International Conference on Information Engineering, Management and Security
2015 [ICIEMS 2015]

ISBN	978-81-929742-7-9	**VOL**	01
Website	www.iciems.in	**eMail**	iciems@asdf.res.in
Received	10 - July - 2015	**Accepted**	31- July - 2015
Article ID	ICIEMS020	**eAID**	ICIEMS.2015.020

MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM

A. Jalpa A Varsur[1], Nikul G Virpariya[2]
[1]Student, M.E.(CSE), Gardi Vidyapith, Gujarat, India
[2]Assistant Professor, CSE Department, Gardi Vidyapith, Gujarat, India

ABSTRACT: *Frequent weighted itemset represent correlation frequently holding in data in which items may weight differently.However,in some contexts,e.g., when the need is to minimize a certain cost function,discovering rare data correlations is more interesting than mining frequent ones. Now a days, some uninteresting items are consider as important.which are appear very less in the database.It is known as Rare or infrequent item.Finding infrequent itemset is useful in many recent field like medical,web,cloud,market basket analysis etc,. for i.e.,in market basket analysis some set of items such as milk and bread customer buy frequently compared to milk and bread gold chain and ring are infrequent items.This paper takes the issue of discovering rare itemset in transaction dataset.*

KEYWORDS: Data Mining,Frequent item,Infrequent item,Threshold,Support,Item weight.

I. INTRODUCTION

Itemset is a set of items.Frequent items are appear very frequently in database,with high support and high confidence. Rare items are reverse from of frequent ones,it has low support and high confidence. [1]That has a vast range of real life application like,
In **Medical** – if we identify the solution of rare disease then we can prevent the person to get affected by it.Here,we don't need this rare disease to become frequent. In **Marketing** strategy knowing the rare items can help us to make them the frequent ones.So,the businessmen gains profit. In **Market basket analysis** for finding which items tend to be purchased together,milk and bread occurs frequently and can be considered as regular case,some items like bad and pillow are infrequently associated itemsets.[1] Recently,the importance is being given for the discovery of infrequent or exceptional patterns[1].threshold plays key role for finding of frequent and infrequent itemsets.The things which are done together for eg.Buying groceries known as association,occurs between items.Association is an implication of the form X→Y.Infrequent itemsets are produced from very vast or enormous datasets.Extraction of frequent itemset is a necessary step in many association techniques[2]. Association rule mining extracts interesting correlation between transactions.In many application some items are appear very frequently in the data,while others rarely appear.if minsup is set too high,those rules that involve both frequent and infrequent items.To find the rule that contain both frequent and rare minsup is set to be very low.This may cause combinatorial explosion for those frequent items will be associated with one another in all possible ways.This problem is called the rare item problem[3]. infrequent itemset do not comprise any infrequent subset. it appears only when threshold is set to very low.The mining algorithm solves the problem of discovering the infrequent itemsets in the given database[2].Infrequent Itemset Mining finds uninteresting items among the huge database.

Cite this article as: A. Jalpa A Varsur, Nikul G Virpariya. "MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM." *International Conference on Information Engineering, Management and Security (2015)*: 121-128. Print.

The Pattern Growth algorithm comes in the early 2000s, for the answer to the problem of generates and test.The main idea is for to avoid the candidate generation step altogether, and to concentrate the search on a specified portion of the initial database.

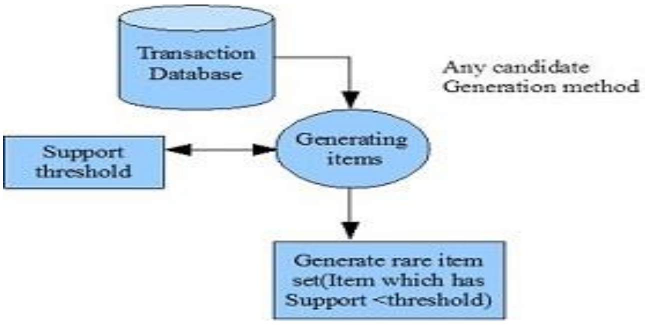
Transaction Database
Any candidate Generation method
Support threshold
Generating items
Generate rare item set(Item which has Support <threshold)

