International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 7, Issue 03 (March 2020) www.ijirae.com
Special Issue – 6th International National Level Conference- "MEEMIC-2020"

COMPETENT TAKING OUT OF COMMON PATTERNS ON HESITANT GRAPH

Prof. Suvathi T.,[1],Gowri Shankar R[2], Madhumidha D [3], Baala Vignesh G[4],
Assistant Professor [1], UG Scholar [2][3][4]
Department of Information Technology,
Sengunthar College of Engineering, Tiruchengode, Tamil Nadu, India
[1]suvathi.007@gmail.com, [2]gowrishankar98@gmail.com,[3]madhulettygirl93@gmail.com,
[4]baalagunasekaran@gmail.com

Manuscript History
Number: **IJIRAE/RS/Vol.07/Issue03/Special Issue/34.MRITSCE10113**
Received: 15, February 2020
Final Correction: 27, February 2020
Final Accepted: 10, March 2020
Published: **14, March 2020**
Editor: Dr.A.Arul Lawrence selvakumar, Chief Editor, IJIRAE, AM Publications, India

Abstract - Frequent itemset mining is a widely exploratory technique that focuses on discovering recurrent correlations among data. The steadfast evolution of markets and business environments prompts the need of data mining algorithms to discover significant correlation changes in order to reactively suit product and service provision to customer needs. Change mining, in the context of frequent itemsets, focuses on detecting and reporting significant changes in the set of mined itemsets from one time period to another. The discovery of frequent generalized itemsets, i.e., itemsets that 1) frequently occur in the source data, and 2) provide a high-level abstraction of the mined knowledge, issues new challenges in the analysis of itemsets that become rare, and thus are no longer extracted, from a certain point. This project proposes a novel kind of dynamic pattern, namely the an Incremental FP- Growth Frequent Pattern Analysis, that represents the evolution of an itemset in consecutive time periods, by reporting the information about its frequent generalizations characterized by minimal redundancy (i.e., minimum level of abstraction) in case it becomes infrequent in a certain time period. To address Frequent Pattern Growth mining, it proposes Frequent Pattern Growth, an algorithm that focuses on avoiding itemset mining followed by post processing by exploiting a support-driven itemset generalization approach. To focus the attention on the minimally redundant frequent generalizations and thus reduce the amount of the generated patterns, the discovery of a smart subset, namely the, is addressed as well in this work.
Keywords: Frequent itemset mining, Incremental FP-Growth Frequent Pattern Analysis.

INTRODUCTION

KNOWLEDGE DISCOVERY IN DATABASES

Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing. Traditionally, data mining and knowledge discovery was performed manually. As time passed, the amount of data in many systems grew to larger than terabyte size, and could no longer be maintained manually. Moreover, for the successful existence of any business, discovering underlying patterns in data is considered essential. The KDD process has reached its peak in the last 10 years. It now houses many different approaches to discovery, which includes inductive learning, Bayesian statistics, semantic query optimization, knowledge acquisition for expert systems and information theory. Thultimate goal is to extract high-level knowledge from low-level data.

KDD includes multidisciplinary activities. This encompasses data storage and access, scaling algorithms to massive data sets and interpreting results. The data cleansing and data access process included in data warehousing facilitate the KDD process. Artificial intelligence also supports KDD by discovering empirical laws from experimentation and observations. The patterns recognized in the data must be valid on new data, and possess some degree of certainty.

CLASSIFICATION OF DISCOVERED KNOWLEDGE FREQUENT PATTERN

The problem of frequent pattern mining has been widely studied in the literature because of its numerous applications to a variety of data mining problems such as clustering and classification. In addition, frequent pattern mining also has numerous applications in diverse domains such as spatiotemporal data, software bug detection, and biological data. The algorithmic aspects of frequent pattern mining have been explored very widely

Applications: Frequent pattern mining have numerous applications to other major data mining problems, Web applications, software bug analysis, and chemical and biological applications. A significant amount of has been devoted to applications because these are particularly important in the context of frequent pattern mining.

Frequent Pattern Mining In Data Streams

In recent years, data stream have become very popular because of the advances in hardware and software technology that can collect and transmit data continuously over time. In such cases, the major constraint on data mining algorithms is to execute the algorithms in a single pass. This can be significantly challenging because frequent and sequential pattern mining methods are generally designed as level-wise methods. Frequent itemsets: In this case, it is not assumed that the number of distinct items is too large. Therefore, the main challenge in this case is computational, because the typical frequent pattern mining methods are multi-pass methods. Multiple passes are clearly not possible in the context of data streams.

