ISSN 2395-1621

Optimization of Frequent Representative Pattern Sets using FP Growth Algorithm

[#1]Preeti C. Tarange, [#2]Prof. Rajesh H. Kulkarni

[1]tarangepreeti@gmail.com
[2]rkpv2002@gmail.com

[#1]Student, Department of Computer Engineering
[#2]Professor, Department of Computer Engineering,JSPM's Narhe Technical Campus, Pune, India

ABSTRACT

Frequent item sets play an essential role in data mining. The time required for generating frequent item sets plays an important role. Some algorithms are designed, considering only the time factor. Some Frequent pattern mining often produces a large number of frequent patterns, which imposes a great challenge on visualizing, understanding and further analysis of the generated patterns. This emerges the need for finding small number frequent occurring patterns. Here, the basic frequent itemset, pattern sets mining problems is explained. To find a minimum representative pattern set with an error free algorithm called MinRPset is used. MinRPset is very memory space-consuming and time-consuming on some dense datasets when the number of frequent closed pattern is more. To solve this problem, another algorithm called FlexRPset is used, which uses one extra parameter K to allow users to make a trade-off between result size and efficiency. It shows that MinRPset and FlexRPset produce fewer representative patterns than RPlocal. This emerges the need for finding small number frequent occurring patterns. The new approach includes analysis of algorithms that are used to find frequent pattern sets by using the technique of FP-Growth i.e. association rule. FP growth allows frequent item set discovery without candidate item set generation. It's two step Approach:

Step 1: Build a compact data structure called the FP-tree Built using 2 passes over the data-set.

Step 2: Extracts frequent item sets directly from the FP-tree Traversal through FP-Tree. We are using FP growth algorithm to determine the frequent item sets bought by user.

Keywords:Representative patterns, frequent pattern summarization, frequent item sets, pattern sets.

ARTICLE INFO

Article History

Received :3rd January 2016

Received in revised form :

4th January 2016

Accepted : 5th January,2016

Published online :

6th January 2016

I. INTRODUCTION

Data mining is an integrative subdivision of a field of computer science. It is the computational mechanism of inventing patterns in large data sets ("big data") containing approaches at the junction of artificial intelligence, machine learning, statistics, and database systems. The comprehensive target of the data mining technique is to draw out knowledge from a data set and alter it into a simple design for farther use.Data mining is a powerful method of extracting knowledge/information from large amount of data. It has great attention in recent years because of growing amount of data and the persistent need of turning the huge data into information. It is widely adapted in many fields like bioinformatics, business analytics, marketing, security and many more. An important task of data mining is discovering frequent patterns that play an important role in clustering, associations mining, correlations, etc. Frequent patterns are patterns that appear in a data set frequently, such as item sets, sub-sequences, or substructures. An item set A (or subsequence, or substructure) is said frequent if it satisfies the predetermined minimum support count, where

$$Supp(A) = \frac{numberofoccurencesofA}{totalno.ofitemsets \in adatabase}$$

An association rule is an implication of the form A Ø. The rule A, D holds the minimum support count of AUB in D. this rule has confidence c in the dataset D.

$$confidence(A - B) = P\left(\frac{B}{A}\right)$$

Association rule mining, in general can be viewed as a two-step process.

- Find all frequent item sets
- Generate strong association rules from the frequent item sets.

Mining frequent patterns from several patterns is one of the most important concepts in data mining. Other data mining concepts can be derived from these concepts. It is the beginning of the data mining technical training because it gives the effective idea about data mining which is not extremely technical.

Pattern Sets:
A pattern is a template, form or model which is used to create or to generate parts of things. In data mining we say that a pattern is a particular behaviour of data, arrangement or formation that might be of a business interest. A frequent pattern sets are item sets, sub-sequences, or substructures that appears in a data set in a frequency manner with no less than a user specified threshold. A substructure can refer to different structural forms such as sub graphs, sub trees or sub lattices which may be combined with item sets. If a substructure is produced frequently in a database is called a frequent pattern. Finding frequent patterns plays an important role in mining associations, correlations and many other interesting relationships among data. Moreover it is useful in data indexing, classifying, clustering and other data mining tasks.

