A Relative Study on Various Techniques for High Utility Itemset Mining from Transactional Databases

Karishma Gaikwad[1], Pratiksha Sambharkar[2],Shubhangi Hinge[3], Sanket Mankar[4], Trushali Raut[5], Prof. Sandip Kamble[6]

[12345]Student, Department of Computer Technology, Rajiv Gandhi College of Engineering & Research Nagpur, India

[6]Assistant Professor, Department of Computer Technology, Rajiv Gandhi College of Engineering & Research Nagpur, India

---***---

Abstract— *Data mining can be characterized as a movement that concentrates some learning contained in expansive transaction databases. Customary information mining strategies have concentrated to a great extent on finding the things that are more frequent in the transaction databases, which is additionally called frequent itemset mining. These information mining strategies depended on bolster certainty display. Itemsets which seem all the more frequently in the database must be of all the more intending to the client from the business perspective. In this paper we show a developing territory called as High Utility Itemset Mining that finds the itemsets considering the recurrence of the itemset as well as utility connected with the itemset. Each itemset have esteem like amount, benefit and other client's advantage. This esteem connected with each thing in a database is known as the utility of that itemset. Those itemsets having utility qualities more noteworthy than given edge are called high utility itemsets. This issue can be distinguished as mining high utility itemsets from transaction database. In numerous regions of professional retail, stock and so on basic leadership is vital. So it can help in mining calculation, the nearness of everything in a transaction database is spoken to by a paired esteem, without considering its amount or a related weight, for example, cost or benefit. However amount, benefit and weight of an itemset are noteworthy for distinguishing certifiable choice issues that require expanding the utility in an association. Mining high utility itemsets from transaction database introduces a more noteworthy test as contrasted and frequent itemset mining, since hostile to monotone property of frequent itemsets is not appropriate in high utility itemsets. In this paper, we display a study on the flow condition of research and the different calculations and systems for high utility itemset mining.*

Keywords— **Data Mining, Frequent Itemset Mining, Utility Mining, High Utility Itemset Mining**

I. INTRODUCTION

Data mining and learning disclosure from information bases has gotten much consideration as of late. Information mining, the extraction of concealed prescient data from substantial databases, is an intense new innovation with awesome potential to help organizations concentrate on the most vital data in their information distribution centers. Learning Discovery in Databases (KDD) is the non-unimportant procedure of recognizing legitimate, already obscure and conceivably helpful examples in information. These examples are utilized to make forecasts or characterizations about new information, clarify existing information, outline the substance of a huge database to bolster basic leadership and give graphical information perception to help people in finding further examples. Information mining is the way toward uncovering nontrivial, beforehand obscure and conceivably valuable data from huge databases. Finding valuable examples covered up in a database assumes a basic part in a few information mining undertakings, for example, frequent example mining, weighted frequent example mining, and high utility example mining. Among them, frequent example mining is a principal inquire about subject that has been connected to various types of databases, for example, transactional databases, spilling databases, and time arrangement databases, and different application spaces, for example, bioinformatics, Web click-stream investigation, and versatile situations. In perspective of this, utility mining develops as a critical theme in information mining field. Mining high utility itemsets from databases alludes to finding the itemsets with high benefits. Here, the significance of itemset utility is intriguing quality, significance, or gainfulness of a thing to clients. Utility of things in a transaction database comprises of two viewpoints:-

- The significance of particular things, which is called outside utility.
- The significance of things in transactions, which is called inner utility.

Utility of an itemset is characterized as the result of its outside utility and its inner utility. An itemset is known as a high utility itemset. In the event that its utility is no not exactly a client determined least utility limit; else, it is known as a low-utility itemset.

Here we are examining some fundamental definitions about utility of a thing, utility of itemset in transaction, utility of itemset in database furthermore related works and characterize the issue of utility mining and after that we will present related systems. Given a limited arrangement of things I= {i1, i2, i3… im} everything ip (1≤p≤m) has a unit benefit pr (ip). An itemset X is an arrangement of k particular things I= {i1, i2, i3… ik}, where ij I, 1≤j≤k. k is the length of X. An itemset with length k is known as a k itemset. A transaction database D = {T1; T2; . . . ; Tn} contains an arrangement of transactions, and every transaction Td (1≤d≤n) has a one of a kind identifier d, called TID. Everything ip in transaction Td is connected with an amount q (ip, Td), that is, the bought amount of ip in Td.

