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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Efficient Hybrid Algorithm for High Utility Itemset Mining

Jeevan Saraf[#1], Vipul Munot[#2], Sahil Patil[#3], Dr. Vijaykumar Bidve[#4]

Department of Information Technology, Marathwada Mitra Mandal College of Engineering, Pune, India

ABSTRACT: Data mining is a computerized process of searching for models in large data sets that involve methods at the intersection of the database system. The popular problem of data mining is the extraction of high utility element sets (HUI) or, more generally, the extraction of public services (UI). The problem of HUI (set of elements of high utility) is mainly the introduction to the set of frequent elements. Frequent pattern mining is a widespread problem in data mining, which involves searching for frequent patterns in transaction databases. Solve the problem of the set of high utility elements (HUI) with some particular data and the state of the art of the algorithms. To store the HUI (set of high utility elements) many popular algorithms have been proposed for this problem, such as "Apriori", FP growth, etc., but now the most popular TKO algorithms (extraction of utility element sets) K in one phase) and TKU (extraction of elements sets Top-K Utility) here TKO is Top K in one phase and TKU is Top K in utility. In this paper, address previous issues by proposing a new frame work for k upper HUI where k is the desired number of HUI to extract. Extraction of high utility element sets is an uncommon term. But we are using it while shopping online, etc. It is part of the business analysis. The main area of application is the analysis of the market basket, where when the customer buys the item he can buy another to maximize the benefit both the customer and supplier profit.

KEYWORDS: - Utility mining, high utility item set, top k Pattern mining, top k high item set mining.

I. INTRODUCTION

Data mining is the efficient discovery of valuable and vivid information from a vast collection of data. Frequent set mining set (FIM) discovers the only frequent elements, but the set of HUI High Utility items. In the FIM profile of the set of elements are not considered. This is because the amount of the purchase does not take into account. Data mining is the process of analyzing data from different points of view and summarizing it in useful data. Data mining is a tool for analyzing data. It allows users to analyze data from different levels or angles, organize them and find the relationships between the data. Data mining is the process of finding patterns between enough fields in the large relational database. A classic algorithm based on Top K models consists of two phases. In the first phase, called phase I, it is the complete set of high transaction weighted utility item set (HTWUI). In the second phase, called phase II, all HUIs are obtained by calculating the exact HTWUI utilities with a database scan. Although many studies have been devoted to the extraction of HUI, it is difficult for users to effectively choose an appropriate minimum threshold. Depending on the threshold, the size of the output can be very small or very large. Also the choice of the threshold significantly impacts the performance of the algorithms if the threshold is too low then too many HUI will be presented to users then it will be difficult for users to understand the results. A large amount of HUI creates data mining algorithms unproductive or out of memory, subsequently the more HUIs create the algorithms, the more resources they consume. Conversely, if the threshold is too high, HUI will not be found.

1.1 Background:-

Frequently generate a huge set of HUIs and their mining performance is degraded consequently. Further in case of long transactions in dataset or low thresholds are set, then this condition may become worst. The huge number of HUIs forms a challenging problem to the mining performance since the more HUIs the algorithm generates, the higher processing time it consumes. Thus to overcome this challenges the efficient algorithms presented. Top k will not work on the parallel mining.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

1.2 Motivation:-

1. Set the value of k which is more intuitive than setting the threshold because k represents the number of Item sets that users want to find whereas choosing the threshold depends primarily on database characteristics, which are often unknown to users.
2. The main point of min utility variable is not given in advance in top k HUI mining In traditional HUI mining the search space can be efficiently increased to algorithm by using a given the min utility threshold value. In scenario of TKO and TKU algorithm min utility threshold value is provided in advance.

1.3 Aim & Objective:-

1. The execution time of TKO algorithm is less but result is incorrect with a garbage value and it is efficient algorithm. The execution time of TKU algorithm is more but result is correct. It is very challenging issue how hybrid algorithm (TKO WITH TKU) is efficient than TKU algorithm. The time factor is very important in that.
2. Need to achieve significantly better performance.
3. The Hybrid Algorithm get HUI fixed Parameter of Rating and view and Number of Buy's

II. RELATED WORK

1. "Efficient tree structures for high-utility pattern mining in incremental databases". Recently, high utility pattern (HUP) mining is one of the most important research issues in data mining due to its ability to consider the non-binary frequency values of items in transactions and different profit values for every item. On the other hand, incremental and interactive data mining provide the ability to use previous data structures and mining results in order to reduce unnecessary calculations when a database is updated, or when the minimum threshold is changed. In this paper, we propose three novel tree structures to efficiently perform incremental and interactive HUP mining. The first tree structure, Incremental HUP Lexicographic Tree (IHUPL-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 (IHUPTF-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 (IHUPTWU-Tree) is designed based on the TWU value of items in descending order. Extensive performance analyses show that our tree structures are very efficient and scalable for incremental and interactive HUP mining.