Frequent Item Sets:
Another concept is a frequent item set which is a type of pattern set. A frequent item set is a parameter that is specified by the user in the database. The parameter is called as a support of an item set. Every subset of a frequent item set is also a frequent pattern. This property is also called as Apriori property or downward closure property. It explains that we do not need to find a count of the item set if subset is not frequent. This will become possible because of the anti- monotone property of support. Frequent item sets should satisfy the minimum support of user threshold (Agrawal, 1993). The support for an item sets never exceeds support for a subset. If we divide the entire database in several partitions then an item set can be frequent only if it is frequent in at least one partition. To find a frequent item set we should go through all sub item sets which themselves are frequent due to the Downward Closure property. The frequent itemsets are found (N.Pasquier, 1999) to reduce the problem of association rules.

II. PREVIOUS WORK

In "A Review on Frequent Pattern Mining" by Vivek B. Satpute proposed that Pattern mining in recent times achieved major importance in the data mining community for the reason of its ability of being used as very important tool for the knowledge discovery and its applicability in the other data mining jobs like classification and clustering. Association rules are always of interest to the both database community as well as data mining users. Here a survey have provided of previous studies made in this area and recognize some vital gaps available in the current knowledge.

In "Analysis of Frequent Item sets and Pattern Sets Mining Algorithms" by Javeriya Naaz Ishtiyaque Syed and Rajeshri R.Shelke, proposed Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters and many more. Many researchers invented ideas to generate the frequent item sets. The time required for generating frequent item sets plays an important role. Some algorithms are designed, considering only the time factor. Their study includes depth analysis of algorithms and discusses some problems of generating frequent item sets (pattern sets) from the algorithm. The unifying feature among the internal working of various mining algorithms is explored. Some Frequent pattern mining often produces a large number of frequent patterns, which imposes a great challenge on visualizing, understanding and further analysis of the generated patterns. This emerges the need for finding small number frequent occurring patterns. In this paper, they explained the basic frequent item set, pattern sets mining problems. They described the main techniques used to solve these problems and give a comprehensive survey of the most influential algorithms that were proposed during the last decade.

In "Efficient Analysis of Pattern and Association Rule Mining Approaches" by Thabet Slimani and Amor Lazzez, proposed the process of data mining produces various patterns from a given data source. The most recognized data mining tasks are the process of discovering frequent item sets, frequent sequential patterns, frequent sequential rules and frequent association rules. Numerous efficient algorithms have been proposed to do the above processes. Frequent pattern mining has been a focused topic in data mining research with a good number of references in literature and for that reason an important progress has been made, varying from performant algorithms for frequent item set mining in transaction databases to complex algorithms, such as sequential pattern mining, structured pattern mining, correlation mining. Association Rule mining (ARM) is one of the utmost current data mining techniques designed to group objects together from large databases aiming to extract the interesting correlation and relation among huge amount of data. In this article, they provide a brief review and analysis of the current status of frequent pattern mining and discuss some promising research directions. Additionally, this paper includes a comparative study between the performances of the described approaches.

III. ALGORITHM USED

A) MinRPset Algorithm:

Here users are allowed to relax the conditions in the problem definition to further reduce the number of representative patterns. Hence this approach is a very flexible approach to find the representative patterns. MinRPset produces the smallest solution that can be possibly taken in practice under the given problem setting, and it takes a reasonable amount of time to finish when the number of frequent closed patterns is below one million. MinRPset is very space-consuming and time-consuming on some dense datasets when the number of frequent closed patterns is large.

