Empirical Methodology and Innovative Algorithm for Mining High Utility Itemset on Temporal Data

Rachith Raj, Ajay Krishnan, C.V. Prasanna Kumar

Abstract: Ever since the humans found the significance of data we are using different methodologies to extricate data patterns in the most useful way to reach at various conclusions based on our needs. The proper utilization of data will leads to meaning full results. Here our focus is to expand the concept of "High utility itemset mining" and also introducing the concept of "timestamp" on it. Here we proposed an effective algorithm to find the best possible data patterns that generate maximum profit within the given "time period" and also proposed a new framework to calculate the "utility support". Our analysis can be utilized to improve profitability in business and also make better decisions for improving sales in the organization.

Key Terms: Data mining, Frequent Itemset Mining, High utility Itemset mining, temporal data.

I. INTRODUCTION

The Method of retrieving vital information or facts from colossal amounts of data is known as ***data mining***. The most crucial study of Data Mining is the retrieval of transactional information. Discovering frequently used & significant information from an expansive database exhibits an imperative part in data mining**. *Frequent itemset mining*** is the core method in data mining, detecting fascinating and frequently appearing items from the dataset. Frequent Itemset Mining could be a strategy for market basket analysis that allows retailers to identify relationships between the items that consumers buy from their business. But when we think in terms of business perspective the sole purpose is to find the pattern that generates maximum profit gaining itemsets by overcoming the limitations of FIM by introducing the new concept on "high utility itemset mining". ***High utility itemset mining*** is an adjunct of the issue of frequent itemset mining**. The** High utility Itemsets allows the merchants to understand consumer's purchase patterns that generate more profit. A ***Temporal data*** is a data related to time instances, time is an important aspect when it comes to a business perspective. Here we introduce the time cube ("TC") which is a new conception for considering time hierarchies in the mining process and we conduct factual experiments on factitious datasets to analyze the algorithm's efficiency. The structure of our paper is as follows-We discuss some related work in conjunction with our research in "section II". We elaborate our study based on the experimental evaluation in "section III". We review our result in sectionIV. We wind up our paper with a conclusion in "sectionV".

Revised Manuscript Received on March 30, 2020.
***** Correspondence Author

Rachith Raj, P.G Student, Department of Computer Science and I.T Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

Ajay Krishnan*, P.G Student, Department of Computer Science and I.T Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

C.V. Prasanna Kumar, Assistant Professor , Department of Computer Science and I.T, Amrita School of Arts and Sciences, Kochi

II. RELATEDWORK

There are lots of studies happened recently in the field of FIM(Frequent Itemset Mining), commonly used traditional algorithm in the area of FIM are A-priori, fp-growth, Lcm etc. Various studies are conducted to optimize and extend FIM algorithm in order to acquire better results. Several works of literature are available on the area of frequent itemset mining and other related fields. Gorbani and abessi have efficiently extended the apriori algorithm by adding the time related information and they also introduced a new value of threshold to resolve the overestimation problem as relates to time, their study was about finding frequent itemset from temporal data[1].HoudaEssalmi et al has proposed the "a-priorimin algorithm" that optimized a-priori algorithm to improve the performance by instigating a strategy for calculating the FI and the above approach has lessen the execution time by improving the performance [2]. Conventional research in mining retail information in any case does not take under consideration the product's cost's and how such settings can affect potential demand, Yen-Liang Chen et al proposed the new methodology by proposing a representation strategy to consolidate cost information into historical transaction data[3]. But in business perspective our concern will be regarding the profit this study has taken us to the most emerging area in data mining called high utility itemset mining. Prashant N et al has proposed the "Tku" & "Tko" that mines from "Tkhuis" without the threshold [4]. YING LIU et al presented a 2-phase algorithm that prunes out the number of items to acquire the highly profited itemset using a different approach [5]. Alva Erwin et al proposed an algorithm to undergo parallel projection strategy to utilize disk space when main memory cannot accommodate large dataset [6], this will give an idea about how we can mine efficiently from colossal dataset. Han et al [7]. Has developed an effective method to mine the "segment wise periodicity" that helps us to understand periodic pattern mining on temporal data. Rakesh Agarwal &RamakrishnanSrikant [8] has presented two algorithm "apriori and aprioriTID" & also presented "apriorihybrid" algorithm that become basis for several studies in the related areas.

