Volume 4, Issue 8, August 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Efficient Framework for Rule Mining Using Eclat on XML Data

Rohini L. Damahe
PG Student, Computer Engg,
Thakur College of Engg & Technology,
Mumbai, India

Veena Kulkarni
Assistant Professor, Computer Engg,
Thakur College of Engineering,
Mumbai, India

Abstract— ***In today's digital world data is been increasing like anything largely in unstructured file format. There are many data mining techniques to mine these files. But it is bit challenging affair to mine semi structured data like XML. Many researchers are engaged in mining rules from XML data using popular algorithms like Apriori. As Apriori creates large amount of candidate sets so it needs more processing space and execution time is also high as it keep creating the candidate sets at every time ,this creates confusion in many researchers to choose Apriori for large data sets. So our idea of extraction of rules from large datasets like Reuters using Eclat rule mining algorithm which uses intersection of transactions for generating candidate itemsets. Our approach enhances the Eclat performance by enforcing comparative vertical power sets for creation of candidate itemsets to enhance the quality of the rules with less processing space and also with less execution time.***

Keywords— ***Frequent patterns, itemsets, rules, Power set, Eclat, Apriori, information retrieval, XQuery.***

I. Introduction

Data mining provides an efficient summary of the data which is actually substitution of original documents. This can be done on following conditions.

- Generally data mining or rule mining techniques get input as the documents in huge number and this data is been filtered by many information retrieval methods.
- Then retrieved data produces summary which can be a substitute of the original documents.

In order to satisfy the second condition the following points are important

Comprehensibility The summarized data should be the content of the given input data
Readability The summary should be in meaningful and readable style
Many of the recent researchers put emphasis on the comprehensibility while keeping readability of the summary to some extent.

Classification is a popular job in data mining that actually predict the class of an unseen instance as accurately as possible. While single label classification, which assigns each rule in the classifier the most obvious label, has been widely studied [1], Another important task in data mining is the discovery of all association rules in data. Classification and association rule discovery are similar, except that there is only one answer to predict in classification, i.e., the class, while association rule can predict any attribute in the data. In recent years, a new approach that integrates association rule with classification, named associative classification, has been proposed [1]. A few accurate classifiers that use associative classification have been presented in the past few years, such as CMAR [9], and CPAR [2].

In existing associative classification techniques, only one class label is associated with each rule derived, and thus rules are not suitable for the prediction of multiple labels. However, multi-label classification may often be useful in practice. Consider for example, a document which has two class labels "Japan" and "Earthquake", and assume that the document is associated 50 times with the "Japan" label and 48 times with the "Earthquake" label, and the number of times the document appears in the training data is 98. A traditional associative technique like CBA generates the rule associated with the "Japan" label simply because it has a larger representation, and discards the other rule. However, it is very useful to generate the other rule, since it brings up useful knowledge having a large representation in the training data, and thus could take a role in classification.

For performing rule mining using any association algorithms feature selection of the data plays a vital role. Selection of the accurate features (terms) can enhance the task more perfectly. Generally the features which can improve the performance of text classification are term frequency, important words, clustering and pruning.

In this paper we are preparing a novel approach of mining association rules by Eclat algorithm coupling with comparative vertical powerset to enhance the quality of the mined rules. In our approach we also mine the rules using Apriori algorithm on the same data set as of the Eclat and trying to justify the efficiency of the Eclat over the Apriori on huge datasets.

In our approach we first extract the data from XML files using java XQuery to store in database. Then this extracted data will be preprocessed in various steps like Stopwords removing, stemming, tokenization, TF-IDF, Shannon info gain and top term identification. This preprocessed data provides an important term in the document on which by applying comparative powerset on vertical frequent patterns, this will then tests on support and confidence given by the user to produce the efficient association rule then Apriori.

The rest of the paper is organized as follows: Section 2 discusses some related work and section 3 presents the design of our approach. The details of the results and some discussions on this approach are presented in section 4 as Results and Discussions. Section 5 elaborates hint of some extension of the approach as future work and conclusion.

II. RELATED WORK

The drastic increasing of internet users across the globe demands the requirement of the pattern recognition and classification in the data. This can be use in many fields like network attack pattern identification, text classification and many more. The major goal of this is to extract the summary in terms of phrases or in terms of some words (rules). So that it can be describe complete given content of the text. There are many various practical applications of the text classification are there, Spam filtering are the most improved one. Search engines may also take advantage of text classification technique to return more accurate results to the query by the user [4].

