Mining of High Value Item Sets with Negative Utility

Aatif Jamshed, Bhawna Mallick

Abstract: ***The patterns generated by frequent pattern mining aims to find the frequent items without considering the utilities of the different items. The traditional association rule mining treats all items to be of equal utility. This is not always the case for a real world application. Utility based data mining is a new area of research and is complementing the frequency based approach. The main objective of Utility Mining is to identify the item sets with highest utilities, by considering profit, quantity, cost or other user preferences as the Utility of the item. Recent approaches developed so far considers the utilities of items to be same over a particular period of time. In our approach we have proposed that the utility of items vary over a period of time. Our work also proposed that the utility of items may also assume negative values. Our work thus treats the data mining in more realistic manner***

Index Terms: Mining, High value Item set, Utility mining, Negative value.

I. INTRODUCTION

Structuredknowledge canbe extractedfrom largedatasetswiththehelpof datamining. Dataminingtasks likeClassificationandClustering also usefrequent pattern mining anditalsoforms thepremiseforassociation rulemining. Clustering also usefrequent pattern mining anditalsoforms thepremiseforassociation rulemining.Relationshipsoftheform X□Y (amongstitem setsthatoccurtogetherina databasewhere X andY aredisjointitem sets) canbefoundusingassociationrulemining. Butfrequentpattern mining algorithmshasa drawback that itworksfor a limited set of applicationswheretheitemsaretreatedtobe of equalutilitytothebusiness.Itdidnot consider quantityand weight that are significantforaddressingreal worlddecision problemsthatrequiremaximizingtheutilityin anorganization.Herecomestheroleof utility based data mining that aims to find the item setsthatareofhighutilitytothebusiness. It treats every item to be of different significanceandutilityof differentitemscan expressasfunctionsofcost,profitetc.Fore.g. sellingamicrowavemayoccurlessfrequently thansaleofbreadandbutterinasuperstore, buttheformergives amuchhigherprofitper unitsold.Thehigh utility item settmining problemistofindallitem setsthathaveutility larger than a user specified value of *min_utility*.Hencefrequencyofitemsisnot the sufficient indicatorof item's interestingness.Othermeasureslike utility shouldalsobetakenintoconsideration.

Mostof thealgorithmsdevelopedsofarfor utility miningeitherfolloweda prioritybased approach which relied on thetechniqueof candidate generate andtestmethod.These approaches sufferedfrom high time costsmoreoverthey alwaysrun underthejeopardy of combinatorial explosion.Moreoverthe algorithms d i d not take into account into this kind of activity.dynamicnatureofutilityandpossibilityofthe negativevalueof utility.Asmoreconstraints onthe utility factors are applied the computational complexityforthealgorithms alsoincreasesthereforeinanutshellwecan say that not only simplification in the calculationof utilityisthechallengebutalso candidate size restrictionis alsoanissue that mostofthealgorithms triedtoaddress.Before HUINIV[5]nootherworkwastaking intoconsideration that utility can also take negativevalues, whentwophasealgorithm[3] wasappliedtothescenarioof negativevalues wesawthatmany ofthehighutilityitem sets whichshouldhavebeen therein theresults weremissingbutsinceHUINIVwas theonly workdonein theareaofnegativevalues therefore therewasaninherentneedto improvise upon the proposed algorithm. Henceforth a challenge to improve the existingHUINIVwas confrontingus.Later sectionofour work describes in detailthat howitmetthesechallenges.

II. RELATEDWORK

High utilityitem settminingisnothingbutan augmentation of frequent pattern mining. Perhaps it can be viewed as a more generalized approach of the later. Business/Fraudulent transaction, retail communities, security,bio-informatics, are someof thereallifeapplicationswhereutility miningcan be employed.SometimesRare item setsaretheresultof utilitymining.These item settprovidesavery usefulinsightto decisionmakers.Forinstance,a supermarket's customers purchasemicrowaveorrefrigerator rarely ascompared tobread,bakingpowder, sauce.Butprofitforthesupermarketincase oflatertransactionsisless thantheformer. Researches thatdefinevariousinterestingness measuresof discoveredpatternisproposedby LiqiangGeng and H.J.Hamilton in[1]. Their work intended to come up with some parametersthatallowreducedtimeandspace costsminingprocess.In thisworkutilityis definedasoneof theinterestingnessmeasures ofdiscoveredpatterns. Afoundationalapproach tomineitem sett utilitiesfrom

Revised Manuscript Received on July 10, 2019.

