ISSN: 2321-7782 (Online)

Volume 2, Issue 3, March 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Paper / Case Study

Available online at: www.ijarcsms.com

Intentional Knowledge to Queries on XML Document

Saumya P[1]
PG Student, Department of CSE
KMCT College of Engineering
Kerala – India

Maya Mathew[2]
Asst. Prof, Department of CSE
KMCT College of Engineering
Kerala – India

Abstract: The research in the field of database has caught its attention on Extensible Markup Language (XML) due to its ability to represent large amount of data in a flexible hierarchical nature without any compulsion on a fixed or absolute schema. Mining data from these document based on query-answering access method is a hard task. In this paper, we describe a new approach as Branch Oriented Rule (BOR) which mines intentional and summarised information about the content of the XML documents. This information is stored in XML format so as to be used for providing quick, approximate answers to queries in future.

Keywords: XML, query-answering, Branch Oriented Rule, intentional knowledge, approximate answers.

I. INTRODUCTION

Extensible Markup Language (XML) [1] has emerged as the dominant standard for describing and exchanging data. The nested and self-describing nature of XML provides a simple, flexible means of application to exchange data. Recently, XML has been used in virtually all areas of internet application programming, producing large amount of information encoded in XML. The ability to extract information from the XML data sources is a very prominent and necessary characteristic with the growth in XML data. One of the toughest problems of finding information from large XML dataset is the fast and concise retrieval of answers. Data mining is used to extract interesting knowledge from large amount of data stored in databases or data warehouses. This knowledge can be represented in various forms such as decision tree, decision rules, clusters etc. Among them association rules have proved to be an effective tool to discover interesting relations in large amount of data.

Intentional knowledge is a summarized representation of the original document. Thus it means that less space and time is required to store and to query it. The XML documents can have an implicit structure, that is, its structure may not have been declared in advance, for example via a DTD or an XML-Schema [2]. Querying such document is quite difficult for users for two main reasons: they are often confused with the huge amount of data available and they are not able to specify reasonable structure to the query conditions. Thus the problems of information overload [3] and information deprivation [4] arises. Hence it is convenient for the users if they know about the structure and the semantic characteristic of the document. For this purpose the researchers incorporated data mining technique called association rule mining [5]. Therefore mining of contents along with the structure of XML provides a new trend in knowledge discovery process [6].

The purpose of this paper is to present a method for mining intentional knowledge from XML datasets by using branch oriented rules. The mined rules are basically used to get view about the structure and the content of XML document and are also used for intentional query answering. The major advantage is that the mining process directly works on the original XML document without translating it into any intermediate format. And also the query is translated from original dataset to BORs set.

© 2014, IJARCSMS All Rights Reserved

II. Objectives

The method in this paper is used to derive summarized intentional knowledge in the form of rules from the xml document, and then store these rules as an alternative dataset to be queried for providing quick and summarized answers. The main objective of this paper is to create a process that allows us to extract branch oriented rules over the XML document which can be used later for query answering and that the process should directly work on the original XML document without converting to any other format. This paper addresses the problem to query over xml documents by mining patterns and knowledge, that is, frequent data. The rule information is stored in XML format making it easy to be handled in the future [7]. XQuery [8] can be used to give query on XML document. In this paper branch organization rules is proposed for representing frequent information in XML document [9]. The construction of indexes and patterns with related constraints can be used for query optimization. The intentional knowledge generated by branch oriented rules can provide answer to understand frequent patterns from data set; it can be used for formulation of query. BOR is not only used when quick answering is required by the user but can also be used when original document is lost, by using extracted information. So BOR provides abstract level of information from XML document [10].

III. Branch Oriented Rules

Data in a XML document is stored in a tree like structure. So mining data from such a document is much difficult than is the case with traditional database. For conventional databases, association rule mining was firstly proposed and it came to be used in XML mining later. Association rules are used to describe the co-occurrence of data in a large size of collected information and are represented in the form X =>Y implications, where X and Y are two arbitrary sets of data, such that $X \cap Y=\varnothing$. The quality of an association rule is measured by help of support and confidence. Support refers to the frequency of the set $X \cap Y$ in the dataset, while confidence corresponds to the conditional probability of finding Y, having found X and is given by supp $(X \cap Y)$/supp(X) [11].

