Babylonian Journal of Networking
Vol.2024, **pp**. 25–30
DOI: https://doi.org/10.58496/BJN/2024/004; ISSN: 3006-5372
https://mesopotamian.press/journals/index.php/BJN

Research Article

Mining Utilities Itemsets based on social network

Sara salman Qasim [1], Lubna Mohammed Hasan [2,*]

[1] *College of Computing & Informatics (CCI), University Tenaga Nasional (UNITEN), Putrajaya Campus, Malaysia.*

[2] *Computer Technology Engineering Department, Al-esraa University, Iraq, Baghdad.*

ARTICLE INFO

Article History
Received 17 Dec 2023
Accepted 15 Feb 2024
Published 03 Mar 2024

Keywords
Facebook
Twitter
LinkedIn
social network

ABSTRACT

Mining utility item sets based on social network data involves extracting meaningful patterns and associations from user interactions. In this paper, the process begins by collecting and preprocessing data from platforms like Facebook, Twitter, or LinkedIn. Utility measures are defined based on frequency of occurrence, user engagement metrics, or other domain-specific criteria. Itemsets that meet certain thresholds are identified using techniques like frequent itemset mining or advanced algorithms like Apriori or FP-growth. Additional analyses, such as association rule mining, uncover relationships between different itemsets or user segments, providing valuable insights for personalized recommendations, targeted advertising, and decision-making processes.

1. INTRODUCTION

Over the last 10 years, the social network has attracted a lot of attention. Through the internet and web 2.0 technology, accessing social network sites like Facebook, LinkedIn, Twitter, and Google+ has become more cheap. Individuals are become increasingly engaged with social media and depending on it for news, information, and other users' opinions on a wide range of topics. Because of their extensive dependence on social media, they produce enormous amounts of data that have three computing problems: scale, noise, and dynamism. These problems often make social network data very difficult to manually examine, necessitating the relevant use of computational methods. Numerous methods, such as trends, patterns, and rules, may be found in large datasets by using data mining [1]. Machine learning, statistical modeling, and information retrieval all make use of data mining methods. During the data analysis process, these strategies make use of pre-processing, analysis, and interpretation of the data.

At this point in time, humans are living in what is often known as the information era. We have been gathering an enormous amount of information in this age of information because we are of the belief that information is the key to power and success, and because we have been able to do so with the help of advanced technologies such as computers, satellites, and other similar devices at our disposal. This data presents a significant number of untapped potential for the discovery of new knowledge. In this section, we will be discussing one of the most fascinating and widely used areas of study, which is data mining for the purpose of locating high utility item sets. Data mining is a technique that aims to either comprehend the past or make predictions about the future [1]. The objective of Knowledge Discovery in Databases, often known as KDD, is to extract information that is both relevant and helpful from massive volumes of data. Frequently occurring itemset mining (FIM) and association rule mining (ARM) are two essential concerns in KDD that have wide applications in a variety of areas [1, 2]. A few fundamental preliminary steps have to be the primary emphasis prior to getting started.

One of the developing tasks in the field of data science is known as high utility pattern mining. This activity involves the discovery of patterns that are of high value in stored databases. When it comes to determining the usefulness of a pattern, there are a number of objective factors that may be used, including its profit, frequency, and weight. Of the many different forms of high utility patterns that may be found in databases, high utility item sets are the ones that have received the most attention from researchers. The term "high utility item set" refers to a collection of values that are stored in a database and are deemed to be of great significance to the user and are evaluated using a utility function. Through the consideration of item numbers and weights, high utility item set mining is able to generalize the difficulty that is associated with frequent item set mining. The discovery of all sets of things that consumers have bought together and that result in a high profit is a

*Corresponding author. Email: eng.lubna.m@gmail.com

common application of high utility item set mining. This chapter gives an introduction to high utility item set mining, covers the methods that are currently considered to be state-of-the-art, explores the extensions and applications of these algorithms, and highlights research possibilities [2].

