

High Utility Item sets From Transactional Databases with efficient Algorithms utility pattern growth (UP-Growth) and UP-Growth+

K RAJA RAJESWARI¹, DDD SURIBABU²

¹PG Scholar In Department of CSE DNR College of Engineering & Technology ,Bhimavaram ,A.P

²HEAD & Assoc.Prof In Department of CSE DNR College of Engineering & Technology ,Bhimavaram ,A.P

Abstract: High utility item sets (HUIs) mining is an emerging topic in data mining, which refers to discovering all itemsets having utility meeting a user specified minimum utility threshold min_util . However, setting min_util appropriately is a difficult problem for users. Generally speaking, finding an appropriate minimum utility threshold by trial and error is a tedious process for users. If min_util is set too low, too many HUIs will be generated, which may cause the mining process to be very inefficient. On the other hand, if min_util is set too high, it is likely that no HUIs will be found. In this paper, we address the above issues by proposing a new framework for top-k high utility item set mining, where k is the desired number of HUIs to be mined. Two types of efficient algorithms named TKU (mining Top-K Utility itemsets) and TKO (mining Top-K utility itemsets in One phase) are proposed for mining such itemsets without the need to set min_util . We provide a structural comparison of the two algorithms with discussions on their advantages and limitations. Empirical evaluations on both real and synthetic datasets show that the performance of the proposed algorithms is close to that of the optimal case of state-of-the-art utility mining algorithms.

Keywords: Backup Privacy, Central Repository, Seed Block, Parity Cloud, Parity cloud service.

I. INTRODUCTION

Frequent item set mining is a fundamental research topic in data mining. The traditional information retrieval process is a valuable item set that requires the utility to associate the transactions. In utility mining, each item is associated with utility (e.g. unit profit) and an occurrence count in each transaction (e.g. quantity). The utility of an itemset represents its importance, which can be measured in terms of weight, value, quantity or other information depending on the user specification.

However, efficiently mining HUIs in databases is not an easy task because the downward closure property used in FIM does not hold for the utility of itemsets. In other words, pruning search space for HUI mining is difficult because a superset of a low utility itemset can be high utility. To tackle this problem, the concept of transaction-weighted utilization (TWU) model was introduced to facilitate the performance of the mining task. In this model, an item set is called high transaction-weighted utilization item-set (HTWUI) if its TWU is no less than min_util , where the TWU of an itemset represents an upper bound on its utility. Therefore, a HUI must be a HTWUI and all the HUIs must be included in the complete set of HTWUIs. A classical TWU model-

based algorithm consists of two phases. In the first phase, called phase I, the complete set of HTWUIs are found. In the second phase, called phase II, all HUIs are obtained by calculating the exact utilities of HTWUIs with one database scan. Although many studies have been devoted to HUI mining, it is difficult for users to choose an appropriate minimum utility threshold in practice. Depending on the threshold, the output size can be very small or very large. Besides, the choice of the threshold greatly influences the performance of the algorithms. If the threshold is set too low, too many HUIs will be presented to the users and it is difficult for the users to comprehend the results. A large number of HUIs also causes the mining algorithm to become inefficient or even run out of memory, because the more HUIs the algorithm generates, the more resources they consume. On the contrary, if the threshold is set too high, no HUI will be found. To find an appropriate value for the min_util threshold, users need to try different thresholds by guessing and re-executing the algorithms over and over until being satisfied with the results. This process is both inconvenient and time-consuming.

To precisely control the output size and discover the itemsets with the highest utilities without setting the threshold, a promising solution is to redefine the task of mining HUIs as mining top-k high utility itemsets (top-k HUIs). The idea is to let the users specify k, i.e., the number of desired itemsets, instead of specifying the minimum utility threshold. Setting k is more intuitive than setting the threshold because k represents the number of itemsets that the users want to find, whereas choosing the threshold depends primarily on database characteristics, which are often unknown to users.