Aatif Jamshed, Information Technology, ABES Engineering College,Ghaziabad, India

Bhawna Mallick, Computer Science, Maverick Quality Advisory Services Pvt.Ltd Ghaziabad, India

databaseshavebeenproposedin [2]by HONGYAOetal.Thispaperdelves into thetheoreticalanalysisoftheresulting utility mining problem and it laid the foundationforvariousfuture utilitymining algorithms.This workintroduces autility bound and support bound property and thereafteritusheredamodel christened as MEU (miningusingexpectedutility)that minedthehighutilityitem sets. Thenatwophasealgorithm[5]byYingLiuet alwasproposed thatclaimedtoefficiently cut downthequantum ofcandidatesandhencea completesetofhigh utilityitem setscomeout moreadroitly. Theproposed twophase algorithm solvestheissueofmaintaining downwardclosureproperty.Besides several othermeasureslike TransactionWeighted utilization (TWU)wereintroducedwhich formsthe premise and works like the benchmarkforseveralother algorithms. Then an algorithmHUINIVwasproposedby Chun-JungChuetal[6]whichmines thehigh utilityitem setsfrom databaseandalsotakes intoaccountheitem withnegativeutility values.Theirworkwas authenticintheway thatitmatchesvery nearly tothereal life shopping basket analytics situation where someitemsmay alsoappearwith negative utilityvalues. Temporal association mining or more precisely calendarbasedminingbyKeshri Vermaetal[4]wassuggestedwhichcanmine the itemsetwhich are frequentin a user definedtimeperiod T.Thesemethods address the problem to excavate items that are frequentforonlysometimeof thedatabase periodically.

III. METHODOLOGY

Thischapterwillbedelvingintominutiaof theapproach thatwecomply toreachtoour motiveofmininghigh utilityitem sett.Firstof all thisportion willexpoundvariousmeasures that areomnipresent throughout the text. It will definevarioustermsandexplainitthat howaretheycalculatedtakinganexample.

Table3.1

TID	A	B	C	D	E	Date	TU
T1	5	3	10	2	2	02/21/05	159
T2	4	1	8	1	10	09/11/06	324
T3	5	6	8	1	2	04/09/05	151
T4	6	4	9	2	1	08/22/07	174
T5	6	2	12	1	1	11/11/02	1256
T6	2	4	15	3	15	05/22/02	492
Sum	28	20	62	10	31	——	2546

The above table denotes various ransactions of the retail store. Each transaction is explicitly given auniquetransactionidas T1, T2 etc. Business storeispresumed to have fiveitemslikeA,B,C,D, andE. Time stamping oftransaction place is on the 7th columnandon8thcolumntransactionutility asdefined later is mentioned. The numbers representsthequantum ofitemsthatwere boughtinthecorrespondingtransaction.

Table3.2

Utility(ITEMS ↓ TIME→)	Quarter 1	Quarter 2	Quarter 3	Quarter 4
A	5	5	8	5
B	2	2	-2	-2
C	10	10	10	100
D	4	8	12	20
E	10	20	20	10

Table3.2describes theutilityvalues ofthe items ofthe retail storeduringthefour quartersoftheyear.

A. Definitions

Definition1.Theutilityofanitemipinthe transactionTdis denotedas u(ip,Td)and definedas p(ip)×q(ip,Td).Forexample,in Table3.1,u({A},T1,q1)=5×5=25.[5]

Definition2.Theutilityofanitem settXinTd is denotedas u(X,Td) anddefinedas ∑(X,Td) Forexample,u({AC},T3,q2)=u({A}, T3,q2) +u({C}, T3,q2)=25+80=105.[5]

Definition3.Theutilityofanitem settXinDis denotedas u(X)anddefinedas ∑(X,Td)For example,u({AD})=u({AD}, T1)+u({AD},T2) +u({AD},T3) +u({AD},T4)+u({AD},T5)) +u({AD},T6))=33+44+33+72+50+34 =266. [5]

Definition4.Thetransactionutilityof a transactionTdis denotedas TU(Td)and definedas u(Td). For example,TU(T4)= u({ABCDE},T4)=48 –8+90+24+20= 174.[5]

Definition5.Thetransaction-weighted utilizationofanitem settXis thesumofthe transactionutilities ofallthetransactions containingX,whichis denotedas TWU(X). Fore.g. TWU(AE)=TU(T1)+TU(T2)+ TU(T3)+TU(T4)+TU(T5)+TU(T6)=159 +324+151+174+1259+492=2546 [5]

Ifanyofthetransactionwouldnothave containedAandC simultaneouslythenthe transactionutilityofthattransactionwould nothavebeentakenintoaccount.

B. U-tree datastructure

For thesakeofsimplicitytounderstandthe constructionofUtreeweconsideronemore sampledatabaseandits correspondingutility value.

