledge from XML data sources turned into a very important and necessary characteristic with the continuous growth in XML data. Data mining, appearing during the late 1980s, has improved during the 1990s especially in transforming vast amounts of data into useful knowledge, and is expected to continue to grow rapidly in the future. (Han and Kamber 2001). Nevertheless, compared to the successful performances in mining well-structured data such as relational databases and object-oriented databases, mining in the semi-structured XML world still remains at a preliminary stage and is confronted with more challenges due to the intrinsic characteristics of XML in both structure and semantics. XML data have a more complex hierarchical structure than a database record. According to these needs, the traditional data mining technology have to be regenerated and reformed for extracting knowledge from XML structure. The aim of XML mining is to integrate the emerging XML technology into data mining technology. Data mining may have simply three major components: Clustering, Classification, Link Analysis (Association Rule Mining or Sequence Analysis). Mining frequent patterns or itemsets is an essential problem in many data mining applications including association rules, correlations, sequential rules, episodes, multi-dimensional patterns and many other important discovery tasks .

Increasing use of XML technology for data storage and data exchange between applications, the subject of mining XML documents has become more researchable and important topic.

The principal purpose of this study is applying association rule algorithms directly to the XML documents with using XQuery which is a functional expression language that can be used to query or process XML data. The web is rich with information. However, the data contained in the web is not well organized which makes obtaining useful information from the web a difficult task. The successful development of eXtensible Markup Language (XML) [1] as a standard to represent semistructured data makes the data contained in the web more readable and the task of mining useful information from the web becomes feasible.Although tools for mining information from XML data are still in their infancy, they are startingto emerge.

The query language XQuery [2] was proposed by the W3C [3] and is currently in "last call" status.

The purpose of XQuery is to provide a flexible way to extract XML data and provide the necessary interaction between the web world and database world. XQuery is expected to become the standard query language for extracting XML data from XML documents.

The topic of mining XML data has received little attention, as the data mining community has focusedon the development of techniques for extracting common structure from heterogeneous XML data.The outline of this paper is as follows: we describe the XQuery implementation that is used to mine XML data ,in order to discover association rules we use the vertical data regular pattern mehod.

II. Literature Review

A. XML Query Languages

Ease of use and performance are the advantages of the XML query languages to programming languages. Most general-purpose programming languages treat XML as any other API, instead of as a first-class part of the language. Instead of providing operators for constructing and navigating XML directly, you have to access it through an API layer. Just as text manipulation is easier in Perl than in, say, Fortran, so a single line of an XML query language like XSLT or XQuery can accomplish the equivalent of hundreds of lines of C, C#, Java, or some other general-purpose language

B. XSLT

The Extensible Stylesheet Language for transformation is an official recommendation of the World Wide Web Consortium which published in 1999. It provides a flexible, powerful language for transforming XML files and uses XML syntax to define transformation rules that are applied to an input XML document to result in a text document that has not to be an XML document. This result can be an HTML document, another XML file, PDF, SVG, java code or a text file. An XSLT transformation is provided as an XML document called a stylesheet, or XSLT stylesheet. A stylesheet is applied to an input XML document, which means that the input XML document is transformed according to the stylesheet into an output document. A stylesheet can be looked at as a set of rewriting rules. These rules are called as a template rule. Each rewriting rule is equipped with a pattern and a body expression. When a stylesheet is applied to an input XML document, a rewriting rule is found, whose pattern matches the root document node of the input XML document. This rule's body expression defines a transformation for the root node, so the rule's body expression is evaluated to compute the output. Subset of XPath can be used to define patterns in template rules that are matched against nodes in an input XML document and full XPath 2.0 expressions can be used inside template's bodies (Hlousek 2005).

C. XPath

XPath is a language for addressing parts of an XML document which is a standard recommended by W3C. XPath defines a library of standard functions but is not itself written in XML because it defines how to locate parts of an XML document, forms the basis for a query language on XML, such as XSLT or Xquery. XPath models an XML document as a tree of nodes of which there are different types, including element nodes, attribute nodes and text nodes (World Wide Web Consortium 1999).XPath 1.0 has been designed to easily identify or match nodes in an XML document with the intention to be used either standalone or in XSLT 1.0 and XPointer.The result of evaluating an XPath 1.0 expression is either an atomic value or a set of nodes from the source XML document usually referred to as nodeset. As the most complicated structure of a result is a set, there is no ordering information about the items in the result set. The absence of order information in a result has been understood as a particular disadvantage. Specifically, information about document order of nodes is lost in a result. The main purpose of the XPath 2.0 is to address nodes in an XML document, which is the same as for the 1.0 version. The main difference between XPath 1.0 andXPath 2.0 is that an XPath 2.0 expression returns a sequence of items instead of a nodeset. Thus, the items in the returned sequence have now their order defined.