Let F be the set of frequent patterns in a dataset D with respect to threshold min_sup, and ˆF be the set of patterns with support no less than min_sup · (1−€) in D. Obviously, F ⊆ ˆF . Given a pattern X ∈ ˆF, we use C(X) to denote the set of frequent patterns that can be _-covered by X. Here, C(X) ⊆ F. If X is frequent, then X ∈C(X). A straightforward algorithm for finding a minimum representative pattern set works as follows.

First mine all patterns in ˆ F, and then generate C(X)—the set of frequent patterns that X covers—for every pattern X ∈ ˆF. Then | ˆF | sets are derived. The elements of these sets are frequent patterns in F. Let S = {C(X) |X ∈ ˆF}. Finding a minimum representative pattern set is now equivalent to finding a minimum number of sets in S that can cover all the frequent patterns in F. This is a set cover problem, and it is NP hard. The well-known greedy algorithm is used to solve the problem, which achieves an approximation ratio of ∑ k i=1(1/i), where k is the maximal size of the sets in S. This simple algorithm is called MinRPset. The greedy algorithm is essentially the best-possible polynomial time approximation algorithm for the set cover problem.

It shows that it usually takes little time to finish. Generating C(X)s is the main bottleneck of the MinRPset algorithm when F and ˆF are large because there is a need to find C(X)s over a large F for a large number of patterns in ˆF .
The following techniques are used to improve the efficiency of MinRPset:

1) Consider closed patterns only;
2) Use a structure called CFP-tree to find C(X)s efficiently; and
3) Use a light-weight compression technique to compress C(X)s. The number of frequent closed patterns can be orders of magnitude smaller than the total number of frequent patterns. Consider only closed patterns improve the efficiency of the MinRPset algorithm in two aspects. On one hand, it reduces the size of individual C(X)s since now they contain only frequent closed patterns. On the other hand, it reduces the number of patterns whose C(X) needs to be generated as now we need to generate C(X)s for closed patterns only.

Algorithm:

1: Mine patterns with support ≥ min_sup · (1−€) and store them in a CFP-tree;
2: DFS_Search_CXs (root);
3: Remove non-closed entries from C(X)s;
4: Apply the greedy set cover algorithm on C(X)s to find representative patterns and output them;
When €=0, the representative patterns are closed frequent patterns.

B) FlexRPSet Algorithm:

Match all query tokens – Only documents that contain all the query tokens are included in the matched list. Here, co-occurrence measures using page counts are defined. How to extract clusters of patterns from snippets to represent numerous semantic relations that exist between two words is shown.

Algorithm:

Input:
- Input: cnode is a CFP-tree node; //cnode is the root node initially.
- K is the minimum number of times that a frequent closed pattern needs to be covered;
- Output: C(X)s;

Description:
1: for each entry E ∈ cnode from left to right do
2: if E is not marked as non-closed then
3: if E.child 6= NULL then
4: Flex Search CXs (E.child);
5: if E is more frequent than its child entries then
6: if (E is frequent AND E is covered less than K times) OR (∃ an ancestor entry E′ of E such that E′is frequent, E′ can be ϱ-covered by E and E′ is covered less than K times) then
7: X=E.pattern;
8: C(X) = Search CX (root, X, E.support);

IV. PROPOSED DESIGN

1) Dataset Training:
To train the dataset to the search based on the pattern to retrieve a collection of documents related to the query pattern. After a query is submitted to a search engine, a list of Web snippets is returned to the user. Assume that if a keyword/phrase exists frequently in the Web-snippets of a specified query, it represents an important concept related to the query because it coexists in close proximity with the query in the top documents.

2) Pattern Generation:
All words from the text are not considered as Pattern. Usually some words occur frequently in almost all of the documents. Because of this property, their discrimination power is negligible. These types of words are called stop words and these words can be filtered out during patternization.

3) Flex RP Set:
Match all query tokens – Only documents that contain all the query tokens are included in the matched list. Here, co-occurrence measures using page counts are defined. How to extract clusters of patterns from snippets to represent numerous semantic relations that exist between two words is shown.

