ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Discovering Frequent Itemsets Using Fast Apriori Algorithm

M. Premalatha[1], T. Menaka[2]

M.Phil Scholar, Dept. of Computer Science (Aided), NGM College, Pollchi, Tamilnadu, India[1]

Assistant Professor, Dept. of Computer Science (SF), NGM College, Pollachi, Tamilnadu, India[2]

ABSTRACT: Utility mining is the application of data mining techniques to discover patterns from the datasets. Itemset extraction and utility mining is characterized by a frequency analysis where the item values correspond to the number of times that term appears in the database. Fast Apriori indexing based hierarchical clustering gives a useful measure which is used to measure the item dependency between data points that is likely to be in terms of their minimum support property. This research proposes an optimal method to estimate the items searching which is measured using indexing method corresponds to transactional database. Each item contains number of transaction functions and its own description which is used to identify the type of database. Further, the proposed methodology of finding the relevant item sets can maintain better quality in terms of relevance utility discovery than the existing methods. Our approach can reduce the number of support count dependencies to be checked in comparison with previous methods.

KEYWORDS: Utility Mining, Frequent Itemsets, Fast Apriori, High utility Itemset, Downward Closure Property (DCP).

I. INTRODUCTION

The objective of this paper is finding the frequent item sets involving rare items without causing frequent items to generate too many meaningless itemsets. To avoid generating a huge set of candidate itemsets as in the Apriori algorithm, this paper presents an efficient Fast Apriori closed high utility itemset discovery with hierarchical clustering algorithm (FAHU-Hierarchical). This algorithm adopts a divide-and-conquer approach to minimize the candidate generation process to only those most likely to be frequent, and employs a compact prefix-tree data structure, frequent-pattern tree (FP-tree), to avoid repetitive scanning of the database.

The FAHU-Hierarchical algorithm performs exactly two scans of the transaction database and mines on the compact data structure, Frequent Pattern-tree, to find all frequent itemsets without generating all possible candidate sets. It shows that FP-growth is about an order of magnitude faster than Apriori in large databases. This gap grows wider when the minimum support threshold reduces. Although the fast apriori algorithm is efficient, sometimes, it is infeasible to construct a main memory-based pattern tree when the database is large, which is very common for the case of clustering. To create a scalable version of Hierarchical-tree growth, we can first partition the database into a set of projected databases, and then construct a closed utility frequent-tree and mine it in each projected database.

II. RELATED WORK

R. Agrawal and R. Srikant [1] proposed the problem of discovering association rules between items in a large database of sales transactions. Key roles of closed sets and minimal generators in concise representations of frequent patterns are proposed by T. Hamrouni. [2]. respect to several aspects and comparative criteria which proves the importance of considering closed sets and minimal generators.J. Han, J. Pei, and Y. Yin. [3] proposed a novel frequent pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree-based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. T. Hamrouni, S. Yahia, and E. M. Nguif, [4] have proposed research

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

analyze on Concise (or condensed) representations of frequent patterns follow the minimum description length (MDL) principle, by providing the shortest description of the whole set of frequent patterns. Fast and memory efficient mining of high utility item sets in data streams is proposed by H.-F. Li et.al. [5]. Efficient mining of high utility itemsets has become one of the most interesting data mining tasks with broad applications. C.-W. Lin, T.-P. Hong, and W.-H. Lu [6] discussed many algorithms were proposed to mine association rules, most of which were based on item frequency values. Considering a customer may buy many copies of an item. G.-C. Lan, T.-P. Hong, and V. S. Tseng [7] proposed the utility mining has widely been discussed in the field of data mining. It finds high utility itemsets by considering both profits and quantities of items in transactional data sets. H. Li, J. Li, L. Wong, M. Feng, and Y. Tan [8] authors interested to test whether a given cause has a given effect. If we cannot specify the nature of the factors involved, such tests are called model-free studies. There are two major strategies to demonstrate associations between risk factors (ie. patterns) and outcome phenotypes (ie. class labels). The first is that of prospective study designs, and the analysis is based on the concept of "relative risk" B. Le, H. Nguyen, T. A. Cao, and B. Vo, [9] illustrated the utility based itemset mining approach which has been discussed widely in recent years. Y. Liu, W. Liao, and A. Choudhary [10] Association rule mining (ARM) identifies frequent itemsets from databases and generates association rules by considering each item in equal value. C. Lucchese, S. Orlando, and R. Perego [11] authors discussed a new scalable algorithm for discovering closed frequent itemsets, a lossless and condensed representation of all the frequent itemsets that can be mined from a transactional database. Y.-C. Li, J.-S. Yeh, and C.-C. Chang [12] discussed the traditional methods of association rule mining consider the appearance of an item in a transaction, whether or not it is purchased, as a binary variable. Efficient mining of association rules using closed itemset lattice is projected by N. Pasquier, Y. Bastide, R. Taouil, and L. Lak. [13]. Discovering association rules is one of the most important task in data mining. Many efficient algorithms have been proposed in the literature. The most noticeable are Apriori, Mannila's algorithm, Partition, Sampling and DIC, that are all based on the Apriori mining method. J. Wang, J. Han, and J. Pei, [14] authors proposed constraint-based frequent itemset mining is necessary when the needs and interests of users are the top priority. In this task, two opposite types of constraint are studied, namely anti-monotone and monotone constraints. B.-E. Shie, H.-F. Hsiao, V. S. Tseng, and P. S [15] discussed the Mining user behaviours in mobile environments is an emerging and important topic in data mining fields.

