Mining Top K High Utility Items for Pharmacy Data

Bharathi R K
PhD, Department of MCA, SJCE,Mysuru

Maithri M
Department of MCA, SJCE,Mysuru

ABSTRACT
Increase in the range of real world applications has led to market data analysis and stock market predictions, thus an emergence of High Utility Itemset (HUI) as one of the most significant research issues.Mining HUI is a technique used to discover itemsets with utility values above a given thresholdin a transaction database. HUI reflects the impact of different items and helps in decision-making process of many applications. Algorithms that can efficiently prune candidates are known to be more efficient. "Mining top K-HUI" can be accomplished by three distinct algorithms such as, Vertical Frequent Format Mining algorithm, Maximum Utility Growth algorithm and Top K High Utility algorithm. An attempt is made to study the behavior of algorithms in terms of efficiency by measuring effectiveness in pruning candidates. To demonstrate the same, in this paper we have considered pharmacy dataset of Mysuru district for the experimentation.

Keywords
High utility mining, Top-K utility item, pattern mining.

1. INTRODUCTION
Data mining is the process of extraction and analysis of relevant data from different perspectives and summing it up into useful information. Various methodologies have been proposed for this purpose. Frequent pattern mining is one of its kind, which enables in finding frequent patterns in transaction databases. Many popular algorithms such as Apriori, FP Growth and DIC have been proposed to address this problem. These algorithms take transaction database and a parameter, minimum support threshold [9] as input and return all set of items (itemsets) that appears in *minsup*transactions. However, the traditional model of FIM may discover a large amount of frequent, yet, low revenue itemsets and lose the information on valuable itemsets having low selling frequencies. Hence, discoveringitemsets with high utilities such as high profits cannot be satisfied by FIM. To overcome this problem high utility itemset mining was proposed.

The problem of frequent pattern mining is extended to HUI mining. HUI mining identifies itemsets whose utility satisfies a given threshold and allows users to quantify the usefulness or preferences of items using different values. But setting an appropriate minimum utility threshold is a difficult problem. Setting minimum threshold low results in huge number of itemsets, on the other hand, setting threshold to higher valuegives very few itemsets. Setting appropriate minimum utility threshold by trial and error is not very efficient. Hence, an algorithm is required which takes the number of result we want as parameter k. Setting k is more intuitive than setting the threshold because k represents the number of itemsets that the user wants to find."Mining top K HUI for pharmacy data" aims at finding the top purchased products from a real pharmacy dataset. Vertical frequent format mining algorithm, Maximum utility growth algorithm and Top K high utility algorithm are used for the implementation and a comparative analysis shows the efficiency of algorithms by effectively pruning candidate itemsets.

2. RELATED WORK
The study of mining the patter for market analysis based on the sales report and transaction databases was inintiated in early 90's. An initial work was proposed in the year 1993 by Agarwalwith the concept of frequent pattern mining for market basket analysis, which is a type of association rule mining [8]. David C. et al. [11] worked on recent advances in parallel frequent pattern mining and analysed them through the Big Data lens and tried to address fewkey challenges such as memory scalability, work partitioning, and load balancing. Sheila A. Abaya proposed a new mechanism in which the Apriori algorithm can be improved with the introduction of key factors such as, set size and set size frequency which in turn were used to eliminate non-significant candidate keys. Jiao Yabing proposed an improved algorithm which is based on classical Apriori algorithm. The results of the improved algorithm proved to be reasonable, effective and could extract more value information. Hua-Fu Li et al. proposed two efficient algorithms namely, MHUI-BIT and MHUI-TID. The algorithms were used for mining High Utility Itemsets in Data Streams within a transaction-sensitive sliding window. Datasets used were Synthetic data generator and IBM generator. Experimental results showed that 8 candidates generated when minimum utility threshold is 1%. .M. Sulaiman Khan et al. [6] conducted experiments on classical association rule mining (ARM) and weighted-ARM revealing 60 frequent items were generated for minimum support of 1%. Chowdhury Farhan Ahmed et al. [5] proposed three novel tree structures to efficiently perform incremental and interactive HUP mining. Kosarak and Chain-store datasets were used. Experimental results showed that 2000k candidates were pruned for minimum utility threshold of 0.35%. Hua-Fu Li et al. [12] proposed efficient sliding window-based algorithms that measures utility of items with and without negative profits. Synthetic dataset was used and 200 items were generated for external utility of 250. Cheng-Wei Wu et al. proposed algorithms for Complex Event Sequences. The work incorporated the concept of utility into episode mining. It addressed a new problem of mining high utility episodes from complex event sequences. UP-Span (Utility episodes mining by spanning prefix) was the algorithm used. Experimental results showed that 1000k candidates generated for minimum threshold of 1%.