Many types of the text representation and classification methods are been proposed in the past. The most popular one is the bag of words that uses the words (terms) in the vector. In [5], the tf-idf methods used for the word representation in the Rocchio classifiers.

To increase the performance of the tf-idf a technique is proposed in [6] where global IDF and weighting entropy drastically improves the performance by 30 percent. Many weighting schemes for the bag of words representation approach were given in [7] [8]. The actual problem of bag of word is to select the most important and limited features (terms) among the enormous set of words. So that they can represent the given text more efficiently to avoid time and space complexity as broadly discussed in [9]. Again a feature selection technique improves to identify the multidimensionality of the features to boost the text classification in appropriate way. Many feature selection techniques are there like Information gain, mutual information, chi-square, Odds ratio and so on. Details of these selection functions were stated in [9], [10].

Pattern mining has been broadly studied and researched in data mining communities for many years. Various efficient algorithms such as Apriori-like algorithms [11], PrefixSpan, FP-tree, SPADE, SLPMiner, and GST [12] have been proposed. These research works have mainly focused on developing efficient mining algorithms for discovering patterns and extracting rules from a large data collection.

However, searching for useful and interesting patterns and rules was still an open problem as discussed in [13], [14], [15]. In the domain of text mining, pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemsets and re -occurring terms, for building up a representation with these new types of features. Any how challenging task is to deal with the large amount of data to extract its pattern or rule to find its summary efficiently.

III. PROPOSED METHOD

In this section, we describe the approach of enriching process of rule mining for XML data using Eclat. The step followed by our proposed system is described as shown in Fig. 1.

Step 1: In this step, an XML file data is been extracted using java XQuery process and then it is been saved in the Database.

Step2: This is the step where all the XML data stored in DB are preprocessing by the following four main activities: Sentence Segmentation, Tokenization, Removing Stop Word, and Word Stemming. Sentence segmentation is boundary detection and separating source text into sentence. Tokenization is separating the input text into individual words. Next, Removing Stop Words, stop words are the words which appear frequently in the text but provide less meaning in identifying the important content of the document such as ‘a’, ‘an’, ‘the’, etc.. The last step for preprocessing is Word Stemming; Word stemming is the process of removing prefixes and suffixes of each word.

Step 3: **Term Weight**

The most repeated word in the document actually plays important role to identify the importance of the sentence. Then rank of the sentences can be calculated as the sum of rank of all words in that sentence. The Rank of any words w_i can be finalized by the calculation of tf-idfs (Inverse document frequency) as shown below in 1 .

W_i= tf x idfi=tf_ix log (N / ni)…… **(1)**

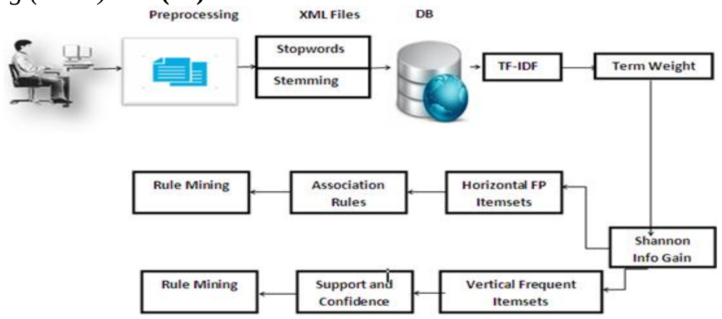

Fig. 1. Overview of our approach

Where tf_i is the tern frequency of word *i* in the document, N is the total number of documents, and n_i is number of documents in which word i occurs.

Step 4: **Information Gain**

In order to summarize each of documents in an IR result, we use Shannon's term weighting based on formation Gain Ratio (IGR).

This method extracts the similarity structure among a set of documents through a hierarchical clustering, then gives higher weights to words that contribute to forming the structure. Important words are calculated based on IGR as follows

$$\mathrm{IGR}(C) = -\sum (|C_i| / |C|) \log (|C_i| / |C|) \quad ....(2)$$

Where C_i is the frequency of the word *w* in Cluster C.

Step 5: Then by using the vertical intersection of the words system identifies the most obvious words for rule mining using powerset. Where all these words are extracting by the comparative recursion of the combination of the words.

Step 6: Then after fetching the important words from all the documents system will perform association rule using Apriori Algorithm with the step stated below.

Let *T* be the training data with *n* attributes *A1*, *A2*, …, *An* and *C* is a list of class labels. A particular value for attribute *Ai* will be denoted *ai*, and the class labels of *C* are denoted *cj*.