Figure 1: Rare itemset mining architecture [9]

II. RELATED WORK

For example, the rule {Bread, Butter} =>{Milk} found in the sales data of a shop would indicate that if a customer buys bread and butter together, he or she is likely to also buy milk. Such information can be used in decision making about marketing policies such as, e.g., product offers, product sales and discount schemes. In addition to the above mentioned example association rules are used today in many application areas including Web usage mining, Intrusion detection, Continuous production, and Bioinformatics [3]. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

The problem of association rule mining [3] is defined as: Let I= {i1, i2,…, in} be a set of n binary attributes called *items*. Let D={t1,t2,…,tm} be a set of transactions called the *database*. Each transaction in database D has a unique transaction identity ID and contains a subset of the items in I [3]. A *rule* is defined as an implication of the form X=>Y where X,Y is subset of I and X intersection Y = Null Set. The sets of items (for short *itemsets*) X and Y are called *antecedent* (if) and *consequent* (then) of the rule respectively.[6]

III PROBLEM DEFINITION

To understand the background of itemset mining, we present different techniques and algorithm in the following subsections.

Goal: Mining infrequent itemset from transaction datasets.

I={i1,i2,..,im}be a set of data items.A transactional dataset T={t1,t2,…,tn}is a set of transactions,where each transaction tq (q € [1,n])is a set of items in I and is characterized by a transaction ID (tid).An itemset I is a set of data items[6].Specifically we denote as k-itemset a set of k-items in I.The support of an itemset is the number of transactions containing I in T.

An itemset I is infrequent if its support is less than or equal to a predefined maximum support threshold ξ.Otherwise,it is said to be frequent[1].

Weighted transactional data set

Let I ={i1,i2,…,im} be a set of items.A weighted transactional data set T is a set of weighted transactions ,where each weighted transaction tq is a set of weighted items <ik,wk>

Weights could be either positive,null or negative numbers.itemsets mined from weighted transactional data sets are called weighted itemsets.

Their expression is similar to the one used for traditional itemsets,i.e., a weighted itemset is a subset of the data items occurring in a weighted transactional data set.The problem of mining itemsets by considering weights associated with each item is known as the weighted itemset mining problem[4].

This approach is focuses on considering item weights in the discovery of infrequent itemsets.To this aim,the problem of evaluating itemset significance in a given weighted transactional data set is addressed by means of a two-step process.

Firstly,the weight of an itemset I associated with a weighted transaction tq € T is defined as an aggregation of its item weights in tq.Secondly,the significance of I with respect to the whole data set T is estimated by combining the itemset significance weights associated with each transaction.

TID	CPU Usage Readings
1	<a,0> <b,100> <c,57> <d,71>
2	<a,0> <b,43> <c,29> <d,71>
3	<a,43> <b,0> <c,43> <d,43>
4	<a,100> <b,0> <c,43> <d,100>
5	<a,86> <b,71> <c,0> <d,71>
6	<a,57> <b,71> <c,0> <d,71>

Figure 3: TABLE Weighted Transactional Data Set

The significance of a weighted transaction, i.e., a set of weighted items, is commonly evaluated in terms of the corresponding item weights. For instance, when evaluating the support of {a,b} in the example data set reported in Table 1, the occurrence of b in tid 1, which represents a highly utilized CPU, should be treated differently from the one of a, which represents an idle CPU at the same instant.

Task (A) entails discovering IWIs and minimal IWIs (MIWIs) which include the item with the least local interest within each transaction. Table 2 reports the IWIs mined from Table 1 by enforcing a maximum IWI-support-min threshold equal to 180 and their corresponding IWI-support-min values. For instance, {a,b} covers the transactions with tids 1, 2, 3, and 4 with a minimal weight 0 (associated with a in tids 1 and 2 and b in tids 3 and 4), while it covers the transactions with tids 5 and 6 with minimal weights 71 and

57, respectively.Hence, its IWI-support-min value is 128.