III. PROPOSED ALGORITHM

A. High Utility Itemsets:

Tremendous research efforts have been devoted to utility mining. These existing studies can be roughly divided into two categories: High utility mining and Closed Item set mining.

- A very large number of high utility itemsets makes it difficult for the users to comprehend the results. It may also cause the algorithms to become inefficient in terms of time and memory requirement, or even run out of memory.
- The HUI mining is not an easy task since the downward closure property [1] in FIM does not hold in utility mining
- A superset of a low utility itemset can be high utility and a subset of a HUI can be low utility.
- The closed itemset mining and properties related to closed itemsets and mention relevant methods.

B. Fast Apriori Based High Utility Itemsets:

Mining high utility itemsets from databases is an important data mining task for discovery of itemsets with high utilities. However, it may present too many HUIs to users, which also degrades the efficiency of the mining process. Frequent itemset mining (FIM) is a one of its popular applications is market basket analysis, which refers to the discovery of sets of items (itemsets) that are frequently purchased together by customers. The proposed system utilize the model for building a Lossless Representation system that suggests high utility itemsets over dynamic datasets using the Fast Apriori closed high utility itemset discovery with hierarchical clustering algorithm (FAHU-Hierarchical). Update the CHUD (Closed High Utility Itemset Discovery) to approach divisive Hierarchical clustering manner. The proposed FAHU-Hierarchical clustering method attempts to address the individual requirements in utility

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

clustering using the notion of frequent itemsets. Its works greedily selects the next frequent itemset, which represents the next cluster, minimizing the overlap of clusters in terms of shared documents.

Advantages of Fast Apriori Based High Utility Itemsets:

- The advantage of using the two constraints (must link and cannot link) can be applied in any order during the mining process.
- FAHU takes advantage of the fact that by using the total order, the complete set of itemsets can be divided into n non-overlapping subspaces.

C. Discovering Frequent Itemsets Using Fast Apriori Infrequent Algorithm:

The proposed of our method is the fast apriori infrequent algorithm based on hierarchical clustering. Our contributions are in providing novel scalable approaches for each building block. We start by counting the support of every item in the dataset and sort them in decreasing order of their frequencies. Next, we sort each transaction with respect to the frequency order of their items. We call this a horizontal sort. We also keep the generated candidate itemsets in horizontal sort. Furthermore, we are careful to generate the candidate itemsets in sorted order with respect to each other. We call this a downward sort. When itemsets are both horizontally and downward sorted, we call them fully sorted. As we show, generating sorted candidate itemsets (for any size *k*), both horizontally and downward, is computationally free and maintaining that sort order for all subsequent candidate and frequent itemsets requires careful implementation, but no cost in execution time. This conceptually simple sorting idea has implications for every subsequent part of the algorithm.

D. Downward Closure Property (DCP):

In a classical Apriori algorithm it is assumed that if the itemset is large, then all its subsets should also be large and is called Downward Closure Property (DCP). This helps algorithm to generate large itemsets of increased size by adding items to itemsets that are already large. In the weighted ARM case where each item is assigned a weight, the DCP does not hold. Because of the weighted support, an itemset may be large even though some of its subsets are not large.