Definition 1: An item is defined by the association of an attribute and its value (*A*i, *ai*), or a combination of between 1 and *n* different attributes values, e.g. < (A1, a1)>, < (A1, a1), (A2, a2)>, (A1, a1), (A2, a2), (A3, a3)>, … etc.

Definition 2: A rule *r* for multi-label classification is represented in the form:

$(A_{i1}, a_{i1}) \wedge (A_{i2}, a_{i2}) \wedge ... \wedge (A_{1m}, a_{im}) \rightarrow c_{i1}....c_{im}$

where the condition of the rule is an item and the consequent is a list of ranked class labels.

Definition 3: The actual occurrence (*ActOccr*) of a rule *r* in *T* is the number of cases in *T* that match *r's* condition.

Definition 4: The support count (*SuppCount*) of *r* is the number of cases in *T* that matches *r's* condition, and belong to a class *ci*. When the item is associated with multiple labels, there should be a different *SuppCount* for each label.

Definition 5: A rule *r* passes the minimum support threshold (*MinSupp*) if for *r*, the *SuppCount*(*r*)/ |*T*| ≥ *MinSupp*, where |*T*| is the number of instances in *T*.

Definition 6: A rule *r* passes the minimum confidence threshold (*MinConf*) if *SuppCount*(*r*)/*ActOccr*(*r*) ≥ *MinConf*.

Definition 7: Any item in *T* that passes the *MinSupp* is said to be a frequent item.

Step 7: In the final step proposed system will perform vertical frequent pattern mining using éclat algorithm as shown below.

Algorithm 1 Eclat Algorithm

```
Input: Alphabet A with ordering ≤ multiset T ⊆ P(A) of sets of Items , Minimum support value minsup Є ℕ.
Output: Set F of frequent Itemsets and their support counts.
F:={(Ø,|T|) }.
CØ:= {(x,T({x}))| x Є A}.
C'Ø:= freq (C_Ø):= {(x,T_x)|(x,T_x) Є C_Ø,
                                   |T_x|≥ minsup }
F:= { Ø }.
Add frequent supersets (Ø, C'_Ø).
```

```
function add Frequent Supersets():
Input: frequent Itemsets p Є P(A) called prefix,
 incidence matrix C of frequent 1-item-extentions of p.
Output: add all frequent extensions of p to global variable F.
for (x, T_x) Є C do
q:= p U {X}.
C_q:={(y,T_x ∩ T_y) | (y,T_y) Є C, y > x}.
C'q := freq(C_q) := {(y,T_y) | (y, T_y) Є C_q,  |T_y| ≥ minsup  }

If  C'_q ≠ Ø then
Add frequent supersets (q,C'_q).
End if
F := F U {(q, |T_x|)}
End for
```

IV. Results and Discussions

To show the effectiveness of the proposed system, some experiments are reported. Selecting a suitable dataset is a critical and important step in designing rule mining system.

There is no condition in data mining for the usage of the specific dataset for the research. Any huge data set can be serve for this purpose. So to perform experiment on our system we use most generalized data set from the Reuters which are in the xml structure. As this data set is huge and having great versatility it provide a good challenge to our task.

A. Practicability of System Demonstration

In our proposed system the user selects the XML dataset and extracts the needed data using XQuery to store in database. After that user need to enter minimum support and confident on the basis of which he wants to extract the rules from Eclat algorithm. Then System performs the series of feature extraction methods like tf-idf and Shannon information gain system. Then by applying a powerset for the intersection of the transaction data system generates the frequent item sets. Then generated frequent item sets will be tested for the minimum support and confidence to get the efficient rule.

B. Relevent Comparisons

Author [17] proposes a method of extracting the rules using Apriori over the XML data using XQuery.For maintaining balance and similarity for the comparison proposed system also uses a dataset which contains about 20 files and average of 6 transactions in each files. And each file is containing more than 12 items.

Then system was tested for various support values to check its feasibility with the Apriori algorithm whose results can be shown in below figure no 2.