Using a parameter k instead of the min_util threshold is very desirable for many applications. For example, to analyze customer purchase behavior, top-k HUI mining serves as a promising solution for users who desire to know "What are the top-k sets of products (i.e., itemsets) that contribute the highest profit to the company?" and "How to efficiently find these itemsets without setting the min_util threshold?" Although top-k HUI mining is essential to many applications, developing efficient algorithms for mining such patterns is not an easy task. It poses four major challenges as discussed below.

II. OBJECTIVE

We propose an efficient algorithm named TKU (mining Top-K Utility itemsets) for discovering top-k HUIs without specifying min_util . We first present its basic version named TKU_{Base} and then

escribetheTKUalgorithm,whichincludesseveralnovelstrategi es.

ThebaselineapproachTKU_{Base}isanextensionofUP-Growth[25],atree-basedalgorithmformingHUIs.TKU-Baseadopts theUP-TreestructureofUP-Growth to maintain the information of transactions and top-kHUIs.TKU_{Base}is executed as constructing theUP-Tree, generating potentialtop-khighutilityitemsets(PKHUIs)fromtheUP-Tree,and(3)identifyingtop-kHUIsfromthesetofPKHUIs.

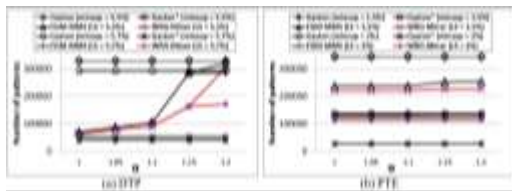
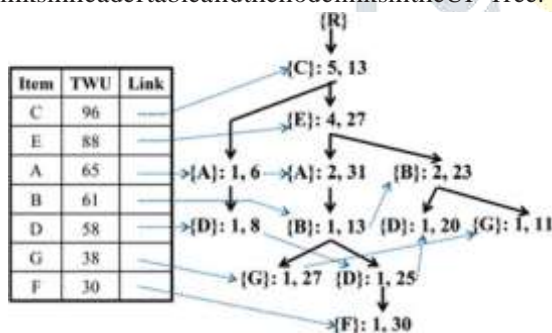


Figure: Frequent itemsets with threshold values

3.1.1 UP-Tree Structure

webrieflyintroducetheUP-Treestructure. Formore detailsaboutit, readers are referred to.EachnodeNofaUP-Treehasfiveentries: N.nameistheitemnameofN; N.countisthesu pportcountofN; N.nuisthenodeutilityofN; N.parentindicateth eparentnodeofN; N.hlinkis anodelinkwhichmaypointtoanode havingthesameitemnameasN.name. TheHeadertableis astructure employedtofacilitatethetraversaloftheUP-Tree.

Aheadertableentrycontainsanitemname,anestimatedutilit yvalue,andalink.Thelinkpointstothe firstnodeintheUP- Treehavingthesameitemnameastheentry. Thenodeswhoseite mnamesarethesamecanbetraversedeficientlybyfollowingthe linksinheadertableandthenodelinksintheUP-Tree.



III. IMPLEMENTATION

TheTKU_{Base}algorithmusesaninternalvariablenamedbor- derminimumutilitythreshold(denotedasmin_util_{Border})whichisin itiallysetto0andraiseddynamicallyafterasufficientnumberofit emsetswithhigherutilityhasbeencaptureduringthegenerat ionofPKHUIs. Thedevelopmentoftheproposedmethodisbase donthefollowingdefini-tionsandlemmas.

Lemma1.LetP%hX₁,X₂,... ,X_Mibeasetofitemsets(M_k),whereX_iisth eithitemsetinP andEU(X_i)EU(X_j) > 0,8 i< j. Inotherwords, X_iist heitemsetwiththeithhighestutilityinP. For any itemset Y, if EU(Y) < EU(X_k), Y isnot a top-k HUI.

Rationale. AccordingtoDefinition10,ifthereexistskitemsets whoseutilitiesarehigherthantheutilityofY, Yisnotatop-kHUI.

Lemma2.LetP%hX₁,X₂,... ,X_Mibeasetofitemsets(M_k),whereX_iisth eithitemsetinPandEU(X_i)EU(X_j) > 0,8 i< j. Ifd_p%EU(X_k),f_{HUI}(D,d)f_{HUI}(D,d_p).