Table 3.3

TID	A	B	C	D	E	F	Transaction Utility ($)
t_1	2	0	1	1	0	0	80
t_2	2	1	1	0	0	0	195
t_3	0	0	1	1	10	0	110
t_4	0	1	0	0	15	0	225
t_5	1	0	1	0	0	1	37
t_6	2	0	0	1	10	0	105
t_7	2	0	0	0	8	1	62
t_8	1	1	0	1	2	0	205
t_9	1	0	0	1	10	0	95
t_{10}	1	1	0	0	5	0	185
total	12	4	4	5	60	2	1299

Table 3.4

ITEM	ITEM ID	UTILITY
A	1	**10**
B	2	**150**
C	3	**25**
D	4	**35**
E	5	**5**
F	6	**2**

IhaveconstitutedtheU-Treedatastructure in awaytocompactlyrepresentthetransaction datarequiredforhighutilitymining.It consistsofanitemheadertableandaU-tree.

C. Item Header Table

TheItemHeader tablecontainsindex,itemid,quantity,TWU, utility,PSTWhereTWUandPSTare transactionweightedutilizationandpointerto subtree respectively.

D. Construction of Item Header Table

Initiallynoitems exceptthe rowheadingsare therein thetable.Thereafterthedatabaseof transactions is scanned once. As the transactionsareread fromthedatabase, theitemheader table is updates using the followingsteps. •Foranyitemwhichisabsentin theitem headertablebutpresentin transaction,anew column in the table is generated and the entries areinitialized,likeitem-id,utility/unit, its quantity,andtheTWU.Atthis stage mapped-idandthepointerremainnull.

•Foreveryitemwhoseentryisthereinthe itemheader tableaswell asthetransaction, update in the corresponding column entries aredonei.e.totalquantityoftheitemisadded toexistingdetail alsoTUisaddedtoexisting entry.

•Itemsin theitemheadertablearestoredin increasing orderof theirTWU and their mappedidisnothing but theircorresponding ranks when placed in ascending order. The itemwhichdoesnotcross theminimumutility isnotincludedintheitemheadertable.

Theitemwiththe ITEMID6i.e.Fis eliminatedasitwas belowtheminimum utilityof120.Theoriginalitems thatwere3, 4,2and1,5arenowhaveanewmappedid i.e.indexas 1,2, 3, 4, 5.

E. *Construction of* U-Tree

Since this work is on temporal mining differenttree willbeconstructedfordifferent timespecificationsbutsince the procedure of construction will remainsame,thefollowing examplewilltalk onlyaboutconstruction of treetakingonlytheutilityvaluesof onlyone timeperiod.Now theveryfirstphaseofthe treeconstructionwhichisinitialization of tree comes into picture. The process of the

Table 3.5 item header table

INDEX	1	2	3	4	5
ITEMID	3	4	2	1	5
UTILITY	25	35	150	10	5
QUANT.	4	5	4	12	60
TWU	422	595	810	964	987
PST	Θ	Θ	Θ	Θ	Θ

initializationoftreeis asfollows

Thetreeisinitializedwithrootnodeofindex 1.ApathofN-1nodewithindexfrom2toN iscreatedsuchthatanynodeiis thesuccessor iof(i-1).Apointertotherootofthesubtree isattachedtoevery itemi presentin theitem header table. i-item sett is attached to every node and its TWU is initialized as 0. Any node i will represent pattern like i,i-1,i-2,i-3..1.Eachpatternalsohasapointerattached withthearray oftotal quantity ofiteminthe transaction.Thisistheprocedureof tree initialization.Thefull treeconstructionisthe nextstepin ourproposedmethodology. The databaseisscannedonceagaintoupdate the Utree.Scanningofeverytransactionisdone inincreasingorderofitem'sTWU.Ifapattern isabsentincurrenttreethennodeiscreated and theappropriatepattern andcorresponding TWU are recorded. Also array of total quantityisupdated.Ifthepatternispresent theonlytheupdateofTWUandtotalquantity isdoneinthepresentUtree. Oneimportant thingtobekeptinmindwhileupdatingthe treeasthatin theitemheadertabletheitems which were absent,theircontribution to the TWUisnotconsideredandexcluded.As the database is scanned once the tree is constructed.The tree correspondingto the aboveexampleis showninabovefigure.

F. *MiningAlgorithm*

Descriptionofourproposedalgorithmtomine therequiredsetof highutilitypatternisgiven in this section. The algorithmis explained using the example dataset/database as described earlier. Item header table's eachitemserveascommencement point tomine HUI.Thisis donebyfirstinspectingtheTWU of everyitemintheitemheadertableinits associatedsubtree.

