Volume No:3, Issue No:7 (December-2017) ISSN No : 2454-423X (Online)

International Journal of Research in Advanced Computer Science Engineering

A Peer Reviewed Open Access International Journal

www.ijracse.com

An Extremely Efficient Algorithm for Mining of High-Utility Itemsets

Kishore Gorle
Department of Computer Science Engineering
Viswanadha Institute of Technology and Management,
Visakhapatnam, Andhra Pradesh 531173, India.

Nemana JayaLakshmi
Department of Computer Science Engineering
Viswanadha Institute of Technology and Management,
Visakhapatnam, Andhra Pradesh 531173, India.

ABSTRACT

High-utility itemset mining (HUIM) is an important data mining task with wide applications. In this paper, we propose a novel algorithm named EFIM (EFficient high-utility Itemset Mining), which introduces several new ideas to more efficiently discovers high-utility itemsets both in terms of execution time and memory. EFIM relies on two upper-bounds named sub-tree utility and local utility to more effectively prune the search space. It also introduces a novel array-based utility counting technique named Fast Utility Counting to calculate these upper-bounds in linear time and space. Moreover, to reduce the cost of database scans, EFIM proposes efficient database projection and transaction merging techniques. An extensive experimental study on various datasets shows that EFIM is in general two to three orders of magnitude faster and consumes up to eight times less memory than the state-of- art algorithms d2HUP, HUI-Miner, HUP-Miner, FHM and UP-Growth+. Keywords: high-utility mining, itemset mining, pattern mining.

INTRODUCTION

Data Mining

Data mining [2] 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 [1]. The primary goal is to discover hidden patterns, unexpected trends in the data. Data mining activities uses combination of techniques from database technologies, statistics, artificial intelligence and machine learning. The term is frequently misused to mean any form of large-scale data or information processing. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns. Over the last two decades data mining has emerged as a significant research area .This is primary due to the interdisciplinary nature of the subject and the diverse range of application domains in which data mining based products and techniques are being employed. This includes bioinformatics, genetics, medicine, clinical research, education, retail and marketing research. Data mining has been considerably used in the analysis of customer transactions in retail research where it is termed as market basket analysis. Market basket analysis has also been used to identify the purchase patterns of the alpha consumer. Alpha consumers are people that play a key role in connecting with the concept behind the inception and design of a product [5].

Frequent Itemset

Mining An itemset can be defined as a non-empty set of items. An itemset with k different items is termed as a k-itemset. For e.g. {bread, butter, milk } may denote a 3-itemset in a supermarket transaction .The notion of frequent itemsets was introduced by Agrawal et al [1].Frequent itemsets are the itemsets that appear frequently in the transactions. The goal of frequent itemset mining is to identify all the itemsets in a transaction dataset . Frequent itemset mining plays an

Cite this article as: Kishore Gorle & Nemana JayaLakshmi, "An Extremely Efficient Algorithm for Mining of High-Utility Itemsets", International Journal of Research in Advanced Computer Science Engineering, Volume 3 Issue 7, 2017, Page 7-12.

essential role in the theory and practice of many important data mining tasks , such as mining association rules [1,2], long patterns [5],emerging patterns[10], and dependency rules .It has been applied in the field of telecommunications [3],census analysis[5] and text analysis. The criterion of being frequent is expressed in terms of support value of the itemsets. The Support value of an itemset is the percentage of transactions that contain the itemset.

Temporal Data Mining (TDM) is defined as the activity of looking for interesting correlations or patterns in large temporal datasets. TDM has evolved from data mining and was highly influenced by the areas of temporal databases and temporal reasoning. Several surveys on temporal knowledge discovery exist [5]. Most temporal data mining techniques convert the temporal data into static representations and exploit existing 'static' machine learning techniques, thus potentially missing some of the temporal semantics. Recently there is a growing interest in the development of temporal data mining techniques in which the temporal dimension is considered more explicitly. Console et al. proposed an extension of the known Decision Trees induction algorithm to the temporal dimension [1]. One advantage of temporal decision trees is that the output of the induction algorithm is a tree that can immediately be used for pattern recognition purposes. However, the method can only be applied to time points, not to time intervals.