Frequent Pattern Mining with Advanced Data Types

Although the frequent pattern mining problem is naturally defined on sets, it can be extended to various advanced data types. The most natural extension of frequent pattern mining algorithms is to the case of temporal data. This was one of the earliest proposed extensions and is referred to as sequential pattern mining. Subsequently, the problem has been generalized to other advanced data types, such as spatiotemporal data, graphs, and uncertain data. Many of the developed algorithms are basic variations of the frequent pattern mining problem. In general, the basic frequent pattern mining algorithms need to be modified carefully to address the variations required by the advanced data types.

OVERVIEW OF UTILITY PATTERN MINING

The older methods of ARM consider the utility of the items by its presence in the transaction set. The frequency of itemset is not sufficient to reflect the actual utility of an itemset. Recently, 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. The utility can be measured in terms of cost, quantity, profit or other expressions of user preferences. For example, a computer system may be more profitable than a telephone in terms of profit. Utility mining model was proposed to define the utility of itemset. The utility is a measure of how useful or profitable an itemset X is. The utility of an itemset X, i.e., u(X), which is the sum of the all utilities of itemset X in all the transactions containing X. An itemset X is called a high utility itemset if and only if u(X) greater than or equal to min_utility, where min_utility is a user defined minimum utility threshold.

High Utility Mining, Categorize of Utility Mining Problems

The objective of utility mining is to discover the itemsets with highest utilities by considering user preferences. In utility mining, the utility of an itemset u(i) is defined as the sum of the utilities of itemset i in all the transactions containing i. An itemset i is called a high utility itemset if and only if u(i)>= min_utility, where min_utility is a user defined minimum utility threshold. Thus the focus of high utility itemset mining is to find all those itemset having utility greater or equal to user defined minimum utility threshold. Measuring the interestingness of discovered patterns is important and the various criteria such as conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, action and utility are used to determine whether or not the pattern is interesting.

Categories of utility mining

Interestingness measures for mining high utility patterns are classified as objective measures, subjective measures and semantic based measures. An objective measure is based only on the raw data. No knowledge about the user or application is required. Most objective measures are based on theories in probability, statistics, or information theory. Conciseness, generality, reliability, peculiarity, and diversity depend only on the data and patterns, and thus can be considered objective. Objective measures such as support or confidence are based only on data. A subjective measure takes into account both the data and the user of these data. To define a subjective measure, access to the user's domain or background knowledge about the data is required.

This access can be obtained by interacting with the user during the data mining process or by explicitly representing the user's knowledge or expectations. Novelty and surprisingness depend on the user of the patterns, as well as the data and patterns themselves, and hence can be considered subjective. Subjective measures such as unexpectedness or novelty take into account the user's domain knowledge. A semantic measure considers the semantics and explanations of the patterns. Since semantic measures involve domain knowledge from the user. Semantic measures also known as utilities consider the data as well as the user's expectation. Utility-based measures, where the relevant semantics are the utilities of the patterns in the domain, are the most common type of semantic measure. To use a utility-based approach, the user must specify additional knowledge about the domain. Unlike subjective measures, where the domain knowledge is about the data itself and is usually represented in a format similar to that of the discovered pattern, the domain knowledge required for semantic measures does not relate to the user's knowledge or expectations concerning the data. Instead, it represents a utility function that reflects the user's goals and this function should be optimized in the mined results.

High utility itemset mining

High Utility Itemset Mining (HUIM) algorithms take care of profits and quantities of the items when transactions take place. Utility based measures use the utilities of the patterns to reflect the user's goals. The objective of High Utility Item set Mining is to identify the item sets that have utility values above a given utility threshold. The various algorithms and techniques for high utility item set mining for pattern prediction are reviewed. Measuring the curiosity of discovered patterns is an active and important area of data mining research. Although much work has been conducted in this area, so far there is no widespread agreement on a formal definition of interestingness in this context. The utility of an item refers to the user's interest for the item. The multiple of item sets external and internal utilities define its utility. External utility is the profit and internal utility is the quantity. For the given utility threshold, if the determined item set utility is greater than the given utility threshold, then it is called as Promising item set, else it is called as Unpromising item set.

CHALLENGES IN UTILITY PATTERN MINING

An item set is called a high utility item set if its utility is no less than a user specified threshold; otherwise, the item set is called a low utility item set. Hence, it cannot satisfy the requirement of the user who is interested in discovering the item sets with high sales profits. Mining high utility item sets from databases is an important task which is essential to a wide range of applications such as website click streaming analysis, cross-marketing in retail stores, business promotion in chain hypermarkets and even biomedical applications. However there may have some challenges

- Even though it produce a lossless representation, it may be meaningless to the users.
- The number of extracted pattern may be high and thus produce problem of large search space.
- They may be slower than the best algorithms for mining high utility item sets.

