International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 2, Mar - Apr 2016

RESEARCH ARTICLE OPEN ACCESS

Precising the Characteristics of IP-Fpm Algorithm

Dr. K.Kavitha
Assistant Professor
Department of Computer Science
Mother Teresa Women's University, Kodaikanal
Tamil Nadu – India

ABSTRACT

Pruning technique in Data Mining is defined as the biggest improvement in the performance of Apriori algorithm. It is used for reduction in candidate set group. Large number of itemsets do not have to be considered for generating candidate itemsets in existing algorithms. Extracting Frequent Patterns from Massive amount of Data is a tedious process which leads time complexity. The research work focuses on the dipping execution time through generating the itemsets increasingly from the database. The algorithm IP-FPM is used to reduce the memory complexity and the execution time. This paper highlights the detailed overview of IP-FPM algorithm.

Keyword:- Support, Confidence, Association Rule, Pruning, Itemsets

I. INTRODUCTION

Earlier to Apriori, AIS approach is used to mine the frequent pattern, which is a direct technique, therefore requires multiple passes over the database in order to generate the frequent item set. AIS algorithm was the first algorithm proposed for mining association rule. It focuses on improving the quality of databases together with necessary functionality to process decision support. In this algorithm only one item consequent association rules are created, which means that the consequent of those rules only contain one item. These drawbacks are removed by Apriori algorithm, in which efficiently, it generates the frequent itemsets. It also added a novel pruning technique to remove the irrelevant and redundant item set that are generated.

Frequent pattern mining is the important step of association rule mining. Although, when minimal support is set small or data set is dense, frequent pattern mining algorithm may generate massive frequent pattern and it is difficult to find knowledge from mining result. Weighted frequent pattern mining can determine more important frequent pattern by considering different weights of each item and with the appearance of a wide range of online data streams applications.

Pruning association rule set contains a pair of directly linked pages out of the association rule set generated by the association rule discovery algorithm. This pruning process can be embedded within the association rule discovery algorithm, which would substantially increase its efficiency. It is based on the schemes for rule clustering and removing irrelevant rule was used. The two steps to remove the infrequent itemsets are (i) find all item set from a huge repository initially, discover the candidates from the dataset and (ii) find the support count for all the corresponding support count. The candidate that satisfied the minimum support is framed as 1-itemset. The process proceeds through generating the higher level item set till it reaches the most frequent item set that satisfies the minimum support count.

Depending on the support value the discovered item sets are pruned to remove the item set that are under the user defined threshold support value. Frequent itemset is used to reduce the number of candidate itemsets, and number of transaction and comparisons. Apriori algorithm drastically decreases the cost required for input and output operation as well as memory requirement. If an itemset is frequent, then all of its subsets must also be frequent. Though it overtakes the AIS approach's drawback it still has problem in scanning the entire databases for many time. Frequent pattern mining has been studied extensively in mining supermarket transactions data and relational data.

Apriori is a great improvement in the history of association rule mining. It is more efficient during the candidate generation process for two reasons; Apriori employs a different candidate's generation method and a new pruning technique. There are out processes to find out all the large itemsets from the database in

Apriori algorithm. First the candidate itemsets are generated, and then the database is scanned to check the actual support count of the corresponding itemsets. During the first scanning of the database the support count of each item is calculated and the large 1-itemsets are generated by pruning technique. It needs scanning the whole data set and examine the itemsets multiple of times, which is very time consuming process. Many frequent pattern algorithms have been proposed.

Pruning algorithm is used to determine the Apriori existence of subsets of the new candidates with in the current generation. All of the new candidates are served through the system again so that all possible candidate pairs can be processed. It is based upon test method and candidate set generation. The problem that always performs during mining frequent relations is its exponential complexity. The advantage of Apriori algorithm is, it uses large itemset property, easily parallelized, and easy to implement. The modified algorithms focus on any one of the following approach.

(1) Reduces the number of passes that are carried over the database or to substitute only a part of database in the place of whole database depending on the current frequent itemsets.

(2) Explore various kinds of pruning techniques, which lessen the number of candidate itemsets.

