ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A Survey on Interactive Representation of Closed+ High Utility Itemset susing Hadoop

Shaikh Hadiya Mohammad Ejaz[1], Prof.P.N.Kathavate[2]

P.G. Student, Dept. of Computer Science &Engineering, Walchand Institute of Technology, Solapur University, Solapur, India[1]

Assistant Professor, Dept. of Computer Science &Engineering, Walchand Institute of Technology, Solapur University, Solapur, India[2]

ABSTRACT: In the past few years a large numbers of algorithms have been design to mining a high utility itemsets from a database.The problem of all those algorithms is to produce a large numbers of high utility itemsets which downgrades the performace of mining process.To achieve the high performance for mining the task we design a novel framework in this paper for provides the mining Closed+ high utility mining itemsets (CHUI),which gives the concise and lossless representation of HUIs. The already existing designed algorithm named CHUD (Closed+ High Utility itemsets Discovery) to find concise representation is further enhanced by implementing the phase-I of CHUD based on AprioriHC-D-DCI (AprioriHC-D algorithm with Distinctive Closed Itemsets) in Hadoop, so system with less data set which required more memory can be analyzed in low memory based system with the help of distributed file system. A DAHU(Drive All High Utility Itemstes) method is introduce to recover all HUIs from set of CHUIs without accessing original databases.

KEYWORDS: Frequent itemset,;closed+ high utility itemset; utility mining; hadoop;data mining

I. INTRODUCTION

Frequent itemset mining(abbreviated as FIM)[2],[3],is a Major research topic in data mining. One of its famous applications is *market basket analysis*, which refers to the discovery of sets of items (itemsets) that are frequently picked up or purchased together by customers. However, in this application, the traditional model of FIM may discover a large amount of frequent or constant itemsets with low profit and drop the information on valuable itemsets having low selling frequencies.There are two problem in FIM such as (1) FIM treats all items as having the same importance/unit profit/weight and (2) it consider that every item in a transaction appears in a binary form, i.e., an item can be either present or absent in a transaction, which doesn't express its purchase quantity in the transaction. Hence, FIM cannot satisfy the requirement of users who need to discover itemsets with high utilities such as high profits.The utility of an itemset represents its importance, which can be measured in word or term of weight, profit, cost, quantity or other information depending on the user choice. An itemset is called a high utility itemset (abbreviated as HUI) if its utility is no less than a user specified or stated minimum utility threshold. Utility mining has a wide range of applications such as website click stream analysis cross-marketing analysis and biomedical domains. However, HUIs mining is not an easy task since the downward closure property[1,2,10] in FIM does not hold in utility mining.The search space cannot be directly pruned or cut backed to find HUIs as in FIM since a superset of a low utility itemset can be a high utility itemset. Many studies [4,5,6,7,8]were suggested for mining HUIs, but they often present a large number of high utility itemsets to users such that awareness of the results becomes difficult.Mean while, the algorithms become inefficient in terms of time and memory need. In particular, the work of the mining task decreases greatly under low minimum utility thresholds or dense databases.

To deminish the computational cost in FIM while presenting less and more important patterns to users, many studies advanced the concise representations, such as *free sets*[9],*nonderivablesets*[10],*maximal itemsets*[28] and *closed itemsets*[11, 12-13 14, 15].These representations successfully reduce the set of itemsets form, but they were developed for frequent itemset mining rather than the high utility itemsetmining.Therefore, an great research question is "*Is it possible*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

toaccept a compact and lossless representation of highutility itemsets encouraged by these representations to addressthe aforementioned problems in HUI mining?"

Responding this question positively tough. Developing a concise and complete representation of HUIs poses several challenges:

1. Combining the concepts of concise or terse representations from FIM into HUI mining, which may not useful to the users.
2. The representation may not provides a powerful reduction in terms of the number of extracted patterns to explain using this representation.
3. Algorithms for obtaining the representation may not be efficient.Algorithms may be slower than the perfect algorithms for mining all HUIs.
4. It may be tough to design and develop an efficient method for regaining all HUIs from the representation.