Definition 1: Utility of an item ip in a transaction Td is denoted as u (ip, Td) and defined as pr(iP)×q(ip,Td)

Definition 2: Utility of an itemset X in Td is denoted as U(x, Td) and defined as Σip∈X∨X⊆Tdu(ip,Td)

Definition 3: Utility of an itemset X in D is denoted as u(X) and ΣX⊆Td∧Td⊆D u(X,Td)

Definition 4: An itemset is called a high utility itemset if its utility is no less than a user-specified minimum utility threshold or low-utility itemset represented by min-util.

Table 1: An Example Database

TID	Transaction	TU
T1	(A,1) (C,10) (D,1)	17
T2	(A,2) (C,6) (E,2) (G,5)	27
T3	(A,2) (B,2) (D,6) (E,2) (F,1)	37
T4	(B,4) (C,13) (D,3) (E,1)	30
T5	(B,2) (C,4) (E,1) (G,2)	13
T6	(A,1) (B,1) (C,1) (D,1) (H,2)	12

Table 2: A Profit Table

Profit	5	2	1	2	3	5	1	1
Item	A	B	C	D	E	F	G	H

From table 1 and 2

U ({A, T1}) =5×1=5

U({AD,T1})=u({A,T1})+u({D,T1})=5+2=7

U ({AD}) =u ({AD, T1}) +u ({AD, T3}) =7+17=24

U ({BD}) =u ({BD, T3}) +u ({BD, T4}) =16+18=34

II. Related Work

A) *Frequent Itemset Mining*

High utility itemset mining finds all high utility itemsets with utility qualities higher than the base utility edge in a transaction database [14]. The utility of an itemset alludes to its related esteem, for example, benefit, amount or some other related measure. Some standard techniques for mining affiliation rules [1, 7] that is finding frequent itemsets depend on the bolster certainty demonstrate. They locate all frequent itemsets from given database. The issue of frequent itemset mining [1, 2] is finding the entire arrangement of itemsets that show up with high event in transactional databases. However the utility of the itemsets is not considered in standard frequent itemset mining calculations. Frequent itemset mining just considers whether a thing has happened frequently in database, however overlooks both the amount and the utility connected with the thing. In any case, the event of an itemset may not be a sufficient pointer of intriguing quality, since it just demonstrates the quantity of transactions in the database that contains the itemset. It doesn't uncover the genuine utility of an itemset, which can be measured regarding cost, amount, benefit, or different articulations of client inclination [17]. In any case, utility of an itemset like benefit, amount and weight are imperative for tending to certifiable choice issues that require expanding the utility in an association. In numerous regions of professional retail, stock, advertising research and so on basic leadership is essential. So it can help in examination of offers, advertising procedures, and outlining diverse sorts of index.

Illustration:

Consider the little case of transaction database, a client purchases numerous things of various amounts in a deal transaction. All in all, everything has a specific level of benefit. For example, expect that in an electronic superstore, the benefit (in INR) of 'Printer Ink' is 5, and that of 'Laser Printer' is 30. Assume 'Printer Ink' happens in 6 transactions, and 'Laser Printer' happens in 2 transactions in a transactional database. In frequent itemset mining, the event recurrence of 'Printer Ink' is 6, and that of 'Laser Printer' is 2. 'Printer Ink' has a higher recurrence. In any case, the aggregate benefit of 'Laser printer' is 60, and that of 'Printer ink' is 30; in this manner, 'Laser Printer' contributes more to the benefit than 'Printer Ink'. Frequent itemsets are essentially itemsets with high frequencies without considering utility. Be that as it may, some infrequent itemsets may likewise contribute more to the aggregate benefit in the database than the frequent itemsets. This case demonstrates the way that frequent itemset mining methodology may not generally fulfill the retail business objective. In all actuality a most profitable clients who may purchase full valued things or high edge things which may not present from substantial number of transactions are vital for retail business since they don't purchase these things frequently.

