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 8, August 2015

An Efficient High Utility Frequent Itemsets Mining Using Fast Apriori Based Hierarchical Clustering Algorithm

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

Research Scholar, Department of Computer Science, NGM College, Pollachi, India[1]

Assistant Professor, Department of Computer Science, NGM College, Pollachi, India[2]

ABSTRACT: 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. In this paper presents a new system to 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 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. Experimental studies on both the synthetic and real-world data streams show the performance of our proposed approach.

KEYWORDS: High utility mining, Fast Apriori algorithm, Hierarchical clustering, frequent mining

I. INTRODUCTION

Frequent itemset mining (FIM) [1], [3], [4], [5], [8] is a fundamental research topic in data mining. 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. However, in this application, the traditional model of FIM may discover a large amount of frequent but low revenue itemsets and lose the information on valuable itemsets having low selling frequencies. These problems are caused by the facts that (1) FIM treats all items as having the same importance/unit profit/weight and (2) it assumes that every item in a transaction appears in a binary form, i.e., an item can be either present or absent in a transaction, which does not indicate its purchase quantity in the transaction. Hence, FIM cannot satisfy the requirement of users who desire to discover itemsets with high utilities such as high profits.

The frequent itemsets identified by ARM does not reflect the impact of any other factor except frequency of the presence or absence of an item. Frequent itemsets may only contribute a small portion of the overall profit, whereas non-frequent itemsets may contribute a large portion of the profit. In reality, a retail business may be interested in identifying its most valuable customers (customers who contribute a major fraction of the profits to the company). These are the customers, who may buy full priced items, high margin items, or gourmet items, which may be absent from a large number of transactions because most customers do not buy these items. In a traditional frequency oriented ARM, these transactions representing highly profitable customers may be left out. For instance, {milk, bread} may be a frequent itemset with support 40%, contributing 4% of the total profit, and the corresponding consumers is Group A, whereas {birthday cake, birthday card} may be a non-frequent itemset with support 8% (assume support threshold is 10%), contributing 8% of the total profit, and the corresponding consumers is Group B.

Nowadays, in any real application, the size of the data set easily goes to hundreds of Mbytes or Gbytes. In order to tackle this to efficiently mine high utility itemsets. Thus, only the combinations of high transaction-weighted utilization itemsets are added into the candidate set at each level. Therefore, the size of the candidate set is substantially reduced during the level-wise search. The memory cost as well as the computation cost is also efficiently reduced.

II. RELATED WORK

In [1] authors considered the problem of discovering association rules between items in a large database of sales transactions. To present two new algorithms for solving this problem those are fundamentally different from the known algorithms. In [2] authors proposed the three novel tree structures to efficiently perform incremental and interactive HUP mining. The first tree structure, Incremental HUP Lexicographic Tree (IHUP_{L}-Tree), is arranged according to an item's lexicographic order. It can capture the incremental data without any restructuring operation. The second tree structure is the IHUP transaction frequency tree (IHUP_{TF}-Tree), which obtains a compact size by arranging items according to their transaction frequency (descending order). To reduce the mining time, the third tree, IHUP-transaction-weighted utilization tree (IHUP_{TWU}-Tree) is designed based on the TWU value of items in descending order. In [3] authors illustrated a structure called free-sets, from which we can approximate any itemset support and we formalize this notion in the framework of ∈-adequate representations. It shows that frequent free-sets can be efficiently extracted using pruning strategies developed for frequent itemset discovery, and that they can be used to approximate the support of any frequent itemset. In [4] authors discussed to identify the redundancies set of all frequent itemsets and to exploit these redundancies in order to reduce the result of a mining operation. Its present deduction rules to derive tight bounds on the support of candidate itemsets. It shows how the deduction rules allow for constructing a minimal representation for all frequent itemsets. It present connections between our proposal and recent proposals for concise representations and we give the results of experiments on real-life datasets that show the effectiveness of the deduction rules. In [5] authors described a practicably interesting mining task to retrieve top-k (closed) itemsets in the presence of the memory constraint. Specifically, as opposed to most previous works that concentrate on improving the mining efficiency or on reducing the memory size by best effort, we first attempt to specify the available upper memory size that can be utilized by mining frequent itemsets. To comply with the upper bound of the memory consumption, two efficient algorithms, called MTK and MTK_Close, are devised for mining frequent itemsets and closed itemsets, respectively, without specifying the subtle minimum support. Instead, users only need to give a more human-understandable parameter, namely the desired number of frequent (closed) itemsets k. In practice, it is quite challenging to constrain the memory consumption while also efficiently retrieving top-k itemsets. In [6] authors proposed a novel idea of top-K objective-directed data mining, which focuses on mining the top-K high utility closed patterns that directly support a given business objective. To association mining, we add the concept of utility to capture highly desirable statistical patterns and present a level-wise item-set mining algorithm. With both positive and negative utilities, the antimonotone pruning strategy in Apriori algorithm no longer holds. In response, we develop a new pruning strategy based on utilities that allow pruning of low utility itemsets to be done by means of a weaker but antimonotonic condition.

