International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 12, Issue 2 (February 2016), PP.41-47

A Novel Method based on AND logic for Frequent, Infrequent and Non-present Item set mining in Transactional Data bases

Sujatha Kamepalli[1], Raja Sekhara Rao Kurra[2] and Sundara Krishna.Y.K[3]
[1]*Research Scholar, CSE Department, Krishna University Machilipatnam, Andhra Pradesh, India*
[2]*Director, Sri Prakash College of Engineering, Thuni, Andhra Pradesh, India.*
[3]*Professor, CSE Department, Krishna University Machilipatnam, Andhra Pradesh, India.*

Abstract:- Association rule mining is one of the most important techniques used for finding the correlations among different fields in large databases.it finds the frequent, infrequent and non-present item sets from the relational data bases. Most of the traditional techniques are useful in finding the frequent item sets. But infrequent patterns and non-present patterns plays vital role in different fields such as marketing, medical, fraud detection and etc. A new method is proposed which finds all frequent, infrequent and non-present item sets. This method contains two steps. In first step the frequency table is constructed based on the given transactional data base. In the second step the algorithm forwarded for finding the frequent, infrequent and non-present item sets. This method is based on AND logic and it works efficiently in finding the patterns when compared to the traditional methods.

Keywords:- Data mining, Association rule mining, frequent, infrequent, non-present item sets.

I. INTRODUCTION

Data mining allows users to understand and discover knowledge in large amounts of data by mining data patterns (or simply called patterns) [1] [2] [3] [4]. Nowadays, with the rapid development of information technology, especially the web service-based application, service-oriented architecture and cloud-computing, continually expanding data are integrated to generate useful information. Many techniques have been used for data mining. Association rules mining (ARM) is one of the most useful techniques. It is the process of finding correlations or patterns among number of fields in large relational databases. It is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. It also enables them to determine the impact on sales, customer satisfaction, and corporate profits.

A pattern can be any type of regularity that appears in data collections, which are considered a kind of summary of the input data [5]. For example, a set of frequent bought together products from a shopping basket analysis, a piece of abnormal gene sequence carried by patients for drug research, a historical record of a visitor's past traveling experiences for planning the next trip, or the reaction of a particular enzyme to the external stimulus for the study of disease treatment. All of these patterns carry useful insights from the collected data and have the potential to solve the problems that occur in practical applications. Pattern mining is a mining process for extracting these valuable data patterns from large amounts of data [3].

The challenges associated with ARM, especially for parallel and distributed data mining, include minimizing I/O, increasing processing speed and reducing communication cost [8]. A major concern in ARM today is to continue to improve algorithm performance. The Apriori-based algorithms find frequent item sets based upon an iterative bottom-up approach to generate candidate item sets. Since the first proposal of association rules mining by R. Agrawal [6, 7],

Often when considering data mining, the focus is on frequent patterns. Although the majority of the most interesting patterns will lie within the frequent ones, there are important patterns that will be ignored with this approach. These are called infrequent patterns.
Take for example the sale of VHS:s and DVD:s. There will be low occurrences of people buying both of them. In terms of data mining, the item set {VHS, DVD} will be infrequent and therefore ignored. However, people that buys DVD:s does not tend to buy VHS:s and vice versa. These items will be competing and an interesting pattern is found [14].

Finding frequent item sets is one of the most investigated fields of data mining. The problem was first presented in paper mining association rules between sets of items in large databases by Agrawal [10]. To analyze the huge amount of data thereby exploiting the consumer behavior and make the correct decision leading to competitive edge over rivals [11]. Frequent item sets are appear in a data set with frequency no less than a user-specified threshold. Frequent item sets play an essential role in many data mining tasks that try to find interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems. Also sequential association rule mining is one of the possible methods to analysis of data used by frequent item sets [12].

II. NEED OF FREQUENT ITEM SET MINING

Studies of frequent item set (or pattern) mining is acknowledged in the data mining field because of its broad applications in mining association rules, correlations, and graph pattern constraint based on frequent patterns, sequential patterns, and many other data mining tasks. Efficient algorithms for mining frequent item sets are crucial for mining association rules as well as for many other data mining tasks. The major challenge found in frequent pattern mining is a large number of result patterns. As the minimum threshold becomes lower, an exponentially large number of item sets are generated. Therefore, pruning unimportant patterns can be done effectively in mining process and that becomes one of the main topics in frequent pattern mining. Consequently, the main aim is to optimize the process of finding patterns which should be efficient, scalable and can detect the important patterns which can be used in various ways [13].