B) *High Utility Itemset Mining*

The constraint of frequent itemset mining lead scientists towards utility based mining approach, which permits a client to helpfully express his or her points of view concerning the value of itemsets as utility and after that find itemsets with high utility qualities higher than given limit [3]. Amid mining process we ought not recognize either frequent or uncommon itemsets but rather distinguish itemsets which are more valuable to us. Our point ought to be in distinguishing itemsets which have higher utilities in the database, regardless of whether these itemsets are frequent itemsets or not. This prompts to another approach in information mining which depends on the idea of utility called as utility mining. High utility itemset mining alludes to the disclosure of high utility itemsets. The principle target of high utility itemset mining is to distinguish the itemsets that have utility values above given utility edge [14]. The term utility alludes to its related benefit or some other related measure [16]. By and by the utility estimation of an itemset can be benefit, amount, weight, ubiquity, page-rank, and measure of some stylish viewpoint, for example, magnificence or plan or some different measures of client's inclination [17].

III. LITERATURE REVIEW

In this area we exhibit a brief audit of the distinctive calculations, methods, ideas and methodologies that have been characterized in different research diaries and distributions. Agrawal, R., Imielinski, T., Swami, A. [1] proposed Frequent itemset mining calculation that uses the Apriori standard. Standard technique depends on Support-Confidence Model. Bolster measure is utilized. An anti-monotone property is utilized to diminish the inquiry space. It produces frequent itemsets and discovers affiliation governs between things in the database. It doesn't recognize the utility of an itemset [1]. Yao, H., Hamilton, H.J., Buzz, C.J. [2] proposed a system for high utility itemset mining. They sum up past work on itemset share measure [2]. This distinguishes two sorts of utilities for things, transaction utility and outside utility. They distinguished and dissected the issue of utility mining. Alongside the utility bound property and the bolster bound property. They characterized the numerical model of utility mining in view of these properties. The utility bound property of any itemset gives an upper bound on the utility estimation of any itemset. This utility bound property can be utilized as a heuristic measure for pruning itemsets as early stages that are not anticipated that would qualify as high utility itemsets [2]. Yao, H., Hamilton, H.J., Buzz, C.J. [3] proposed a calculation named Umining and another heuristic based calculation UminingH to discover high utility itemsets. They apply pruning techniques in view of the scientific properties of utility limitations. Calculations are more productive than any past utility based mining calculation. Liu, Y., Liao, W.K., Choudhary A. [4] proposed a two stage calculation to mine high utility itemsets. They utilized a transaction weighted utility (TWU) measure to prune the inquiry space. The calculations in light of the hopeful era and-test approach. The proposed calculation experiences poor execution when mining thick datasets and long examples much like the Apriori [1]. It requires least database examines, substantially less memory space and less computational cost. It can without much of a stretch handle extensive databases. Erwin, A., Gopalan, R.P., N.R. Achuthan [5] proposed an effective CTU-Mine Algorithm in view of Pattern Growth approach. They present a reduced information structure called as Compressed Transaction Utility tree (CTU-tree) for utility mining, and another calculation called CTU-Mine for mining high utility itemsets. They indicate CTU-Mine works more effectively than Two Phase for thick datasets and long example datasets. In the event that the limits are high, then Two Phase runs moderately quick contrasted with CTU-Mine, however when the utility limit gets to be lower, CTUMine beats Two-phase. Erwin, A., Gopalan, R.P., N.R. Achuthan [7] proposed a proficient calculation called CTU-PRO for utility mining utilizing the example development approach. They proposed another minimized information representation named Compressed Utility Pattern tree (CUP-tree) which develops the CFP-tree of [11] for utility mining. TWU measure is utilized for pruning the inquiry space yet it keeps away from a rescan of the database. They demonstrate CTU-PRO works more effectively than Two-phase and CTU-Mine on thick information sets. Proposed calculation is likewise more proficient on scanty datasets at low bolster thresholds. TWU measure is an overestimation of potential high utility itemsets, in this way requiring more memory space and more calculation when contrasted with the example development calculations. Erwin, R.P. Gopalan, and N.R. Achuthan [14] proposed a calculation called CTU-PROL for mining high utility itemsets from vast datasets. They utilized the example development approach [6]. The calculation first finds the extensive TWU things in the transaction database and if the dataset is little, it makes information structure called Compressed Utility Pattern Tree (CUP-Tree) for mining high utility itemsets. On the off chance that the information sets are too huge to be in any way held in fundamental memory, the calculation makes subdivisions utilizing parallel projections that can be along these lines mined autonomously. For every subdivision, a CUP-Tree is utilized to mine the total arrangement of high utility itemsets. The counter monotone property of TWU is utilized for pruning the hunt space of subdivisions in CTU-PROL, yet not at all like Two-phase of Liu et al. [4], CTU-PROL calculation keeps away from a rescan of the database to decide the real utility of high TWU itemsets. The execution of calculation is looked at against the Two-phase calculation in [4] furthermore with CTU-Mine in [5]. The outcomes demonstrate that CTU-PROL beats previous calculations on both scanty and thick datasets at most bolster levels for long and short examples.