Although the topic of HUM is not new and many studies have addressed the topic of FIM and HUM from transaction database, the approach for mining high utility items in our work is different.

3. MINING TOP-K HIGH UTILITY ITEMS
In this section, we introduce a methodology followed to implement the proposed work. Mining top-k high utility items (MTKHUI)is accomplished by three distinct algorithms namely, Vertical frequent format mining (VFFM) algorithm, Maximum utility growth (MUG) algorithmand Top K high utility (TKU) algorithm.

3.1 VFFM Algorithm

Vertical Format Frequent Mining (VFFM) algorithm is designed to find frequent items from the dataset. The algorithm calculates the row sum values (count) for each row. Count values of 1-itemsets are checked against minimum supports which satisfy minimum support produces frequent 1-itemsets. The items that are less than the minimum support are considered as infrequent. It can be pruned, thereby save the effort of unnecessarily obtaining their counts during the subsequent process. From the frequent 1-itemset, 2-itemsets are generated without scanning the original database. Frequent 2-itemsets are identified by calculating the count values of each item and checking the obtained value against the minimum support. The same procedure applied to find the frequent k-itemsets from (k-1) frequent itemsets.

Steps involved

1. Scan the transaction database and perform transpose operation.

2. Store attributes values in the form of binary data.

3. Calculate the row sum

4. Count values of 1-itemsets till k-itemsets are calculated.

5. Compare count with support value and identify frequent items.

3.2 MUG Algorithm

In high utility itemset mining techniques, reducing the number of candidates is a crucial challenge because identifying actual high utility itemsets from candidates is very time consuming task. It means the more number of candidate itemsets leads to the more execution time, and thus it is needed to decrease the number of candidates for efficiently mining high utility itemsets. MU-Growth (Maximal Utility Growth) can reduce the number of candidate itemsets effectively with real item utilities, minimum and maximum item utilities, item utilities, and supports of local items. If estimated maximum utility of a candidate itemset is less than *minutil*, the candidate itemset is pruned. In other words, the candidate itemset is not generated in mining process

Steps involved

1. Construct MIQ tree

2. Select item from item dataset

3. Scan each transaction to find the selected item and record the quantity associated with the item in every transaction.

4. Calculate TWU value for each item

5. Reconstruct the tree by arranging items in TWU descending order

6. Calculate real item utilities

7. Using real item utility and TWU value of item, calculate estimated maximum itemset utility.

8. If the estimated maximum itemset utility is greater than threshold, the item is considered as HUI.

3.3 Top-K High Utility Algorithm

Top-k pattern mining algorithm sets minimum support threshold *minsup*to 0 to ensure that all the top k patterns will be found. Then, the algorithm starts searching for patterns by using a search strategy. As soon as a pattern is found, it is added to a list of patterns L ordered by the support of patterns. The list L is used to maintain the top-k patterns found until now. The value of *minsup*is raised to the support of the least interesting pattern in L,once k patterns are found. Raising *minsup*is used to prune the search space when searching for more patterns. Thereafter, each time when the pattern is found that meets the support threshold, it is inserted into L and the patterns in L that does not satisfy the threshold are removed from L, further the threshold is raised to the support of the least frequent patterns in L. This procedure is continued until no pattern is found by the search strategy.

Steps involved

1. Scan the transaction database

2. Generating all the itemsets

3. Calculate MIU and TWU

4. Choose the entire potential candidate for high utility itemsets with the increasing threshold method.

5. Choose the top-k high utility itemset

4. DATASET

To understand the behavior of algorithms, experimental analysis is performed using pharmacy dataset. An attempt was made to collect the information regarding the sales of medicines through the distributor of mysore district. The Mysore District distributor distributes the medicine to the pharmacy centers of various localities. The setup is tested against three variants of the dataset [Table 1]. Demographic information about the pharmacy products include, item id, item name, quantity, unit price and Transaction information include transaction id, area, date, items and total number of items.