IWI	IWI-Support-min	IWI	IWI-Support-min
{c}	172 (Minimal)	{a,b,c}	0 (Not Minimal)
{a,b}	128 (Minimal)	{a,b,d}	128 (Not Minimal)
{a,c}	86 (Not Minimal)	{a,c,d}	86 (Not Minimal)
{b,c}	86 (Not Minimal)	{b,c,d}	86 (Not Minimal)
{c,d}	172 (Not Minimal)	{a,b,c,d}	0 (Not Minimal)

TABLE IWIs Extracted from the Data Set from above Table

Maximum IWI-support-max threshold = 390

IWI	IWI-Support-max	IWI	IWI-Support-max
{a}	286 (Minimal)	{a,c}	0 (Not Minimal)
{b}	285 (Minimal)	{b,c}	128 (Not Minimal)
{c}	172 (Not Minimal)		

TABLE IWIs Extracted from the Data Set from above Table

Base Algorithm

Input: T,a weighted transaction dataset ,

Cite this article as: A. Jalpa A Varsur, Nikul G Virpariya. "MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM." *International Conference on Information Engineering, Management and Security (2015)*: 121-128. Print.

maximum IWI support threshold
Output: set of IWIs
1. initialization of items.
2. count IWI-support
3. tree -a new empty tree
4. for all weighted transaction tq in T do,
5. TEq -equivalence transaction
6. for all transaction insert tej in tree
7. end for
8. end for
9. IWI Mining
10. return set of IWIs

IV. PROPOSED ALGORITHM

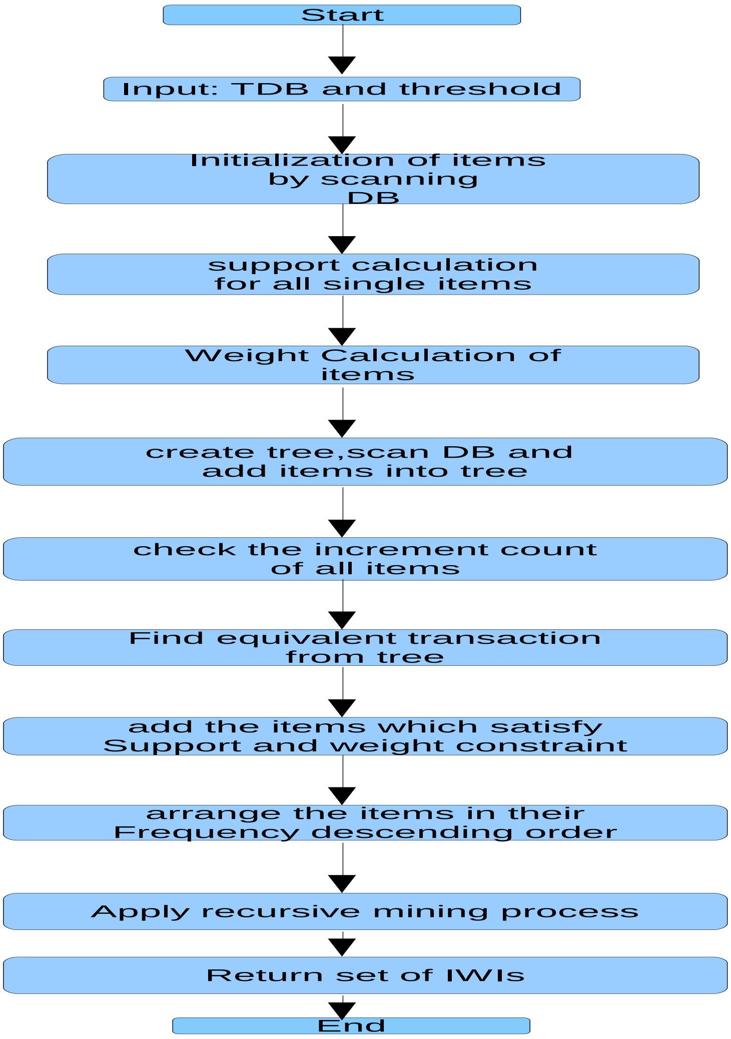
Start
Input: TDB and threshold
Initialization of items by scanning DB
support calculation for all single items
Weight Calculation of items
create tree,scan DB and add items into tree
check the increment count of all items
Find equivalent transaction from tree
add the items which satisfy Support and weight constraint
arrange the items in their Frequency descending order
Apply recursive mining process
Return set of IWIs
End