III. NEED OF INFREQUENT ITEM SET MINING

Some infrequent patterns may also suggest the occurrence of interesting rare events or exceptional situations in the data. For example, if {Fire = Yes} is frequent but {Fire = Yes, Alarm = On} is infrequent, then the latter is an interesting infrequent pattern because it may indicate faulty alarm systems. To detect such unusual situations, the expected support of a pattern must be determined, so that, if a pattern turns out to have a considerably lower support than expected, it is declared as an interesting infrequent pattern.

Mining infrequent patterns is a challenging endeavor because there is an enormous number of such patterns that can be derived from a given data set. More specifically, the key issues in mining infrequent patterns are: (1) how to identify interesting infrequent patterns, and (2) how to efficiently discover them in large data sets. To get a different perspective on various types of interesting infrequent patterns, two related concepts are negative patterns and negatively correlated patterns [15].

IV. RELATED WORK

Frequent pattern mining was first proposed by Agrawal et al.[10] form market basket analysis in the form of association rule mining. It analyses customer buying habits by finding associations between the different items that customers place in their "shopping baskets". For instance, if customers are buying milk, how likely are they going to also buy cereal (and what kind of cereal) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and arrange their shelf space. Jiawei Han at all explains about frequent pattern mining and its future directions. Ashish Gupta at all proposed algorithms for Minimally Infrequent Itemset Mining using Pattern-Growth Paradigm and Residual Trees. Alex Tze Hiang Sim at all explains about Mining Infrequent and Interesting Rules from Transaction Records. Laszlo Szathmary at all describes about Generating Rare Association Rules Using the Minimal Rare Item sets Family.

V. PROPOSED METHOD

The proposed method is based on AND logical operation. This method contains two steps. In first step the frequency table is constructed based on the given transactional data base. In the second step the algorithm forwarded for finding the frequent, infrequent and non present item sets. For explaining the proposed method the following transactional data base is taken as an example.

TID	ITEMSET
T100	I1,I4,I3
T200	I1,I5
T300	I1,I2,I3
T400	I2,I4,I5
T500	I3,I5
T600	I3,I4

Consider T= {T1, T2, T3, T4….} is the set of transactions and I= {I1, I2, I3…….} is the set of items.

Algorithm:
Input: given a transactional database which contains m transactions m>=1 and each transaction contain some set of items. Maximum negative count (MNC).
Output: frequent, infrequent and non present item sets (FI, IFI, and NPI).

Method:
//construction of a frequency table
For each (t ϵ T)
{
For each (i ϵ I)
{
If (i ϵ t)
Return 1;
Else
Return 0;
}
}
//finding frequent, infrequent, non present item sets.
//Initialize FI, IFI, NPI.
FI=ϴ, IFI= ϴ and NPI=ϴ
For each (k-item set) (1<=k<=n)
{
If ((i1^i2^i3^i4^…….^ik)==0)
Count=count+1;
If(count==0)
NPI= NPI Ụ {k-item set}
Else if (count<=MNC)
IFI= IFI Ụ {k-item set}
Else
FI=FI Ụ {k-item set}
}

Explanation:
Step1: consider the maximum negative count
Step2: construct the frequency table corresponding to each item I in the given item set. This frequency table is a Boolean table in which the value of corresponding becomes one if it is in the given transaction in the given set of transactions otherwise if it is not present then the value corresponding to item becomes zero. For the given example the frequency table is given below.

	T1	**T2**	**T3**	**T4**	**T5**	**T6**
I1	1	1	1	0	0	0
I2	0	0	1	1	0	0
I3	1	0	1	0	1	1
I4	1	0	0	1	0	1
I5	0	1	0	1	0	1

Step3: identify the support counts of 1-item sets which can be obtained by counting the number of ones in each row in the frequency table.

1-item set	**Support count**
I1	3(frequent)
I2	2(infrequent)
I3	4(frequent)
I4	3(frequent)
I5	3(frequent)

Compare the support counts with maximum negative count.if the support is equal to zero then it is considered as non –present item set.if the support count is less than the maximum negative count then it is considered as infrequent item set otherwise the item set is frequent item set.

Step4: perform the AND operation among the corresponding Boolean variables of transactions of item sets to find non-present, infrequent and frequent k-itemsets.in every time a new frequency table is constructed by its previous frequency table using AND operation. New frequency table for 2-itemsets constructed from the frequency table of 1-item sets.