Initial research in financial and stock trading issues lead to the identification of some factors that are considered among experts to influence the price of a stock. At first, it is a reasonable thought that the behavior of an investor depends on the size of the owner company. Temporal data mining provides some additional capabilities required in cases where the evolution of the existing data and their interactions need to be observed through the time dimension. The stock market is one of them. We propose a tool that collects stock data and after analyzing and interpreting them, it will be able to act on the basis of these interpretations. The capabilities of this tool are based on temporal data mining patterns, extracted from stock market data [4]

RELATED WORK

The FUP algorithm updates the association rules in a database when new transactions are added to the database. Algorithm FUP [6-9] is based on the framework of Apriori and is designed to discover the new frequent itemsets iteratively. The idea is to store the counts of all the frequent itemsets found in a previous mining operation. Using these stored counts and examining the newly added transactions, the overall count of these candidate itemsets are then obtained by scanning the original database. An extension to the work in the authors propose an algorithm FUP2 for updating the existing association rules when transactions are added to and deleted from the database. UWEP (Update with Early Pruning) is an efficient incremental algorithm, that counts the original database at most once, and the increment exactly once. In addition, the number of candidates generated and counted is minimized. In recent years, processing data from data streams becomes a popular topic in data mining. A number of algorithms FTP-DS and RAM-DS have been proposed to process data in data streams. Lossy Counting divided incoming stream conceptually into buckets. It uses bucket boundaries and maximal possibleerror to update or delete the itemsets with frequency for mining frequent itemsets. FTP-DS is a regressionbased algorithm for mining frequent temporal patterns from data streams [10].

The principle of the aPriori algorithm:- One of the most common approaches to mining frequent patterns is the apriori method and when a transactional database represented as a set of sequences of transactions performed by one entity is used , the manipulation of temporal sequences requires that some adaptations be made to the apriori algorithm. The most important modification is on the notion of support: support is now the fraction of entities, which had consumed the itemsets in any of their possible transaction, i.e. an entity could only contribute one time to increment the support of each itemset, beside it could had consumed that itemset several times. After identifying the large itemsets, the itemsets with support greater than the minimum support allowed, they are translated to an integer, and each sequence is transformed in a new sequence, whose elements are the large itemsets of the previous-one. The next step is to find the large sequences. For achieve this, the algorithm acts iteratively as apriori: first it generates the candidate sequences and then it chooses the large sequences from the candidate ones, until there are no candidates.One of the most costly operations in apriori-based approaches is the candidate generation [11].

A proposal to frequent pattern mining states that it is possible to find frequent patterns avoiding the candidate generation-test. Extending this to deal with sequential data is presented in . The discovery of relevant association rules is one of the most important methods used to perform data mining on transactional databases. An effective algorithm to discover association rules is the apriori . Adapting this method to deal with temporal information leads to some different approaches. Common subsequences can be used to derive association rules with predictive value, as is done, for instance, in the analysis of discretized, multi-dimensional time series .

A possible approach consists on extending the notion of a typical rule X → Y (which states if X occurs then Y occurs) to be a rule with a new meaning: X →T Y (which states: if X occurs then Y will occur within time T). Stating a rule in this new form, allows for controlling the impact of the occurrence of an event to the other event occurrence, within a specific time interval. Another method consists on considering cyclic rules. A cyclic rule is one that occurs at regular time intervals, i.e. transactions that support specific rules occur periodically, for example at every first Monday of a month. In order to discover these rules, it is necessary to search for them in a restrict portion of time, since they may occur repeatedly at specific time instants but on a little portion of the global time considered. A method to discover such rules is applying an algorithm similar to the apriori, and after having the set of traditional rules, detects the cycles behind the rules. A more efficient approach to discover cyclic rules consists on inverting the process: first discover the cyclic large itemsets and then generate the rules [13].

A natural extension to this method consists in allowing the existence of different time units, such as days, weeks or months, and is achieved by defining calendar algebra to define and manipulate groups of time intervals. Rules discovered are designated calendricassociation rules. A different approach to the discovery of relations in multivariate time sequences is based on the definition of N-dimensional transaction databases. Transactions in these databases are obtained by discretizing, if necessary, continuous attributes .This type of databases can then be mined to obtain association rules. However, new definitions for association rules, support and confidence are necessary. The great difference is the notion of address, which locates each event in a multi-dimensional space and allows for expressing the confidence and support level in a new way.