Figure 4: Proposed Algorithm Design

Proposed Algorithm Steps:

Input : - T (transaction database TDB) ,
ξ (maximum IWI-support threshold)
Output :- Ɉ (set of IWIs)

Step 1: Ɉ = Ø /* initialization of items by scanning DB */
Step 2: count the support for all single items

$$\text{Support} = \frac{(X \cup Y).\text{count}}{n}$$

Cite this article as: A. Jalpa A Varsur, Nikul G Virpariya. “MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM.” *International Conference on Information Engineering, Management and Security (2015)*: 121-128. Print.

$$\text{Confidence} = \frac{(X \cup Y).\text{count}}{\text{c.count}}$$

Step 3: weight calculation of all items.

- » Generation of MIS equation will be used for weight calculation,we take MIW instead of MIS
- » $MIW(i) \equiv \begin{cases} M(i) & \text{where } M(i) > LS \\ LS & \end{cases}$ LS otherwise
- » $M(i) \equiv \beta f(i)$
- » here, f (i) is the actual frequency of item i in data or the support expressed in percentage of the data set size
- » LS = user specify lowest minimum weight
- » β = a parameter to control MIW value for items
- » If $\beta = 0$ we have only one minimum weight.
- » If $\beta = 1$ and $f(i) \geq LS$
- » $f(i)$ is the MIW value for i.

Step 4: create initial fp-tree

- » add items into tree
- » for all transaction ti ∈ T
- » for all tej ∈ ti
- » Insert tej in tree

Step 5: If tej.sup ≤ tsup && tej.weight ≥ tweight

- » If item does not satisfy the Weight Constrain and support constrain then remove it from transaction

Step 6: Order items in their frequency. descending order.
Step 7: Ɉ ← Mining process
Step 8: return set of infrequent items
Step 9: end

Advantages

It will Reduce time.
It Require less memory.
It removes the frequent items from tree.

Limitation

Number of database scan are higher
It performs save procedure each time when particular item is found.
So execution time grows higher due to this limitation.

V. STUDY OF TOOL

Java Technology

JAVA is an object oriented, platform independent and middle level language.It contains JVM (Java Virtual Machine) which is able to execute any programmore effciently.The feature of Platform Independence makes it different from the other Technologiesavailable today.

Eclipse Tool

Eclipse is an integrated development environment (IDE).It contains a base workspace and an extensible plug-in system for customizingthe environment.Eclipse is written mostly in Java and thus can be used to develop applications.
Eclipse started as a proprietary IBM product (IBM Visual age for Smalltalk/Java)

SPMF

SPMF is an **open-source data mining mining library** written in **Java**, specialized in **pattern mining.**It is distributed under the **GPL v3 license.**It offers implementations of **78 data mining algorithms** for:

- **sequential pattern mining,**
- **association rule mining,**
- **frequent itemset mining,**

Cite this article as: A. Jalpa A Varsur, Nikul G Virpariya. "MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM." *International Conference on Information Engineering, Management and Security (2015)*: 121-128. Print.

- **high-utility pattern mining,**
- **sequential rule mining,**
- **clustering.**

The source code of each algorithm can be integrated in other Java software.Moreover, SPMF can be used as a standalone program with a simple user interface or from the command line.The current version is **v0.96r16** and was released the **28th April 2015.**

V. RESULT ANALYSIS

Aggregate function

Aggregate function is a function where the values of multiple rows are grouped together as input or certain criteria to from a single value or more significant meaning or measurement such as a set,a bag or a list.for eg. Function like average( ),count ( ),maximum ( ). It will returns a single value.