II. RELATED WORK

Data mining process can be carried out using various techniques namely, association rule mining, classification, summarization, clustering, sequence discovery, and time-series analysis. Depending on the application and requirement suitable technique was used. As well as these techniques have generated a large amount of patterns were extracted. To achieve better results in less time, number of patterns that are generated using the above-mentioned techniques should be lesser. Few works that were related to the mining techniques were discussed in this section. Paper [1] discovered a technique to reduce the generation of the number of pattern. For that they have studied the interestingness measure proposed in machine learning, statistics, and data mining techniques. In addition to that, they have also investigated the correlation among the patterns.Moreover, they expressed that the pruning based on support count have significantly eliminated the negatively correlated and uncorrelated pattern generation. On considering with the interesting association patterns they have also introduced a new linear metric with respect to coefficients of correlation, named IS.

In Paper [2], a new approach developed for mining the association rules from a graph data set. The approach was named as AGM, which is extended from the basket analysis. The transaction can be represented through the adjacency matrix and frequent pattern appearing in the matrices. Paper [3] have focused on the discovery of frequent pattern from a web log data with the aim to obtain the hidden information. The association rules can use any of the techniques such as the rule structure cover, rule clustering, informative cover method, interesting measure, etc. to group the rules (Pruning).

In Paper[10], a system was implemented to discover the association rules from the web log containing web usage of users. They have pruned the rules that contain pages that are linked directly. Their results showed that, interestingness measure like support count can be used to sort the association rules that were generated. Paper[6] used interesting measures to select the association rule. They have also suited the way to cluster the patterns that were distributed over different interesting measures. The association rules are also used for clustering.

III. IP-FPM ALGORITHM

Association rule mining (ARM) is the most popular knowledge discovery technique used in several areas of applications. In ARM, huge number of Association rules are generated from the large volume of dataset.

Association rules are created by analyzing data for frequent and infrequent patterns. But association rules have redundant information and thus all of them cannot be used directly for an application. After mining rules , it is necessary to prune or group rules according to interestingness measure. Interestingness measure is a measure, used to control the generation of association rules(AR) called support,confidence, lift, conviction and so on.

Compute most prioritized frequent pattern is another important problem of mining frequent pattern. To overcome all problems and limitations, to propose a new and effective approach to mine frequent patterns called Priority based frequent item generation.

The candidate generation in the proposed method is similar to that of Apriori candidate generation. If any length k-itemset is not frequent in the database, its length k+1 superset can never be frequent. It is used to reduce the number of candidates significantly. The other class of algorithm called pattern growth technique does not require candidate itemsets generation.Although, minsup is used to reduce the itemset identification to the finding of only those itemsets that exceed a specified presene within the database, but tuning an appropriate minsup for different datasets is not an easy task.

There may not be necessary for many applications to process the full database in order to obtain the frequent pattern of that database. For instance an enterprise may involve finding frequent pattern for only a particular month. Therefore, it is enough to process the transactions that belong to that particular month.

In such cases the users may specify the upper and lower limits of the truncations in a database that are needed to be processed. In Figure 2, it shows the desired transactions which are extracted based on the User' response. After getting the Upper and Lower Limit from the user's decision. It has two choices either selects all the itemsets based on the priority or selecting particular frequent itemsets and also listed infrequent itemsets.

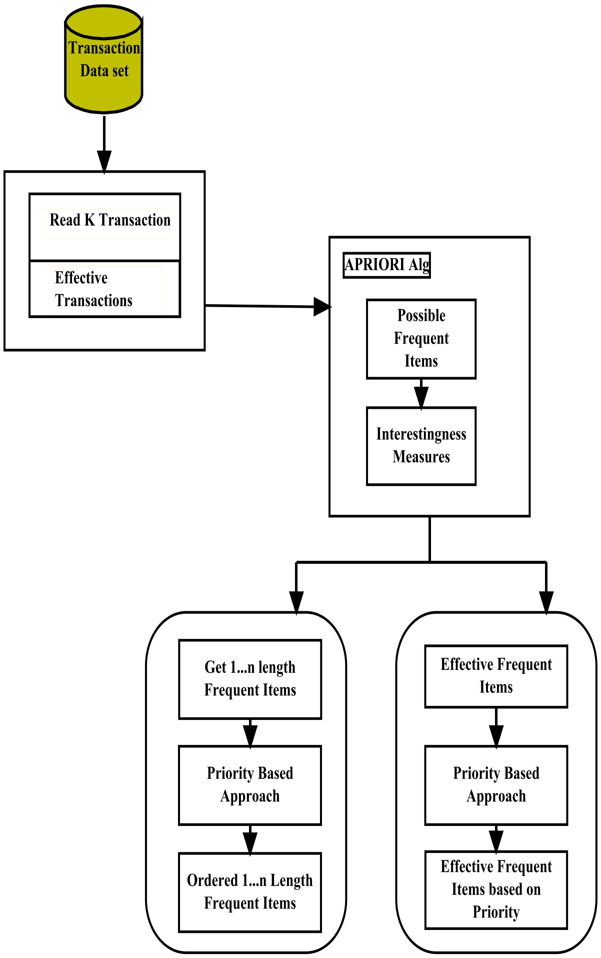

Figure 2
Retrieve itemset from repository

Itemset processing

In order to process the itemset authors have framed four different boxes namely,

- Anticipated infrequent (AI), Anticipated frequent (AF), Definite infrequent (DI) Definite frequent (DF)

