International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 5 Issue 2, February-2016, Impact Factor: 1.544

A Survey of Maximal Frequent Item set mining Algorithm

Mohammed Asad Qureshi[1], Prof. Abhishek Raghuvanshi[2]

[1,2]Department of Information Technology, Mahakal Institute of Technology, Ujjain (MP) India

ABSTRACT

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies and equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining. In this paper, we are presenting a survey of maximal frequent item set mining techniques.

Keywords: Data Mining, Data Warehousing, Association Rules, Frequent Item Set.

1. INTRODUCTION

Frequently, the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart. There is some real benefit if your data is already part of a data warehouse. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. Furthermore, we will have already addressed many of the problems of data consolidation and put in place maintenance procedures. The data mining database may be a logical rather than a physical subset of our data warehouse, provided that the data warehouse DBMS can support the additional resource demands of data mining. If it cannot, then you will be better off with a separate data mining database [4].

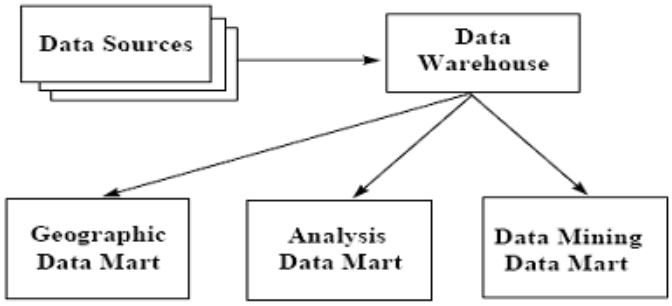

Figure 1: Data Warehouse and its Relations with Other Streams

Types of data mining techniques:

Neural Networks/Pattern Recognition - Neural Networks are used in a black box fashion. One creates a test data set, lets the neural network learn patterns based on known outcomes, and then sets the neural network loose on huge amounts of data. For example, a credit card company has 3,000 records, 100 of which are known fraud records. Neural networks are known for not being very helpful in teaching analysts about the data, just finding patterns that match. Neural networks have been used for optical character recognition to help the Post Office automate the delivery process without having to use humans to read addresses [11].

Cluster Detection/Market Basket Analysis - Association rules identify collections of data attributes that are statistically related in the underlying data. An association rule is of the form X => Y where X and Y are disjoint conjunctions of attribute-value pairs. The confidence of the rule is the conditional probability of Y given X, Pr(Y|X), and the support of the rule is the prior probability of X and Y, Pr(X and Y). Here probability is taken to be the observed frequency in the data set**.**

Visualization - Data volumes have grown to such huge levels; it is going to be impossible for humans to process it by any text-based method effectively, soon. We will probably see an approach to data mining using visualization appear that will be something like Microsoft's Photosynthesis. The technology is there, it will just take an analyst with some vision to sit down and put it together.

Decision Tree/Rule Induction - Based on the loan data a bank has, the outcomes of the loans (default or paid), and limits of acceptable levels of default, the decision tree can set up the guidelines for the lending institution. These decision trees are very similar to the first decision support (or expert) systems.

Classification: The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

The approach affects the explanation capability of the system. Once an effective classifier is developed, it is used in a predictive mode to classify new records into these same predefined classes. For example, a classifier capable of identifying risky loans could be used to aid in the decision of whether to grant a loan to an individual.

Genetic Algorithms - GAs are techniques that act like bacteria growing in a petri dish. You set up a data set then give the GA ability to do different things for whether a direction or outcome is favorable. The GA will move in a direction that will hopefully optimize the final result. GAs are used mostly for process optimization, such as scheduling, workflow, batching, and process re-engineering. Think of GA as simulations run over and over to find optimal results and the infrastructure around being able to both run the simulations and the ways to set up which results are optimal.