2. "Mining high-utility item sets" Traditional association rule mining algorithms only generate a large number of highly frequent rules, but these rules do not provide useful answers for what the high utility rules are. We develop 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 anti-monotone 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 item sets to be done by means of a weaker but anti-monotonic condition. Our experimental results show that our algorithm does not require a user specified minimum utility and hence is effective in practice.

3. "Mining top-k frequent closed patterns without minimum support" In this paper, we propose a new mining task: mining top-k frequent closed patterns of length no less than min_/spllscr/, where k is the desired number of frequent closed patterns to be mined, and min_/spllscr/ is the minimal length of each pattern. An efficient algorithm, called TFP, is developed for mining such patterns without minimum support. Two methods, closed-node-count and descendant-sum are proposed to effectively raise support threshold and prune FP-tree both during and after the construction of FP-tree. During the mining process, a novel top-down and bottom-up combined FP-tree mining strategy is developed to speed-up support-raising and closed frequent pattern discovering. In addition, a fast hash-based closed pattern verification scheme has been

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

employed to check efficiently if a potential closed pattern is really closed. Our performance study shows that in most cases, TFP outperforms CLOSET and CHARM, two efficient frequent closed pattern mining algorithms, even when both are running with the best tuned min-support. Furthermore, the method can be extended to generate association rules and to incorporate user-specified constraints.

4. "Mining frequent patterns without candidate Generation" Mining frequent patterns in transaction databases, times Series databases, and many other kinds of databases have been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there existproli c patterns and/or long patterns. In this study, we propose a novel frequent pattern tree (FP-tree) structure, which is an extended pre xtree structure for storing compressed, crucial information about frequent patterns, and develop an e client FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. Exigency of mining is achieved with three techniques: (1) a large database is compressed into a highly condensed, much smaller data structure, which avoids costly, repeated database scans, (2) our FP-tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of candidate sets, and (3) a partitioning-based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining con need patterns in conditional databases, which dramatically reduces the search space. Our performance study shows that the FP-growth method is ancient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent pattern mining methods.

5. "Novel Concise Representations of High Utility Item sets Using Generator Patterns" Mining High Utility Item sets (HUIs) is an important task with many applications. However, the set of HUIs can be very large, which makes HUI mining algorithms suffer from long execution times and huge memory consumption. To address this issue, concise representations of HUIs have been proposed. However, no concise representation of HUIs has been proposed based on the concept of generator despite that it provides several benefits in many applications. In this paper, we incorporate the concept of generator into HUI mining and devise two new concise representations of HUIs, called High Utility Generators (HUGs) and Generator of High Utility Item sets (GHUIs). Two efficient algorithms named HUG-Miner and GHUI-Miner are proposed to respectively mine these representations. Experiments on both real and synthetic datasets show that proposed algorithms are very efficient and that these representations are up to 36 times smaller than the set of all HUIs.

6. "Mining Top-K Sequential Rules" Mining sequential rules requires specifying parameters that are often difficult to set (the minimal confidence and minimal support). Depending on the choice of these parameters, current algorithms can become very slow and generate an extremely large amount of results or generate too few results, omitting valuable information. This is a serious problem because in practice users have limited resources for analyzing the results and thus are often only interested in discovering a certain amount of results, and fine-tuning the parameters can be very time-consuming. In this paper, we address this problem by proposing TopSeqRules, an efficient algorithm for mining the top-k sequential rules from sequence databases, where k is the number of sequential rules to be found and is set by the user. Experimental results on real-life datasets show that the algorithm has excellent performance and scalability.

7. "Direct Discovery of High Utility Itemsets without Candidate Generation" Utility mining emerged recently to address the limitation of frequent itemset mining by introducing interestingness measures that reflect both the statistical significance and the user's expectation. Among utility mining problems, utility mining with the itemset share framework is a hard one as no anti-monotone property holds with the interestingness measure. The state-of-the-art works on this problem all employ a two-phase, candidate generation approach, which suffers from the scalability issue due to the huge number of candidates. This paper proposes a high utility itemset growth approach that works in a single phase without generating candidates. Our basic approach is to enumerate itemsets by prefix extensions, to prune search space by utility upper bounding, and to maintain

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

original utility information in the mining process by a novel data structure. Such a data structure enables us to compute a tight bound for powerful pruning and to directly identify high utility itemsets in an efficient and scalable way. We further enhance the efficiency significantly by introducing recursive irrelevant item filtering with sparse data, and a lookahead strategy with dense data. Extensive experiments on sparse and dense, synthetic and real data suggest that our algorithm outperforms the state-of-the-art algorithms over one order of magnitude.