2. PRELIMINARIES

2.1 Association Rule Mining (ARM)

The process of association rule mining, often known as ARM, is highly regarded for its ability to uncover co-occurrences, linkages, and common patterns among things that are included inside a database or a series of transactions. Association Rules is a notion that may be used to detect regularities between items that are stored in several databases. One may break the process of mining association rules down into two stages: the first stage involves the generation of frequent item sets, and the second stage involves the generation of association rules. When it comes to association rule mining, the most difficult problem is identifying itemsets that occur often. Association rule mining involves a number of important phases, one of which is the discovery of a frequent itemset. Due to the fact that the answer to the second sub-problem is straightforward, the majority of the researchers had focused their attention on the process of generating frequent item sets. The market-basket analysis process makes extensive use of ARM. By way of illustration, common item sets may be discovered via the examination of market basket data. Subsequently, association rules can be formulated by predicting the acquisition of further products through the utilization of conditional probability [1, 2]. One example of an association rule might be the following: "If a consumer purchases a computer, there is an eighty percent chance that he will also purchase a thumb drive." As a result, association rule mining is the most significant and extensively researched data mining technique. It is utilized by the majority of organizations for the purpose of decision making, with the goal of enhancing their performance in terms of sales and product quality, as well as improving their profits.

2.2 Frequent Itemsets Mining (FIM)

One of the most interesting subfields of data mining is known as frequent itemsets mining. The basis data for frequent itemset mining is organized in the form of sets of instances, which are often referred to as transactions. Each of these instances has a number of characteristics (also called items). It was the first algorithm ever developed for mining frequent item collections. This method works by first searching the database for all of the frequent 1-itemsets, then moving on to find all of the frequent 2-itemsets, then continuing on to discover all of the frequent 3-itemsets, and so on. Joining frequent item sets of length n minus one results in the generation of candidate item sets of length n at each iteration. The frequency of each candidate item set is examined before it is added to the collection of frequent item sets. The purpose of the Frequent Itemset Mining technique is to locate all of the itemsets that are considered to be frequent in a transaction database.

3. HIGH UTILITY MINING ITEMSET (HUMI)

An introduction to the fundamental ideas of Data Mining, Association Rule Mining, and Frequent Itemsets Mining was presented in the part that came before this one. The purpose of this section is to provide a concise summary of the many methods, ideas, and techniques that have been outlined in a variety of academic articles pertaining to High Utility Mining. The notion of Utility Mining is an expansion of the conventional kind of mining known as frequent itemset mining. The classic methods to ARM take into account the practicality of the items based on the number of times they appear in the transaction set. There is not enough information available to accurately indicate the real usefulness of an itemset based on its frequency. For instance, the sales manager may not be interested in frequent item sets that do not make a large amount of profit. The effective mining of high utility item sets has emerged as one of the most difficult issues in the field of data mining in recent developments. Utility Mining refers to the process of locating item sets that have a high utility to each individual. There are a number of ways in which the utility may be evaluated, including in terms of cost, profit, or individual preferences. It is possible, for instance, that a computer system will generate a higher profit than a telephone would [3]. In light of this, utility mining has emerged as a subject of significant importance within the discipline of data mining. Finding the item sets that have the highest earnings is what is meant by the term "mining high utility itemsets from databases." When it comes to this context, the term "itemset utility" refers to the degree to which an item is interesting, important, or profitable for users. The relevance of objects in a transaction database may be broken down into two categories: the first is the significance of individual items, which is referred to as external utility, and the second is the significance of items inside transactions, which is referred to as internal utility. According to one definition, the utility of an itemset is equal to the product of its internal usefulness and its exterior utility. An itemset is referred to be a high utility itemset provided that its utility is more than or equal to a minimum utility threshold that has been determined by the user; otherwise, it is referred to as a low-utility itemset [4].

$$Utility\ of\ Itemset\ (U)\ =\ Internal\ Utility\ (i)\ *\ External\ Utility\ (e)$$

Example Let Table 2 be a database containing five transactions. Each row in Table 2 represents a transaction, in which each letter represents an item and has a purchase quantity (internal utility). Table 3 represents the unit profits associated with each itemset.