D. XQuery

The purpose of XQuery is extracting data from entire XML documents, collections of XML documents, or only document fragments. XQuery is derived from an XML query language called Quilt, which in turn borrowed features from several other languages, including XPath 1.0, XQL, XML-QL, SQL, and OQL. XQuery 1.0 is the superset of XPath 2.0 both in syntax and semantics. XQuery is a functional expression language that can be used to query or process XML data or any data that can be represented within the same model as XML. Being purely an expression language, XQuery programs are easier to understand and maintain than XSLT, because they do not include the complexities or management of templates (rule-based system) . This is especially true for highly structured data, and for longer programs. XQuery will still be able to effectively process semi-structured data. The query language is small and powerful. Moreover, XQuery is a full-fledged programming language. It provides if/then statements, loops, variables, quantified expressions and a set with the most important functions .Applications are made simpler by performing a single XQuery request over these views and receiving satisfactory results in one step also it has both an easy, human readable form and an XML representation .There are several operators provided to filter the documents for conditions which match either content or structure . The core of queries is a FLWOR expression. FLWOR is coming from 'for, let, where, order by, return'. With for and letXML fragments can be bound to a variable which then can be further processed in the

III. Methodology

In this paper we mine XML data using XQuery and Vertical Data Regular patterns method .The main use of VDRP method is to capture the database contents in full with one database scan to find regular patterns.

A. System Architecture

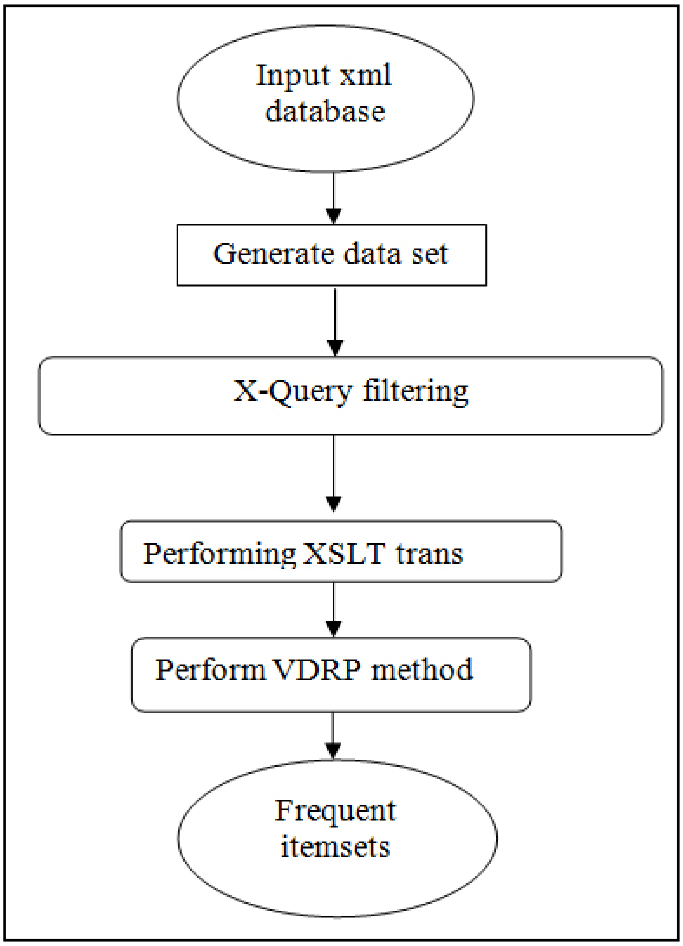

Fig.1: System Architecture.

B. Generate Dataset

During this phase the actual XML transaction data files are stored into transaction data base, in which each XML transaction data file is uniquely identified with transaction identifiers. The XML document structure is given:

```
 <transactions>
<transaction>
                    <items>
                          <item> abc</item>
                         <item> bred</item>
                          <item> jam</item>
                  </items>
  </transaction>
 <transaction>
                    <items>
                          <item> car</item>
                         <item> gloves</item>
                          <item> helmet</item>
                   </items>
  </transaction>
   </transactions>
```

Fig2.Input XML File

C. X-Query Filtering and Xslt Trans Demo

Select the generated XML document and write X Query "for $x

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

in (items) return <item>{$x}</item>" to filter the data. After this transform the XML data to csv (comma separated values) format and then store these values for finding frequent pattern

D. VDRP Method

```
VDRP – method:
Input    : DB, λ= 3
Output : Complete regular Patterns
Procedure :
Let $X_i \subset L$ be a k-itemset
$P^{X}i = 0$ for all $X_i$
For each $X_i$
        Find the next transaction $T_j$
        $P^{X}i = j - P^{X}i$
       Max_reg (R) = max($P^{X}i$ )
repeat
    If max_reg > λ
        Delete the itemset
Else
  $X_i$ is a regular item set
Increase the k value using "and" operation until no candidate is generated.
```