2. BACKGROUND AND RELATED WORK

Association Rules and Frequent Item sets

The market-basket problem [21] assumes we have some large number of items, e.g., bread," "milk." Customers their market baskets with some subset of the items, and we get to know what items people buy together, even if we don't know who they are. Marketers use this information to position items, and control the way a typical customer traverses the store. In addition to the marketing application, the same sort of question has the following uses:

1. Baskets = documents; items = words.
   Words appearing frequently together in documents may represent phrases or linked concepts. Can be used for intelligence gathering.
2. Baskets = sentences, items = documents.
   Two documents with many of the same sentences could represent plagiarism or mirror sites on the Web.

Data Mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process [25]. Data mining functions include clustering, classification, prediction and link analysis (associations). One of the most important data mining applications is that of mining association rules. Association rules [23], first introduced in 1993 are used to identify relationships among a set of items in a database. These relationships are not based on inherent properties of the data themselves (as with functional dependencies), but rather based on co-occurrence of the data items. Association rule mining finds interesting association or correlation relationship among a large set of data items. Association rule are used to predict the associatively of two or more things together on the basis of analysis of available facts and figures.

DEFINITION:

Let I ={I1, I2, … , Im} be a set of m distinct attributes, also called literals. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that T⊆I. An association rule [5] is an implication of the form X =>Y, where X, Y ⊆I, are sets of items called item sets, and X ∩ Y=Ø. Here, X is called antecedent, and Y consequent.

Two important measures for association rules support (s) and confidence (α) can be defined as follows.

SUPPORT

The support (s) of an association rule is the ratio (in percent) of the records that contain XUY to the total number of records in the database. Therefore, if we say that the support of a rule is 5% then it means that 5% of the total records contain XUY. Support is the statistical significance of an association rule. Grocery store managers probably would not be concerned about how peanut butter and bread are related if less than 5% of store transactions have this combination of purchases. While a high support is often desirable for association rules, this is not always the case. For example, if we were using association rules to predict the failure of telecommunications switching nodes based on what set of events occur prior to failure, even if these events do not occur very frequently association rules showing this relationship would still be important.

CONFIDENCE

For a given number of records, confidence (α) is the ratio (in percent) of the number of records that contain XUY to the number of records that contain X. Thus, if we say that a rule has a confidence of 85%, it means that 85% of the records containing X also contain Y. The confidence of a rule indicates the degree of correlation in the dataset between X and Y. Confidence is a measure of a rule's strength. Often a large confidence is required for association rules. If a set of events occur a small percentage of the time before a switch failure or if a product is purchased only very rarely with peanut butter, these relationships may not be of much use for management.

Association rule mining is a two-step process:

1) Find all sets of items which occur with a frequency that is greater than or equal to the User-specified threshold support,s.
2) Generate the desired rules using the large item sets, which have user-specified threshold confidence **α**.

3. LITERATURE SURVEY

Mining frequent item sets is an important problem in data mining and is also the first step of deriving association rules [2]. Hence many efficient item set mining algorithms (e.g., Apriori [2] and FP-growth [10]) have been proposed. While all these algorithms work well for databases with precise values but it is not clear how they can be used to mine probabilistic data For uncertain databases the Aggarwal [1] and Chui [9] developed efficient frequent pattern mining algorithms based on the expected support counts of the patterns. However Bernecker et al. [3] Sun[14] and Yiu [16] found that the use of expected support may render important patterns missing. Hence they proposed to compute the probability that a pattern is frequent and introduced the notion of PFI. In work done in [3] the dynamic programming based solutions were developed to retrieve PFIs from attribute uncertain databases. However their algorithms compute exact probabilities and verify that an item set is a PFI in O(n2) time. The proposed model-based algorithms avoid the use of dynamic programming and are able to verify a PFI much faster. In [16] the approximate algorithms for deriving threshold-based PFIs from tuple-uncertain data streams were developed. The Zhang et al. [16] only considered the extraction of singletons (i.e., sets of single items) our solution discovers patterns with more than one item. Recently Sun [14] developed an exact threshold based PFI mining algorithm. However it does not support attribute-uncertain data considered in this paper. In a preliminary version of this paper [15] we examined a model-based approach for mining PFIs. we study how this algorithm can be extended to support the mining of evolving data.