Table 1. Characteristics of Dataset

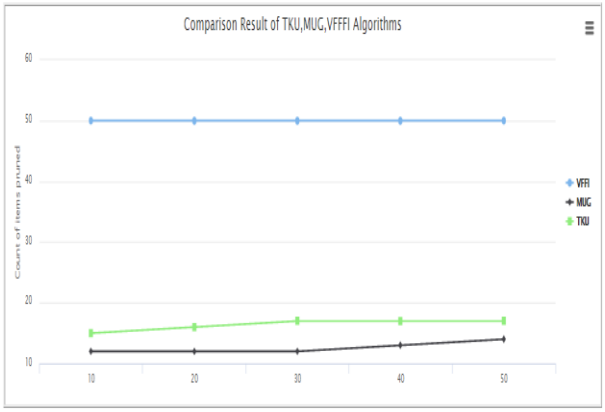
Dataset	# Trans.	Avg. length of trans.	# Items
PH1	520	4.4	50
PH2	1250	6.2	100
PH3	3822	7.3	200

5. EXPERIMENTAL EVALUATION

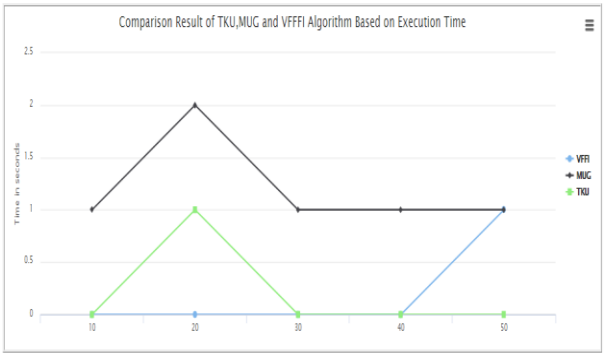
In this section, proposed work is to evaluate the performance of three algorithmsviz: VFMM, MUG and Top-K high utility dataset. The performance of the algorithms was evaluated by experimenting on three variants of pharmacy dataset (PH1, PH2 and PH3). The variants differ by number of items and transactions. Table 1 shows the characteristics of the dataset. The performance of algorithms is evaluated based on number of items pruned for certain threshold and time take by each algorithm. Experiments were performed on a computer with a 2.20 GHz Intel Core Processor and 4 GB of memory, running Windows 7. All the algorithms are implemented in Java.

Figure 1, 2 and 3 shows performance of algorithms on PH1, PH2 and PH3 respectively. Figure 1(a), 2(a) and 3(a) shows variations in number of items pruned by three algorithms. X-axis indicates the minimum utility threshold and Y-Axis indicates the number of items pruned. It can be observed that, number of items pruned by TKU is comparatively more than MUG. Figure 1(b), 2(b) and 3(b) shows variations in time taken by three algorithms. X-axis indicates the minimum utility threshold and Y-Axis indicates time in seconds. It can be observed that, although, time taken by TKU algorithm is

more compared to MUG and VFFM, its efficiency in pruning candidates is relatively high.