Table 1 shows four large itemsets of size 2 (AB, AC, BC, BD) and two large itemsets of size 3 (ABC, BCD), generated using tables 3 and 4. In classical ARM, when the weights are not considered, all of the six itemsets are large. But if we consider items weights and calculate the weighted support of itemsets according to definition 3 and 7, a new set of support values are obtained. Although the classical support of all itemsets is large, if ABC and BCD are frequent then their subsets must be large according to classical ARM. But considering the weighted support, AB and BC are no longer frequent.

Table 1: Frequent itemsets with invalid DCP (weighted settings)

Large Itemsets	Support (40%)	Large?	Weighted Support (0.4)	Large
AB	40%	Yes	0.162	No
AC	60%	Yes	0.42	Yes
ABC	40%	Yes	0.4	Yes
BC	40%	Yes	0.36	No
BD	60%	Yes	0.72	Yes
BCD	40%	Yes	0.72	Yes

E. Discovering Itemset Membership Function:

The traditional way to discover the itermsets needed for a certain data set is to consult a domain expert who will define the sets and their membership functions. This requires access to domain knowledge which can be difficult

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

or expensive to acquire. In order to make an automatic discovery of item sets possible, an approach has been developed which generates itemsets automatically by hierarchical clustering. This method can be used to divide quantitative attributes into sets, which deals with the problem that it is not always easy do define the sets of apriori. The proposed method uses a clustering method to find the *k* clusters. The whole process of automatically discovering items can be subdivided into four steps:

- Transform the database to make clustering possible (the value of all the attributes has to be positive integer).
- Find the clusters of the transformed database using a clustering method.
- For each quantitative attribute, itemsets are constructed using the medoids.
- Generate the item membership functions.

After discovering *k* medoids, we can compute *k* sets out of them. We define {*m1* ,*m2* , ... ,*mk*} as the *k* medoids from a database. The *i* -th medoid can be defined as *mi*={*ai1*,*ai2*, ... ,*ain*} . If we want to discover the itemsets for the *j* -th attribute, ranging from min j to max j , our mid-points will be {*ai1*, *ai2*,... , *ain*} . The itemsets will then show the following ranges: {*minj – a2j*}, {*a1j−a3j*}, {*a(i-1)j − a(i+1)j*}, ... ,{*a(k-1)j −maxj* } . Finally, the membership functions for the transaction sets have to be computed.

IV. SIMULATION RESULTS

Compared with Apriori-based algorithm for mining High utility Closed + itemsets) the proposed fast utility mining algorithm requires an additional iterative procedure to compute the infrequent weights of all transactions. The database is scanned exactly once in each iteration. Therefore, the convergence rate of the utility weights is critical to the performance. It is clear that fast apriori converges fast on transaction databases. Generally, three or four iterations are enough to achieve a good estimation, which means that our link-based method works at the cost of three or four additional database scans over the traditional techniques.

This research paper is, to the best of our knowledge, the first attempt to perform closed itemset mining from weighted data. However, other algorithms are able to mine utility itemsets from unweighted data. Hence, to also analyze the efficiency of the proposed approach when tackling the frequent itemset mining from unweighted data, we compared normal apriori rule mining execution time with that of a benchmark algorithm. Hierarchical based Fast Apriori infrequent closed utility mining algorithm is, to the best of our knowledge, the latest algorithm that performs both minimal and non-minimal (unweighted) infrequent itemset mining from unweighted data.

Table 2: Execution Time (Seconds)

K	Two Phase	Apriori- HC	Apriori- HC-D	Fast Apriori- HC
10	0	0	5000	4000
20	13200	13000	100	90
30	8500	9000	10	10
40	8600	1500	10.5	10
50	810	800	9.6	9.4
60	700	700	9.6	9.3
70	700	700	9.5	9.2
80	60	60	9.4	9
90	11	10	8	8

Table 2 represents the comparison of previously Algorithm like Two-phase, Apriori-HC (High utility Closed) and Apriori-HC-D (High utility Closed Discarding) with proposed Fast Apriori-HC (Hierarchical Cluster) matrix

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

manipulation and our proposed method utility Weighted Itemset Mining with leverage techniques when run on transaction Database records.