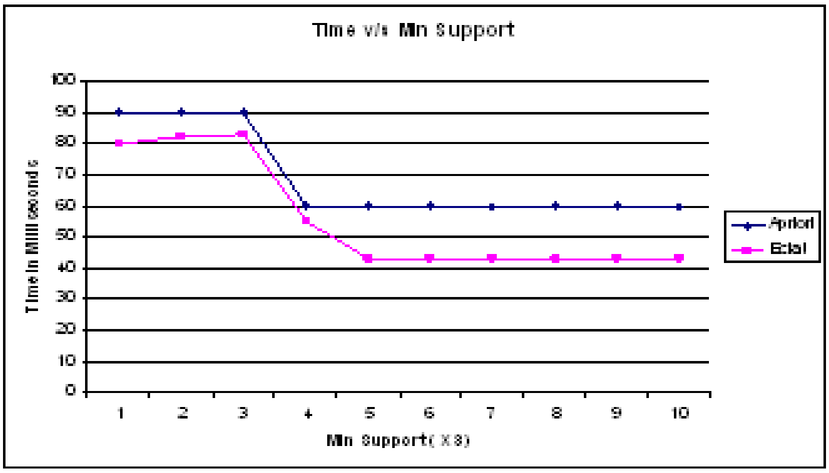

Fig 2 .Time comparison of Apriori and Eclat Algorithms

It is clearly observed from the figure 2 that as the support increases the processed time of both the algorithms leaps for same value. The proposed system of Eclat has achieved better precision as compared to system proposed by the author [17] which uses Apriori as the mining algorithm. This shows frequent items fetched by intersection of transaction perform well in time and also gives good quality of rules.

V. Conclusions and Future Work

In the proposed approach of mining association rules system efficiently enhance the feature of Eclat algorithm with comparative powerset.Comparitive powerset extract the maximum frequent itemsets from important words which are been decided by tf-idf and Shannon information gain. Proposed system enforces the powerset with multi recursion methodology to get as maximum as possible of intersection transactions. This method actually enhances the Eclat algorithm to create frequent itemsets on intersection and thereby to reduce the space and time complexity efficiently.

System efficiently takes comparatively less processing time to get the rules for the given minimum support than the other mining algorithms like Apriori. Which are creating more frequent items on each run even on small datasets; this actually doubts the selection of Apriori algorithm for huge datasets. The comparison of both algorithms were discussed in the last section, where éclat is over coming Apriori clearly in all possible given minimum support, This justifies Eclat over Apriori for huge datasets.

As the feature work of this proposed method, frequent itemsets can be extracting on the basis of group of distinct terms with recursive multithreading methodology to enhance the time complexity to perform the rule mining in exponentially less time.

.

References

[1] W. Li, J. Han and J. Pei. CMAR: Accurate and efficient classification based on multiple class association rule. In ICDM'01, San Jose, CA, Nov. 2001, pp. 369-376.

[2] X. Yin and J. Han. CPAR: Classification based on predictive association rule. In SDM 2003, San Francisco, CA, May 2003.

[3] An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm.R.Porkodi,V.Bhuvaneswari, R.Rajesh,T.Amudha,-IACC 2009

[4] Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski, and Michael W. Mahoney. Feature selection methods for text classi cation. In KDD '07: proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 230{239, New York, NY, USA, 2007

[5] X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

[6] S.T. Dumais, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.

[7] K. Aas and L. Eikvil, “Text Categorisation: A Survey,” Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[8] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval,” Information Processing and Management: An Int’l J., vol. 24, no. 5, pp. 513-523, 1988.

[9] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[10] D.D. Lewis, “Feature Selection and Feature Extraction for Text Categorization,” Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[11] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 478-499, 1994.

[12] Y. Huang and S. Lin, “Mining Sequential Patterns Using Graph Search Techniques,” Proc. 27th Ann. Int’l Computer Software and Applications Conf., pp. 4-9, 2003.

[13] Y. Li, W. Yang, and Y. Xu, “Multi-Tier Granule Mining for Representations of Multidimensional Association Rules,” Proc. IEEE Sixth Int’l Conf. Data Mining (ICDM ’06), pp. 953-958, 2006.

[14] Y. Li and N. Zhong, “Interpretations of Association Rules by Granular Computing,” Proc. IEEE Third Int’l Conf. Data Mining (ICDM ’03), pp. 593-596, 2003.

[15] Y. Xu and Y. Li, “Generating Concise Association Rules,” Proc. ACM 16th Conf. Information and Knowledge Management (CIKM ’07), pp. 781-790, 2007.

[16] M. Wasson, “Using leading text for news summaries: Evaluation results and implications for commercial summarization applications” In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL. Pp.1364-1368. 1998.

[17] R.Porkordi,V Bhuvaneshwari, R. Rajesh and T. Amudha “An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm “, IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009