ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Mining High Utility Item-sets for Assigning Sections to Indian Laws: An Overview

Suvarna Asade, Prof. R. V. Argiddi

P.G Student, Department of CSE, WIT, Solapur University, Solapur, MH, India

Associate Professor, Department of CSE, WIT, Solapur University, Solapur, MH, India

ABSTRACT: In utility mining, each item is associated with a utility that could be profit, quantity, cost or other user preferences. Objective of Utility Mining is to identify the item sets with highest utilities. Basically the utility of an item set represents its importance, which can be measured in terms of information depending on the user specification and requirements. Item set is termed to be high utility item set if its utility is greater than min_util i.e. user specified minimum utility threshold. Practically in many applications high utility item sets consists of rare items. Different decision making domains such as business transactions, medical, security, fraudulent transaction, retail etc. make use of rare item sets to get useful information. Such as in supermarket customer purchase washing machine or fridge rarely as compared to butter or sugar. But previous transaction provides more profit for the supermarket. In this paper, will give an overview of high utility mining problem. Organization of this survey is given as follows: In first section we introduced basic terms like Data mining, frequent pattern mining, Association Rule mining, Utility Mining and Rare Item set Mining. In second section we summarize some important previous research work related to utility mining. A brief overview of various algorithms has been presented in this section. In last section we concluded the survey work by discussing some applications of Utility.

KEYWORDS: Utility Mining, High-utility item sets, rare item sets, Frequent Item set mining, Transaction, Weighted Utilization. Component**.**

I. INTRODUCTION

Data Mining: Data mining is about the extraction of hidden predictive information from large databases and is a powerful new technology with great potential to help companies focus on the important information in their data warehouses. Data mining tools used to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The prospective analyses offered by data mining technology move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve and answer. They search databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations and expertise. There are two general classification of Data Mining: Descriptive Mining and Predictive Mining. Clustering, Association Rule Discovery, Sequential Pattern Discovery are the Descriptive Mining techniques. They are used to find human-interpretable patterns that describe the data. Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables are Predictive Mining techniques.

1.1 Association rule Mining

One of the important areas of research is Association Rule Mining (ARM) in data mining. It is a prominent part of Knowledge Discovery in Databases (KDD).That's why it requires more concentration to explore. Association rule mining (ARM) is a technique for discovering co-occurrences, correlations, and frequent patterns, associations among items in a set of transactions or a database. We find rules having confidence and support above user defined threshold. The process of Association Rule Mining is divided into two steps: The first is to find all frequent item sets in data base

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

then to generate association rules. Market based analysis widely used ARM. For example, Market basket data are analyzed, frequent item sets are found then association rules can be generated by predicting the purchase of other items by conditional probability. Given a set of transactions where each transaction is a set of items, an association rule is an expression of the form X - >Y, where X and Y are sets of items. The problem of mining association rules was first introduced in [1] and later broadened in [2], for the case of databases consisting of categorical attributes alone.

1.2 Frequent Itemset Mining

Frequent pattern mining is the key area in the data mining concept which reveals the interesting pattern in the large database. Frequent pattern discovers the item set frequently occurs in a dataset and this information can be used in variety of applications such as market dataset analysis, Indexing and retrievals, detection of software bug, web link analysis etc. Frequent pattern mining considers only whether an item is present or not in a transaction. It reveals the pattern which appears more than the user specified support count. The problem of frequent itemset mining is popular. But it has some important limitations when it comes to analyzing customer transactions. An important limitation is that purchase quantities are not taken into account. Thus, an item may only appear once or zero time in a transaction. Thus, if a customer has bought five breads, ten breads or twenty breads, it is viewed as the same. A second important limitation is that all items are seen as having the same importance, utility of weight. For example, if a customer buys a very expensive bottle of wine or just a piece of bread, it is seen as being equally important. Thus, frequent pattern mining may find many frequent patterns that are not interesting. For example, one may find that {bread, milk} is a frequent pattern. However, from a business perspective, this pattern may be uninteresting because it does not generate much profit. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit such as perhaps {caviar, wine}.Let I = {a1, a2… an} be a set of n distinct literals called items. An item set is a non-empty set of items. An item set X = (a1, a2, …, ak) with k items is referred to as k-item set, A transactionT = consists of a transaction identifier (TID) and a setof items (a1, a2, …, ak), where ij Î I, j = 1, 2, …, k.The frequency of an item set X is the probability of X occurring in a transaction T. A frequent item set is the item set having frequency support greater a minimum user specified threshold.