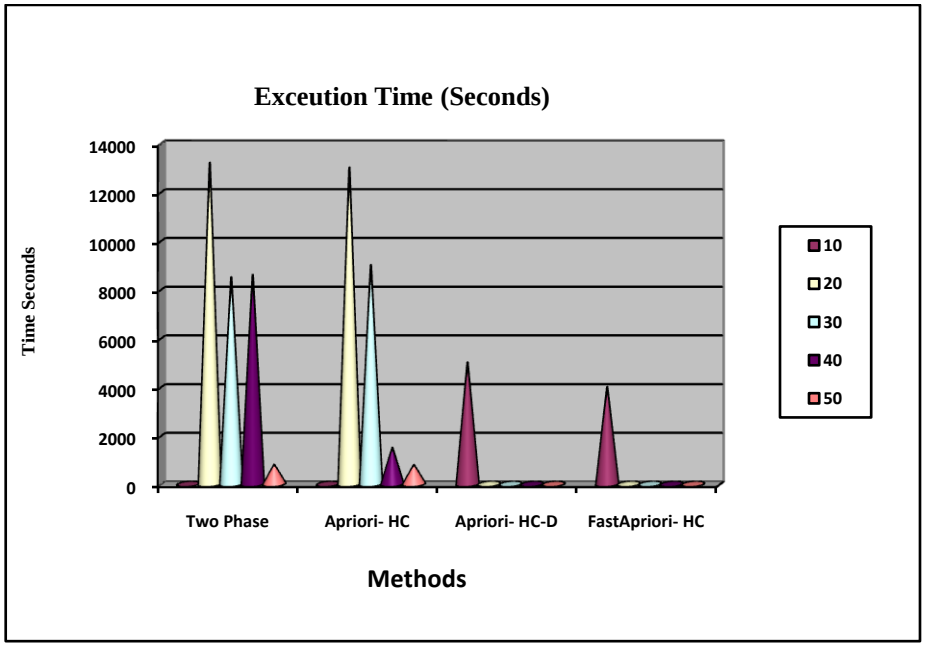

Figure 1 Execution Time

Figure 1 represents the comparison of previously Algorithm like Two-phase, Apriori-HC (High utility Closed) and Apriori-HC-D (High utility Closed Discarding) with proposed Fast Apriori-HC (Hierarchical Cluster) matrix manipulation and our proposed method utility Weighted Itemset Mining with leverage techniques when run on transaction Database records.

Table 3: Comparison of Precision value of existing methods

Min.sup value	Two Phase	Apriori- HC	Apriori- HC-D	Fast Apriori- HC
2	66.67	70	50	75
3	55.1	59	68.4	80.1
4	59	62	67.75	84
5	72	66	68.77	88

Tables 3 shows the precision values of algorithm Fast Apriori based Hierarchical clustering on transaction database, under different minimum expected supports, respectively.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

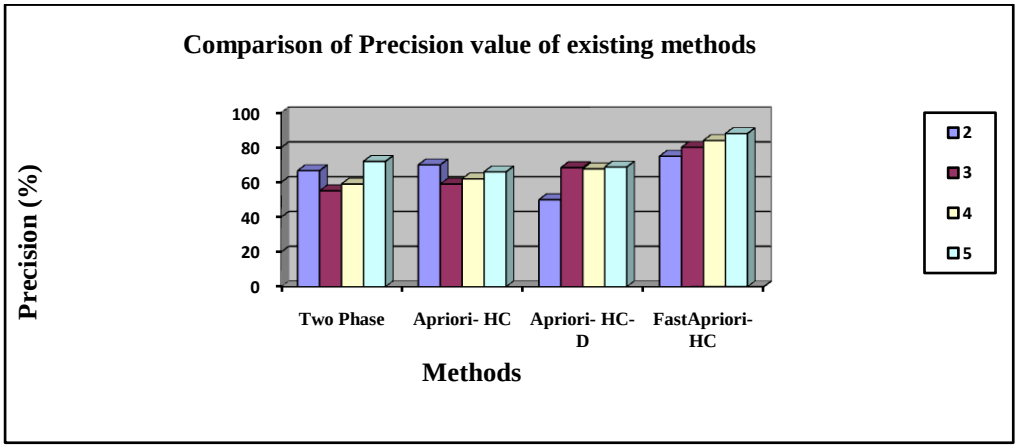

Figure 2: Precision value

Figure 2 shows the precision values of algorithm Fast Apriori based Hierarchical clustering on transaction database, under different minimum expected supports, respectively.

Table 4: Comparison of Recall value of existing methods

Min.sup value	Two Phase	Apriori- HC	Apriori- HC-D	Fast Apriori- HC
2	78	78.69	81.55	83.33
3	61.55	60.99	65.11	66.67
4	77	78.07	85.21	88.34
5	71.21	84.36	86.66	87.6

Table 4 shows the recall values of algorithm Fast Apriori based Hierarchical clustering on transaction database, under different minimum expected supports, respectively.

