Journal of Theoretical and Applied Information Technology
15th November 2012. Vol. 45 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NEW ONTOLOGY BASED ASSOCIATION RULES MINING ALGORITHM

[1]PENG ZHU, [2]FEI JIA

[1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China

[2]Division of Education Affairs, Nanjing Forest Police College, Nanjing 210094, Jiangsu, China

ABSTRACT

For traditional data mining techniques cannot be directly applied to the semi-structured XML data mining problem, this paper proposes a novel ontology and association rules based XML mining algorithm. The algorithm firstly introduces the domain ontology and hash technology to improve the operation of emerging frequent item sets and generating association rules, then uses a hash table to store the domain ontology, and at last the algorithm transforms the operation of the database into memory tree based on XML. Simulation results show that the algorithm can effectively reduce the size of XML documents and the association rules is easier to understand, so the advantages of the algorithm are shown.

Keywords: *Ontology, Data Mining, XML, Association Rules*

1. INTRODUCTION

Data mining also called Knowledge Discovery in Database-KDD [1], which has been used in more and more areas as well achieve good results. Currently, data mining has become an international leading-edge research field.

Web technology has changed the way of people's publish, access and use information dramatically since its emergence in 1990s. Especially in recent years, the emergence of a new generation of Web environment based on XML [2], could compatible with the existing Web applications well, and can better achieve information sharing and exchange in Web. The convenience and semi-structured characteristics of its text-based make XML has been widely applied in many fields, like, in information management, electronic commerce, personalized publishing, mobile communication, online education, exchange of electronic documents, and is still expanding its range of applications. XML has become the de facto standard of Data representation and exchange on the Internet [3]. For these more and more data using XML document format to store, exchange and performance, there is a growing need for further, in-depth knowledge except the existing information extraction, Web search and information processing methods. This means that data mining is necessary to be done.

However, precisely because XML is a kind of semi-structured text data, it has many weaknesses as text documents and semi-structured data, like, analysis the document must read it in the order form, the visit is inefficient; irregular organize the information, or may be constantly changing its structure or even may be incomplete, etc. The traditional data mining technology is the main face to the structured data-based relational database, transaction database and data warehouse [4]. Thus we cannot apply the traditional relational data-based mining methods, such as Apriori, to the semi-structured data mining directly. Therefore, to develop effective methods for XML data mining become an important issue in the field of data mining and XML technology research areas.

2. RELATED WORK

Today, Internet has become an indispensable tool for everyone, Web usage mining correspondingly becomes a hotspot, which uses large amounts of data in the Web server log and other relevant data sets for mining analysis and gains valuable knowledge model about usage of relevant Web site. Nagi et al. [5] describe a work utilizes association rules mining integrated with fuzziness factor in order to analyze weblog data. The target is to find pages that are accessed together by majority of the users and hence should be linked in a proper way in order to maximize user satisfaction by providing to the users the access flow they expect. This way the number of visitors to the analyzed website will be maximized and hence the target will be achieved.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Existing systems that are currently in use, such as AxisLogMiner and WebMiner, will be analyzed. Zhu et al. [6] propose a new vector space retrieval algorithm based on association diagram extension of key words. By using key words and the related words appearing simultaneously in a large scale, the algorithm allows generating the association diagram which indicates the simultaneous relationship between key words. The degree of association between any two key words is represented by mutual information. In addition, the algorithm derives the weight of key words in retrieval vector via association diagram, thus the vector space retrieval based on association diagram extension of key words is realized. Martinez-de-Pison et al. [7] propose an experience based on the use of association rules from multiple time series captured from industrial processes. The main goal is to seek useful knowledge for explaining failures in these processes. An overall method is developed to obtain association rules that represent the repeated relationships between pre-defined episodes in multiple time series, using a time window and a time lag. First, the process involves working in an iterative and interactive manner with several pre-processing and segmentation algorithms for each kind of time series in order to obtain significant events. In the next step, a search is made for sequences of events called episodes that are repeated among the various time series according to a pre-set consequent, a pre-established time window and a time lag. Extraction is then made of the association rules for those episodes that appear many times and have a high rate of hits. Finally, a case study is described regarding the application of this methodology to a historical database of 150 variables from an industrial process for galvanizing steel coils. Ho et al. [8] propose artificial intelligence methodology provides investors with the ability to learn the association among different parameters. After the associations are extracted, investors can apply the rules in their decision support systems. In this work, the model is built with the ultimate goal of predicting the level of the Hang Seng Index in Hong Kong. The movement of Hang Seng Index, which is associated with other economics indices including the gross domestic product (GDP) index, the consumer price index (CPI), the interest rate, and the export value of goods from Hong Kong, is learnt by the proposed method.