In this paper, we resolves all of these challenges by introducing a concise and meaningful representation of HUIs named *Closed+ High Utility Itemsets*(Closed+ HUIs),which combine the concept of closed itemset into HUI mining. The improvements are four-fold in correspondence to answering the four challenges already mentioned:

1. The introduce representation is *lossless* by using a new structure or framework named *utility unit array* that allows regaing all HUIs and their utilities efficiently.
2. The proposed representation is also tight or compact.
3. We recommend an efficient or adequate algorithm, named *CHUD* (Closed+ High Utility itemsetDiscovery) to find conciserepresentation based on AprioriHC-D-DCI(AprioriHC-D algorithm with Distinctive Closed Itemsets) [24]as phase-I of CHUD algorithm to generate a candidate itemsets in Hadoop framework. So system with less memory or data set which required more memory can be analyzed in low memory based system with the help of distributed file system.
4. We introduces a top-down mechanism named *DAHU* (*D*erive *AllH*igh *U*tility itemsets) for efficiently restoring all HUIs from the set of Closed^{+} HUIs. The combination of CHUD and DAHU gives a new way to obtain all HUIs and it outperforms *UPGrowth*[16], the state-ofthe- art algorithm for mining HUIs.

II. RELATED WORK

C.-W Wu, P. Fournier-Viger, P. S. Yu. and V. S. Tseng, [17] this paper provides mining closed+ high utility itemsets, which serves as a concise and lossless representation of high utility itemsets. They present an efficient algorithm named CHUD (Closed+ High Utility itemset Discovery) for mining closed+ high utility itemsets.A method named DAHU (Derive All High Utility itemsets) is designed to repair complete high utility itemsets from the set of closed+ high utility itemsets without achieving the original database. Further enhance the performance of CHUD algorithm they includes three effective and effecient strategies named REG, RML and DCM.

C.-W. Wu, B.-E.Shie, V. S. Tseng and P. S. Yu, [25]they address problem by proposing a new framework named top-k high utility itemset mining, where k is represent the number of high utility itemsets to be mined.They designed an efficient algorithm named TKU (Top-K Utility itemsets mining) for mining such itemsets without setting min_util.

B.-E. Shie, V. S. Tseng, and P. S. Yu, [18] this paper, they proposed a novel algorithm, namely GUIDE, for efficiently mining temporal maximal utility itemsets from the landmark time to the present in data streams. They also proposed a new data structure, namely TMUI-tree, for storing information in the processes of mining utility patterns from data streams. The main contributions of GUIDE and TMUI-tree are that GUIDE is the first one-pass algorithm for mining maximal utility itemsets in data streams and TMUI-tree is easy to maintain and it can help GUIDE finding TMUIs efficiently.

V.S. Tseng et al,[19] in this paper focuses on temporal high utility itemset mining(THUI).Which discovered the temporal high utility itemsets with less candidate itemsets and higher performance. THUI-Mine employs a filtering threshold in each partition to generate a progressive set of itemsets.They are two problem with THUI-mine algorithm first it require huge memory and second its generate a lot of false candidate itemsets.

C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong ,and Y.-K. Lee ,[20]authors propose three novel tree structures to efficiently to perform interactive and incremental HUP mining. The first tree structure provides Incremental HUP Lexicographic Tree (IHUPL-Tree), is arranged according to an item's lexicographic order.The incremental data can be capture without any reconstructing operation.The second tree structure is the IHUP Transaction Frequency Tree (IHUPTF-Tree), which provides a compact size by arranging items according to their transaction frequency

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

(descending order).The third tree, IHUP-Transaction-WeightedUtilization Tree (IHUPTWU-Tree) is designed based on the TWU value of items in descending order to reduce the mining time.

B. Vo, H. Nguyen, T. B. Ho, and B. Le, [21] they designed the parallel method for mining HUIs from vertically partitioned distributed databases, and the efficient algorithm is also proposed. The mining algorithm in distributed databases is more efficient than that in centralized database.The algorithm scans only local databases once and only item that its twu satisfies minutil must be sent to MasterSite by using WIT-tree technique. Therefore, it spends a ;imited time for communication between MasterSite and SlaverSites.

III. PROPOSED ALGORITHM

The CHUD algorithm using the AprioriHC-D-DCI(AprioriHC-D algorithm with Distinctive Closed Itemsets) to determine Candidate itemsets in phase-I of CHUD algorithm using Hadoop, so system with less data set which required more memory can be analyzed in low memory based system with the help of distributed file system.The CHUD algorithm uses an Itemset-Tidsetpair Tree (IT-Tree) to determine CHUIs.Each node N(X) in an IT-Tree, subsists of an itemset X, its Tidset g(X), and two ordered sets of items named PREV-SET(X) and POST-SET(X).The algorithm is recursively generate all closed high utility itemstes. To storing the transaction utilities of transactions it using the data structure called transaction utility table (TU-Table). It gives a list of pairs <R, TU (TR)> where the first value indicate a transaction ID R and the second value indicate the transaction utility of TR. And after that we obtain the all CHUIs. A DAHU(Derive All High Utility itemsets) method is proposed to regain all HUIs and their absolute utilities from the set of CHUIs without accessing the original database.