	T1	T2	T3	T4	T5	T6
I1,I2	0	0	1	0	0	0
I1,I3	1	0	1	0	0	0
I1,I4	1	0	0	0	0	0
I1,I5	0	1	0	0	0	0
I2,I3	0	0	1	0	0	0
I2,I4	0	0	0	1	0	0
I2,I5	0	0	0	1	0	0
I3,I4	1	0	0	0	0	1
I3,I5	0	0	0	0	0	1
I4,I5	0	0	0	1	0	1

2-item set	Support count
I1,I2	1(infrequent)
I1,I3	2(infrequent)
I1,I4	1(infrequent)
I1,I5	1(infrequent)
I2,I3	1(infrequent)
I2,I4	1(infrequent)
I2,I5	1(infrequent)
I3,I4	2(infrequent)
I3,I5	1(infrequent)
I4,I5	2(infrequent)

3-item set	Support count
I1,I2,I3	1(infrequent)
I1,I2,I4	0(non-present)
I1,I2,I5	0(non-present)
I1,I3,I4	1(infrequent)
I1,I3,I5	0(non-present)
I1,I4,I5	0(non-present)
I2,I3,I4	0(non-present)
I2,I3,I5	0(non-present)
I2,I4,I5	1(infrequent)
I3,I4,I5	1(infrequent)

This process can be continued till all the k-item sets are found. This process is very efficient since with one data base scan we can find all the k-item sets.

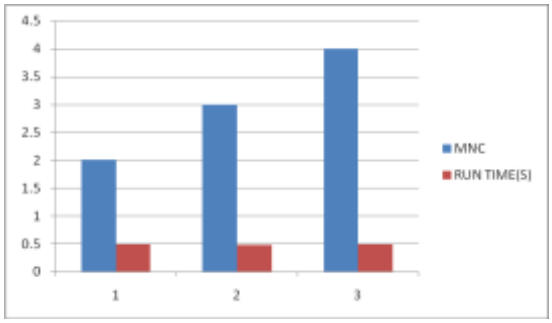
Experimental results:

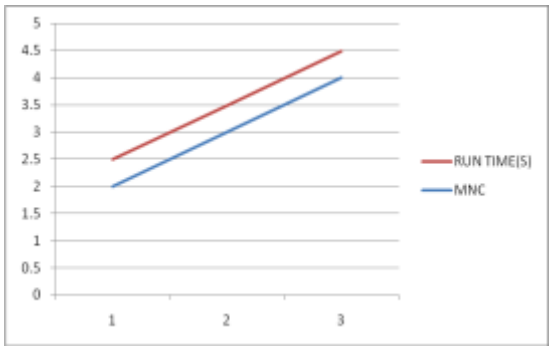
Experiments are conducted on synthetic dataset to study the performance of the proposed algorithm. The two synthetic databases termed as DB1, DB2are considered for the experiment purpose. The number of items and the number of transactions of these databases are shown in the following table.

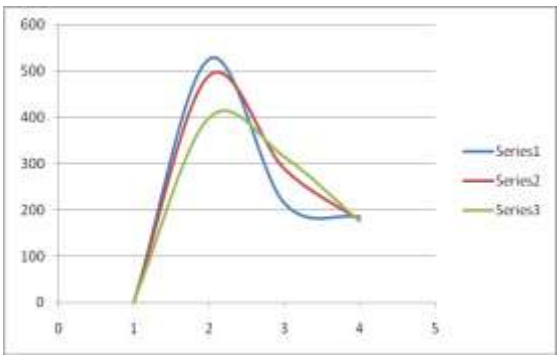
	DB1	DB2
Number of Items	10	10
Number of Transactions	1000	1500

	Frequent Items	Infrequent Items	Non-Present Items	Run Time(S)
DB1	400	240	198	0.61
DB2	475	255	163	0.66

MNC	Frequent Items	Infrequent Items	Non-Present Items	Run Time(S)
2	526	214	185	0.49
3	491	289	178	0.48
4	400	315	176	0.49







VI. CONCLUSION

More and more researchers have realized the importance of infrequent patterns and non-present patterns with the increasing demands in the applications of anomaly detection, fraud detection, and medical field and also in marketing. The proposed algorithm is used to find the frequent infrequent and non-present item sets from large transactional data bases efficiently. This approach uses the AND logic. This method works with one

data base scan. In future this algorithm can be extended to find the frequent, infrequent and non-present item sets in weighted transactional data bases also.