8. "Mining High Utility Itemsets in Big Data" In recent years, extensive studies have been conducted on high utility itemsets (HUI) mining with wide applications. However, most of them assume that data are stored in centralized databases with a single machine performing the mining tasks. Consequently, existing algorithms cannot be applied to the big data environments, where data are often distributed and too large to be dealt with by a single machine. To address this issue, we propose a new framework for mining high utility itemsets in big data. A novel algorithm named PHUI-Growth (Parallel mining High Utility Itemsets by pattern-Growth) is proposed for parallel mining HUIs on Hadoop platform, which inherits several nice properties of Hadoop, including easy deployment, fault recovery, low communication overheads and high scalability. Moreover, it adopts the MapReduce architecture to partition the whole mining tasks into smaller independent subtasks and uses Hadoop distributed file system to manage distributed data so that it allows to parallel discover HUIs from distributed data across multiple commodity computers in a reliable, fault tolerance manner. Experimental results on both synthetic and real datasets show that PHUI-Growth has high performance on large-scale datasets and outperforms state-of-the-art non-parallel type of HUI mining algorithms.

9. "Isolated items discarding strategy for discovering high utility item sets" 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. However, customers may purchase more than one of the same item, and the unit cost may vary among items. Utility mining, a generalized form of the share mining model, attempts to overcome this problem. Since the Apriori pruning strategy cannot identify high utility item sets, developing an efficient algorithm is crucial for utility mining. This study proposes the Isolated Items Discarding Strategy (IIDS), which can be applied to any existing level-wise utility mining method to reduce candidates and to improve performance. The most efficient known models for share mining are ShFSM and DCG, which also work adequately for utility mining as well. By applying IIDS to ShFSM and DCG, the two methods FUM and DCG+ were implemented, respectively. For both synthetic and real datasets, experimental results reveal that the performance of FUM and DCG+ is more efficient than that of ShFSM and DCG, respectively. Therefore, IIDS is an effective strategy for utility mining.

10. "ExMiner: An efficient algorithm for mining top-k frequent patterns" Conventional frequent pattern mining algorithms require users to specify some minimum support threshold. If that specified-value is large, users may lose interesting information. In contrast, a small minimum support threshold results in a huge set of frequent patterns that users may not be able to screen for useful knowledge. To solve this problem and make algorithms more user-friendly, an idea of mining the k-most interesting frequent patterns has been proposed. This idea is based upon an algorithm for mining frequent patterns without a minimum support threshold, but with a k number of highest frequency patterns. In this paper, we propose an explorative mining algorithm, called ExMiner, to mine k-most interesting (i.e. top-k) frequent patterns from large scale datasets effectively and efficiently. The ExMiner is then combined with the idea of "build once mine anytime" to mine top-k frequent patterns sequentially. Experiments on both synthetic and real data show that our proposed methods are more efficient compared to the existing ones.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

III. PROPOSED SYSTEM

In the proposed framework, we address the problems mentioned above by proposing another system for calculating the means and means responsible for a high utility configured in parallel extraction using TKU and TKO. Two types of production calculations called TKU (extraction of sets of utility elements Top-K) and TKO (sets of themes of extraction Top-K are proposed in one phase) to extract these series of elements without the need to establish a utility minimum. But the TKO algorithm have the main disadvantage of not mainly accumulating the result of TKO given the value of the garbage in the set of high utility items isthe result of the TKU algorithm is increased but the execution time is high, so the alternative solution is to find the efficient algorithm in the proposed combination of the TKO and TKU algorithm system. It can be said that the result of TKO Top K in one phase is given at the entrance of TKU Top K in the utility result of TKO and TKU is increased and the execution time is low. In the proposed system, a new algorithm is generated for combining the name TKO and TKU as TKO WITH TKU or TKMHUI Top k Main set of utility elements.

Module:

Module 1 - Administrator (Admin)

The administrator preserve database of the transactions made by customers. In the daily market basis, each day a new product is let go, so that the administrator would add the product or items, and update the new product view the stock details.

Module 2 - User (Customer)

Customer can purchase the number of items. All the purchased items history is stored in the transaction database.

Module 3 - Construction of Up Tree

In Up Tree Dynamic Table is generated by algorithms. Mainly the Up growth is considerable to get the PHUI item set.

Module 4 TKO and TKU Algorithms

In Combination of TKO and TKU algorithms first the TKO (Top k in one phase) algorithms is called and then output of TKO is given as the input of TKU (Top k in utility phases) then the actual result is TKU Result.