4) FP-Growth:
FP-growth is a method that mines the whole set of frequent item sets with no candidate generation. FP-growth is founded on theory of the divide and-conquer. The first scan of the database gives a list of frequent items. In that list the items are arranged by descending order of frequency.

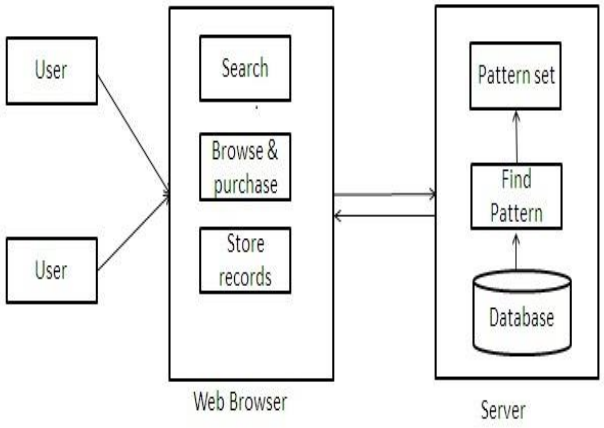

Fig 1: System Architecture Diagram

V. OBJECTIVE OF PROPOSED WORK

- To build an efficient scalable algorithm an efficient and scalable algorithm for incremental mining problem for frequent item sets. In prior approaches like Apriori algorithm and FP-Growth, whole process is to be re-initialized from base and is not suitable for dynamic analysis.
- To study various pattern mining algorithms that can be used for discovering knowledge (patterns) from human interactions, also to study and compare various association rule mining algorithms.
- To minimize the cost of finding frequent item set using the new approach of FP-Growth i.e. association rule.
- Reduce running time and memory usage.
- Have good efficiency.

VI. CONCLUSION

A large number of frequent pattern sets is used to find the frequent closed pattern sets. A frequent closed pattern generates a representative pattern sets. For finding a representative pattern set two algorithms MinRPset and FlexRPset are used. MinRPset and FlexRP set is used to keep record of the frequent closed pattern set which provides the best approximate result. MinRPset becomes slow when large amount of frequent closed pattern sets that consumes more memory space. MinRPset is expensive than RPlocal set. To overcome this problem, FlexRPset is used. FlexRPset adds extra parameter value to find the minimum number of representative pattern sets by covering all frequent closed pattern sets. Thus, to study and use FP-growth method that mines the whole set of frequent itemsets with no candidate generation and find frequent pattern sets by using the technique of FP-Growth i.e. association rule. We are using FP growth algorithm to determine the items searched by user frequently, using these searches or items bought we are finding the pattern set. Many researchers' unreal the idea to get the frequent item sets/ pattern sets. Some frequent pattern mining often produces a large number of frequent patterns, which imposes a great challenge on visualizing, understanding and further analysis of the generated patterns. This emerges the need for finding small number frequent occurring patterns. The time needed for generating frequent pattern sets plays associate important role. Some algorithms are designed, considering solely the time issue.

VII. REFRENCES

[1] Vivek B. Satpute, "A Review on Frequent Pattern Mining", in International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014, ISSN 2091-2730.

[2] Javeriya Naaz Ishtiyaque Syed, Rajeshri R.Shelke, "Analysis of Frequent Item sets and Pattern Sets Mining Algorithms", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 3 Issue: 2, page no: 249-253.

[3] Thabet Slimani, Amor Lazzez , "Efficient Analysis of Pattern and Association Rule Mining Approaches", College of Computer Science and Information Technology , Taif University , KSA.

[4] Guimei Liu, Haojun Zhang, and Limsoon Wong, "A Flexible Approach to Finding Representative Pattern Sets", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.

[5] K. Kavitha, C. Anand, "A Novel Approach in Data Mining for Representative Pattern Sets", International Conference on Science, Technology, Engineering and Management [ICON-STEM'15], Journal of Chemical and Pharmaceutical Sciences, ISSN: 0974-2115.