REFERENCES

[1]. K. Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons., 2003.
[2]. J. Han and M. Kamber. Data Mining: Concpets and Techniques. Academic Press, 2nd edition edition, 2006.
[3]. J. Vreeken. Making Pattern Mining Useful. PhD thesis, the Dutch Research School for Information and Knowledge Systems, Netherlands, 2009.
[4]. F. Hadzic, H. Tan, and T.S. Dillon. Mining of Data with Complex Structures, volume 333. Springer, 2010.
[5]. J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future direction. Data Mining and Knowledge Discovery, 15:55–86, 2007.
[6]. www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.html.
[7]. searchbusinessanalytics.techtarget.com/definition/association-rules-in-data.html
[8]. Agrawal R,Srikant R. "Mining sequential patterns" In the Proc.1995Int Conf.on Data Engineering,Taibei,Taiwan,March1995
[9]. Piatetsky-Shapiro G. Discovery, "Analysis and Presentation os Strong Rules, Knowledge Discovery in Databases", AAAI/MIT, Menlo Park, CA, 1991, pp.229-248.
[10]. Agrawal R., Imielienski T. and A. Swami, Mining Association Rules between Sets of Items in Large Databases, Proc. Conf. on Management of Data, 207–216 (1993)
[11]. Raorane A.A., Kulkarni R.V. and Jitkar B.D., Association Rule – Extracting Knowledge Using Market Basket Analysis, Res. J. Recent Sci., 1(2), 19-27 (2012)
[12]. Shrivastava Neeraj and Lodhi Singh Swati, Overview of Non-redundant Association Rule Mining, Res. J. Recent Sci., 1(2), 108-112 (2012)
[13]. Pramod S. and Vyas O.P., Survey on Frequent Item set Mining Algorithms, In Proc. International Journal of Computer Applications (0975 - 8887), 1(15), 86–91 (2010)
[14]. Mining Infrequent Pat terns johan b jarnle (johbj551) peterzhu (petzh912) lilin köpin university, 2009 tnm 033 data mining Chapter- 5 Infrequent Patterns shodha ganga.

AUTHOR'S PROFILE:

K. Sujatha is pursuing her Ph.D. in Krishna University,Machilipatnam, A.P. She is interested doing research in data mining. Present she is carrying her research in Infrequent Pattern Mining. She has 14 international journal publications with high impact factors and indexing. She has two national journal publications. She attended for Two AICTE sponsored 2-week workshops and attended for a number of FDPs. She is member in Indian Association of Engineers (IAE). She is Life Time Member in International Association of Engineering & Technology For Skill Development (IAETSD). She has a total of 11 years' experience in teaching. She is working as an Associate Professor in CSE Department, Malineni Lakshmaiah Engineering College, Singaraya konda, Prakasam District. A.P.

Prof. K. Rajasekhara Rao is director of Sri Prakash College of Engineering, Thuni, East Godavari District, Andhra Pradesh. He held several key positions in K.L.University, as Dean (Administration) & Principal, K L College of Engineering (Autonomous). Having more than 26 years of teaching and research experience, Prof. Rao is actively engaged in the research related to Embedded Systems, Software Engineering and Knowledge Management. He had obtained Ph.D in Computer Science & Engineering from Acharya Nagarjuna University (ANU), Guntur, Andhra Pradesh and produced 58 publications in various International/National Journals and Conferences. Prof.KRR was awarded with "Patron Award" for his outstanding contribution, by India's prestigious professional society Computer Society of India (CSI) for the year 2011 in Ahemadabad. He has been adjudged as best teacher and has been honored with "Best Teacher Award", seven times. Dr. Rajasekhar is a Fellow of IETE, Life Member's of IE, ISTE, ISCA & CSI (Computer Society of India). Dr.Rajasekhar is nominated as sectional committee member for Engineering Sciences of 100th Annual Convention of Indian Science Congress Association. He has been the past Chairman of the Koneru Chapter of CSI.

Dr. Y. K. Sundara Krishna qualified in Ph.D. in Computer Science and Engineering from Osmania University, Hyderabad. Now, he is working as Professor in the Department of Computer Science, Krishna University, Machilipatnam and presently holding several key positions in Krishna University. His research interests are Mobile Computing, Service Oriented Architecture and Geographical Information Systems and having practical work experience in the areas of Computing Systems including Developing Simulators for Distributed Dynamic Cellular Computing Systems, Applications of Embedded and Win32 clients, Maintenance of Multi-user System Software. He has about 24 international publications and about 2 national publications. Has attended about 5 international conferences and 25 national conferences.