All the other works on the retrieval of frequent patterns from imprecise data includes [4], it studied approximate frequent patterns on noisy data then the [11], it examined association rules on fuzzy sets and [13], proposed the notion of a vague association rule. However none of these solutions are developed on the uncertainty models studied here.

For evolving databases there are a few incremental mining algorithms that work for exact data have been developed. Just For example in [6] the Fast Update algorithm (FUP) was proposed to efficiently maintain frequent item set & for a database to which new tuples are inserted. The proposed incremental mining framework is inspired by FUP. In [7] the FUP2 algorithm was developed to handle both addition and deletion of tuples. The work done by ZIGZAG [1] also examines the efficient maintenance of maximal frequent item sets for databases that are constantly changing. In [8] a data structure called (CATS Tree) was introduced to maintain frequent item sets in evolving databases. Another data structure called CanTree [12] arranges tree nodes in an order that is not affected by changes in item frequency. This data structure is used to support mining on a changing database.

The developments of computed technology in last few decades are used to handle large scale data that includes large transaction financial data, bulletins, emails etc. Hence information has become a power that made possible for user to voice their opinions and interact. As a result revolves around the practice, data mining [17] come into sites. Association rule mining is one of the Data Mining techniques used in distributed database. In distributed database the data may be partitioned into fragments and each fragment is assigned to one site. The issue of privacy arises when the data is distributed among multiple sites and no other party wishes to provide their private data to their sites but their main goal is to know the global result obtained by the mining process. However privacy preserving data mining came into the picture. As the database is distributed, different users can access it without interfering with one another. In distributed environment, database is partitioned into disjoint fragments and each site consists of only one fragment.

Data can be partitioned in three different ways that is, like horizontally partitioned data, vertically partitioned data or mixed partitioned data.

Horizontal partitioning: - The data can be partitioned horizontally where each fragment consists of a subset of the records of relation R. Horizontal partitioning [20] [22] [23] [24] divides a table into several tables. The tables have been partitioned in such a way that query references are done by using least number of tables else excessive UNION queries are used to merge the tables sensibly at query time that can affect the performance.

Vertical partitioning: - The data can be divided into a set of small physical files each having the subset of the original relation, the relation is the database transaction that normally requires the subsets of the attributes.

Mixed partitioning: - The data is first partitioned horizontally and each partitioned fragment is further partitioned into vertical fragments and vice versa.

The market basket analysis used association rule mining [20][21] in distributed environment. Association rule mining [18][19][17] is used to find rules that will predict the occurrence of an item and based on the occurrences of other items in the transaction, search patterns gave association rules where the support will be counted as the fraction of transaction that contains an item X and an item Y and confidence can be measured in a transaction the item i appear in transaction that also contains an item X

Privacy preserving distributed mining of association rule [21][17] for a horizontally partitioned dataset across multiple sites are computed. The basis of this algorithm [21][17] is the apriori algorithm that uses K-1 frequent sets. The problem of generation size of one item set may be carried out with secure computation on multiple sites by generating the candidate set, the pruning method, finding the union of large item set. In [25], the authors conducted a comparative study to analyze the performance of FP-Growth & other frequent item set mining algorithms.

CONCLUSION

Mining frequent item sets is an important problem in data mining and is also the first step of deriving association rules. Hence many efficient item set mining algorithms have been proposed. Maximal frequent item set mining is a very important and popular research topic in the field of data mining. This paper contains a comprehensive survey over the methods available for the maximal frequent item set mining.

