International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 7, pp 54-58, July-2019

Potential Mining of High-Utility Itemsets Using an Effective Algorithm

Naresh Katkuri

Associate Professor, Department of Computer Science and Engineering, Malla Reddy College of Engineering for Women, Hyderabad -500100, Telangana, India

Abstract - High-utility itemset mining (HUIM) or more generally utility mining. To give an overview explains why it is interesting, and provide source code of Java implementations of the state-of-the-art algorithms for this problem, and datasets. To address these limitations, the problem of frequent itemset mining has been redefined as the problem of high-utility itemset mining. In this problem, a transaction database contains transactions where purchase quantities are taken into account as well as the unit profit of each item. high-utility itemset mining is to find the itemsets (group of items) that generate a high profit in a database when they are sold together. It is considered to be a high-utility itemset. In general, the utility of an itemset in a transaction is the quantity of each item from the itemset multiplied by their unit profit.

Keywords: high- utility mining, FUP, Apriori mining, pattern mining.

INTRODUCTION

Data Mining

Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions. and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection; Data mining techniques are useful in many research projects, including mathematics, cybernetics, genetics and marketing.

Specific data mining benefits vary depending on the goal and the industry. Sales and marketing departments can mine customer data to improve lead conversion rates or to create one-to-one marketing campaigns. Data mining information on historical sales patterns and customer behaviors can be used to build prediction models for future sales, new products and services.

Companies in the financial industry use data mining tools to build risk models and detect fraud. The manufacturing industry uses data mining tools to improve product safety, identify quality issues, manage the supply chain and improve operations.

Frequent Itemset

Frequent Itemset searches for frequent items in the data-set. In frequent mining usually the interesting associations and correlations between item sets in transactional and relational databases are found. In short, Frequent Mining shows which items appear together in a transaction or relation.

Need of Frequent Itemset Mining:

Frequent mining is generation of association rules from a Transactional Dataset. If there are 2 items X and Y purchased frequently then its good to put them together in stores or provide some discount offer on one item on purchase of other item. This can really increase the sales.

For example it is likely to find that if a customer buys Milk and bread he/she also buys Butter. So the association rule is [‘milk]^[‘bread’]=>[‘butter’]. So seller can suggest the customer to buy butter if he/she buys Milk and Bread.

Temporal Data Mining

Temporal data mining refers to the extraction of implicit, non-trivial, and potentially useful abstract information from large collections of temporal data. Temporal data are sequences of a primary data type, most commonly numerical or categorical values

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 7, pp 54-58, July-2019

and sometimes multivariate or composite information. Examples of temporal data are regular time series (e.g., stock ticks, EEG), event sequences (e.g., sensor readings, packet traces, medical records, weblog data), and temporal databases (e.g., relations with timestamped tuples, databases with versioning). The common factor of all these sequence types is the total ordering of their elements. They differ on the type of primary information, the regularity of the elements in the sequence, and on whether there is explicit temporal information associated to each element (e.g., timestamps). There are several mining tasks that can be applied on...

Temporal Data Mining (TDM) is an active and rapidly evolving area in Big Data science. In 2006, Laxman and Unnikrishnan firstly gave a complete survey on TDM theories and developed new algorithms for discovering frequency episodes in the event stream. Their new algorithms are, both space-wise and time-wise, significantly more efficient than the earlier algorithms reported. These new TDM techniques soon found their applications in the real world, such as neuronal network studies and the automobile industry.

RELATED WORK

Fast Update Algorithm: FUP (fast update) [10] is the first algorithm of incremental association rule mining. Cheung et al thus proposed the FUP algorithm (Cheung et al.1996) to effectively handle new transactions for maintaining association rules.Update With Early Pruning (UWEP): Occurrence of potentially huge set of candidate itemset and multiple scans of the database is the issue. This can significantly reduce the number of candidate itemsets, with the trade-off that an additional set of unchecked itemsets has to be maintained.

Apriori algorithm:- Apriori was proposed by Agrawal and Srikant in 1994. bottom search algorithm moving upward level-wise in the lattice. This is when a transactional database represented it. Name of the algorithm is Apriori because Algorithm uses prior knowledge itemset properties. To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.

If an itemset is infrequent then it supersets will be infrequent. Apriori Algorithm can be slow. The main limitation is time required to hold a vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets i.e. it is not an efficient approach for large number of datasets[11].

Forexample:

1 — Milk | 2 — Bread | 3 — Butter | 4 — Ice cream | 5 — Coffee | 6 — Tea

Sets before and after pre-processing would look like the following : {Milk, Bread, Butter} → {1, 2, 3}

{Butter, Ice cream} →{3, 4}

{Coffee, Tea, Butter, Bread} →{5, 6, 3, 2}

{Milk, Bread, Tea, Ice cream} →{1, 2, 6, 4}

The reason behind assigning these numbers, is to speed up the process of discovering subsets. In DataBase Management System (DBMS), this kind of approach is known as normalization.

The implementation of this method works to allow the existence of different time units days, months, years and is gathered by working on calendar algebra in manipulating groups with frequent intermission. Calendric approach which works based on the definition of N-dimensional transaction databases to multivariate time sequences association rules is designated with different. Transactions from these databases are performed by discretizing using continuous attributes .this is to be mined to obtain association rules. However, new definitions for association rules, support and confidence are necessary, which locates each event in a multi- dimensional space.