In the second database examine, the calculation discovers all the two component transaction-weighted usage itemsets and it brings about three component transactions weighted use itemsets. The disadvantage of this calculation is that it experiences level astute hopeful era and test philosophy [18].

J Hu et al built up a calculation for frequent thing set mining that distinguish high utility thing mixes. The objective of this calculation is to discover sections of information, characterized through blends of a few things (rules), which fulfill certain conditions as a gathering and boost a predefined target work. The high utility example mining issue considered is not the same as previous methodologies, as it behaviors control disclosure concerning singular traits and additionally regarding the general standard for the mined set, endeavoring to discover gatherings of such examples that together adds to the most to a predefined target work [19].

Y-C. Li, J-S. Yeh and C-C. Chang proposed a disengaged thing disposing of technique (IIDS). In this paper, they found high utility itemsets furthermore diminished the quantity of hopefuls in each database examine. They recovered productive high utility itemsets utilizing the mining calculation called FUM and DCG+. In this system they demonstrated a superior execution than all the past high utility example mining procedure. Be that as it may, their calculations still endure with the issue of level astute era and test issue of Apriori and it require various database filters [20].

Liu Jian-ping, Wang Ying, Yang Fan-ding et al proposed a calculation called tree based incremental affiliation manage mining calculation (Pre-Fp). It depends on a FUFP (quick redesign frequent example) mining technique. The significant objective of FUFP is the re-utilization of beforehand mined frequent things while moving onto incremental mining. The benefit of FUFP is that it decreases the quantity of hopeful set in the overhauling strategy. In FUFP, all connections are bidirectional while in FP-tree, connections are just unidirectional. The benefit of bidirectional is that it is anything but difficult to include, evacuate the youngster hub without much recreation. The FUFP structure is utilized as a contribution to the pre-extensive tree which gives positive check contrast at whatever point little information is added to unique database. It manages few changes in database if there should arise an occurrence of embedding new transaction. In this paper the calculation arranges the things into three classifications: frequent, infrequent and pre-expansive. Pre-vast itemsets has two backings limit esteem i.e. upper and lower edge. The downside of this approach is that it is tedious [21].

Ahmed CF, Tanbeer SK, Jeong BS et al created HUC-Prune. In the current high utility example mining it produce a level astute applicant era and test philosophy to keep up the hopeful example and they require a few database examines which is specifically reliant on the competitor length. To conquer this, they proposed a novel tree based applicant pruning strategy called HUC-tree, (high utility competitor tree) which catches the critical utility data of transaction database. HUC-Prune is totally free of high utility applicant example and it requires three database sweeps to compute the outcome for utility example. The downside of this approach is that it is extremely hard to keep up the calculation for bigger database check locales [22].