EXISTING SYSTEM

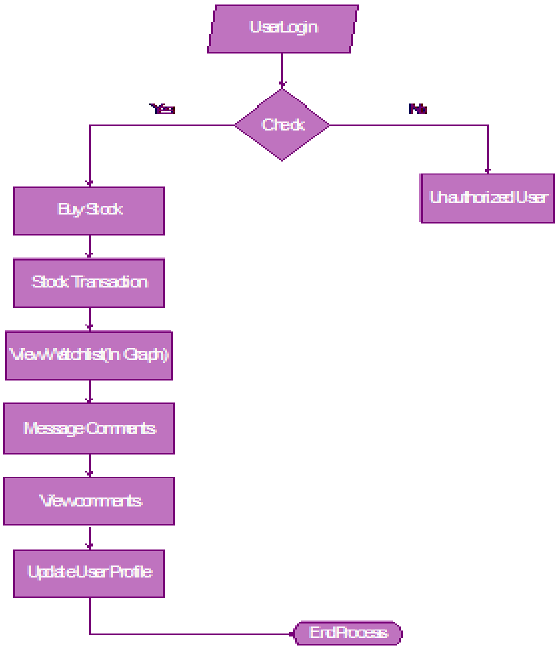
The rationale behind mining frequent itemsets is that only itemsets with high frequency are of interest to users. However, the practical usefulness of frequent itemsets is limited by the significance of the discovered itemsets. A frequent itemset only reflects the statistical correlation between items, and it does not reflect the semantic significance of the items. In this paper, we propose a utility based itemsetmining approach to overcome this limitation.

PROPOSED SYSTEM

The rationale behind mining frequent itemsets is that only itemsets with high frequency are of interest to users. However, the practical usefulness of frequent itemsets is limited by the significance of the discovered item sets. A frequent itemset only reflects the statistical correlation between items, and it does not reflect the semantic significance of the items. In this paper, we propose a utility based item set mining approach to overcome this limitation. The proposed approach permits users to quantify their preferences concerning the usefulness of item sets using utility values. The usefulness of an itemset is characterized as a utility constraint. That is, an item set is interesting to the user only if it satisfies a given utility constraint. We show that the pruning strategies used in previous item set mining approaches cannot be applied to utility constraints [12]. In response, we identify several mathematical properties of utility constraints. Then, two novel pruning strategies are designed. Two algorithms for utility based item set mining are developed by incorporating these pruning strategies. The algorithms are evaluated by applying them to synthetic and real world databases. Experimental results show that the proposed algorithms are effective on the databases tested.

IMPLEMENTATION

Process flow diagram:-



Stock Trading:-The main function of a stock market is the dealings of stock between investors. Stocks are grouped into industry groups according to their primary business focus. A transaction is willing of an investor to sell some stock is not only characterized by its price but also buy many others variables. There is an interaction among all these variables and only s deep study could show the behaviour of a stock over time.

Temporal Data Mining:-Temporal data mining is a research field of growing interest in which techniques and algorithms are applied on data collected over time. The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequence's of events. The discovery of relations between sequences of events involves mainly three steps: the representation and modelling of the data sequence in a suitable form; the definition of similarity measures between sequences; and the application of models and representations to the actual mining problems. Other authors have used a different approach to classify data mining problems and algorithms.

Temporal data mining Is a single step in the process of Knowledge Discovery in Temporal Databases that enumerates structures (temporal patterns or models) over the temporal data. Examples of temporal data mining tasks are classification and clustering of time series, discovery of temporal patterns or trends in the data, associations of events over time, similarity-based time series retrieval, time series indexing and segmentation. In the stock market domain, temporal data mining could indeed play an essential role.

Time Series:

A Time Series is an ordered sequence of data points. Typically it's measured at successive times spaced at uniform time intervals. A huge amount of data is collected everyday in the form of event time sequences. Common examples are recording of different values of stock shares during a day, each access to a computer by an external network, bank transactions, or events related to malfunctions in an industrial plant. These sequences

represent valuable sources of information not only to search for a particular value or event at a specific time, but also to analyze the frequency of certain events, discover their regularity, or discover set of events related by particular temporal relationships. These types of analyses can be very useful for deriving implicit information from the raw data, and for predicting the future behaviour of the process that we are monitoring.
Principle Of Apriori Algorithm [7]:

One of the most common approaches to mining frequent patterns is the apriori method and when a transactional database represented as a set of sequences of transactions performed by one entity is used, the manipulation of temporal sequences requires that some adaptations be made to the apriori algorithm. The most important modification is on the notion of support: support is now the fraction of entities, which had consumed the itemsets in any of their possible transaction, i.e. an entity could only contribute one time to increment the support of each item set, beside it could had consumed that item set several times. After identifying the large itemsets, the itemsets with support greater than the minimum support allowed, they are translated to an integer, and each sequence is transformed in a new sequence, whose elements are the large itemsets of the previous-one. The next step is to find the large sequences. For achieve this, the algorithm acts iteratively as apriori: first it generates the candidate sequences and then it chooses the large sequences from the candidate ones, until there are no candidates. One of the most costly operations in apriori-based approaches is the candidate generation.

CONCLUSION

Frequent itemset mining is based on the rationale that the itemsets which appear more frequently in the transaction databases are of more importance to the user .However the practical usefulness of mining the frequent itemset by considering only the frequency of appearance of the itemsets is challenged in many application domains such as retail research.

It has been that in many real applications that the itemsets that contribute the most in terms of some user defined utility function (for e.g. profit) are not necessarily frequent itemsets. Utility mining attempts to bridge this gap by using item utilities as an indicative measurement of the importance of that item in the user's perspective. Utility mining is a comparatively new area of research and most of the literature work is focused towards reducing the search space while searching for the high utility itemsets.

In this paper we have presented a brief review of the various approaches and algorithms for mining of high utility itemsets. In the next paper we will present a deeper insight into the different pruning techniques used to detect and prune unnecessary candidate itemsets early in the search for high utility itemsets.

REFERENCES

[1] R. Agrawal , T. Imielinski, A. Swami, 1993, mining association rules between sets of items in large databases, in: proceedings of the ACM SIGMOD International Conference on Management of data, pp. 207-216

[2] R. Agrawal, R Srikant, Fast algorithms for mining association rules,in : Proceedings of 20th international Conference on Very Large Databases ,Santiago, Chile, 1994, pp.487-499

[3] K. Ali , S.Manganaris, R.Srikant , Partial classification using association rules, in:Proceeedings of the 3rd International Conference on Knowledge Discovery and Data Mining , Newport Beach, California, 1997, pp. 115-118

[4] C.F.Ahmed , S.K.Tanbeer, Jeong Byeong-Soo, Lee Young-Koo, Efficient tree structures for high utility pattern mining in incremental databases, in: IEEE Transactions on Knowledge and Data Engineering 21(12) (2009)

[5] R.J.Bayardo, Efficiently mining long patterns from databases, in:Prodeedings of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, 1998, pp.85-93.

[6] R.J.Bayardo, R.Agarwal ,D.Gunopulos, Constraint based rule mining in large databases , in:Proceedings of the 15th International Conference on Data Engineering, Sydney, Australia, 1999,pp.188- 197.

[7] B.Barber, H.J Hamilton , Extracting share frequency itemsets with infrequent subsets, Data Mining and Knowledge Discovery 7(2) (2003)153-185.

[8] C.H. Cai , A.W.C Fu, C.H.Cheng , w.W. Kwong, Mining association rules with weighted items,in:Proceedings of IEEE International Database Engineering and Applications Symposium, Cardiff, United kingdom, 1998, pp.68-77

[9] Chan , Q.Yang,Y.D Shen, Mining high utility itemsets, in:Proceedings of the 3rd IEEE International Conference on Data Mining , Melbourne , Florida, 2003, pp.19-26

[10] G.Dong , J.Li, Efficient mining of emerging patterns :discovering trends and differences, in:Proceedings of the 5th international Conference on Knowledge Discovery and Data Mining ,San Diego, 199, pp.43-52

[11] A.Erwin, R.P.Gopalan,N.R.Achuthan, Efficient mining of high utility itemsets from large datasets, in: Advances in Knowledge Discovery , Springer Lecture Notes in Computer Science , volume 5012/2008, pp. 554-561

[12] J Han, J.Pei, Y.Yin ,R. Mao Mining frequent Patterns without candidate generation:a frequent -pattern tree approach , Data Mining and Knowledge Discovery 8(1)(2004) 53-87

[13] J.Hu, A. Mojsilovic , High-utility pattern mining :A method for discovery of high-utility ietmsets,in :Pattern Recognition 40(2007) 3317-3324