Rationale.LetKHbethethecompletesetoftop- kHUIs.Ifj KH jk,d%min{EU(X)j X2KH}(byDefinition11).Beacu sed%min{EU(X)j X2KH}min{EU(X_i)j X₂P, li k}%EU(X_k)%d_p,ddpandf_{HUI}(D,d)f_{HUI}(D,d_p).

Example3.Considerthatk%4andabs_min_util%0.LetPbetheset ofall1- itemsets{{A}:20,{D}:20,{B}:16,{E}:15,{C}:13,G:7,F:5}inD,wh ere thenumberbesideeachitem- setisitsabsoluteutility.ByLemma1,{C},{G},{F}areunpromisi ngtothetop- 4HUIs. Thereforeabs_min_utilcanberaisedtothefourthhigh estutilityvalueinP(i.e.,15)andnotop-kHUIswillbemissed.

Afterraisingabs_min_util,TKU_{Base}appliestheUP- Growthsearchprocedurewithabs_min_util%min_util_{Border}to generatePKHUIs.ThoughLemma1providesawaytoraisemin_uti l_{Border},itcannotbeappliedduringthegenerationofPKHUIsinpha seI. Thisisbecause theutilitiesofPKHUIsareunknownduringph aseI. Asolutiontothisproblemistousealowerboundontheutilit yofPKHUIsduringphaseltoraisemin_util_{Border}. Alowerboundo ntheutilityofPKHUIsisprovidedbythefollowingdefinitions.

Definition12(Minimumutilityofanitem).Theminimumutilit yo fanitemI2Iisdenotedasmiu(I)anddefinedasthevalueEU(I,T_j)for wh ich: 9T_j2Dsuchthat0< EU(I,T_j)< EU(I,T_j).Anequivalentdefin itionisthatmiu(I)%min{EU(I,T_j)j T_j2Dandr2g(I)}.

Lemma3.LetC%hX₁,X₂,... ,X_Mibeasetofitemsets(M_k),whereX_iisth eithitemsetinCandMIU(X_i)MIU(X_j) > 0,8 i< j. Inotherwords, X_iistheitemsetwiththeithhighestMIUvalueinC. ForanyitemsetY, ifT WU(Y)< d_{MC}%min{MIU(X_i)j X₂C, lik}, thenYisnotatop- kHUI.

Rationale. AccordingtoDefinition8,EU(Y)TWU(Y). IfTWU(Y) < d_{MC}, wehaveEU(Y)<d_{MC}. Besides, 0< EU(Y)< MIU(X_i)EU(X_i),8 X_i2C, lik. AccordingtoDefinition10,ifthereexistkitemsetswhoseutili tiesarehigherthantheutil-ityofY, Yisnotatop-kHUI.

Lemma4.LetC%hX₁,X₂,... ,X_Mibeasetofitemsets(M_k),whereX_iisth eithitemsetinCandMIU(X_i)MIU(X_j) > 0,8 i< j. Ifd_{MC}%MIU(X_k), f_{HUI}(D,d)f_{HUI}(D,d_{MC}). Rationale.LetKHbethethecompletesetoftopk HUIs.Ifj KH jk,d%min{EU(X)j X2KH}(byDefinition11).Beacu sed%min{EU(X)j X2KH}min{EU(X_i)j X₂C, lik}min communication ability to backup locations in order to increase data integration.

Resource storage allocation:

Heterogeneous clouds consist many different hardware and software such as hybrid storage and diverse disks. In cloud-based enterprises, entire business data are stored in the cloud storage. So, data protection, safety and recovery are critical in these environments. Using fastest disk technology in the event of a disaster for replication of data in storage location.

min_util_{Border}couldberaisedbeforetheconstructionoftheUP Treeandprunemoreunpromisingitemsintransactions,thenum berofnodesmaintainedinmemorycouldbereducedandthemin inalgorithmcouldachievebetterperformance. Basedonthiside a, weproposeastrategy namedPE(PreevaluationStep)toraisemin _util_{Border}duringthefirstscanofthedatabase.