APPLICATIONS OF HIGH UTILITY PATTERN MINING

An emerging topic in the field of data mining is Utility Mining. The main objective of Utility Mining is to identify the itemsets with highest utilities, by considering profit, quantity, cost or other user preferences. Mining High Utility itemsets from a transaction database is to find itemsets that have utility above a user-specified threshold. Itemset Utility Mining is an extension of Frequent Itemset mining, which discovers itemsets that occur frequently. In many real-life applications, high-utility itemsets consist of rare items. Rare itemsets provide useful information in different decision-making domains such as business transactions, medical, security, fraudulent transactions, retail communities. For example, in a supermarket, customers purchase microwave ovens or frying pans rarely as compared to bread, washing powder, soap. But the former transactions yield more profit for the supermarket. Similarly, the high-profit rare itemsets are found to be very useful in many application areas. For example, in medical application, the rare combination of symptoms can provide useful insights for doctors. A retail business may be interested in identifying its most valuable customers i.e. who contribute a major fraction of overall company profit. Several researches about itemset utility mining were proposed. In this paper, a literature survey of various algorithms for high utility rare itemset mining has been presented.

OBJECTIVES OF THIS WORK

The practical problem of frequent-item set discovery in data-stream environments which may suffer from data overload. The main issues include frequent-pattern mining and data-overload handling. Therefore, a mining algorithm together with two dedicated overload-handling mechanisms is proposed. The algorithm extracts basic information from streaming data and keeps the information in its data structure. The mining task is accomplished when requested by calculating the approximate counts of itemsets and then returning the frequent ones. When there exists data overload, one of the two mechanisms is executed to settle the overload by either improving system throughput or shedding data load finding frequent itemsets over the transactional data stream.

Unlike most of existing algorithms, our method works based on the theory of Approximate Inclusion and Exclusion. Without incrementally maintaining the overall synopsis of the stream, we can approximate the itemsets' counts according to certain kept information and the counts bounding technique. Some additional techniques are designed and integrated into the algorithm for performance improvement.

PROPOSED WORK

In the Proposed work here develop the two new algorithms, collectively called Fp-Growth algorithm, effectively avoids the problem of "best moving product prediction", and when combined with the pruning and validating methods, achieves even better performance. Here also propose a fast validating method to further speed up our Fp-Growth algorithm. The efficiency and effectiveness of Fp-Growth are verified through extensive experiments on both real and synthetic datasets. Fp-Growth adopts the prefix-projection recursion framework of the Prefix Span algorithm in a new algorithmic setting, and effectively avoids the problem of "best moving product prediction". The contributions are summarized as follows: Two general uncertain sequence data models that are abstracted from many real-life applications involving uncertain sequence data: the sequence-level uncertain model, and the element-level uncertain model. Transaction DB and Profit table are input to the system to discover potential highly utilized Item sets.

Create UP-tree: Fp-Growth algorithm is created using discarding unfavorable global items and reducing global node utility. The Fp-Growth algorithm has fields as Node.name which contain name of the item and Parent Node. After calculating transaction utility and transaction weighted utility, the item sets having less utility than predefined minimum threshold utility are disposed. After disposing the unfavorable items the global node utilities is reduced. And nodes are inserted into UP tree using create Fp-Growth algorithm. The local unpromising Item and node utility. Discarding local unpromising items: Construct conditional pattern base of bottom item entry in header table Retrieve the entire path related to that item CPB. Conditional UP tree created by two scans over CPB. Local unfavorable items removed using path utility of each item in CPB paths are organized in descending order. The reorganized path is inserted into conditional utility pattern tree using reduce local node utility strategy. Identify potential high utility item sets and their utilities form Fp-Growth algorithm will eliminate the local unfavorable items and Reduce local node utility.Pruning techniques and a fast validating method are developed to further improve the efficiency of Fp-Growth algorithm, which is verified by extensive experiments.

ADVANTAGES OF PROPOSED SYSTEM

- On single-level projection, since the advantage of bi-level projection may not be significant when the pseudo-projected database is stored in main memory.
- Low in memory usage.
- High in performance and data retrieval latency time.
- It can measure the efficiency of the uncertain stream clustering method.
- The running time of all the algorithms increases almost linear.