With the rapid growth of computer and Internet technologies, c-learning has become a major trend in the computer assisted teaching and learning fields. Most past researches for web-based learning focused on the issues of adaptive presentation, adaptive navigation support, curriculum sequencing, and intelligent analysis of student's solutions. These systems commonly neglect to consider whether learner can understand the learning courseware and generate misconception or not. To neglect learner's learning misconception will lead to obviously reducing learning performance, thus generating learning difficult. In order to discover common learning misconceptions of learners, Chen et al. [9] study employs the association rule to mine the learner profile for diagnosing learners' common learning misconceptions during learning processes. The association rules that occurring misconception A implies occurring misconception B can be discovered utilizing the proposed association rule learning diagnosis approach. Meanwhile, this study applies the discovered association rules of the common learning misconceptions to tune courseware structure through modifying the difficulty parameters of courseware in the courseware database so that learning pathway is appropriately tuned. Besides, they also present a remedy learning approach based on the discovered common learning misconceptions to promote learning performance.

At present, a lot of works have to do with the positive association rules in WEB usage mining, but negative association rules is more important, Yang et al. [10] have applied negative association rules technology to WEB usage mining in the course of the experiment they have proved that the negative association rules have a more important role on access pattern to web visitors, give the mining algorithms, to solve the deficiencies in which positive association rules are referred to. A variety of the web-mining techniques are now being extensively utilized to extract useful knowledge about customer behaviors on the Internet. However, the naive interpretation of the web-mining results would lead to poor decision on customer behaviors, which is likely to cause waste of time and efforts on managing electronic commerce strategy. To overcome this pitfall, Lee et al. [11] propose using the cognitive map-based interpretation of the web-mining results. Conventional approach to obtaining the web-mining results is based on the association rule approach (ARA), while the cognitive map approach (CMA) is believed to provide more robust support in interpreting the web-mining results. Therefore, to compare the interpretation capability of the two approaches, the four constructs such as perceived usefulness, causality, information richness, users' attitude and intention to use the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

approaches are adopted in the research model and tested against the questionnaire data.

To avoid returning irrelevant web pages for search engine results, technologies that match user queries to web pages have been widely developed. Du et al. [12] propose a new study, in which web pages for search engine results are classified as low-adjacence or high-adjacence sets. To match user queries with web pages using formal concept analysis, a concept lattice of the low-adjacence set is defined and the non-redundancy association rules defined by Zaki for the concept lattice are extended. OR- and AND-RULEs between non-query and query keywords are proposed and an algorithm and mining method for these rules are proposed for the concept lattice. The time complexity of the algorithm is polynomial. The amount of ontologies and semantic: annotations available on the Web is constantly increasing the type of complex and heterogeneous graph-structured data raises new challenges lot the data mining community. Nebot et al. [13] present a novel method for mining association rules from semantic instance data repositories expressed in RDF/S and OWL. They take advantage of the schema-level knowledge encoded in the ontology to derive just the appropriate transactions which will later feed traditional association rules algorithms. This process is guided by the analyst requirements, expressed in the form of a query Pattern experiments performed on real world semantic data enjoy Promising results and slimy the usefulness.

In this paper, we propose a novel ontology and association rules based XML mining algorithm. The algorithm firstly introduces the domain ontology and hash technology to improve the operation of emerging frequent item sets and generating association rules, then uses a hash table to store the domain ontology, and at last the algorithm transforms the operation of the database into memory tree based on XML.

3. ALGORITHM DESCRIPTION

3.1 Ontology

Ontology is a representation of knowledge formal. It provide a clear and consistent representation of terminology and methods that help people to observe the problems and dealing with affairs, provide public vocabulary of areas and define different levels of formal meanings of terms and relationships between terms. It is organized by taxonomy, and includes the typical model of the original language of the ontology and can provide a public and consistent understanding of the field. It overcomes the semantic content of the communication mismatch problem.

Ontology structure is divided into the following five stages [14]: a) Identify the purpose and scope of the ontology application: establish the field of study: establish the corresponding domain ontology or process ontology. b) Ontology analysis: define the relationship between all terms and ontology meaning. c) Represent ontology: to select a proper method of ontology according to the system need. d) Ontology test: main test the clarity, consistency, integrity, scalability of ontology. e) Ontology building: test the ontology according to the above criteria, to meet the requirements store the file form, otherwise switch to b).