IV. Results

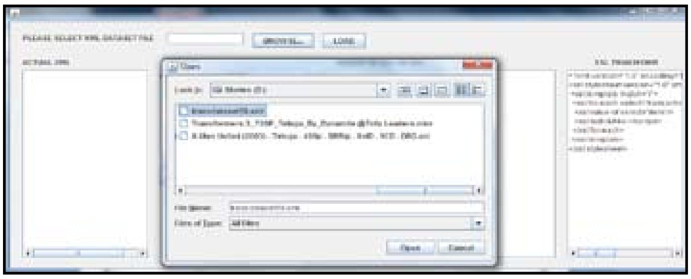

Fig. 3: select database from drive

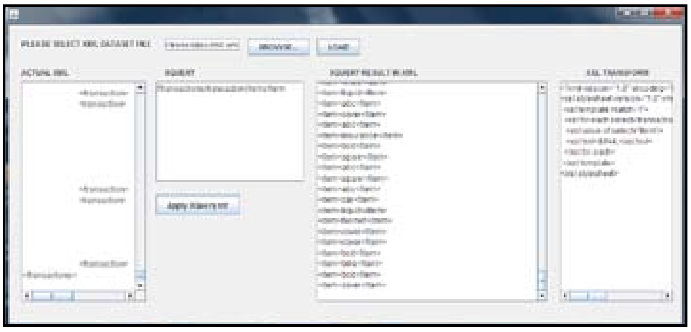

Fig. 4: Apply XQuery.

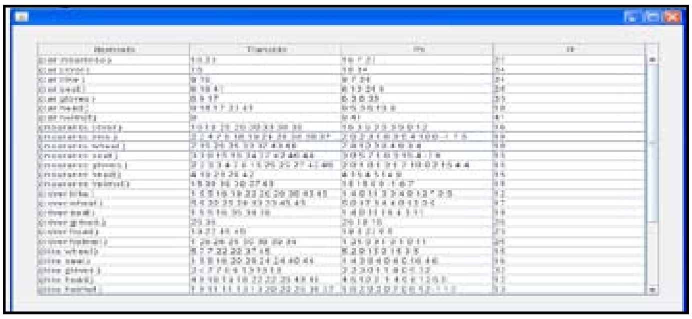

Fig. 5: Frequent Items

V. Conclusion

In this paper we presented a VDRP method. This method is better than the existing RP – tree algorithm because it utilizes the advantages of Vertical Transaction Database format and require only one database scan. This table (method) is efficient and scalable over large databases and faster than the RP-table. This method is very simple and works without complicated data structures. It needs only simple operations like union, intersection, subtraction etc.,When regular k-item set to generate regular (k+1)-item set, the mode of intersection of any two sets is used. Pruning is done first in this paper, namely finding max_reg (R). If R is greater than user-given regularity threshold ( ), we'll delete that itemsets.

References

[1] *World Wide Web Consortium. Extensible Markup Language (XML) 1.0 (Second Edition) W3C Recommendation. http://www.w3.org/XML.*

[2] *World Wide Web Consortium. XQuery 1.0: An XML Query Language (W3C Working Draft). http://www.w3.org/TR/2002/WDxquery- 20020816, Aug. 2002*

[3] *World Wide Web Consortium. http://www.w3.org.*

[4] *Han, J., Yin, Y. Yin, "Mining Frequent Patterns without candidate generation", In Proc. ACM SIGMOD international Conference on management of Data, PP. 1-12 (2000).*

[5] *Hiawei Han, Michelins Kamber, "Data Mining : Concepts and Techniques", 2nd ed. An Imprint of Elsevier, Morgan Kaufmann publishers, pp. 232-248, 2006.*

[6] *R. Agarwal, and R. Srikanth, "Fast algorithms for mining association rules", In Proc. 1994 Int. Conf. Very Large Databases (VLDBA"94), pages 487- 499, Santiago, Chile, Sept. 1994.*

[7] *S. K. Tanbeer, C. F. Ahmed, B.S. Jeong, and Y.K. Lee, "Mining Regular Patterns in Transactional Databases", IEICE Trans. On Information Systems, E91-D, 11, pp. 2568-2577, 2008.*

[8] *G. Yi-ming, W. Zhi-jun, "A Vertical format algorithm for mining frequent item sets", IEEE Transactions, pp. 11-13, 2010.*

[9] *Mohammed J. Zaki, karam Gouda. "Fast Vertical Mining using Diffsets", SIGKDD "03, August 24 - 27, 2003, Copyright 2003 ACM 1-58113-737-0/03/0008.*

M. Suneetha is presently working as Assistant Professor in Information Technology Department at GMRIT, Rajam. Pursuing her Ph.D from Acharya Nagarjuna University. She received her B.Tech degree in Computer Science and Information Technology from AITAM College, Tekkali, India and M.Tech degree in Software Engineering from JNTUH, Hyderabad, India. Her area of interest is Data mining. Email:sunita.merugula@gmail.com

M. Jyothi is presently working as Assistant professor in Information Technology Department at GMRIT, Rajam. Pursuing her Ph.D from Acharya Nagarjuna University. She received her B.Tech degree in Computer Science and Information Technology from AITAM College, Tekkali, India and M.Tech degree in Computer Science and Engineering from JNTUK, Kakinada, India. Her area of interest information security and data mining. Email:jyothirajb4u@gmail.com