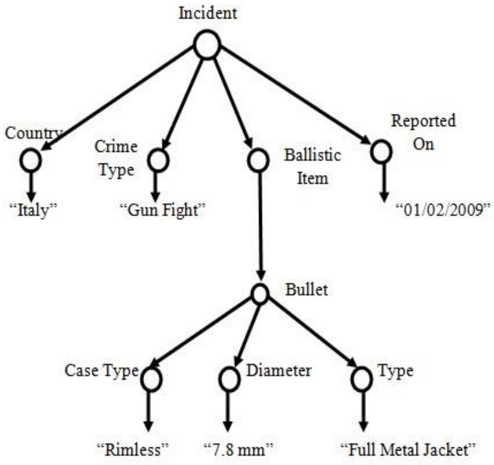

Fig 1. A Sample XML File: "Incident.xml"

The idea of mining association rules in order to provide summarized knowledge about the XML document had been studied either in proposals by using XQuery and techniques in XML context or by implementing graph or tree based algorithms. Fig 1 shows an example of an XML document incident.xml which consists of nodes like country, crime type, ballistic item used for the crime and date of reporting. The main steps involved in association rule mining are finding the frequent subtrees having support greater than a given threshold support from the XML document and finding the rules from the frequent subtrees which satisfies a minimum confidence. Algorithm 1 is used for finding frequent subtrees and they are handed over to the function that computes all possible rules.

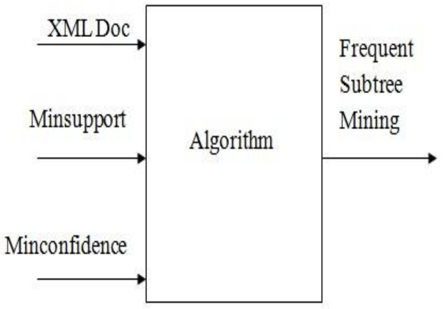

Fig 2. Frequent Sub tree Mining

Input for algorithm 1 is a XML document Q with threshold support i.e. minsupport and threshold confidence i.e. minconfidence. The output will be the frequent subtree which has the support and confidence greater than the predefined user given minsupport and minconfidence.

Algorithm 1:

```
Generate-Rules (Q, minsupport, minconfidence)
    $F_{filter}$=FilterFrequentSubtree(Q, minsupport)
    Rule_List=∅
    For all s ϵ $F_{filter}$ do
        Temporary_List=Get_Rules(s, minconfidence)
        Rule_List= Rule_List ∪ Temporary_List
    endFor
    return Rule_List
```

```
Get_Rules(s, minconfidence)
    Rule_List=∅
    Black_List=∅
    For all c, subtree of s do
        if c is not a subtree of any element in Black_List then
            confidence =supp(s)/supp(c)
            if confidence ≥ minconfidence
                NewRule=(c, s, confidence, supp(s))
                Rule_List= Rule_List ∪ {NewRule}
            else
                Black_List= Black_List ∪ c
            endif
        endif
    endfor
    return Rule_List
```

Branch Oriented Rules are generated by selecting an item having the support and confidence value above a predefined support and confidence value. Consider an example of rule generated as A=>B with support as 0.09 and confidence as 0.75, then it is said that if there is a node labeled A in the original document, with 75 percent probability that the node has a child or sibling labeled B. Each extracted rule is stored inside a tag element <rule> which consists of three attributes such as ID, support and confidence of the rule. Such extracted rules are stored into file with XML format called rule file, which can be used further as a source to get idea about original XML document. Answer to the given user query can be found from this file containing rules by matching the conditions specified in the query.

IV. Rule Indexing and updation

Branch Oriented Rule provides intentional answer to the user query which is more concise. It gives the properties which data frequently satisfies instead of describing the data in terms of its properties. To each path present in at least one rule, indexes are assigned. Index file containing set of references to each node in the rules is an XML document. XML documents that contain data may go on changing. In case of documents which frequently undergo changes, previously obtained rule and index files are updated instead of creating new rules.

Once rule files and index files are saved they are used as a source of data to be queried. User enters the query which is on original document and this query is translated into intentional query and then fired on extracted datasets. Due to this user get intentional answer to the queries and not the exact answers. These answers will give the property which is frequently satisfied and will be precise. References to the rules are obtained by matching the conditions in the query with the nodes in index file. The index file returns rule IDs matching the conditions in the query and only those rules are searched for answer. Thus answers are returned from mined knowledge not from original document and it is also useful in case the original document is lost. Fig 3. Shows the system architecture.