REFERENCES

[1] C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent Pattern Mining with Uncertain Data,” Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
[2] R. Agrawal, T. Imieli_nski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 1993.
[3] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle, “Probabilistic Frequent Itemset Mining in Uncertain Databases,” Proc. 15th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
[4] H. Cheng, P. Yu, and J. Han, “Approximate Frequent Itemset Mining in the Presence of Random Noise,” Proc. Soft Computing for Knowledge Discovery and Data Mining, pp. 363-389, 2008.
[5] R. Cheng, D. Kalashnikov, and S. Prabhakar, “Evaluating Probabilistic Queries over Imprecise Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2003.
[6] D. Cheung, J. Han, V. Ng, and C. Wong, “Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique,” Proc. 12th Int’l Conf. Data Eng. (ICDE), 1996.
[7] D. Cheung, S.D. Lee, and B. Kao, “A General Incremental Technique for Maintaining Discovered Association Rules,” Proc. Fifth Int’l Conf. Database Systems for Advanced Applications (DASFAA), 1997.

[8] W. Cheung and O.R. Zaï¨ane, "Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint," Proc. Seventh Int'l Database Eng. and Applications Symp. (IDEAS), 2003.
[9] C.K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), 2007.
[10] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000.
[11] C. Kuok, A. Fu, and M. Wong, "Mining Fuzzy Association Rules in Databases," SIGMOD Record, vol. 27, no. 1, pp. 41-46, 1998.
[12] C.K.-S. Leung, Q.I. Khan, and T. Hoque, "Cantree: A Tree Structure for Efficient Incremental Mining of Frequent Patterns," Proc. IEEE Fifth Int'l Conf. Data Mining (ICDM), 2005.
[13] A. Lu, Y. Ke, J. Cheng, and W. Ng, "Mining Vague Association Rules," Proc. 12th Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2007.
[14] L. Sun, R. Cheng, D.W. Cheung, and J. Cheng, "Mining Uncertain Data with Probabilistic Guarantees," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2010.
[15] L. Wang, R. Cheng, S.D. Lee, and D. Cheung, "Accelerating Probabilistic Frequent Itemset Mining: A Model-Based Approach," Proc. 19th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2010.
[16] Q. Zhang, F. Li, and K. Yi, "Finding Frequent Items in Probabilistic Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008.
[17] Han, J. Kamber, M, "Data Mining Concepts and Techniques". Morgan Kaufmann, San Francisco, 2006.
[18] Agrawal, R., et al "Mining association rules between sets of items in large database". In: Proc. of ACM SIGMOD'93, D.C, ACM Press, Washington, pp.207-216, 1993.
[19] Agarwal, R., Imielinski, T., Swamy, A. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-210, 1993.
[20] Srikant, R., Agrawal, R "Mining generalized association rules", In: VLDB'95, pp.479-488, 1994.
[21] Kantarcioglu, M., Clifton, C, "Privacy-Preserving distributed mining of association rules on horizontally partitioned data", In IEEE Transactions on Knowledge and Data Engineering Journal, IEEE Press, Vol 16(9), pp.1026-1037, 2004.
[22] Sugumar, Jayakumar, R., Rengarajan, C "Design a Secure Multi Site Computation System for Privacy Preserving Data Mining". International Journal of Computer Science and Telecommunications, Vol 3, pp.101-105. 2012.
[23] N V Muthu Lakshmi, Dr. K Sandhya Rani ,"Privacy Preserving Association Rule Mining without Trusted Site for Horizontal Partitioned database", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, pp.17-29, 2012.
[24] N V Muthu lakshmi, Dr. K Sandhya Rani, "Privacy Preserving Association Rule Mining in Horizontally Partitioned Databases Using Cryptography Techniques", International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 3 (1) , PP. 3176 – 3182, 2012.
[25] Pramod S, O P Vyas, "Survey on Frequent Item Sets Mining Algorithms", International Journal of Computer Applications (IJCA), Vol 1(15), PP. 86-91, 2010.