Strategy2(PE:PreEvaluation). ThestrategyPEusesastructure namedPreEvaluationMatrix(PEM)to store lower bounds of theut

ilities of certain 2-item sets. Each entry in PEM is denoted as $PEM[x][y]$ and corresponds to the lower bound of $EU(\{x, y\})$, where $x, y \in I$. Initially, each value in PEM is set to 0. When a transaction $T_r \in \{I_1, I_2, \dots, I_M\}$ ($I_j \in I, 1 \leq j \leq M$) is retrieved during the first database scan, the utility of $\{I_i\} \setminus \{I_j\}$ ($1 \leq i \leq M$) in T_r is added to the value of the corresponding entry of $PEM[I_i][I_j]$ in PEM. After scanning all the transactions, if the k th highest value in PEM is higher than min_util_{Border} , min_util_{Border} can be raised to the k th highest value in PEM. The space complexity of the strategy is $O(|I|^2/2)$, where $|I|$ is the number of distinct items in the database.

Example 5. Consider the database of Table 1. When $T_1 = \{(A, 1), (C, 1), (D, 1)\}$ is retrieved, the corresponding entries $PEM[A][C]$, $PEM[A][D]$ are accumulated with $EU(\{AC\}, T_1) = 6$ and $EU(\{AD\}, T_1) = 7$. The remaining transactions in the database are processed by the same procedure. After that, if min_util_{Border} is lower than the k th highest value in PEM, min_util_{Border} is set to the k th highest value in PEM. 3 shows the value of each entry in PEM after scanning the database. If $k = 4$, the fourth highest value in PEM is $PEM[B][E] = 18$. If min_util_{Border} is less than this value, min_util_{Border} is raised to 18.

Notice that in TKU_{Base} , the strategy DGU proposed in [25] cannot be applied, because min_util_{Border} is set to 0 before the construction of the UP-Tree. However, if we apply the strategy PE to raise min_util_{Border} during the first database scan, DGU can be further applied to prune those items whose TWUs are less than min_util_{Border} , which reduces the size of the UP-Tree and the number of candidates produced in phase 1.

We also propose a strategy called NU (raising the threshold by Node Utilities), which is applied during the construction of the UP-Tree. The strategy NU is developed based on the following lemmas. **Lemma 8.** Let $PATH = \{N_1, N_2, \dots, N_M, R\}$ be a path from a node N_1 to the root R in UP-Tree and I_2 be the item name of $N_i, 1 \leq i \leq M$. $PATH = \{N_1, N_2, \dots, N_M, R\}$ represents a unique itemset $X = \{I_1, I_2, \dots, I_M\}$ in the database. Besides, the node utility of N_1 is a lower bound on the utility of X . **Rationale.** The UP-Tree is constructed by applying the strategy DGN [25]. According to the rationale described in [25], the utility of the itemset $X = \{I_1, I_2, \dots, I_M\}$ is guaranteed to be higher than the node utility of N_1 . Therefore, $N_1 \cdot nuEU(\{I_1, I_2, \dots, I_M\})$.

Lemma 9. If there are M nodes in the UP-Tree, there are at least M distinct itemsets whose utilities are higher than 0.

Rationale. By Lemma 8, each path from a node in the UP-Tree to the root forms a unique path, which represents a unique itemset whose utility is higher than zero in the database. Therefore, M distinct nodes in the UP-Tree yield M distinct itemsets whose utilities are higher than zero.

Lemma 10. Let $SetNode = \{N_1, N_2, \dots, N_M\}$ be an ordered set containing all nodes in the UP-Tree (M nodes). Let N_i be the i th node in $SetNode$ and $N_i \cdot nuN_j \cdot nu > 0, 8i < j$. If $d_{NU} \geq N_k \cdot nu$, then $f_{HUI}(D, d) \geq f_{HUI}(D, d_{NU})$.

Rationale. By Lemma 8, each path from a node $N_i \in SetNode$ to the root R represents a unique itemset $N_i, 1 \leq i \leq M$. Let $SetItemset = \{X_1, X_2, \dots, X_M\}$ be an ordered set of itemsets that are represented by the nodes in $SetNode$, where $EU(X_i) > 0, 8i < j$. Let KH be the complete set of top- k HUIs in the database D . If $|KH| \geq k$, then $d_{NU} \geq \min\{EU(X) \mid X \in KH\}$.