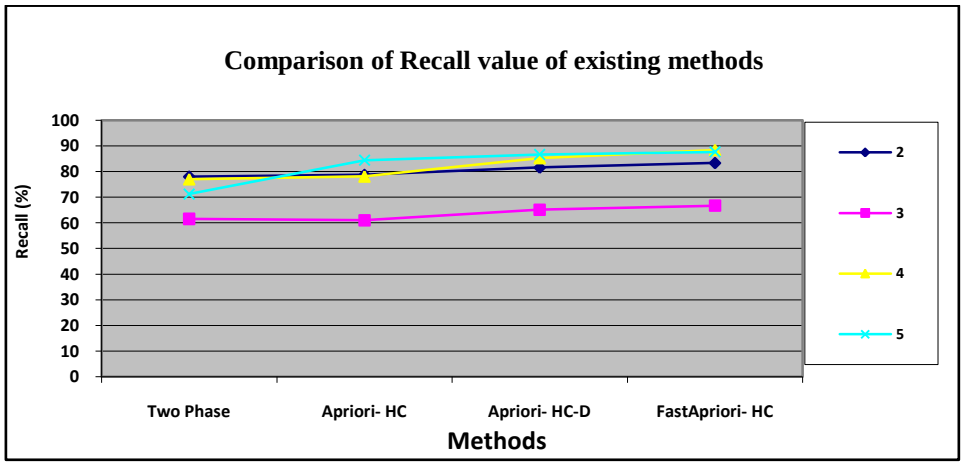

Figure 3: Comparison of Recall value of existing methods

Figure 3 shows the recall values of algorithm Fast Apriori based Hierarchical clustering on transaction database, under different minimum expected supports, respectively

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Table 5: Comparison of Accuracy value of existing methods

Min.sup. value	Two Phase	Apriori- HC	Apriori- HC-D	Fast Apriori- HC
2	88.54	89.44	90.06	92.8571
3	66	70.98	78.60	80
4	80.29	84.377	86	87.5
5	89.99	90.74	90.95	91.57

Table 5 shows the Accuracy values of algorithm Fast Apriori based Hierarchical clustering on transaction database, under different minimum expected supports, respectively

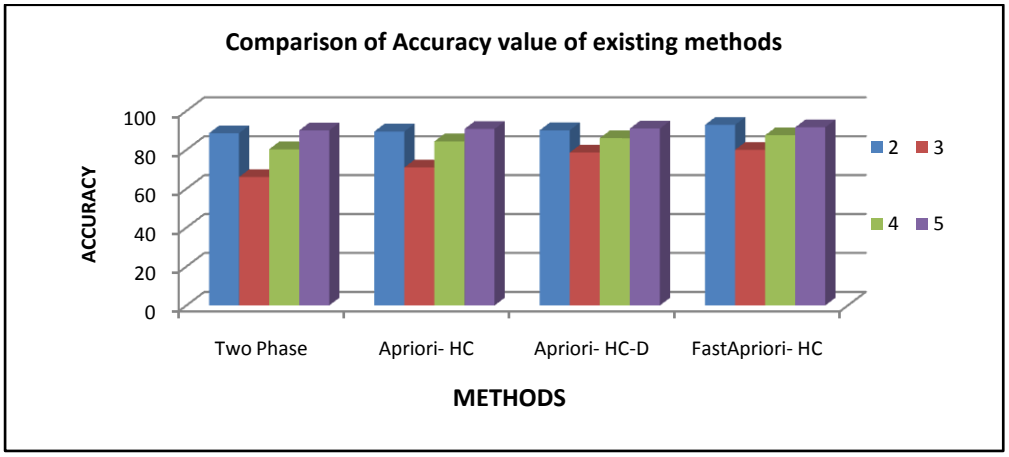

Figure 4: Comparison of Accuracy value of existing methods

Figure 4 shows the Accuracy values of algorithm Fast Apriori based Hierarchical clustering on transaction database, under different minimum expected supports, respectively

V. CONCLUSION AND FUTURE WORK

This research work faces the issue of discovering utility itemsets by using hierarchical clustering weights for differentiating between relevant items and not within each transaction. Two algorithms that accomplish Fast Apriori closed utility itemset mining and Hierarchical Clustering Weighted Itemset mining efficiently is also proposed. The usefulness of the discovered patterns has been validated on data coming from a real-life context with the help of a domain expert. Discovering utility rules is at the heart of data mining. Mining for closed itermset rules between items in large database of sales transactions has been recognized as an important area of database research. These rules can be effectively used to uncover unknown relationships, producing results that can provide a basis for forecasting and decision making. In this paper we proposed an efficient Fast Apriori hierarchical Weighted Itemset utility mining technique to find all co-occurrence relationships among data items.