Reducing the execution time

First scan of data will remove the frequent items and reduce the no. of scans .by tree pruning strategy it will find the prunable items and reduce the time

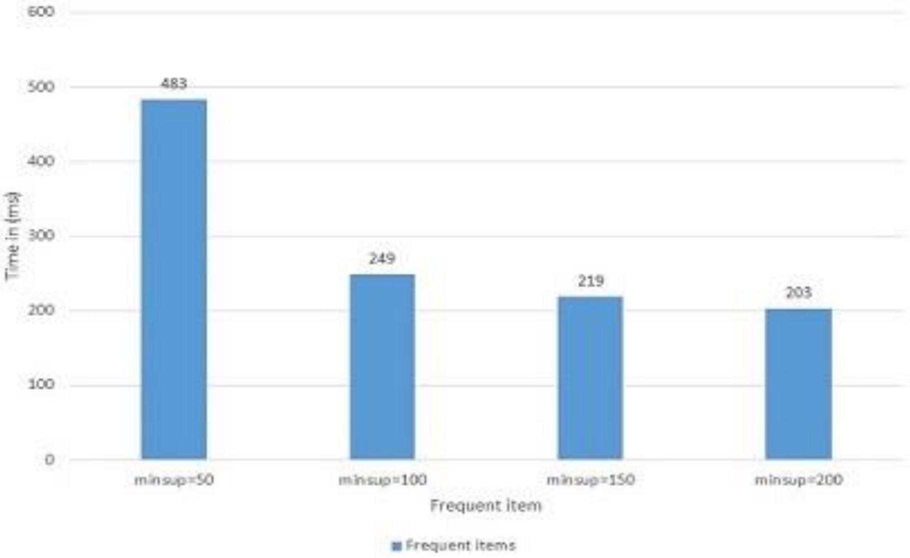

Figure 5: Performance of different threshold values

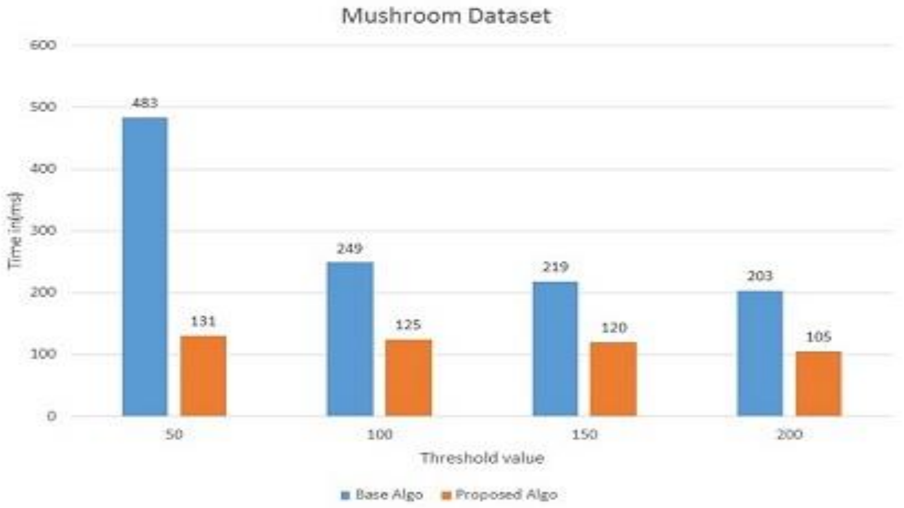

Figure 6: Execution time comparison over mushroom dataset

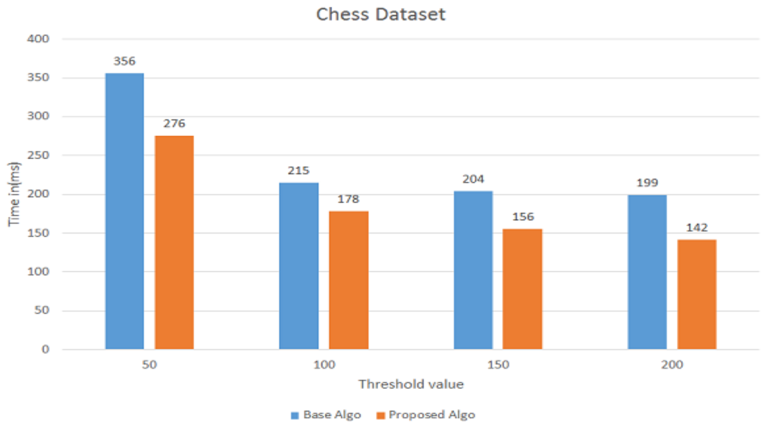

Figure 7: Execution time comparison over chess dataset

The algorithm first constructs the tree out of the original data set and then grows the frequent patterns.For a faster execution,the data should be preprocessed before applying the algorithm.