Shih-Sheng Chen et al (2011) proposed a strategy for frequent intermittent example utilizing different least backings. This is a proficient way to deal with find frequent example since it depends on numerous base limit bolster in light of ongoing occasion. Every one of the things in transaction is masterminded by least thing support (MIS), and it doesn't hold download conclusion property, rather it utilizes sorted conclusion property in light of climbing request. At that point PFP (intermittent frequent example) calculation is connected which is same as that of FP-development where restrictive example base is utilized to find frequent examples. This calculation is more proficient as far as memory space, subsequently diminishing the quantity of database outputs [23].

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Young-Koo Lee, and Ho-Jin Choi et al proposed a Single-pass incremental and intuitive digging for finding weighted frequent examples. The current weighted frequent example (WFP) digging can't be connected for incremental and intelligent WFP digging furthermore for stream information mining since they depend on a static database and its require various database examines. To defeat this, they proposed two novel tree structures IWFPTWA (Incremental WFP tree in light of weight rising request) and IWFPTFD (Incremental WFP tree in view of plunging request) and two new calculations IWFPWA and IWFPFD for incremental and intuitive mining utilizing a solitary database filter. IWFPFD guarantees that any non-applicant thing can't show up before competitor things in any branch of IWFPTFD and in this manner accelerates the prefix tree. The downside of this approach is that extensive memory space, tedious and it is exceptionally hard to bolster the calculation for bigger databases [24] [25].

IV. PROPOSED SYSTEM

The Proposed techniques can diminish the overestimated utilities of PHUIs as well as enormously decrease the quantity of hopefuls. Diverse sorts of both genuine and engineered information sets are utilized as a part of a progression of examinations to the execution of the proposed calculation with cutting edge utility mining calculations. Exploratory results demonstrate that UP-Growth and UP-Growth+ beat different calculations considerably in term of execution time, particularly when databases contain bunches of long transactions or low least utility limits are set.

Advantages:

- Two calculations, named Utility example growth(UP Growth)and UP-Growth+, and a reduced tree structure, called utility example tree(UP-Tree),for finding high utility thing sets and keeping up critical data identified with utility examples inside databases are proposed.
- High-Utility thing sets can be created from UP-Tree proficiently with just two sweeps of unique databases. A few systems are proposed for encouraging the mining procedure of UP-Growth+ by keeping up just fundamental data in UP-Tree.

- By these Strategies, overestimated utilities of applicants can be all around decreased by disposing of utilities of the things that can't be high utility or are not included in hunt space.

V. CONCLUSIONS

In this paper, a distributed and dynamic method is proposed to produce finish set of high utility itemsets from vast databases. Mining high utility itemsets from databases alludes to finding the itemsets with high benefit. In dispersed, it arranges the unpromising things in light of the base utility itemsets from transactions database. This approach makes appropriated environment with one ace hub and two slave hubs examines the database once and numbers the event of everything. The huge database is disseminated to all slave hubs. The worldwide table has the last resultant. Incremental Mining Algorithm is utilized where consistent overhauling continues showing up in a database. At last incremental database is revised and the high utility itemsets is found. Subsequently, it gives speedier execution, that is diminished time and cost.

REFERENCES

[1] [1] Agrawal, R., Imielinski, T., Swami, A., "Mining Association Rules between Sets of Items in Large Database", In: ACM SIGMOD International Conference on Management of Data (1993).

[2] [2] Yao, H., Hamilton, H.J., Buzz, C. J., "A Foundational Approach to Mining Itemset Utilities from Databases", In: 4th SIAM International Conference on Data Mining, Florida USA (2004).

[3] [3] Yao, H., Hamilton, H.J., "Mining itemset utilities from transaction databases", Data & Knowledge Engineering 59(3), 603–626 (2006).

[4] [4] Liu, Y., Liao, W.K., Choudhary, A., "A Fast High Utility Itemsets Mining Algorithm", In: 1st Workshop on Utility-Based Data Mining, Chicago Illinois (2005).

[5] [5] Erwin, A., Gopalan, R.P., N.R. Achuthan, "CTUMine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach", In: IEEE CIT 2007.Aizu Wakamatsu, Japan (2007).