Ontology representation: In order to describe and represent ontology, in recent years appeared a variety of ontology language. this paper choose one of the ontology languages——OWL (Web Ontology Language) [15]. OWL is designed by the World Wide Web Consortium Web Ontology Working Group. Its syntax is very similar to DAML + OIL, and can easily be converted to the latter. OWL can be used to clearly express the meaning of the vocabulary entries as well as the relationship between these entries. This express of vocabulary entries and relationship between them is called Ontology. OWL has more mechanism to express semantics than XML、RDF and RDF Schema, so it exceed capability of XML, RDF and RDF Schema that can only expressed machine-readable document content online.

3.2 Optimization And Transplantation Of Apriori Algorithm

The proposed mining algorithm is obtained based on the improved Apriori algorithm. Although Apriori has been optimization in a certain amount, but is not satisfactory in practical applications。 The proposed migration and optimization program introduce the field ontology and hash (Hash) technology and improve the operation of the frequent item sets and association rules of operation generated, and use the hash table to store the relevant domain ontology; this will make the classic Apriori algorithm database operations into operations on the XML tree memory. The advantage is better to play the XML's strengths, can operated from the relational database, and also when search ontology, generate candidate item sets and frequent item sets can direct access to memory when the hash table. It reduces disk I/O operation, greatly improving the efficiency of the algorithm. Of course, this program has some limitations: For example, large memory space, high space

Journal of Theoretical and Applied Information Technology
15th November 2012. Vol. 45 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

complexity, etc. The algorithm flowchart is shown in Fig. 1.

As shown in Fig. 1, in the improvement program, first of all, use the Apriori algorithm to find the transaction itemsets from the records idea, after find frequent item sets, search for all to meet the minimum support and minimum confidence of the strong association rules, and calculate its confidence. Apriori algorithm here was taken to improve the Apriori algorithm, which introduces the domain ontology. Program implementation is to use Java to apportionment XML data source into the XML tree and then calculate.

Store the candidate itemsets and frequent itemsets are in the local disk or memory. When the candidate set is too large, to save memory, the candidate set is stored as a local temporary file, removed frequent itemsets when it is end, and no longer occupy the storage space. In-memory XML tree XML_TREE_SUPPORT used to store frequent itemsets. As when calculating the confidence level need the corresponding services confidence, build index of XML_TREE_SUPPORT, applicate hash technology on each transaction Storage.

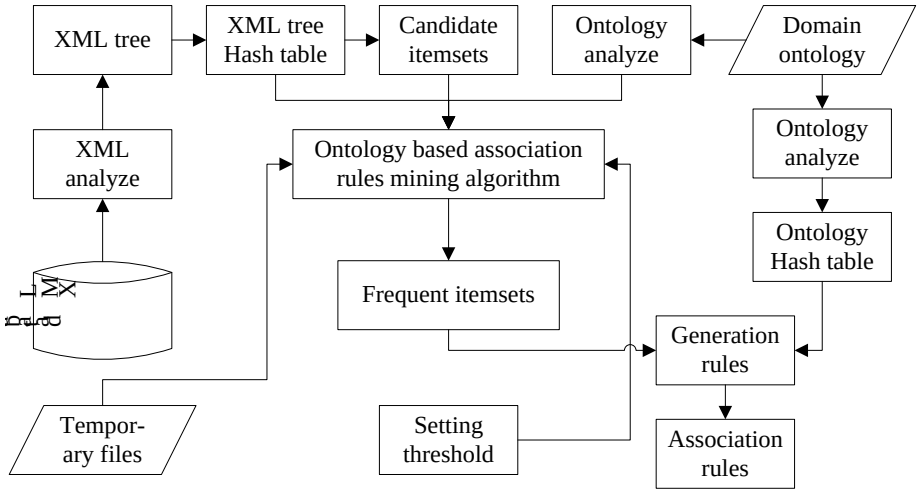

Figure 1: Algorithm flowchart

3.3 Ontology-Based Association Rules Xml Mining Approach

Apriori is one of the first mining association rules algorithms, it always has four stages: For a given database D, minimum support min-sup and the minimum confidence min-confidence. Steps are as follows:

Step 1: Scan the database D, find the items collection is greater than or equal to minimum support, set the project L1.

Step 2: Use large set length *k*-1 of projects to generate the candidate item sets.

Step 3: Calculation of all candidate item set support, to determine whether it is greater than or equal to the minimum support , the candidate will meet the conditions set excavation project, form a frequent set of *k* length.