System Architecture

This section presents the strategy of implementation. Fig. 1 shows the overall steps undertaken in extracting utility itemsets by Fp-Growth approach. Data Assembly also includes collecting of data, in these types of experiments for testing different types of datasets are collected from different website. The data is challenging due to the number of characteristics, the number of records, and the sparseness of the data (each records contains only small portion of items). In this experiment different dataset e.g. live, distributed, transactional, with different properties are selected to prove the efficiency of algorithm. e.g. Census data, Land registry, Retail, Zoo, Mashroom, pima.D38.N768.C2. The algorithm implies that frequent itemsets are mined through an iterative level-wise approach, based on candidate generation. Mining high utility itemsets with a set of techniques for pruning candidate itemsets. Minimum item utility table is used here to reduce an excessively high estimated utility. In UP-Growth+ algorithm negligible amount of node utilities in each path are used to make the approximate pruning values closer to real utility values of the pruned items in database. IHUP a tree based structure is used to keep the information about item sets and their utilities. Each node of an IHUP-Tree consists of

1. An item name
2. A TWU value
3. A support count.

A table named header table is employed to facilitate the traversal of UP-Tree. Each entry records an item name, an excessively high estimate utility, and a link. The link points to the last occurrence of the node which has the same item as the entry in the UP-Tree. Fig. 2 shows classified and unclassified elements generated by both algorithms.

This stage is concerned with High utility itemsets, that may be classified or unclassified are taken into account which is an output from both Apriori and UP-Growth Algorithm, take it as an input for this Fp-Growth algorithmic strategy, which works on both categories i.e. classified and unclassified itemsets and gives utility pattern extracted from given datasets.

CONCLUSION

Several strategies are proposed to decrease overestimated utility and enhance the performance of utility mining,. The Fp-Growth strategy is used to improve the performance by reducing both the search space and time with number of candidates. A Incremental FP-Growth approach will take the advantage of both algorithms. This system is aimed to reduce the size of normal implementation of any technique that has been used. Also, use of new data structure may recreate the tree by deleting all nodes of non-frequent itemsets after a scanning a specific percentage of database. We have proposed mining method for frequent items using Fp-Growth approach. Same method has been utilized for classification of various datasets with respective features provided by specific domain.

REFERENCES

1. Ahmed C. F., Tanbeer S. K., Jeong B.-S., and Lee Y. -K., "Efficient tree structures for high utility pattern mining in incremental databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 12, pp. 1708– 1721, 2009.
2. Agrawal R., Imielinski T., and Swami A., "Mining association rules between sets of items in large databases," In Special Interest Group on Knowledge Discovery in Data. Association for Computing Machinery, pp. 207–216, 1993.
3. Anusmitha A., Renjana Ramachandran M., "Utility pattern mining: a concise and lossless representation using up growth", International Journal of Advanced in Computer and Communication Engineering, Vol. 4, No. 7, pp. 451–457, 2015.
4. Chun-Wei Lin J., Wensheng Gan., Fournier-Viger P., and Yang L., Liu Q., Frnda J., Sevcik L., Voznak M., " High utility itemset-mining and privacy-preserving utility mining," Vol. 7, No. 11, pp. 74–80, 2016.
5. Dawar S., Goya V. I., "UP - Hist tree: An efficient data structure for mining high utility patterns from transaction databases," In Proceedings of the 19th International Database Engineering & Applications Symposium. Association for Computing Machinery, pp. 56–61, 2015.
6. De Bie T., "Maximum entropy models and subjective interestingness: an application to tiles in binary databases," Data Mining and Knowledge Discovery, Vol. 23, No. 3, pp. 407–446, 2011.
7. Erwin A., Gopalan R. P and. Achuthan N. R., "Efficient mining of high utility itemsets from large datasets," In Proceeding of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 554–561, 2008.
8. Fournier-Viger P., Wu C.-W., Zida S., and Tseng V.S., "Fhm: Faster high-utility itemset mining using estimated utility Co-occurrence pruning," In Proceedings of the 21th International Symposium on Methodologies for Intelligent Systems. Springer, pp.83-92, 2014.
9. Geng L., Hamilton H.J, "Interestingness measures for data mining: A survey," Association for Computing Machinery. Vol. 38, No. 3, pp.1–9, 2006.
10. Han J., Pei J., Yin Y., Mao R., "Mining frequent patterns without candidate generation: a frequent-pattern tree approach," Data Mining Knowledge Discovery in Data.Vol. 8, No. 1, pp. 53–87, 2004.
11. Junqiang Liu., Ke Wang., Benjamin., Fung C.M.,"Mining High Utility Patterns in One Phase without Generating Candidates", IEEE Transactions on Knowledge and Data Engineering, Vol. 28, No. 5, pp.1–14, 2016.
12. Jyothi Pillai., Vyas O.P., "Overview of itemset utility mining and its applications," International Journal of Computer Applications, Vol. 5, No. 11, pp. 9 –13, 2010.
13. Liu J., Wang K., and Fung B., "Direct discovery of high utility itemsets without candidate generation," In Proceedings of the 12th International Conference. IEEE, pp. 984–989, 2012.
14. Liu M., Qu J., "Mining high utility itemsets without Candidate generation," Conference on Information and Knowledge Management. Association for Computing Machinery, pp. 55–64, 2012.
15. Liu J., Pan Y., Wang K., and Han J., "Mining frequent item sets by opportunistic projection," In Special Interest Group on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp.229–238, 2002.
16. Liu V., Liao W., and Choudhary A., "A fast high utility itemsets mining algorithm," in utility – Based Data Mining Workshop in Special Interest Group on Knowledge Discovery in Data. Association for Computing Machinery, pp. 253 – 262, 2005.
17. Li Y.-C., Yeh J.-S., and Chang C.-C., "Isolated items discarding Strategy for discovering high utility itemsets," Data & Knowledge Engineering, Vol. 64, No. 1, pp. 198–217, 2008.