III. MINING HIGH UTILITY ITEMSET ON TEMPORAL DATA

A. New Framework

Let **X** = { x1,x2, x3…….xn} be a group of items in the dataset and **Ds** be the Dataset containing the transactions. Each transaction has an unique transactional ID called "**TID**" and also the time in which it occur (tt). When we deal with data related to time we define a new term called Time Cube (TC).

Refer Fig Below

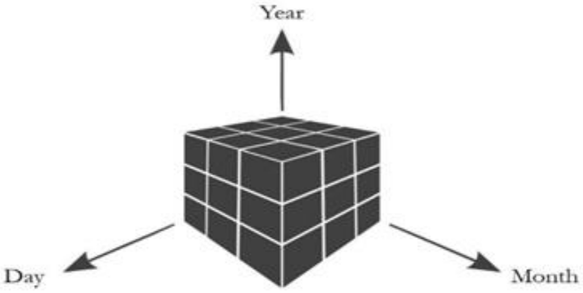


Fig. 1. The cubic model with 3 time hierarchies

Frequent itemset mining has few limitations, as the quantity purchased is not taken into account, all the items have the same importance and profit is not calculated, to resolve these limitations the problem of "frequent itemset mining on temporal data" is redefined as the problem of "high utility itemset mining on temporal data".

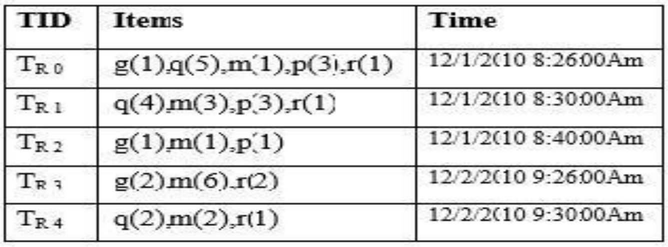
Let "**X** "be the itemset, and **xi** is the items in the itemset and value of **I** can varies from "**0 to n**".

U= $\sum_{i=0}^{m}(P(xi) * Q(xi))$ **/m**

P(x) = unit profit of an item, **Q(x)** = quantity of an item in the transaction and **m**= number of items in the itemset.

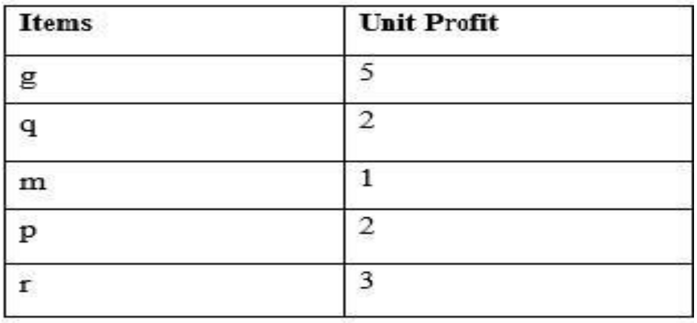
(Refer Table 1 and Table2).

(Table.1 Illustration of transactional data with quantity)



TID	Items	Time
T_{R0}	g(1),q(5),m(1),p(3),r(1)	12/1/2010 8:26:00Am
T_{R1}	q(4),m(3),p(3),r(1)	12/1/2010 8:30:00Am
T_{R2}	g(1),m(1),p(1)	12/1/2010 8:40:00Am
T_{R3}	g(2),m(6),r(2)	12/2/2010 9:26:00Am
T_{R4}	q(2),m(2),r(1)	12/2/2010 9:30:00Am

(Table.2 Items and their Unit Profit)



Items	Unit Profit
g	5
q	2
m	1
p	2
r	3

Utility(X) = $\sum_{n=0}^{k}$ U/k

Where Utility(X) is the utility for the items and **k** refers to the number of transactions in which an item is present.