Step 4: Repeat the above steps until no new candidate item sets generated then stop.

In this paper, the traditional algorithm is as the basis to improve the proposed algorithm. Mining object of traditional Apriori algorithm is transaction-oriented database, the Ontology-based XML mining association rules mining algorithm is an object-oriented XML data source. In this article, the XML data source stored in to the hash table during the data pre-processing time to speed up the speed of traversing the entire XML tree.

Meanwhile, the algorithm in this paper does three improvements on the traditional Apriori algorithm: First, when stored the XML data source in to the hash table and to each transaction and each data item is also added to all the parent of this transaction, then add the parent to candidate item sets as normal data, when traversing XML trees, also count on the parent class. Second, according to the pre-existence table hierarchy of domain ontology, Can get parent of each item sets and remove the parent class does not appear in any concentration of the candidate; Third, pruning also include key items of the item *x* and its ancestor set.

3.4 Steps Of The Algorithm

Resolve document XML to Tree XML in the project of Fig. 1. And after resolving ontology to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Hash table, execute the following steps to find the frequent item sets.

Step 1: Based on the thought of Apriori, creating candidate item set C_1. Count each item point with the simple ergodicity of XML Tree. Setting the min support and usually it is assigned by the user. Assure the frequent the assemble L_1 in itemset 1, and any of the frequent count of the program in L_1 should not less than the Minsupport.

Step 2: Execute candidate itemset 2. $L_1 \infty L_1$ is used to produce candidate itemset C_2. And each itemset in C_2 is produced by a connection of two frequent item set which belong to L_1. Search Hashtable and delete the itemset which has the term *x* and its ancestor item. And also the parent classes which do not appear in any candidate itemset should be deleted. Scan the transaction data XML Tree and compute each support of itemset in C_2. And use a temporary file XML for C_2.

Step 3: Select the transaction, the frequent count of whose is not less that the MinSupport, so to make sure the Assemble L_2 in frequent itemset 2. And at the same time, compute the Hash function corresponding to the transaction and add to the frequent itemset XML with the XML_TREE_NODE_ ID combined with item 2.

Step 4: Produce a collection of candidate sets C_3, using the characteristics of Apriori algorithm and pruning technique to remove all candidate set its subsets is not frequent itemsets. Thus greatly reduce the 3-set the size of the candidate items, so to improve efficiency to generate frequent item sets L_3, query ontology hash table. And then delete any candidate which does not appear in the parent class. Generate frequent item set L_3, calculate the degree of support for each transaction.

Step 5: So the cycle continues to generate candidate sets, and forms this generate frequent item sets, until the candidate set is empty.

4. EXPERIMENTAL RESULTS

To verify the method validity of the excavation proposed in this paper, using Java implemented test the algorithm. XML documents using Microsoft's MSXML parser to quickly resolve in order to generate a document tree. The experimental use of domain ontology by the University of Sheffield Natural Language Processing group released Animals used for experimental research ontology [16], it is described in British children's books appear in the animal species of the domain ontology, its purpose is that the primary standard for research to experiment use. As currently there are no dedicated mining XML documents, experimental test data used is from in the simulation library data. Data set contains 5000 multiple services, and after transformation, the XML document size is about 11.5 MB.

As the proposed algorithm uses hash table to store the XML document tree, so we can traverse the entire data set in a short time, so there is no need to repeat the scan XML documents, for the mining process saves a lot of overhead. Experiments show that the method by introducing domain ontology of the XML document frequent generalized data can effectively reduce the size of XML documents, mining the association rules and make the algorithm easier to be understood.

Algorithm drawback is memory for larger, higher space complexity, in addition the introduction of a parent class of each, algorithm to traverse the XML tree, the item count when a corresponding increase in performance will have some negative impact. Fig. 2 shows the diversification between the number of the frequent sets which be mined by the algorithm and the threshold of minimum support.