Initially all 1-itemset $I_1, I_2, \dots \dots, I_n$ that belong to the user specified transactions are set to AI. Then the DB is read K transactions at a time in the defined transaction set. The computation of the largest itemset is continued until the number of transaction read is less than the upper limit transaction number. The above process is repeated until the AI and AF are present. If the itemsets $I_1, I_2, \dots \dots, I_n \in T_j$, then increment the counter value for the corresponding item. For each $I \in AI$ and the counter value is satisfied for the minimum weight (one of the interestingness measure which are discussed in the subsequent subsection) then set that item to AF(Anticipated Frequent), i.e. if $W_{min} \geq TW_{min}$ of an item then$I \in AF$.

Once the item belongs to either AF or DF, i.e. $I \in AF \parallel DF$ then their immediate superset is set to AI. Consider that if $I_1, I_2 \in AF \parallel DF$then $I_1 I_2 \in AI$. Here, $I_1 I_2$ is the superset of the items $I_1 \& I_2$. Once the entire transactions are read then the items that belongs to AF are assigned to DF. Likewise, items present in the box AI is moved to DI box, which shows that items of this box does not participate in the frequent pattern. Whereas, the items that belongs to the DF box participate in the frequent pattern. It expresses the steps taken to discover the frequent patterns. The first step is to find single rule composed of frequent itemsets. The second step is to extract interesting combined and composite association rule sets. In order to compute the interestingness measures, the support count of all frequent itemset is recorded in the frequent itemset generation step.

Rule Generation and Interestingness measure

Interestingness measures are used to detect the items where they occur frequently in the transactions of a database. It deals with a new algorithm that focuses on generating the frequent itemsets increasingly. Initially, the user specified itemsets are retrieved from the dataset, which are then managed to compute the following interestingness measures.

Let S be a dataset with |S| instances. From the dataset, association rules of the form X→Y is generated using Apriori algorithm. The item-sets X and Y are called antecedent and consequent of the rule. Generation of association rule is controlled by the value of support and confidence. The itemset that are less than the user specified threshold value for weight are not taken for further processing still it satisfies the processed value and frequent pattern is established.

Weight

This value plays a major role in deciding the frequent pattern. This can be calculated through taking the average value of lift and confidence. Mathematically it can be represented through the equation .

$$Weight(I_1|I_2) = \frac{Conf(I_1|I_2)+Lift(I_1|I_2)}{2}$$

The weight values are calculated for the itemset that are more than 1-itemset since confidence is not computed for 1-itemset. Hence, for 1-itemset, their corresponding support value computed from equation 1 is assigned. Following section illustrates the concept of generating the frequent pattern from a data base containing six transactions.

Transaction Based on IP-FPM

In customers' transaction databases, frequent itemset is used. A customers' transaction database is a sequence of transactions which is having group of ietms (T = t1,… tn), where each transaction in the database is an itemset $ti \subseteq I$. K-itemset is defined as an itemset with k elements. The support of an itemset X in T, denoted as sup T(X), is the number of those transactions that contain X.

An itemset is frequent if its support is greater than a support threshold, originally denoted by minsup. The frequent itemset mining problem is to find all frequent itemset and counts item occurrences to determine large 1-itemsets. This process is repeated until no new large 1-itemsets are identified. (k+1) length candidate itemsets are generated from length k large itemsets.

Candidate itemsets containing subsets of length k that are not large are pruned. Support of each candidate itemset is counted by scanning the database, reducing candidate itemsets that are small. Most of the algorithms accept an Apriori method which generates a candidate pattern by extending currently frequent pattern and then test the candidate. During this process, many infrequent patterns are produced.