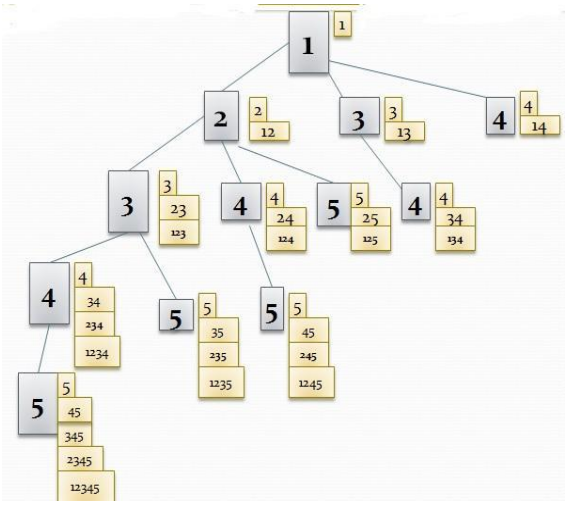

Figure3.1

Algorithm starts with subtree that has item withindex ormappedid1asitsroot,then the correspondingitem'sactual utilityischecked whetheritliesabove orbelow the threshold. Thisresultinpattern1withtheactualutility of100 which certainly lies below the threshold. Hence this pattern 1s is not consideredamongthe HUI.Now the current node's childsubtrees arecheckedfortheitems thatare leftintheitemheadertablestarting from rightmostitem anduptocurrentitem's secondrightneighbor.Atthesametime an arraytostore theintermediateresultsofTWU while traversingthe treeisinitialized.This array istermedassum_TWU.Nowwestart from therightmostitemintheitemheader tablewhichis5.Alltheoccurrenceof 5is inspected under 1's subtree and every pattern'sTWUisaddedtothearray,thisis howtheTWUof thepattern15ischecked.

Thereare4occurrencesof5under1'ssubtree but each occurrence apart from 125 is neglected because those patterns are not presentin thedatabaseitselfi.e.they havethe TWUof 0.TheTWUof125is110whichis less than minimum utility threshold. And henceitsrealutilitycalculationwillnot be done.Also5 being therightmostitem,the augmentationtopattern15isnotrequired.

occurrence of 4 under 1's subtree is considered and which results in pattern 14, 124,134, 1234. Butsince1234is absentinthe database itself it is not considered. The patterns14,134and124occurwith the correspondingTWUof 35,195and80which addsuptogiveaTWUvalueof310whichis >thresholdandthereforebeingthehigh TWU itemwe'llcontinuetofindtheactual utility contribution of 14,124,134 which is 35, 45 and45respectivelywhich addupto125and lessthanthethresholdvalueof 150.Nowthe longerpattern with 14asthepreffixare exploredbut then thereisnopattern present hencethealgorithmagainiterates tominethe pattern of13.Thereisaonlypresenceof13 with TWU of195 and since the value lies abovetheutility threshold thereforetheactual utilitycalculationofthispattern13isdone anditcomes upto175whichis alsothehigh utility item settsoitisamong theHUI.Again longerpatternslike 134and135are tobe searched andmined with 13asthepreffix and weseethatthepattern 135isabsentand134 results in the HUI withutility of195. The original id of pattern 134 is item sett 123. Again 134isfurtherextendedto minefor patternslike 1345butnopatternlike thisis presentin thedatabasethereforetheiteration stops here,andsincethe3issecondright neighborof 1thereforepath12isnotmined andisleftforlater computation.Moreover sincethereisapath 13in thetreetherefore this'pattern TWUandquantity isattachedto thecorrespondingnodeintheleftmostbranch.

Thiscompletestheminingof HUIitemswith subtreehavingrootnode1.Nowthesubtree of2 ischecked.Patternslike2and12 are derivedfrom theroot.Itshouldbenoted that from thisnodepattern12canbederived.And nowtheprogramwillcheckfortheoccurrence ofrightmostitemundernotonly2butalso12.