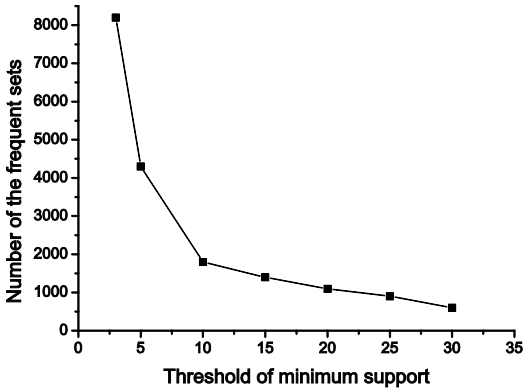

Figure 2: The Diversification Between The Number Of The Frequent Sets And The Threshold Of Minimum Support

5. CONCLUSION

XML documents of past mining algorithms are specific items of data mining, this paper introduced to the mining process domain ontology, allowing users a higher level in the excavation, mining results generated layers, help users better decision-making. The method of this article firstly introduces the domain ontology and hash technology to improve the operation of emerging frequent item sets and generating association rules, then uses a hash table to store the domain ontology, and at last the algorithm transforms the operation of the database into memory tree based on XML.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Experiment results show that the method can effectively reduce the size of XML documents and the association rules is easier to understand.

ACKNOWLEDGEMENTS

This work was supported by Ministry of Education of the People's Republic of China, Humanities and Social Sciences project. (No. 12YJC870036)

REFRENCES:

[1] Y. Sebastian, H. H. Then Patrick, "Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses", Knowledge-Based Systems, Vol. 24, No. 5, 2011, pp. 609-620.

[2] J. H. Feng, G. L. Li, "Efficient fuzzy type-ahead search in XML data", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 5, 2012, pp. 882-895.

[3] Y. Lou, Z. H. Li, Q. Chen, "Semantic relevance ranking for XML keyword search", Information Science, Vol. 190, 2012, pp. 127-143.

[4] Y. H. Chang, C. Y. Wu, C. C. Lo, "Processing XML queries with structural and full-text constraints", Journal of Information Science and Engineering, Vol. 28, No. 2, 2012, pp. 221-242.

[5] M. Nagi, A. ElSheikh, I. Sleiman, P. Peng, M. Rifaie, K. Kianmehr, P. Karampelas, M. Ridley, J. Rokne, R. Alhajj, "Association rules mining based approach for web usage mining", Proceedings of 2011 IEEE International Conference on Information Reuse and Integration, IEEE Conference Publishing Services, August 03-05, 2011, pp. 166-171.

[6] P. Zhu, F. Jia, P. Wu, "A novel information retrieval algorithm based on association diagram extension of key words", Information - An International Interdisciplinary Journal, Vol. 15, No. 10, 2012, pp. 4065-4079.

[7] F. J. Martinez-de-Pison, A. Sanz, E. Martinez-de-Pison, E. Jimenez, D. Conti, "Mining association rules from time series to explain failures in a hot-dip galvanizing steel line", Computers & Industrial Engineering, Vol. 63, No. 1, 2012, pp. 22-36.

[8] G. T. S. Ho, W. H. Ip, C. H. Wu, Y. K. Tse, "Using a fuzzy association rule mining approach to identify the financial data association", Expert Systems With Applications, Vol. 39, No. 10, 2012, pp. 9054-9063.

[9] C. M. Chen, Y. L. Hsieh, S. H. Hsu, "Mining learner profile utilizing association rule for web-based learning diagnosis", Expert Systems With Applications, Vol. 33, No. 1, 2007, pp. 6-22.

[10] B. Yang, X. J. Dong, F. F. Shi, "Research of WEB usage mining based on negative association rules", Proceedings of 2009 International Forum on Computer Science Technology and Applications, IEEE Conference Publishing Services, December 25-27, 2009, pp. 196-199.

[11] K. C. Lee, S. Lee, "Interpreting the web-mining results by cognitive map and association rule approach", Information Processing & Management, Vol. 47, No. 4, 2011, pp. 482-490.

[12] Y. J. Du, H. M. Li, "Strategy for mining association rules for web pages based on formal concept analysis", Applied Soft Computing, Vol. 10, No. 3, 2010, pp. 772-783.

[13] V. Nebot, R. Berlanga, "Mining association rules from semantic web data", Proceedings of Trends in Applied Intelligent Systems, SPRINGER-VERLAG BERLIN, June 01-04, 2010, pp. 504-513.

[14] J. Xie, F. Liu, S. U. Guan, "Tree-structure based ontology integration", Journal of Information Science, Vol. 37, No. 6, 2011, pp. 594-613.

[15] V. Milea, F. Frasincar, U. Kaymak, "TOWL: A temporal web ontology language", IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, Vol. 42, No. 1, 2012, pp. 268-281.

[16] F. A. Batzias, C. G. Siontorou, "Creating a specific domain ontology for supporting R&D in the science-based sector - The case of biosensors", Expert Systems With Applications, Vol. 39, No. 11, 2012, pp. 9994-10015.