This work is further extended to calculate the weight value using the automatic expert system. It is also extended with prognosis, treatment of diseases and also for ranking the intrusion threats in wireless networks

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
[2] T. Hamrouni, "Key roles of closed sets and minimal generators in concise representations of frequent patterns," Intell. Data Anal., vol. 16, no. 4, pp. 581–631, 2012.
[3] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1–12.
[4] T. Hamrouni, S. Yahia, and E. M. Nguifo, "Sweeping the disjunctive search space towards mining new exact concise representations of frequent itemsets," Data Knowl. Eng., vol. 68, no. 10, pp. 1091–1111, 2009.
[5] H.-F. Li, H.-Y. Huang, Y.-C. Chen, Y.-J. Liu, and S.-Y. Lee, "Fast and memory efficient mining of high utility itemsets in data streams," in Proc. IEEE Int. Conf. Data Mining, 2008, pp. 881–886.
[6] C.-W. Lin, T.-P. Hong, and W.-H. Lu, "An effective tree structure for mining high utility itemsets," Expert Syst. Appl., vol. 38, no. 6, pp. 7419–7424, 2011.
[7] G.-C. Lan, T.-P. Hong, and V. S. Tseng, "An efficient projectionbased indexing approach for mining high utility itemsets," Knowl. Inf. Syst, vol. 38, no. 1, pp. 85–107, 2014.
[8] H. Li, J. Li, L. Wong, M. Feng, and Y. Tan, "Relative risk and odds ratio: A data mining perspective," in Proc. ACM SIGACT-SIGMOD-SIGART Symp. Principles Database Syst., 2005, pp. 368–377.
[9] B. Le, H. Nguyen, T. A. Cao, and B. Vo, "A novel algorithm for mining high utility itemsets," in Proc. 1st Asian Conf. Intell. Inf. Database Syst., 2009, pp. 13–17.
[10] Y. Liu, W. Liao, and A. Choudhary, "A fast high utility itemsets mining algorithm," in Proc. Utility-Based Data Mining Workshop, 2005, pp. 90–99.
[11] C. Lucchese, S. Orlando, and R. Perego, "Fast and memory efficient mining of frequent closed itemsets," IEEE Trans. Knowl. Data Eng., vol. 18, no. 1, pp. 21–36, Jan. 2006.
[12] Y.-C. Li, J.-S. Yeh, and C.-C. Chang, "Isolated items discarding strategy for discovering high utility itemsets," Data Knowl. Eng., vol. 64, no. 1, pp. 198–217, 2008.
[13] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient mining of association rules using closed itemset lattice," J. Inf. Syst., vol 24, no. 1, pp. 25–46, 1999.
[14] J. Wang, J. Han, and J. Pei, "Closeţþ: Searching for the best strategies for mining frequent closed itemsets," in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003, pp. 236–245.
[15] B.-E. Shie, H.-F. Hsiao, V. S. Tseng, and P. S. Yu, "Mining high utility mobile sequential patterns in mobile commerce environments," in Proc. Int. Conf. Database Syst. Adv. Appl., 2011, vol. 6587, pp. 224–238.

BIOGRAPHY

T.Menaka received her MCA., Degree from Meenakshi Government College for Women, Madurai, Tamilnadu, India in 2005. She completed her M.Phil. Degree in Computer Science from Bharathiar University, Coimbatore, India in 2008. She served as a Faculty of Computer Science Department at Saraswathi Thyagaraja College, Pollachi from 2006 to 2010. Presently, she has been working as an Assistant Professor in the department of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 2010. She has published papers in international/national journal and conferences. Her research focuses on Data Mining and Digital Image Processing.

M.Premalatha is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India. She received Master of Science Degree (Computer Science) in 2013 from Bharathiar University, Coimbatore, India. She has presented papers in International/National Conferences and attended Workshops, Seminars and published papers in International Journal. Her research focuses on Data Mining.