[6] [6] Han, J., Wang, J., Yin, Y., "Mining frequent patterns without candidate generation", In: ACM SIGMOD International Conference on Management of Data (2000).

[7] [7] Erwin, A., Gopalan, R.P., Achuthan, N.R, "A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets", In: International Workshop on Integrating AI and Data Mining. Gold Coast, Australia (2007).

[8] [8] CUCIS. Center for Ultra-scale Computing and Information Security, North-western University.

[9] [9] Yao, H., Hamilton, H.J., Geng, L., "A Unified Framework for Utility Based Measures for Mining Itemsets", In: ACM SIGKDD 2nd Workshop on Utility-Based Data Mining (2006).

[10] [10] Pei, J., "Pattern Growth Methods for Frequent Pattern Mining", Simon Fraser University (2002).S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

[11] [11] Sucahyo, Y.G., Gopalan, R.P., CT-PRO: "A Bottom- Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tree Data Structure", In: IEEE ICDM Workshop on Frequent Itemset Mining Implementation (FIMI), Brighton UK (2004).

[12] [12] G. Salton, Automatic Text Processing, Addison- Wesley Publishing, 1989.

[13] [13] J. Pei, J. Han, L.V.S. Lakshmanan, "Pushing convertible constraints in frequent itemset mining", Data Mining and Knowledge Discovery 8 (3) (2004) 227–252.

[14] [14] A. Erwin, R.P. Gopalan, and N.R. Achuthan, "Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 554–561, 2008. © Springer- Verlag Berlin Heidelberg 2008.

[15] [15] Bin Chen, Peter Hass, Peter Scheuermann, "A New Two-Phase Sampling Based Algorithm for Discovering Association Rules", SIGKDD '02 Edmonton, Albetra, Canada © 2002 ACM 1 58113 567 X/02/2007.

[16] [16] Ming-Yen lin, Tzer-Fu Tu, Sue-Chen Hsueh, "High utility pattern mining using the maximal itemset property and, lexicographic tree structures", Information Science 215(2012) 1-14.

[17] [17] Sudip Bhattacharya, Deepty Dubey, "High utility itemset mining, International Journal of Emerging Technology and advanced Engineering", ISSN 2250-2459, Volume 2, issue 8, August 2012.

[18] [18] Y.Liu, W.K. Liao and A. Choudhary, ―A two phase algorithm for fast discovery of high utility itemset‖, Cheng, D. and Liu. H. PAKDD, LNCS. PP: 689-695, 2005.

[19] [19] J.Hu, A. Mojsilovic, ―High utility pattern mining: A method for discovery of high utility itemssets‖, in: pattern recognition. PP: 3317-3324, 2007.

[20] [20] Y.-C. Li,j,-s. Yeh, and C.-C. Chang, ―Isolated Items Discarding Strategy for Discovering High Utility Itemsets,‖ Data and Knowledge engg., pp: 198-217, 2008.

[21] [21] Liu Jian-Ping, Wang Ying Fan-Ding, ‖Incremental Mining algorithm Pre-FP in Association Rule Based on FP-tree‖, Networking and Distributed Computing, International Conference, pp: 199-203, 2010.

[22] [22] Ahmed CF,Tanbeer SK,Jeong B-S, Lee Y-K (2011) ―HUC-Prune: An Efficient Candidate Pruning Technique to mine high utility patterns‖ Appl Intell PP: 181–198, 2011.

[23] [23] Shih-Sheng Chen, Tony Cheng-Kui Huang, Zhe-Min Lin, ―New and efficient knowledge discovery of partial periodic patterns with multiple minimum supports‖, The Journal of Systems and Software 84, pp. 1638–1651, 2011, ELSEVIER.

[24] [24] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Young-Koo Lee a,Ho-Jin Choi(2012) ―Single-pass incremental and interactive mining for weighted frequent patterns‖, Expert Systems with Applications 39 pp.7976–7994, ELSEVIER 2012.

[25] [25] Vincent S Tseng, Bai-En Shie, Cheng-Wu, Philip S,‖ Efficient algorithms for mining high utility itemsets from transactional databases‖, IEEE Transactions on knowledge and data engineering, 2013.