A k-itemset is frequent only if all of its sub-itemsets are frequent. This implies that frequent itemset 1-itemsets can be mined by first scanning the database. This itemset is used to find the frequent 2-itemsets. This process repeats until no more frequent k-itemsets can be generated for some k. This is the essence of the Apriori algorithm and its alternative.

This example contains six transactions with three items in a DB, i.e.

$$I = \{A, B, C\},\ T = \{T_1, T_2, T_3, T_4, T_5, T_6\}.$$

Table 3.2 represents the data base along with the transactions.

Table 3.3 : Database with transactions

	A	B	C
T_1	1	1	0
T_2	0	0	1
T_3	1	0	1
T_4	1	0	0
T_5	1	0	1
T_6	1	0	1

In this example, the predefined threshold values for support count and weight are considered as 0.4. Three transactions are read at a single time (i.e. K=3). Initially all the items in the DB are assigned to AI and the boxes are denoted as shown in Figure 3.3.

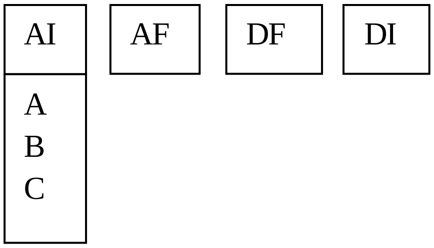

Figure 3.3: Initial A=B=C=0 and after 1K process

In equation 3.1, items of A, B, C values are calculated by the 1-itemset that support count values. As the computed values are less than the predefined threshold value they are not moved from the AI box. Therefore, 1K process maintains the itemset as same as in Figure 1. 1K scans first three transactions, i.e. T_1, T_2, T_3 in the DB.

For this stage also support count values is taken from the weight. It also contains the 1-itemset. The items A and C have the support value that is higher than the user defined threshold value, whereas the item B is less than the predefined value. Therefore, after computing the 2K, the items in the above transactions A and C are transferred from AI to AF, and the item B is moved from AI to DI. Figure 3.4 denotes the items in four different boxes after 2K process.

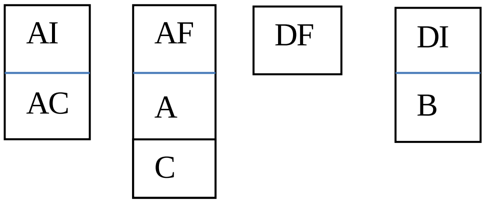

Figure 3.4: Items in corresponding boxes after 2K process

Once the A and C items are transformed to the AF box, the next 2-itemset are produced through combining the items that are present at the AF box. This process differs from the general Apriori algorithm since Apriori finds 2-itemset for all the items current in the DB. This process decreases the time for additional computations. From the initial stage these values are reduced.

During 3K process, the confidence, lift and weight parameters are computed for the 2-itemset present in the AI box using the equation 2, 3, and 4 respectively. Based on the calculated values the itemsets are further managed to transfer from their present box to the new boxes. Figure 3.5 expresses the itemset in corresponding boxes after 3K process.

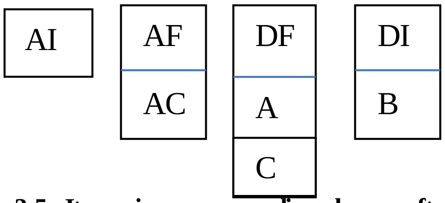

Figure 3.5: Items in corresponding boxes after 3K process

Further, formation of 3-itemset becomes difficult due to the less threshold support value presented for the item B. Therefore, the itemset present in AF box is transferred to DF box as shown in figure 3.6.

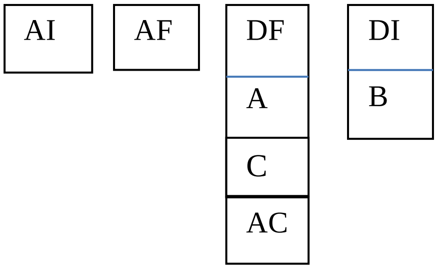

Figure 3.6: Final Frequent pattern

From figure 3.6 it is clearly shown that the most frequent pattern is AC. Once the entire algorithm is implemented completely all the items and itemsets will be either in DI or DF.

The largest itemset present in the DF is measured as the most frequent pattern. This algorithm protects time and exponential difficulty of Apriori algorithm since it produced the itemset from only the items that satisfy the minimum support count or weight, which decreases the generation of further itemsets.