Thepatternslike25,125ans 24and124are checkedforhigh TWU.AndifTWU>thresholdthanactualutilitycalculationof the patternisdone. Fromthispoint24and25 comesouttobeof HUIandtheymapto originalitem sett14 and 45respectively.The extensionof 25isnotpossiblebut24is extended tocheckfor245 and1245.And algorithmchecks for bothofthese item setsandfinds245tobeaHUIwithutilityof 255. Alsoitisnoted thatboth24and245existsin the tree therefore their quantity value and TWUvalueisadded tocorrespondingnode occurrenceattheleftmostbranchof thetree. Again thealgorithmsloopsforchecking the subtreewithroot3.In thiscasetherightmost item's occurrences arecheckedunderthe occurrenceof root123and23 and3.The wholealgorithmsfollowthesameprocedure asdesribedaboveandfindtheHUIof35and 235.Similarlysubtreewithroot4willgive 234and34asHUI.And5willgive5,45, 345,2345ashigh utilityitem sets.Afterthisis donemappingof indexestooriginalitemis donewhichgivestheresultof ourproposed approachi.e.setof Highutilityitem sett, they areas follows:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 3.6

Item sets	Utility values	Item Sets	Utility Values
B	600	BE	560
E	300	DE	300
AB	490	ABC	195
AD	200	ABD	195
AE	245	ABE	355
BC	175	ADE	255
BD	185	BDE	195
RESULTS		ABDE	205

IV. PERFORMANCE STUDY

In this section the performance of our proposedalgorithmisempiricallystudied.The Algorithmisdevelopedinjava and tested on IntelPentiumcorei3processor,3GbRAM on transaction was50000.Wegenerateutility value randomly following a lognormal distributionasin[5].TheAlgorithmtookmore timeinexecution when utility threshold were higher as comparedtothecasewhenutility thresholdarelower anditoutperformedmany other previous algorithms when utility thresholds arefurtherlowered.

V. CONCLUSION.

In thispaper,anovel algorithm thatminesthe high utility item setsfrom a transaction temporaldatabase is proposed whichminesthecompletesetofhigh utilityitem setsovera userdefinedtimeperiod.Thealgorithm most closely relates tothereal world scenario wheretheitem utilitiesarenotconstantand keepon changing.Thisalgorithm addresses many problemslikemaintainingdownward closureproperty andalsoaddressingthe problem toconsidernegativevalues.This algorithm usesatreebasedapproach anddoes notfollow candidate generationandtest methodwhichefficientlyminestherequired setofitems.

REFERENCES

1. Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), 38(3), 9.
2. Yao, H., Hamilton, H. J., & Butz, C. J. (2004, April). A foundational approach to mining itemset utilities from databases. In Proceedings of the 2004 SIAM International Conference on Data Mining (pp. 482-486). Society for Industrial and Applied Mathematics.
3. Li, Y., Ning, P., Wang, X. S., & Jajodia, S. (2003). Discovering calendar-based temporal association rules. *Data & Knowledge Engineering*, *44*(2),193-218.
4. Keshari Verma,O.P.Vyas,Efficient calendar based temporal Association Rule. ACMSIGMOD,ISSNNo.0163-5808,vol34,No3September2005.
5. Liu, Y., Liao, W. K., & Choudhary, A. (2005, May). A two-phase algorithm for fast discovery of high utility itemsets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*(pp. 689-695). Springer, Berlin, Heidelberg.
6. Chu, C. J., Tseng, V. S., & Liang, T. (2009). An efficient algorithm for mining high utility itemsets with negative item values in large databases. *Applied Mathematics and Computation*, *215*(2), 767-778.Computation Volume 215, issue 2 , 15September2009
7. Pillai, J., Soni, S., Vyas, O. P., & Muyeba, M. (2010). A Conceptual Approach to Temporal Rare Item set Utility Mining. *International Journal of Computer Applications*, *1*, 65-72.

AUTHORS PROFILE

Mr. Aatif Jamshed is pursuing Ph.d from Uttarakhand Technical University (A State Govt),Dehradun. He has completed his M.Tech in Computer science & Engineering from Jamia Hamdard University. He has experience in real life projects from leading IT company in India with proficiency in Java, J2EE. Mr.Jamshed has about 10 years of Teaching and Industrial experience across multiple institutes like Galgotias, Vidya, Beganto Inc. etc. He has published more than 20 Research papers in reputed International Refereed Journals/ Conferences. He has published 5 Books, 2 out of them are available on Amazon. He has contributed as Session Chair, organizing Committee, Member of Reviewer Committee and Technical Program Committee for multiple International Conferences (IEEE). He is the member of UACEE, IAENG, IACSIT, CSTA etc. His

Dr Bhawna Mallick is working as senior consultant at Maverick Quality Advisory Services Pvt. Ltd Ghaziabad, India,She was Professor and Head of the Department at Galgotia College of engineering and Technology.She is having 20+ years' experience. Degree(s) PhD, M.Tech and B.E (Computer Technology). She has published more than 30 Research papers in reputed International Refereed Journals/ Conferences. She has organized 2 International Conferences and many workshops & Seminars.