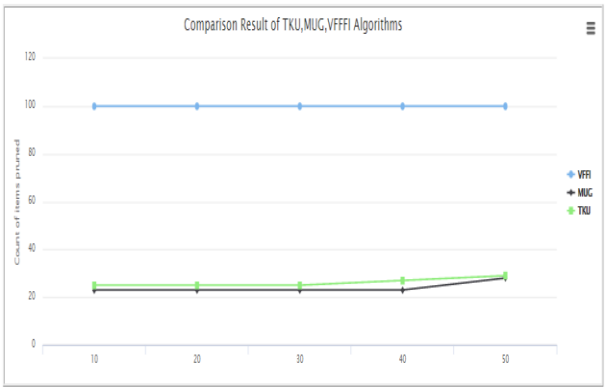


(a) Number of items pruned

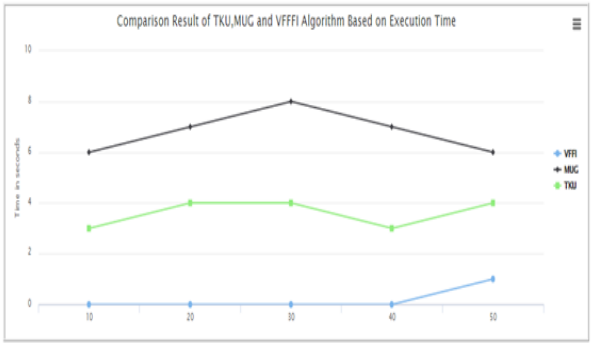


(b)Time Taken

Fig 1: Performance of algorithms on PH1

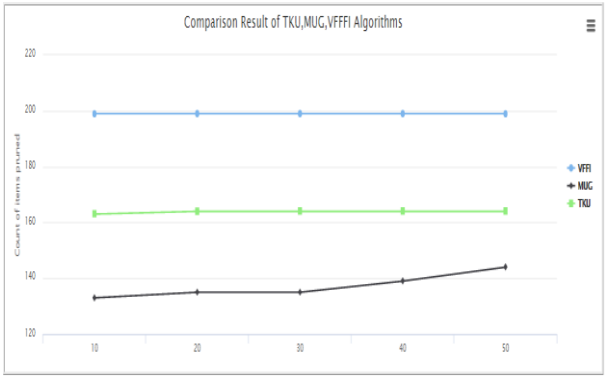


(a) Number of items pruned

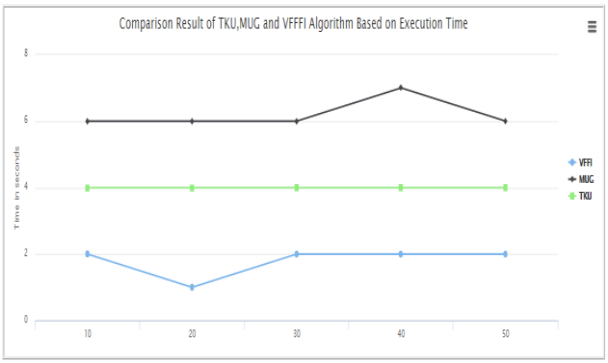


(b) Time Taken

Figure 2.Performance of algorithms on PH2



(a) Number of items pruned



(b) Time Taken

Figure 3: Performance of algorithms on PH3

6. CONCLUSION

Frequent pattern mining and high utility mining techniques provide a way to find the items with high utility taking item quantity and profit into account. The proposed work provides a better approach in finding the top purchased items by providing the desired number of items to be displayed and therefore avoiding the problem of setting the threshold. This study also helps in analyzing the market states of pharmacy and gives an insight about the diseases spreading in various areas of the city.

Mining top K high utility items for pharmacy data can be of use in sales analysis in finding top purchased products. The algorithmic result shows slight variations with the actual result as profit of each item is taken into account. Algorithms can be further enhanced retrieve result based only quantity of each item. The implementation takes area as a parameter based on which top purchased items for the given area is being displayed. Providing generic names of items can be helpful in identifying the kind of product being purchased in that specific area.

7. REFERENCES

[1] Jiao Yabing, "Research of an Improved Apriori Algorithm in Data Mining Association Rules" International Journal of Computer and Communication Engineering Vol. 2, No. 1, January 2013

[2] Borgelt, C: "An implementation of fp-growth algorithm". In proceedings of the 1st international workshop on open source data mining: Frequent pattern mining implementations, OSDM 2005, NY, USA, pp. 1-5.

[3] GostaGrahne and Jianfei Zhu "Fast Algorithms for Frequent Itemset Mining Using FP-Trees" IEEE

Transactions On Knowledge And Data Engineering, Vol. 17, No. 10, October 2005 pp 1347-1362

[4] David C. Anastasiu and Jeremy Iverson and Shaden Smith and George Karypis "Big Data Frequent Pattern Mining" Springer, (2014) pp 225-259.

[5] ChowdhuryFarhan Ahmed, Syed KhairuzzamanTanbeer, Byeong-SooJeong, and Young-Koo LeeEfficient, "Tree Structures for High Utility Pattern Mining in Incremental Databases".IEEE transactions on knowledge and data engineering, vol. 21, no. 12, December 2009. Pp 17008-1721

[6] M. Sulaiman Khan, Maybin Muyeba1, FransCoenen School of Computing "A Weighted Utility Framework for Mining Association Rules" EMS 2008: 87-92

[7] HeungmoRyanga, UnilYuna, and Keun Ho Ryu, "Discovering high utility itemsets with multiple minimum supports" Intelligent Data Analysis, vol. 18, no. 6, pp. 1027-1047, 201

[8] Sudip Bhattacharya and DeeptyDubey, "High Utility Itemset Mining" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 8, August 2012)

[9] "Fast Frequent Pattern Mining Using Vertical Data Format for Knowledge Discovery". International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-5, Issue-5) Expert Systems with Applications 41 (2014) 3861–3878.

[10] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases (International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc, 1994), pp. 487–499.

[11] Borgelt C. Frequent item set mining (Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2012, 2(6)), pp. 437–456. https://doi.org/10.1002/widm.1074

[12] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu "Efficient Algorithms for Mining Top-K High Utility Itemsets" IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 1, January 2016, Pages 54-67.

[13] Data Mining Concepts and Techniques by Han & Kamber (2nd edition) pages 9-23

[14] Data Mining techniques by Arun K Pujari (1st edition) pages (69-100)