TABLE I. TRANSACTION DATABASE

Trans id	***Transaction***	***Transaction Utility***
T1	A(1),B(1),E(1),W(1)	5
T2	A(1),B(1),E(3)	8
T3	A(1),B(1),F(2)	8
T4	E(2),G(1)	5
T5	A(1),B(1),F(3)	11

TABLE II. UNIT PROFITS ASSOCIATED WITH ITEMS

Item Name	A	B	E	F	G	W
Unit Profit	1	1	2	3	1	1

4. RELATED WORK

In this paragraph, we will review the range of research that has been utility mining using user identification over social networks as shown below:

In 2003 **M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg [5],** has been an approach of calculating a similarity score between a pair of entities has been studied and applied a variety of areas in the past, including string similarity (or distance) ,document similarity used in document clustering and information filtering. These approaches have been extended into further applications, such as genetics, natural language processing, image processing and others.

In 2005 **X. Dong and A. Halevy. [6],**Work has been done in the field of personal information reconciliation, which includes identifying duplicated person references among personal data (documents). The stepwise comparison method proposed it may be applied to social networks, however, our approach allows for more flexibility and complexity, which is important in profile comparison on social networks. A general reference reconciliation method proposed in utilizes the attributes of entities as well as relationships between entities for the identification process. mainly focuses on the special application of profile matching in social networks and the empirical study of the importance of profile fields.

In 2008 **R. Thiagarajan, G. Manjunath, and M. Stumptner,[7],**Vector-based comparison algorithms have been used for document and query matching in traditional search engines and more recently in ontology-based search engines. The notion of user profile vector comparison has been introduced in relation to collaborative information retrieval (CIR) systems. These methods remain focused on the query-matching/document-retrieval problems. In contrast, our work focuses on the social network user profile. The profile matching problems together with the graph-based algorithm are overcome by our field matching techniques and the vector-based comparison algorithm incorporating weights.

In 2003 **Chan et al,[8],** has presented the extension to the Apriori algorithms in terms of the Novel algorithm OOA mining with the top-K utility frequent closed patterns, he also observes that the candidate set pruning strategy exploring the antimonotone property used in apriori algorithm do not hold for utility mining.

In 2005 **Liu et al,[9],** proposed a Two-phase algorithm for finding high utility itemsets that can discover high utility itemsets more efficiently.It works in two phases, in Phase I, a term transaction-weighted utilization is defined, and proposed the transaction-weighted utilization mining model that holds Transaction-weighted Downward Closure Property. That is, if a k-itemset is a low transaction-weighted utilization itemset, none of its supersets can be a high transaction-weighted utilization itemset. The transactionweighted utilization mining not only effectively restricts the search space, but also covers all the high utility itemsets. Although Phase I may overestimate some itemsets due to the different definitions, only one extra database scan is needed in Phase II to filter out the overestimated itemsets.

In 2012 **Liu & Qu ,[10],**proposed HUI-MINER algorithm. In this paper Utility List is created. It first creates an initial utility list for itemsets of the length 1 for promising items. Then HUIMINER constructs recursively a utility list for each itemset of the length k using a pair of utility lists for itemsets of the length k-1.For mining high utility itmesets, each utility list for an itemset keeps the information of TIDs for all of transactions containing the itemset, utility values of the item set in the transactions, and the sum of utilities of the remaining items that can be included to super itemsets of the itemset in the transactions. The distinct advantage of HUIMiner is that it avoids the costly candidates generation and utility computation.

In 2014 **Philippe Fournier-Viger,[11],** proposed FHM algorithm [40]. It extends the Hui-Miner Algorithm. It is a Depth-first search Algorithm. It relies on utility-lists to calculate the exact utility of itemsets. This algorithm integrates a novel strategy named EUCP (Estimated Utility Co-occurrence Pruning) to reduce the number of joins operations when mining high utility itemsets using the utility list data structure. Estimated Utility Co-Occurrence Structure (EUCS) stores the transaction weighted utility (TWU) of all 2-itemsets. It built during the initial database scans. FHM is up to 6 times faster than HUI - Miner.