III. PROPOSED ALGORITHM

A. *Preprocessing:*

The initial process for high utility mining is data pre-processing that will arrange the data sets for use in data mining processes. Researchers have to set clear criteria to filter all data sets suitable to the research objectives. The first step in data pre-processing is the data cleaning process that gets rid of noise and anomalies. The data items has been reduced and altered into the format that is appropriate for mining process to analyze and gathering.

The pre-processing procedure is utilized for the initial high utility mining in the original database. For the generation of mining high utility itemsets in the dataset, and so on, the incremental procedure is employed. In pre-processing steps the original transaction database is partitioned into three partitions, i.e., {Column1, Column2, Column3}, in the pre processing procedure. Each partition is scanned sequentially for the generation of candidate 2-itemsets in the first scan of the database items.

DOI: 10.15680/IJIRCCE.2015. 0308042

B. *Fast Apriori closed high utility itemset discovery*

The High utility item set feature selection will use the hierarchical manner with fast Apriori-based algorithm to generate the frequent sets of attribute relation rules. With Fast Apriori-based algorithm used to recognize and create features that are associated and change to other features sets in the group, more successful action in hierarchical technique is required. We have to filter the rules that appropriate to research objective. Fast Apriori is a formation to count candidate item sets efficiently. It generates candidate item sets of length *k* from the *k*-1 item sets and keeps away from expanding all the item sets. Then it removes the candidates which have an infrequent sub pattern. The candidate set contains all frequent *k*-length item sets. After that, it scans the entire transaction database to determine frequent item sets among the candidates. With fast Apriori technique the algorithm can reduce time processing in generating fewer groups of item sets and avoid infrequent candidate item sets expansion.

The Fast Apriori closed high utility itemset discovery Algorithms steps is given below,

```
Algorithm 1: Fast Apriori closed high utility itemset discovery
Input: $I$: Dataset Training set count; $m$: Minimum confidence value;
Output: frequent itemsets in $L_k$
$C1$: Generate the candidate itemsets; $L1$: Save the frequent itemsets
for $k$ = 1 to $I$ do
 Step1: Join $L_{k-1}$ $r$ with $L_{k-1}c$, as follows:
     insert into $C_k$
     select $p.item_1$, $p.item_2$,…, $p.item_{k-1}$, $q.item_{k-1}$
     from $L_{k-1}$ $p$, $L_{k-1}q$
     where $p.item_1 = q.item_1,…, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$
 Step 2: Generate all (k-1)-subsets from the candidate itemsets in Ck
 Step 3: Prune all candidate itemsets from $C_k$ where some ($k$-1)-subset of the candidate itemset is not in the
 frequent itemset $L_{k-1}$
end for
```

C. *Hierarchical clustering based Closed High Utility Itemset Discovery*

The Hierarchical clustering is a post processing of high utility frequent item set data to find clusters or groups of similar data in hierarchical structure. In each division, the item membership levels have some similarity in type of data. The principles of hierarchical clustering are searching value of score (conditions i.e, <,> and =) in similarity, and assigning each memberships to be in the different group of other members that have similar or same score.

The cluster belong to a new method for close itemest discovery labeling of hierarchical clusters using transaction database as labels which we find to be more interesting and efficient. Using fast apriori based frequent itemsets as labels to clusters is a best approach to find the sets.

IV. Conclusion and Future Work

In this paper proposed the Hierarchical clustering based high utility itermset mining has for users to apply to suitable data types and usage. From this paper, we present one of fast apriori based item set based on unsupervised data mining technique that integrated other utility mining technique and synchronized processing. The post processing added the closed utility based apriori algorithm in feature selection to get better feature set. In the hierarchal clustering process we used result of frequent itemset data to generate initial centroids that works better for data sets.

In future work, we intend to enhance the hierarchal algorithm to develop the experimental methods for non-linear database to control the growth of infrequent attributes of the result data.

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. C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, “Efficient tree structures for high utility pattern mining in incremental databases,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708–1721, Dec. 2009.

3. J.-F. Boulicaut, A. Bykowski, and C. Rigotti, “Free-sets: A condensed representation of Boolean data for the approximation of frequency queries,” Data Mining Knowl. Discovery, vol. 7, no. 1, pp. 5–22, 2003.
4. T. Calders and B. Goethals, “Mining all non-derivable frequent itemsets,” in Proc. Int. Conf. Eur. Conf. Principles Data Mining Knowl. Discovery, 2002, pp. 74–85.
5. K. Chuang, J. Huang, and M. Chen, “Mining top-k frequent patterns in the presence of the memory constraint,” VLDB J., vol. 17, pp. 1321–1344, 2008.
6. R. Chan, Q. Yang, and Y. Shen, “Mining high utility itemsets,” in Proc. IEEE Int. Conf. Data Min., 2003, pp. 19–26.
7. A. Erwin, R. P. Gopalan, and N. R. Achuthan, “Efficient mining of high utility itemsets from large datasets,” in Proc. Int. Conf. Pacific-Asia Conf. Knowl. Discovery Data Mining, 2008, pp. 554–561.
8. K. Gouda and M. J. Zaki, “Efficiently mining maximal frequent itemsets,” in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 163–170.

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.