With this, frequent pattern for a specific itemsets are retrieved separately. For example, if the user needs 1-itemset's frequent set then from DF A and C alone are retrieved. Whereas, if user needs the 2-itemsets pattern, then AC is retrieved from DI box. Figure 3.7, expresses the retrieval of the 2-itemset alone from the DF box.

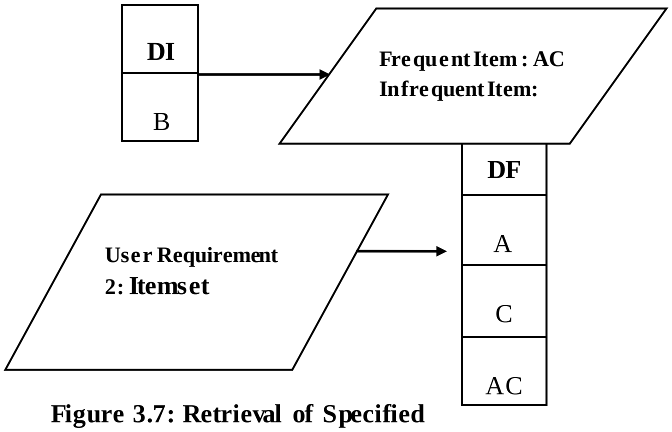

Figure 3.7: Retrieval of Specified itemset

IV. CONCLUSION

This proposed work focused on the reduction of the exponential complexity of the Apriori algorithm. To overcome the problem in Apriori algorithm, introduced a new algorithm named IP- FPM. This algorithm will generate the frequent patterns in the incremental way. Only the items that have supported the minimum threshold weight will participate in the further generation of itemsets. Therefore, this concept reduces the exponential complexity present in the APRIORI algorithm. In addition to the transaction that are to be used in the frequent pattern generation are also specified explicitly by the user. Moreover, extra feature of this proposed work is that users can view the frequent pattern of the required item set separately.

REFERENCES

[1] Pang-Ning Tan, and Vipin Kumar, "Interestingness Measure for Association Patterns: A perspective," in the proceedings of workshop on Machine Learning and Data Mining, 2000.

[2] Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In PKDD'00, pp. 13-23, 2000.

[3] [3] Iváncsy R, and Vajk I, "Frequent pattern mining in web log data," Acta Polytech. Hungarica, Vol. 3, issue 1, pp. 77-90, 2006

[4] Hassam H. Malik, and John R Kender, "Clustering web images using association rules, interestingness measures, and hypergraph partitions," in the proceedings of the sixth International Conference on web engineering, pp. 48-55, 2006

[5] Adnan M, Ajhaji R, and Rokne J, "Identifying Social Communities by Frequent Pattern Mining," in the proceedings of thirteenth International Conference on Information Visualisation, pp. 413-418, 2009

[6] Kannan S, and Bhaskaran R., "Association Rule Pruning based on Interestingness Measures with Clustering," in the proceedings of International Journal of

Computer Science, Vol. 6, issue 1, pp. 35-43, 2009

[7] Chowdhury F. A, Syed K. T, and Byeong S. J., "An Efficient Method for Incremental Mining of Share-Frequent Patterns," 12th International Asia-Pacific Web Conference, pp. 147-153, 2010

[8] Sunil Joshi, and Dr. Jain R. C., "A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database," in the proceedings of second international Conference on Communication Software and Networks, pp. 498-501, 2010

[9] Afshar Alam M, Sapna Jain, and Ranjit Biswas, "DataAprori algorithm: Implementation of scalable Data Mining by using Aprori algorithm," in the proceedings of International Journal of InnovativeTechnology and Creative Engineering, Vol. 1, No. 11, pp. 25-34, 2011

[10] Maja Dimitrijevic, and Zita Bosnjak, "Web Usage Association Rule Mining System," in the proceedings of Interdisciplinary Journal of Information, Knowledge, and Management, Vol. 6, pp. 137-150, 2011

[11] Wang Jie, and Zeng Yu, "DCEFP-Miner: Mining Closed Weighted Frequent Patterns over Data Streams," in the proceedings of ninth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 637-641, 2012

[12] Feng-gang Li, Ying-Jia Sun, Zhi-wei Ni,Yu Liang, and Xue-ming Mao, "The Utility Frequent Pattern Mining Based in Slide Window in Data Stream," in the International Conference on Intelligent Computation Technology and Automation, pp. 414-419, 2012

[13] K.Kavitha, Dr. E.Ramaraj, "IP-FPM-Intensified Priority Based Frequent Pattern Mining", European Journal of Scientific Research, March 2013.