18. Sarode, Nutan, and Devendra Gadekar, " A review on efficient algorithms for mining high utility itemsets, "International Journal of Science and Research, Vol. 3, No. 12, pp.708 –710, 2014.

19. Shankar S., Purusothoman T.P, Jayanthi S., Babu N., "A fast agorithm for mining high utility itemsets" ,In Proceedings of IEEE International Advance Computing Conference (IACC), Patiala, India, pp.1459-1464, 2009.

20. Tan P.N., Kumar V., and Srivastava J., "Selecting the right objective measure for association analysis," Information Systems, Vol. 29, No. 4, pp. 293–313, 2004.

21. Tseng V. S., Shie B.-E., Wu C.-W., and Yu P. S., "Efficient algorithms for mining high utility itemsets from transactional databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 8, pp. 1772–1786, 2013.

22. Yao H., Hamilton H. J., Butz C.J., "A foundational approach to mining itemset utilities from databases," ICDM 2004, pp. 482-486.

23. Yao H., Hamilton H. J., Geng L., "A unified framework for utility-based measures for mining itemsets," in Utility-Based Data Mining," In Special Interest Group on Knowledge Discovery in Data. Association for Computing Machinery, pp. 28–37, 2006.

24. Zaki M.J., "Scalable algorithms for association mining," IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 3, pp. 372–390, 2000.

25. Dr.R.Satish Kumar and Dr.K.Umadevi " Torque Improvement for an Exterior Rotor PermanentMagnet Brushless DC Motor", International journal of Innovative research in Advanced Engineering, Vol. 1, 2014, pp. 1-5 (Impact factor 1.311).

26. Dr.R.Satish Kumar and Dr.K.Umadevi " Novel Technique for Measurements of Dielectric Properties and Microwave Heating of In-Shell Eggs without Explosions in Microwave Oven for Pasteurization", International journal of Innovative research in Advanced Engineering, Vol. 2, , 2015, pp. 69-77 (Impact factor 1.311).

27. Dr.R.Satish Kumar and Dr.M. Y. Sanavullah" Theoretical and experimental study of cooking regions for shell eggs in a domestic Microwave oven", International Conference on Electronics Computer Technology, 2011, https://doi:10.1109/ICECTECH.2011.5941909

28. Raja, G P& Mangai, S 2018, 'Investigation On Optimization, Prioritizing and Weight Allocation Techniques for Load Balancing and Controlling Multimedia Traffic in Wireless Mesh Network', International Journal of Business Information Systems, SCOPUS Indexed Journal (Inderscience) - (P ISSN No: 1746-0972). Published Online: 10th Feb 2020, DOI: 10.1504/IJBIS.2020.105161.IF: 0.72.

29. Raja, G P& Mangai, S 2017, 'Firefly Load Balancing Based Energy Optimized Routing for Multimedia Data Delivery in Wireless Mesh Network', Cluster Computing-The Journal of Networks Software Tools and Applications, SCOPUS Indexed Journal (Springer) - (E ISSN No: 1573-7543).Published Online: 27th Dec 2017, https://doi.org/10.1007/s10586-017-1557-1. IF: 2.040.

30. Geetha. E & Nagarajan. C , 2019, 'Stochastic Rule Control Algorithm Based Enlistment of Induction Motor Parameters Monitoring in IoT Applications', Springer, Wireless Personal Communications. October 2018, Volume 102, Issue 4, pp 3629 - 3645.