1.3 Utility Mining

In real life, the merchant are interested in selling the itemset which generate more profits, but the frequent pattern mining generates only frequent itemset without considering quantity of the item sold or the profit of the item. Even though frequent itemset mining discovers crucial frequent patterns, it leaves the profit and quantity of the item is the disadvantage. To address this issue weighted association rule mining has been developed. It tries to find the association of itemset in a database by considering the profit/weight of the itemset.To address these, utility mining has been introduced utility mining considers both the profit and the number of items purchased. In this the utility of an itemset is calculated as the product of the profit of the item and the number of item purchased.Utility mining model was proposed in [3] to define the utility of item set. The utility is a measure of how useful or profitable an item set X is. The utility of an item set X, i.e., u(X), is the sum of the utilities of item set X in all the transactions containing X. An item set X is called a high utility item set if and only if u(X) >= min_utility, where min_utility is a user-defined minimum utility threshold [11].Frequent item set mining follows the download closure property, if K- item set is generated, K+1 itemset can be generated by considering only K-item set or in other words K+1 itemset will contain only the item set present in the K-itemset. This downward closure property is not satisfied, if k-itemset is low utility itemset K+1 can be high Utility item set and vice versa. Both the monotonic and anti monotonic property is not supported by high utility item set. This issue is addressed by the over estimation method [4].

II. LITERATURE REVIEW

More and more courts around the world are providing online access to judgments of cases, both past and present. With this exponential growth of online access to legal judgments, it has become increasingly important to provide improved mechanisms to extract information quickly and present rudimentary structured knowledge instead of mere information to the legal community. Automatic text summarization attempts to address this problem by extracting information

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

content, and presenting the most important content to the legal user. The other major problem we address is that of retrieval of judgments relevant to the cases a legal user is currently involved in. To facilitate this we need to construct a knowledge base in the form of a legal ontology. In this chapter, we present the methodologies related to single document summarization based on the method of extraction of key sentences from the documents as a general approach. This chapter also explains the importance of statistical approach to automatic extraction of sentences from the documents for text summarization. We also outline the different approaches to the summarization for a legal domain, and the use of legal ontology for knowledge representation of legal terms.

Data Mining refers to extracting or mining knowledge from large amounts of data. In large databases finding of frequent patterns task is very important use full in many applications over the past few years. The primary goal is to discover hidden patterns, unexpected trends in the data. Data mining is concerned with analysis of large volumes of data to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes. Data mining activities uses combination of techniques from database artificial intelligence, statistics, technologies machine learning . Utility Mining is one of the most challenging data mining tasks is the mining of high utility itemsets efficiently. Identification of the itemsets with high utilities is called as Utility Mining.

Association rules mining (ARM) [1] is one of the most widely used techniques in data mining and knowledge discovery and has tremendous applications like business, science and other domains. Make the decisions about marketing activities such as, e.g., promotional pricing or product placements. A high utility itemset is defined as: A group of items in a transaction database is called itemset. This itemset in a transaction database consists of two aspects: First one is itemset in a single transaction is called internal utility and second one is itemset in different transaction database is called external utility. The transaction utility of an itemset is defined as the multiplication of external utility by the internal utility. By transaction utility, transaction weight utilizations (TWU) can be found. To call an itemset as high utility itemset only if its utility is not less than a user specified minimum support threshold utility value; otherwise itemset is treated as low utility itemset.

To generate these high utility itemsets mining recently in 2010, UP - Growth (Utility Pattern Growth) algorithm [2] was proposed by Vincent S. Tseng et al. for discovering high utility itemsets and a tree based data structure called UP - Tree (Utility Pattern tree) which efficiently maintains the information of transaction database related to the utility patterns. Four strategies (DGU, DGN, DLU, and DLN) used for efficient construction of UP - Tree and the processing in UP - Growth [3]. By applying these strategies, can not only efficiently decrease the estimated utilities of the potential high utility itemsets (PHUI) but also effectively reduce the number of candidates. But this algorithm takes more execution time for phase II (identify local utility itemsets) and I/O cost Efficient discovery of frequent itemsets in large datasets is a crucial task of data mining. In recent years, several approaches have been proposed for generating high utility patterns; they arise the problems of producing a large number of candidate itemsets for high utility itemsets and probably degrade mining performance in terms of speed and space. Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant approaches have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets.

Liu et al. proposed an algorithm named Two- Phase [4] which is mainly composed of two mining phases. In phase I, it employs an Apriori-based level-wise method to enumerate HTWUIs. Candidate itemsets with length k are generated from length k-1 HTWUIs, and their TWUs are computed by scanning the database once in each pass. After the above steps, the complete set of HTWUIs is collected in phase I. In phase II, HTWUIs that are high utility itemsets are identified with an additional database scan.