In 2015 **Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin,Cheng Wei Wu,and Vincent S. Tseng,[12].** Here several new ideas have introduced to more efficiently discovers high utility itemsets both in terms of execution time and memory **.**EFIM relies on two upperbounds named sub-tree utility and local utility to more effectively prune the search space. It also introduces a novel array-based utility counting technique called Fast Utility Counting to calculate these upper-bounds in linear time and space. Transaction merging is obviously desirable. However, a key problem is to implement it efficiently. To find identical transactions in O(n) time, sort the original database according to a new total order T on transactions. Sorting is achieved in time, and is performed only once. Projected databases generated by EFIM are often very small due to transaction merging.

In 2015 **Jerry Chun-Wei Lin et al,[13],** has proposed a novel framework for mining potential high-utility itemsets (PHUIs) over uncertain databases. This is the first paper to address the issue of mining potential high-utility itemsets from uncertain databases. The upper-bound-based algorithm (PHUI-UP) and the list-based algorithm (PHUI-List) are respectively proposed to consider the mining of not only high-utility but also high probability itemsets from uncertain databases. The designed PHUI-UP algorithm is based on the proposed downward closure property to level-wisely generate-and-test candidates for mining PHUIs. The second PHUI-List algorithm is further developed to improve the performance based on the designed vertical PU-list structure for directly mining PHUIs without candidate generation.

In 2016 & 2017 the recent MUHUI and PHUIMUS have also proposed [14],[15].
Overall the most extensive and remarkable contributions in HUIM can be seen in the below Table 4.

TABLE III. SUMMARY OF REMARKABLE CONTRIBUTION IN HUM

Studied By	*Algorithms*	*Year of Publication*	*Outcomes*
Chan Q.,YangY., and Shen D.	OOA Algorithm	2003	Mining the top-K utility frequent closed patterns. Antimonotone property used in apriori algorithm do not hold for utility mining.
Liu Y., Liao W. And Choudhary A.	Two Phase Algorithm	2005	In two phases, Phase Itransaction-weighted utilization is defined. Phase II - one extra database scan to filter out the overestimated itemsets
Liu M. and Qu J.	HUI-MINER Algorithm	2012	It avoids the costly candidates generation and utility computation and generate high utility itemsets.
Philippe Fournier Viger ,ChengWeiWu,Souley mane Zida and Vincent S. Tseng	FHM Algorithm	2014	It is a Depth-first search Algorithm. Reduces the number of join operations using the utility list data structure. It is up to 6 times faster than HUIMiner.
Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin,Cheng Wei Wu,and Vincent S. Tseng	EFIM Algorithm	2015	Two upper-bounds named sub-tree utility and local utility to more effectively prune the search space is used. Array-based utility counting technique is proposed.
LinJCW,Gan, W,FournierViger P, Hong TP and Tseng VS	PHUI Algorithm	2015	Addressed the issue of mining potential highutility itemsets from uncertain databases. Upper-bound-based algorithm (PHUI-UP) and the list-based algorithm (PHUI-List) are proposed.
Lin JCW,Gan W,FournierViger P, Hong TP and Tseng VS	MUHUI Algorithm	2016	Based on the probability-utility-list (PU-list) structure,It directly mine PHUIs without candidate generation and can reduce the construction.
Ju Wang, Fuxian of PU-lists and thus Liu, and Chunjie Jin	PHUIMUS Algorithm	2017	Represents the itemsets with high utilities and high existential probabilities over uncertain data stream based on sliding windows

5. CONCLUSION

A Utility mining is an apparent topic in data mining. The main focus in the field of Utility Mining is not only FIM but also the consideration of utility. Practically it has been found that the utility is of great interest in industry if considers with high utility itemsets. Different decision making domains such as business transactions, medical, security, fraudulent transaction, retail etc. make use of high item sets to get useful information. Survey on different high utility item set mining algorithms which were proposed is presented in this paper. This survey will be helpful for developing new efficient and optimize techniques for high utility item set mining. The open research opportunities in this field can be in the form of Novel Applications development by applying existing pattern mining algorithms in new ways, the performance can be enhanced in terms of memory and time utilization and can discover more complex and meaningful type of patterns. As the concept of High Utility Itemset Mining has some vast opportunities to be researched, the future work will incorporate soft computing methodologies for high utility itmesets mining such as the intuitionistic fuzzy logic can be explored in the field of High Utility Itemset Mining further, the associations of the student ideas were explored by employing the Apriori algorithm and, as can be seen from the results obtained, the contribution of Facebook to communication between classmates is more than to communication between students and teachers. Moreover, students who hold these views believe that Facebook is a good medium for accessing rich resources. More of these types of rules can be revealed by using the Apriori algorithm and the use of social network sites for educational ends can be reformed in the light of these rules. If the increasing trend in social network sites usage is considered, the importance of applications and approaches related to social networks can be easily understood. Targeting specific ages or sex may strategically affect the success of developed applications. As a consequence, data mining methods can be successfully employed on social network usage data.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Funding