EXISTING SYSTEM

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 7, pp 54-58, July-2019

Itemsets that meet a minimum support threshold are referred to as frequent itemsets. The rationale behind the use of support is that a retail organization is only interested in those itemsets that occur frequently. However, the support of an itemset tells only the number of transactions in which the itemset was purchased. The exact number of items purchased is not analyzed and the precise impact of the purchase of an itemset cannot be measured in terms of stock, cost or profit. This shortcoming of the support measure prompted the development of a measure called itemset share, the fraction of some numerical value, such as total quantity of items sold or total profit, that is contributed by the items when they occur in an itemset (Carter et al., 1997). Overall, share-based measures provide more information whenever items are purchased in multiples.

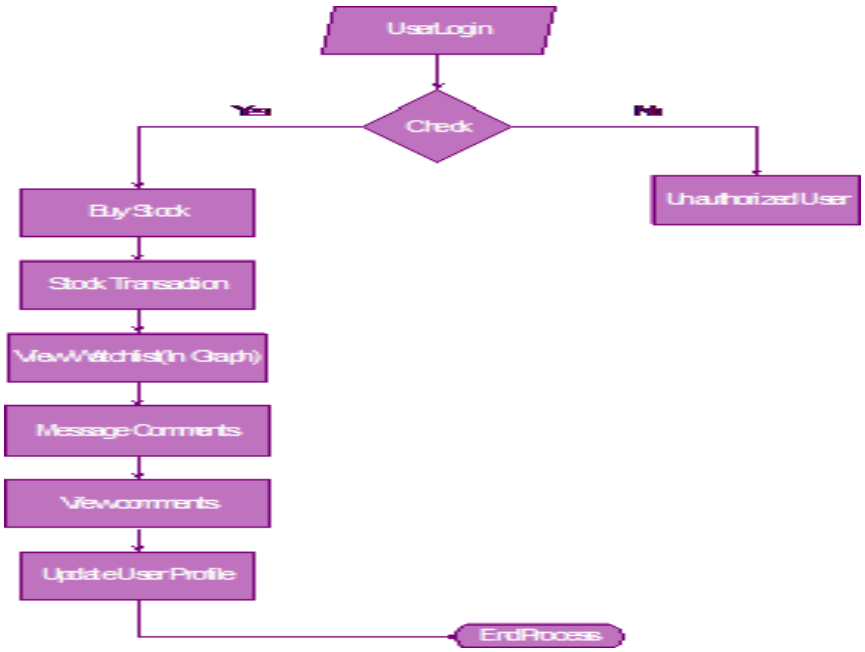
PROPOSED SYSTEM

In this paper we propose rationale behind mining frequent itemsets since the users prior for itemsets with high frequency. In practical point the usage of itemset mining is limited comparatively with the proposed system.In this paper, to overcome this limitation we propose a utility based item set mining approach. FUP algorithm works only for incremented database not for decremented database.

To overcome the problem of FUP a new extension FUP2 has been developed. The algorithms are evaluated by applying real world databases and synthetic databases. Experimental results show that the proposed algorithms are effective on the databases tested.

IMPLEMENTATION

Process flow diagram:



Stock Trading:- The best and most agreeable aspect of the new business is that one can become rich without risk. Indeed, without endangering your capital, and without having anything to do with correspondence, advances of money, warehouses, postage, cashiers, suspensions of payment, and other unforeseen incidents, you have the prospect of gaining wealth if, in the case of bad luck in your transactions, you will only change your name. The stock market — the daytime adventure serial of the well to do would not be the stock market if it did not have its ups and downs and it has many other distinctive characteristics.

Temporal Data Mining:- Time Series: a sequence of the occurrence of some data pattern in time Temporal Pattern: the structure of the time data over a period of time series, perhaps represented as a vector in a Q-dimensional metric space, Temporal Pattern Cluster used to characterize and/or predict events the set of all vectors within some specified similarity distance of a Phase Space: a state space of metrics that describe the temporal pattern Event temporal pattern (e.g., Fourier space, wavelets, …) Characterization Function: connects events to temporal patterns; characterizes the event in phase space.

ISSN (online): 2581-3048
Volume 3, Issue 7, pp 54-58, July-2019

The theory of a model may be formulated in a logical formalism able to express quantitative knowledge and approximate truth. In addition, temporal data mining needs to include an investigation of tightly related issues such as temporal data warehousing, temporal OLAP, computing temporal measurements, and so on.

Time Series:

Time series is a record of the values of any fluctuating quantity measured at different points of time. One characteristic feature which distinguishes time series data from other types of data is that, in general, the values of the series at different time instants will be correlated. Application of time series analysis techniques in temporal data mining is often called Time Series Data Mining. Future behavior of the process that we are monitoring are used and finding the right way to view it to discover useful temporal patterns.

Principle Of Apriori Algorithm [7]:

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori. This algorithm uses two steps "join" and "prune" to reduce the search space. It is an iterative approach to discover the most frequent itemsets. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. Join and Prune steps are easy to implement on large itemsets in large databases. One of the most costly operations in apriori-based approaches is the candidate generation.

CONCLUSION

In this paper, we analyzed and studied various existing improved apriori algorithm to mine frequent itemsets. Mainly common drawbacks are found in various existing apriori algorithm which is improved by using different approaches. It can be applied to many different applications like market basket analysis, telecommunication, network analysis, banking services and many others. In the future work, the problem of large number of candidate sets generated can still be improved and constraints can be applied.

This model presented a brief review of a deeper insight into the different pruning techniques, the various approaches and algorithms for mining of high utility itemsets. In the next we 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] 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.

ISSN (online): 2581-3048
Volume 3, Issue 7, pp 54-58, July-2019

[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.

****