Ahmed et al. [5] proposed a tree-based algorithm, named IHUP. A tree based structure called IHUP-Tree is used to maintain the information about itemsets and their utilities.

Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu, proposed [6] , " Efficient Algorithms for Mining Top-K High Utility Item sets". In this paper introduce high utility mining. High utility item sets (HUIs) mining is an emerging topic in data mining, which refers to discovering all item sets having a utility meeting a user-specified minimum utility threshold min_util. However, setting min_util appropriately is a difficult problem for users

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Regina Castro, and Peter W. Li [7] proposed "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus" .In this paper introduce detection of patients with elevated risk of developing diabetes mellitus is critical to the improved prevention and overall clinical management of these patients.
Cheng-Wei Wu, and Philip S. Yu, proposed [8] "Efficient Algorithms for Mining High Utility Item sets from Transactional Databases". In this paper introduce Mining high utility item sets from a transactional database refers to the discovery of item sets with high utility like profits.
O Srinivasa Rao, and MHM Krishna Prasad, proposed [9], "An Improved UP-Growth High Utility Item set Mining". In This paper introduce the UP-Growth is one of the efficient algorithms to generate high utility item sets depending on construction of a global UP-Tree.
S. Jayanthi and Nishanth Babu proposed [10] "A Fast Algorithm for Mining High Utility Item sets". In This Paper introduce Utility based data mining is a new research the economic usefulness of item sets as utility values and then area entranced in all types of utility factors in data mining find item sets with utility values higher than a minimum processes and focused at integrating utility considerations in threshold utility value as set by the user.

III. DISCUSSIONS

Base on above survey we got the actual idea of how to retrieve high utility mining item set from transaction database. so , we can expand the system any kind of relational database, with real time example. Data Mining refers to extracting or mining knowledge from large amounts of data. In large databases finding of frequent patterns task is very important use full in many applications over the past few years. The primary goal is to discover hidden patterns, unexpected trends in the data. Data mining is concerned with analysis of large volumes of data to automatically discover interesting regularities or relationships which in turn leads to better understanding of the underlying processes. Data mining activities uses combination of techniques from database artificial intelligence, statistics, technologies machine learning . Utility Mining is one of the most challenging data mining tasks is the mining of high utility itemsets efficiently. Identification of the itemsets with high utilities is called as Utility Mining

IV. CONCLUSION

A Utility mining is an evident topic in data mining. It focuses on utility consideration while item set mining. All aspects of economic utility in data mining are covered in utility mining. Practically in many applications high utility item sets plays an important role. Different decision making domains such as business transactions, medical, security, fraudulent transaction, retail etc. make use of rare item sets to get useful information. Some of the Applications of Utility Ming includes: In supermarket customer purchase washing machine or fridge rarely as compared to butter or sugar. But previous transaction provides more profit for the supermarket. Also in medical application the unique or rare combination of symptoms can give a useful information about disease of patient to doctors [13]. Retails business is able to find out their high investing customers with utility mining [14].Survey on different high utility item set mining algorithms which were proposed are presented in this paper. This survey will be helpful for developing new efficient and optimize technique for high utility item set mining.

REFERENCES

[1] 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.
[2] C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee. Efficient tree structures for high utility pattern mining in incremental databases. In IEEE Transactions on Knowledge and Data Engineering, Vol. 21, Issue 12, pp. 1708-1721, 2009.
[3] Y. Liu, W. Liao, and A. Choudhary, "A Fast High Utility Itemsets Mining Algorithm," Proc. Utility-Based Data Mining Workshop, 2005
[4] B.-E. Shie, V. S. Tseng, and P. S. Yu. Online mining of temporal maximal utility itemsets from data streams. In Proc.of the 25th Annual ACM Symposium on Applied Computing, Switzerland, Mar., 2010
[5] J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases," Proc. 21th Int'l Conf. Very Large Data Bases, pp. 420-431, Sept. 1995.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

[6] Terry M. Therneau, Steven S. Cha, M. Regina Castro, and Peter W. Li," Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus", VOL. 27, NO. 1, JANUARY 2015.
[7] Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu," Efficient Algorithms for Mining Top-K High Utility Item sets", DOI 10.1109/TKDE.2015.
[8] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu," Efficient Algorithms for Mining High Utility Item sets from Transactional Databases", VOL. 25, NO. 8, AUGUST 2013.
[9] Adinarayanareddy B,O Srinivasa Rao, MHM Krishna Prasad," An Improved UP-Growth High Utility Item set Mining", Volume 58– No.2, November 2012
[10] Hong Yao, Howard J. Hamilton," Mining item set utilities from transaction databases", S4S 0A2 18 November 2005.