No financial contributions or endorsements from institutions or sponsors are mentioned in the author's paper.

Acknowledgment

The author expresses appreciation to the institution for their continuous support and access to relevant research materials.

References

[1] Fournier-Viger P., Chun-Wei Lin J., Truong-Chi T., Nkambou R. (2019) A "Survey of High Utility Itemset Mining". In: Fournier-Viger P., Lin JW., Nkambou R., Vo B., Tseng V. (eds) High-Utility Pattern Mining. Studies in Big Data, vol 51. Springer, Cham.
[2] H.Yao, H.J.Hamilton, C.J.Butz, "A foundation approach to mining itemset utilities from databases", in: Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida , pp.482-486, 2004.
[3] Pillai J. and Vyas O.P. "Overview of itemset Utiltiy Mining and its Applications",August International Journal of Computer Applications (0975 - 8887) Volume 5 – No. 11, 2010.
[4] Vincent S.Tseng,Bai-En shie,Cheng-Wei Wu and Pjillip S.Yu, "Efficient Algorithms for Mining High Utility Itemset from Transactional Databases",8 August 2013,IEEE Transactions on Knowledge and Data Engineering ,Vol 25 pp 1172-1786,2013.
[5] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. IEEE Intelligent Systems, 18(5):16–23, 2003.
[6] X. Dong and A. Halevy. A platform for personal information management and integration. In CIDR, pages 119–130, 2005.
[7] R. Thiagarajan, G. Manjunath, and M. Stumptner. Finding experts by semantic matching of user profiles. Technical Report HPL-2008-172, HP Laboratories, October 2008.
[8] Chen C.L., Tseng F.S., Liang T., "Mining fuzzy frequent itemsets for hierarchical document clustering", Inf Process Manage,46:193–211,2010.
[9] Liu Y., Liao W., and Choudhary A., "A Fast High Utility Itemsets Mining Algorithm," Proc. Utility-Based Data Mining Workshop, 2005.
[10] Liu M. and Qu J., "Mining High Utility Itemsets without Candidate Generation" ,CIKM'12 ,Maui,HI,USA, ACM, October29-November 2,2012.
[11] Philippe Fournier Viger ,Cheng-Wei Wu,Souleymane Zida,Vincent S. Tseng, "FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning", Proc. 21st International Symposiumon methodologies for Intelliegnet Systems (ISMIS 2014),Springer,LNAI,pp 83-92,2014.
[12] Souleymane Zida, Philippe Fournier-Viger, Jerry ChunWei Lin,Cheng-Wei Wu,Vincent S. Tseng, "EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining", 30 December 2015,Mexican International Conference on Artificial Intelligence Advances in Artificial Intelligence and Soft Computing pp 530- 546,2015.
[13] Lin JCW, Tin L, Fournier-Viger P, Hong TP., "A fast algorithm for mining fuzzy frequent itemsets", J Intell Fuzzy Syst, 9:2373–2379, 2015.
[14] Lin JCW,Gan W,Fournier-Viger P, Hong TP, Tseng VS, "Efficiently mining uncertain high-utility itemsets". Springer International Publishing Switzerland 2016,WAIM 2016, Part I, LNCS 9658, pp. 17–30, 2016.
[15] Ju Wang, Fuxian Liu, and Chunjie Jin, "PHUIMUS: A Potential High Utility Itemsets Mining Algorithm Based on Stream Data with Uncertainty",Hindawi Mathematical Problems in Engineering Volume, Article ID 8576829, 13 pages 2017.