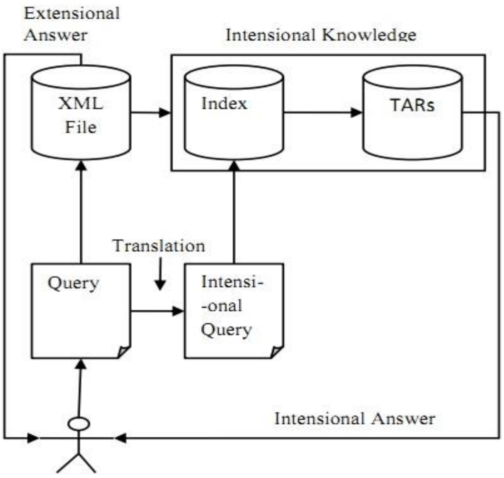

Fig 3. System Architecture

V. CONCLUSION

In this paper an XML query answering framework has been proposed which extracts branch oriented rules from a given XML file to support xml queries. The main goal is to mine all frequent association rules without the need to specify the structure and content of the rules. This process has been characterized by key aspects such as it works directly on the XML documents, without transforming the data into any other format. The rules extracted are stored in XML format so that by using these rules information can be extracted from XML dataset. In this paper the proposed framework takes XML document as input with predefined threshold support and confidence to generate BORs. Hence we can use mined knowledge to gain information effectively from the original datasets.

Acknowledgement

Miss. Saumya P and Miss. Maya Mathew wish to thanks International Journal of Advance Research in Computer Science and Management Studies. This work was supported in part by the department of Computer Science and Engineering from KMCT College of Engineering.

References

1. World Wide Web Consortium, Extensible Markup Language (XML) 1.0, http://www.w3C.org/TR/REC-xml/, 1998
2. S. Gasparini and E. Quintarelli, “Intensional query answering to xquery expressions”. In Proc. of the 16th Int. Conf. on Database and Expert Systems Applications, pages 544–553, 2005
3. Yongming Guo Dehua Chen Liangxu Liu, Jiajin Le ,“A Frame of Per Personalized Information Filtering System Based on XML”, Networked Computing and Information Management, Volume: 1, 2008. NCM '08.
4. R.Sree Lekshmi, B. Sasi kumar, “ Extracting Information from Semistructured XML using TARs“, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-1, October 2012.
5. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc. 20th Int’l Conf. Very Large Data Bases, pp. 478-499, 1994.

6. Mirjana Mazuran, Elisa Quintarelli, Letizia Tanca, “Data Mining for XML Query-Answering Support”, IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 8, August 2012.
7. Mirjana Mazuran, Elisa Quintarelli, and Letiza Tanca, “Mining Tree-Based Association Rules from XML Documents,” technical report, Politecnico di Milano, http://home.dei.polimi.it/quintare/Papers/MQT09-RR.pdf, 2009.
8. S. Gasparini and E. Quintarelli, “Intensional Query Answering to XQuery Expressions,” Proc. 16th Int’l Conf. Database and Expert Systems Applications, pp. 544-553, 2005
9. E. Baralis, P. Garza, Elisa Quintarelli, and Letiza Tanca, “Answering XML Queries by Means of Data Summaries,” ACM Trans. Information Systems, vol. 25, no. 3, p. 10, 2007.
10. J. Paik, H.Y. Youn, and U.M. Kim, “A New Method for Mining Association Rules from a Collection of XML Documents,” Proc. Int’l Conf. Computational Science and Its Applications, pp. 936-945, 2005
11. D. Braga, A. Campi, S. Ceri, M. Klemettinen, and P. Lanzi, “Discovering Interesting Information in XML Data with Association Rules,” Proc. ACM Symp. Applied Computing, pp. 450-454, 2003

AUTHOR(S) PROFILE

Miss. Saumya received her B-tech degree in Information Technology from AWH Engineering College, Calicut, Kerala, India and is presently doing her final year Master of Technology from KMCT College of Engineering Calicut, Calicut University, Kerala, India. Her research interest includes Data Mining and Knowledge Engineering.

Miss. Maya Mathew received her M.Tech in Computer Science and Engineering from New Horizon College of Engineering, Bangalore and is presently working as an Assistant Professor in Department of Computer Science and Engineering, KMCT College of Engineering, Calicut, India.Her area of interest includes Computer Networking and Cloud Computing.