B. Recommended Algorithm for mining high utility itemset on temporal data

The Suggested Algorithm Intended to find itemset X That generates maximum profit above the utility threshold considering the time cube on temporal data. Candidate having utility threshold above the minimum utility threshold within that time cube is considered as the high profitable itemset. By Implementing below 3 algorithms we able to generate most profitable itemsets along with their purchased date and their respective profit.

Algorithm 1.1: High utility itemset Mining
Algorithm with time cube

```
Input: [Data (Ds), minth, TC] Output: $L_K$
L1(Ds,minth,TC)
k=2
While ($L_{k-1}$! = NULL)
$C_K$=CandidateGen ($L_{K-1}$)
  For (i=0: n, X=$C_K(i)^{TC}$)
     CalcUtility ($X^{TC}$)
  Endfor
  For (every$Y_i$→$M_j$→$D_K$|$Y_i$,$M_j$,$D_K$∈TC)
     If (utility ($^{CTC}$) >= minth)
     .L $^{TC}$=LK U$C^{TC}$
  Endfor
  k++
End while
```

```
Input: [Data (Ds), minth, TC] Output: L1
$L_1$=∅
For(Xi, i→0 :n, Xi∈I)
    For( $Y_i$→$M_j$→$D_K$| $Y_i$, $M_j$, $D_K$∈TC)
        IF( Xi contains)
           Sum += Utility(Xi)
           Count Support Xi
    Endfor
    Utility(Xi) = Sum/count Support
    If( Utility(Xi) >=minth)
        $L_1$= $L_1$U $X^{TC}$
    Endif
Endfor
```

Algorithm 1.3: Candidate generation Algorithm

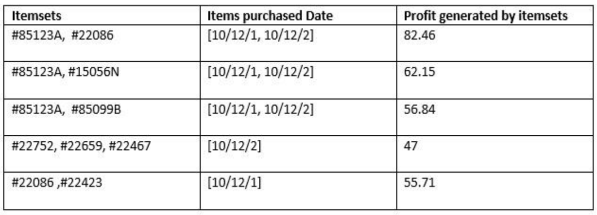
```
Input: [LK-1] Output: $C_K$

$C_K$=∅
For all pairs of $L_{i1}$, $L_{j1}$∈ $L_{K-1}$
        Cr = $L_{i1}$⋈$L_{j1}$
        If |Cr|=K
           Put Cr into $C_K$
        Endif
Endfor
```

Algorithm (1.1) is the procedure to mine the high utility itemset within the time cube "TC". Algorithm (1.2) will get input parameters such as Data, minimal utility threshold, TC which is passed from algorithm 1.1, utility weight of each itemset within the time cube is calculated by our new framework. All the itemset above the minimal utility threshold will be added to L1. L1 is returned to algorithm (1.1). Algorithm (1.3) is a procedure to generate candidate itemset (until Lk-1 not equal to null). The algorithm (1.3) use "joint operator" that generates higher level itemset from the lower level. The result of candidate itemset will return to algorithm (1.1) and algorithm (1.1) will compute the "utility" of each candidate itemset, all the itemset above minimal threshold within the time cube that provides maximum profit is considered as our outcome.