Fig1.show the overview of the proposed Closed High Utility Itemsets(CHUIs) representation.

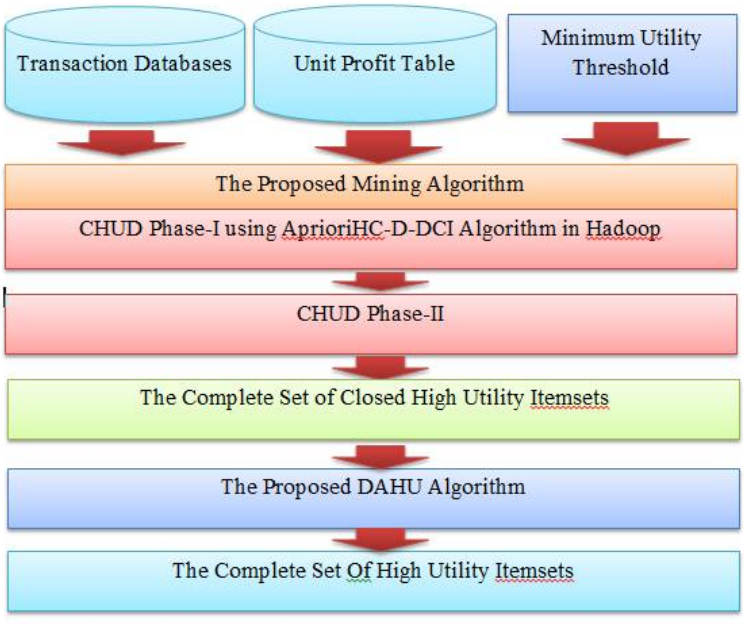

Fig1.Overview Of Proposed Closed High Utility Itemsets(CHUIs) representation

In Fig1 shows the inputs Transaction Database ,Unit Profit Table and Minimum Utility Threshold are given to the phase-I of CHUD algorithm,which using AprioriHC-D-DCI algorithm to generate the Candidate Itemsets in hadoop.Then after that this generated Candidate Itemsets is given to phase-II of CHUD algorithm,which generate the Complete Set Of High Utility Itemsets.Futher a method named DAHU(Derive All High Utility Itemsets) is proposed to generate the Complete Set High Utility itemsets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

IV. EXPECTED RESULT

- Using AprioriHC-D-DCI in phase-I of CHUD algorithm generate frequent candidate items from transactional database.
- The proposed CHUD algorithm using hadoop is may efficient than the existing CHUD algorithm in terms of memory space.

V. CONCLUSION

This paper discussed the various data mining algorithms for mining High Utility Itemsets. Mainly efficient algorithm named as CHUD(Closed High Utility Itemsets Discovery) based on AprioriHC-D-DCI (AprioriHC-D algorithm with Distinctive Closed Itemsets) in Hadoop.The TU-Table (Transaction Utility Table) [23] data structure is used for storing the transaction utilities of transactions.The DAHU(Derive All High Utility itemsets) method is proposed to regain all HUIs from the set of CHUIs without accessing the original database.