VI. CONCLUSION AND FUTURE WORK

The Propose System has improves the performance of IWI mining algorithm by using FP-Growth Structure.
It reduce the execution time.
at mining time tree will remove the frequent items and we get only rare items.
Thus,we conclude that the proposed system has better performance and it will require less memory.

Future Work:

As future,we plan for discovering rare itemset ,weight calculation of items will be done by user.IWI algorithm can also be implemented in advanced decision making system and business intelligence.

ACKNOWLEDGEMENT

We are deeply indebted & would like to express gratitude to our thesis guide Prof. Nikul Virpariya , B. H. Gardi College of Engineering & Technology for his great efforts and instructive comments in the dissertation work.

We would also like to extend our gratitude to Prof.Hemal Rajyaguru, Head of the Computer Science & Engineering Department, B. H. Gardi College of Engineering & Technology for his continuous encouragement and motivation.

We would also like to extend our gratitude to Prof. Vaseem Ghada, PG Coordinator, B. H. Gardi College of Engineering & Technology for his continuous support and cooperation.

We should express our thanks to our dear friends & our classmates for their help in this research; for their company during the research, for their help in developing the simulation environment.

We would like to express our special thanks to our family for their endless love and support throughout our life. Without them, life would not be that easy and beautiful.

REFERENCES

[1]K.S.Sadhasivam,Tamilarasi,“Mining Rare ItemSet with Automated Support Thresholds”,*Journal of Computer Science*,pp.394-399,2011.
[2]Lugi Troiano,Cosimo Birtolo,“A Fast Algorithm for Mining Rare ItemSets”,*IEEE Ninth International Conference on Intelligent Systems Design and Applications*,2009.
[3]Mehdi Adda,Lei Wu,“Rare ItemSet Mining”,*IEEE Sixth International Conference on Machine Learning and Applications*,2007.

Cite this article as: A. Jalpa A Varsur, Nikul G Virpariya. “MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM.” *International Conference on Information Engineering, Management and Security (2015)*: 121-128. Print.

[4]Petko Valtchev,Amedeo Napoli,“Towards Rare ItemSet Mining”,*19th IEEE International Conference on Tools with Artificial Intelligence*,2007.

[5]K.Sun,Fengshan Bai,“Mining weighted Association Rules without Preassigned Rules”,*IEEE Transactions On Knowledge and Data Engineering*,vol.20,No.4,April 2008.

[6]Luca Cagliero,Paolo Garza,“Infrequent Weighted ItemSet Mining Using Frequent Pattern Growth”,*IEEE Transactions On Knowledge and Data Engineering*,vol.26,No.4,April 2014.

[7]Gou Masuda,Norihiro Sakamoto,“A Framework for Dynamic evidence based medicine using data mining”,*CBMS’02:Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems*,2002.

[8]H.Yun,D.Ha,“Mining association rules on significant rare data using relative support”,*The Journal of Systems and Software*,pp.181-191,2003.

[9]Nidhi Sethi,Pradeep Sharma,“Efficient Algorithm for Mining Rare Itemsets over Time Variant Transactional Database”,*International Journal of Computer Science and Information Technologies*,vol.5,2014.

[10]J.Jenifa,Dr.V.Sampath Kumar,“Study on predicting various Mining Techniques Using weighted Itemsets”,*IOSR*,vol.9,pp.30-39,Mar-Apr 2014.

Cite this article as: A. Jalpa A Varsur, Nikul G Virpariya. “MINING RARE ITEMSET BASED ON FP-GROWTH ALGORITHM.” *International Conference on Information Engineering, Management and Security (2015)*: 121-128. Print.