IV. RESULT AND DISCUSSION

We have implemented the proposed algorithm with the real dataset, on our implementation we go through various steps like data loading, Data preprocessing, candidate key generation and calculated the utility support of the "candidate itemset" using our new framework and from that, we able to generate L1 and process goes through until "Lk-1 not equal to null" and we able to find the most profitable "itemsets pattern" in the given "time interval" with at most accuracy. The tabular representation of the result is given below-



Itemsets	Items purchased Date	Profit generated by itemsets
#85123A, #22086	[10/12/1, 10/12/2]	82.46
#85123A, #15056N	[10/12/1, 10/12/2]	62.15
#85123A, #85099B	[10/12/1, 10/12/2]	56.84
#22752, #22659, #22467	[10/12/2]	47
#22086 ,#22423	[10/12/1]	55.71

(Tabular representation of experimental result)

V. CONCLUSION

In our study, we have gone through the mining of "high utility Itemsets" together with "Temporal pattern" for that we proposed an algorithm that extracts itemsets generating high profit along with their "time-related information", that help us to identify the high profit generating purchase pattern on specific "time interval". and we able to overcome few limitations existed in frequent pattern mining and extended " the concept of high utility itemset mining" by adding " time cube (TC) " & also we able to introducing the new framework to calculate the 'utility support' by considering the quantity, unit profit, etc. to obtain the result. Our study can be utilized in the business domain to improve profitability as well as sales, and from the manager's point of view, they able to make better decisions to improve their organization's business strategy.

REFERENCES

1. MazaherGhorbani and MasoudAbessi, "A New Methodology for Mining Frequent Itemsets on Temporal Data"IEEE Transactions on Engineering Management, Volume: PP, Issue: 99, Year-June 2017
2. HoudaEssalmi, Mohamed El Far, Mohammed El Mohajir, Mohamed Chahhou, "A Novel Approach for mining frequent itemsets: AprioriMin" 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) Pages: 286- 289, Year: 2016.
3. Y.-L. Chen, T. C.-K. Huang and S.-K. Chang, "A novel approach for discovering retail knowledge with price information from transaction databases"
4. Dr. Prashant N. Chatur, Snehal D. Ambulkar "Efficient Algorithms for mining HighUtilityItemset".
5. Y. Liu, W. Liao, and A. Choudhary, "A two-phase algorithm for fast discovery of high utility itemsets", In Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Vol. 3518, pp. 689-695, 2005.
6. Alva Erwin, Raj P. Gopalan, N.R. Achuthan, A Bottom-Up Projection Based Algorithm for Mining High UtilityItemsets.
7. Jiawei Han, Wan gong, yiwen yin, Mining Segment-Wise Periodic Pattern in Time Related Databases.
8. Rakesh Agarwal and RamakrishnanSrikant, 1994, Fast Algorithm for Mining Association rules. In proc.1994 int. conf, Very Large Databases.
9. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Amsterdam, the Netherlands: Elsevier, 2011.
10. Philippe Fournier-Viger, Jerry Chun-Wei Lin, Roger Nkambou, Bay VO and Vincent Tseng. High-Utility Pattern Mining: Theory, Algorithms and Applications.
11. T. Mitsa, Temporal Data Mining, Boca Raton, FL, USA: CRC Press, 2010.
12. Charu C. Aggarwal, Jiawei Han, Frequent Pattern Mining, 2014.

AUTHORS PROFILE

Rachith Raj, currently pursuing Master of Computer Applications from Amrita Vishwa Vidyapeetham, Coimbatore India at Amrita Kochi Campus. He Completed Bachelor of Computer Applications from Mahatma Gandhi University, Kerala, India. His areas of Interest Includes Data mining, Information Retrieval and Project Management.

Ajay Krishnan, currently pursuing 5 Year Integrated BCA-MCA from Amrita Vishwa Vidyapeetham headquarters at Coimbatore India at Amrita Kochi Campus. His areas of interest include Data mining, Network Security and Project Management.

C.V .Prasanna Kumar, currently Assistant professor at Amrita Vishwa Vidyapeetham, India at Amrita Kochi Campus. Qualifications: MCA, PGDCA. The Areas of Interest are Cryptography, Database, Networking and Data mining. Also 22 years academic experience, including 10 years in South Korea, Cambodia, Laos, Vietnam.