REFERENCES

1. Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu, *Fellow, IEEE*, Vincent S. Tseng " Efficient Algorithms for Mining the Concise and Lossless Representation of Closed High Utility Itemsets ",IEEE Transactions on Knowledge and Data Engineering.1041-4347(c)2013 IEEE.
2. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proc. of the 20th Int'l Conf. on Very Large Data Bases, pp. 487-499, 1994.
3. J. -F. Boulicaut, A. Bykowski, and C. Rigotti, "Free-sets: A Condensed Repre-sentation of Boolean Data for the Approximation of Frequency Queries," Data Mining and Knowledge Discovery, Vol. 7, Issue 1, pp. 5–22.
4. C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee."Efficient Tree Structures for High utility Pattern Mining in Incremental Databases,"inIEEE Transactions on Knowledge and Data Engineering, Vol. 21, Issue 12,pp. 1708-1721, 2009.
5. Y. Liu, W. Liao, and A. Choudhary."A fast high utility itemsets mining algorithm," in Proc. of the Utility-Based Data Mining Workshop, pp. 90-99,2005.
6. Y.-C. Li, J.-S. Yeh, and C.-C. Chang,"Isolated Items Discarding Strategy for Discovering High utility Itemsets,"in Data & Knowledge Engineering, Vol. 64, Issue 1, pp. 198-217, 2008.
7. V. S. Tseng, C.-W. Wu, B.-E. Shie, and P. S. Yu," UP-Growth: an efficient algorithm for high utility itemsetmining,"in Proc. of Int'l Conf. on ACM SIGKDD, pp. 253–262, 2010.
8. B. Vo, H. Nguyen, T. B. Ho, and B. Le. "Parallel Method for Mining High utility Itemsets from Vertically Partitioned Distributed Databases,"in Proc. of Int'l Conf. on Knowledge-based and Intelligent Information and Enginnering Systems, pp. 251-260, 2009.
9. J. -F. Boulicaut, A. Bykowski, and C. Rigotti. "Free-sets: a condensed representation of boolean data for the approximation of frequency queries,"in Data Mining and Knowledge Discovery, Vol. 7, Issue 1, pp. 5–22.
10. T. Calders and B. Goethals." Mining all non-derivable frequent itemsets,"in Proc. of the Int'l Conf. on European Conference on Principles of Data Mining and Knowledge Discovery, pp. 74-85, 2002.
11. G.-C. Lan, T.-P. Hong, V. S. Tseng, "An Efficient Projection-based Indexing Approach for Mining High Utility Itemsets". Knowledge and Information System, Vol. 38, Issue 1, pp. 85-107, 2014.
12. B. Le, H. Nguyen, T. A. Cao, and B. Vo, "A Novel Algorithm for Mining High utility Itemsets," in Proc. of the First Asian Conference on Intelligent Infor-mation and Database Systems, pp.13-17, 2009.
13. C. Lucchese, S. Orlando and R. Perego, "Fast and Memory Efficient Mining of Frequent Closed Itemsets," IEEE Transactions on Knowledge and Data Engi-neering, Vol. 18, Issue 1, pp. 21-36, 2006.
14. B.-E. Shie, H.-F. Hsiao, V. S. Tseng and P. S. Yu, "Mining High Utility Mobile Sequential Patterns in Mobile Commerce Environments," in Proc. of the Intl. Conf. on Database Systems for Advanced Applications and Lecture Notes in Com-puter Science (LNCS), Vol. 6587/2011, pp. 224-238, 2011.
15. J. Wang, J. Han, and J. Pei, "Closet+: Searching for the Best Strategies for Mining Frequent Closed Itemsets," in Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 236–245, 2003.
16. V. S. Tseng, C.-W. Wu, B.-E. Shie, and P. S. Yu." UP-Growth: an efficient algorithm for high utility itemsetmining.,"in Proc. of Int'l Conf. on ACM SIGKDD, pp. 253–262, 2010.
17. C.-W Wu, P. Fournier-Viger, P. S. Yu. and V. S. Tseng, "Efficient Mining of a Concise and Lossless Representation of High Utility Itemsets," in Proc. of the IEEE Int'l Conf. on Data Mining, pp. 824-833, 2011.
18. B.-E. Shie, V. S. Tseng, and P. S. Yu, "Online Mining of Temporal Maximal Utility Itemsets from Data Streams," in Proc. of Annual ACM Symposium on Applied Computing, pp. 1622-1626, 2010.
19. Tseng V.S, C.W. Wu, B.E. Shie, and P.S. Yu, "UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining," in Proc. 16th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD"10), pp. 253-262, 2010.
20. C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, "Efficient Tree Struc-tures for High utility Pattern Mining in Incremental Databases," IEEE Trans-actions on Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, 2009.
21. V. S. Tseng, C.-W.Wu, B.-E.Shie, and P. S. Yu, "UP-Growth: An Efficient Algorithm for High Utility Itemset Mining," in Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 253–262, 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

22. S.Shankar, T.P.Purusothoman, S. Jayanthi,N.Babu,"A fast agorithm for mining high utility itemsets," in: Proceedings of IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pp.1459-1464.
23. K. Gouda and M. J. Zaki," Efficiently mining maximal frequent itemsets," in Proc. of IEEE Int'l Conf. on Data Mining, pp. 163-170, 2001.
24. Immanuel K, E.Manohar, Dr.D.C.Joy Winnie Wise"An Effective Mining theConcise Representation Based on the Odd Ratio Pattern and Distinctive Closed HighUtilityItemsets,"International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)
25. K.Chuang, J. Huang, M. Chen, "Mining Top-K Frequent Patterns in the Presence of the Memory Constraint," VLDB Journal, Vol. 17, pp.1321-1344, 2008.