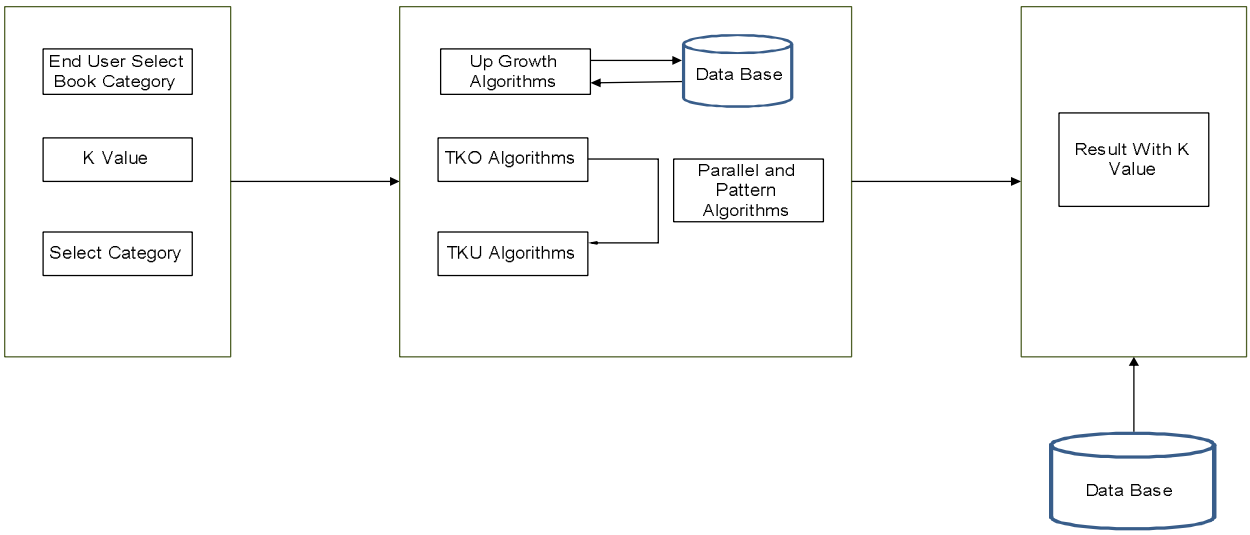

Fig 1: System Overview

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

IV. CONCULSION

In this paper, we looked at the question of the best sets of high-use mining mines, where k is the coveted number of highly useful sets of things to extract. The most competent combination of TKO WITH TKU of the TKO and TKU calculations is proposed to extract such sets of objects without establishing utility limits. Instead TKO is the first single phase algorithm developed for top-k HUI mining called PHUI (high potential set of utility elements) and PHUI is given to TKU in the utility phases. Empirical evaluations on different types of real and synthetic data sets display the proposed algorithms have good scalability in large data sets and the performance of the proposed algorithms are close to the optimal case of the state of the combination of both phases in an algorithm

REFERENCES

1. C. Ahmed, S. Tanbeer, B. Jeong, and Y. 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.
2. R. Chan, Q. Yang, and Y. Shen, “Mining high-utility itemsets,” in Proc. IEEE Int. Conf. Data Mining, 2003, pp. 19–26.
3. J. Han, J. Wang, Y. Lu, and P. Tzvetkov, “Mining top-k frequent closed patterns without minimum support,” in Proc. IEEE Int. Conf. Data Mining, 2002, pp. 211–218.
4. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc. ACM SIGMOD Int. Conf. Manag. Data, 2000, pp. 1–12.
5. P. Fournier-Viger, C. Wu, and V. S. Tseng, “Novel concise representations of high utility itemsets using generator patterns,” in Proc. Int. Conf. Adv. Data Mining Appl. Lecture Notes Comput. Sci., 2014, vol. 8933, pp. 30–43.
6. P. Fournier-Viger and V. S. Tseng, “Mining top-k sequential rules,” in Proc. Int. Conf. Adv. Data Mining Appl., 2011, pp. 180–194.
7. J. Liu, K. Wang, and B. Fung, “Direct discovery of high utility itemsets without candidate generation,” in Proc. IEEE Int. Conf. Data Mining, 2012, pp. 984–989.
8. Y. Lin, C. Wu, and V. S. Tseng, “Mining high utility itemsets in big data,” in Proc. Int. Conf. Pacific-Asia Conf. Knowl. Discovery Data Mining, 2015, pp. 649–661.
9. Y. Li, J. Yeh, and C. Chang, “Isolated items discarding strategy for discovering high-utility itemsets,” Data Knowl. Eng., vol. 64, no. 1, pp. 198–217, 2008.
10. T. Quang, S. Oyanagi, and K. Yamazaki, “ExMiner: An efficient algorithm for mining top-k frequent patterns,” in Proc. Int. Conf. Adv. Data Mining Appl., 2006, pp. 436 – 447.