Definition 11. Because $d_{NU} \geq \min\{EU(X) \mid X \in KH\}$

$\min\{EU(X) \mid X \in SetItemset, |K| \leq k\} \geq \min\{N_i \cdot nu \mid N_i \in SetNode, |K| \leq k\}$, we have $d_{NU} \geq d_{HUI}(D, d)$ and $f_{HUI}(D, d) \geq f_{HUI}(D, d_{NU})$.

By Lemma 8, 9 and 10, if there are no less than k nodes in the UP-Tree during its construction and the k th highest node utility in the UP-Tree is higher than the current min_util_{Border} , min_util_{Border} can be safely raised to the k th highest node utility in the UP-Tree.

Example 6. Let the notation N_a represent a node of the UP-Tree such that a is the item stored in N_a . If $k = 4$, when the first reorganized transaction $T_1 = \{(C, 1), (A, 1), (D, 1)\}$ is inserted into the UP-Tree, then nodes

$N_{\{C\}}, N_{\{A\}}$ and $N_{\{D\}}$ are created with node utilities 1, 6 and 8, which are respectively lower bounds on the utilities of itemsets $\{C\}, \{AC\}$ and $\{DAC\}$. When the second reorganized transaction $T_2 = \{(C, 6), (E, 2), (A, 2), (G, 5)\}$ is inserted into the UP-Tree, there are more than four nodes in the tree. By Lemma 10, min_util_{Border} can be raised to the fourth highest node utility in the current UP-Tree.

Strategy 3 (NU: raising the threshold by Node Utilities). The strategy NU is applied during the construction of the UP-Tree (during the second database scan). If there are more than k nodes in the current UP-Tree and the k th highest node utility value NU_k is higher than min_util_{Border} , min_util_{Border} can be raised to NU_k . After inserting all reorganized transactions, the size of the constructed UP-Tree can be further reduced by pruning items whose TWU values are less than min_util_{Border} in the UP-Tree.

IV. CONCLUSION

In this paper, we have studied the problem of top- k high utility itemsets mining, where k is the desired number of high utility itemsets to be mined. Two efficient algorithms TKU (mining Top- k Utility itemsets) and TKO (mining Top- k Utility itemsets in One phase) are proposed for mining such itemsets without setting minimum utility thresholds. TKU is the first two-phase algorithm for mining top- k high utility itemsets, which incorporates five strategies PE, NU, MD, MC and SET to effectively raise the border minimum utility thresholds and further prune the search space. On the other hand, TKO is the first one-phase algorithm developed for top- k HUI mining, which integrates the novel strategies RUC, RUCZ and EPB to greatly improve its performance. Empirical evaluations on different types of real and synthetic datasets show that the proposed algorithms have good scalability on large datasets and the performance of the proposed algorithms is close to the optimal case of the state-of-the-art two-phase and one-phase utility mining algorithms.

Although we have proposed a new framework for top- k HUI mining, it has not yet been incorporated with other utility mining tasks to discover different types of top- k high utility patterns such as top- k high utility episodes, top- k closed high utility itemsets, top- k high utility web access patterns and top- k mobile high utility sequential patterns. These leave wider rooms for exploration as future work.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
 [2] C. Ahmed, S. Tanbeer, B. Jeong, and Y. Lee, "Efficient tree structures for high-

- utilitypatternmininginincrementaldatabases,"IEEETrans.Knowl.DataEng.,vol.21,no.12,pp.1708–1721,Dec.2009.
- [3]K.Chuang,J.Huang,andM.Chen,"Miningtop-kfrequentpatternsinthepresenceofthememoryconstraint,"VLDBJ.,vol.17,pp.1321–1344,2008.
- [4]R.Chan,Q.Yang,andY.Shen,"Mininghigh-utilityitemsets,"inProc.IEEEInt.Conf.DataMining,2003,pp.19–26.
- [5]P.Fournier-VigerandV.S.Tseng,"Miningtop-ksequentialrules,"inProc.Int.Conf.Adv.DataMiningAppl.,2011,pp.180–194.
- [6]P.Fournier-Viger,C.Wu,andV.S.Tseng,"Miningtop-kassociationrules,"inProc.Int.Conf.Can.Conf.Adv.Artif.Intell.,2012,pp.61–73.
- [7]P.Fournier-Viger,C.Wu,andV.S.Tseng,"Novelconciserepresentationsofhighutilityitemsetsusinggeneratorpatterns,"inProc.Int.Conf.Adv.DataMiningAppl.LectureNotesComput.Sci.,2014,vol.8933,pp.30–43.
- [8]J.Han,J.Pei,andY.Yin,"Miningfrequentpatternswithoutcandidategeneration,"inProc.ACMSIGMODInt.Conf.Manag.Data,2000,pp.1–12.
- [9]J.Han,J.Wang,Y.Lu,andP.Tzvetkov,"Miningtop-kfrequentclosedpatternswithoutminimumsupport,"inProc.IEEEInt.Conf.DataMining,2002,pp.211–218.
- [10]S.Krishnamoorthy,"Pruningstrategiesformininghighutilityitemsets,"ExpertSyst.Appl.,vol.42,no.5,pp.2371–2381,2015.
- [11]C.Lin,T.Hong,G.Lan,J.Wong,andW.Lin,"Efficientupdatingofdiscoveredhigh-utilityitemsetsfortransactiondeletionindynamicdatabases,"Adv.Eng.Informat.,vol.29,no.1,pp.16–27,2015.
- [12]G.Lan,T.Hong,V.S.Tseng,andS.Wang,"Applyingthemaximumutilitymeasureinhighutilitysequentialpatternmining,"ExpertSyst.Appl.,vol.41,no.11,pp.5071–5081,2014.
- [13]Y.Liu,W.Liao,andA.Choudhary,"Afasthighutilityitemsetsminingalgorithm,"inProc.Utility-BasedDataMiningWorkshop,2005,pp.90–99.
- [14]M.LiuandJ.Qu,"Mininghighutilityitemsetswithoutcandidategeneration,"inProc.ACMInt.Conf.Inf.Knowl.Manag.,2012,pp.55–64.
- [15]J.Liu,K.Wang,andB.Fung,"Directdiscoveryofhighutilityitemsetswithoutcandidategeneration,"inProc.IEEEInt.Conf.DataMining,2012,pp.984–989.
- [16]Y.Lin,C.Wu,andV.S.Tseng,"Mininghighutilityitemsetsinbigdata,"inProc.Int.Conf.Pacific-AsiaConf.Knowl.DiscoveryDataMining,2015,pp.649–661.
- [17]Y.Li,J.Yeh,andC.Chang,"Isolateditemsdiscardingstrategyfordiscoveringhigh-utilityitemsets,"DataKnowl.Eng.,vol.64,no.1,pp.198–217,2008.
- [18]J.Pisharath,Y.Liu,B.Ozisikyilmaz,R.Narayanan,W.K.Liao,A.Choudhary,andG.Memik,NU-MineBenchversion2.0datasetandtechnicalreport[Online].Available:<http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html>,2005.
- [19]G.PyuanandU.Yun,"Miningtop-kfrequentpatternswithcombinationreducingtechniques,"Appl.Intell.,vol.41,no.1,pp.76–98,2014.
- [20]T.Quang,S.Oyanagi,andK.Yamazaki,"ExMiner:Anefficientalgorithmforminingtop-kfrequentpatterns,"inProc.Int.Conf.Adv.DataMiningAppl.,2006,pp.436–447.
- [21]H.RyangandU.Yun,"Top-khighutilitypatternminingwitheffectivethresholddraisingStrategies,"Knowl.-BasedSyst.,vol.76,pp.109–126,2015.
- [22]H.Ryang,U.Yun,andK.Ryu,"Discoveringhighutilityitemsetswithmultipleminimumsupports,"Intell.DataAnal.,vol.18,no.6,pp.1027–1047,2014.