Volume 2, Issue 11, November 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Hash-Set Technique of Association Rules

Rachna Somkunwar*
Computer Department, Nagpur University
Nagpur, India

Abstract— Data mining is the analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful for the owner. The data mining of association rules is an essential research aspect in the data mining fields. The association rules provide an effective means to found the potential link between the data, reflecting a built-in association between the data. Association rules are usually required to satisfy a user-specified minimum support and user-specified confidence at the same time. Association rule generation is usually split up into two separate steps: First, Minimum support is applied to find all frequent itemsets in a database. Second, these frequent itemsets and the minimum confidence constraints are used to form rules. While the second step is straight forward and first needs more attention. Apriori algorithm scans the database too many times. When the database storing a large number of data services, the limited memory capacity, and the system I/O load considerable time scanning the database will be a very long time, so efficiency is very low. In this paper, we introduce a new technique known as Hash-set technique that will help to increase the efficiency of algorithm by avoiding multiple scanning the database. Our Hash Techniques based algorithm is used to reduce the frequently scanning of item set also it helps to mine interesting association rules on the basis of lift and conviction technique.

Keywords— Association rules, Apriori Algorithm, lift and conviction technique, Has-set technique, Data Mining.

I. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyse local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. In order to support this analysis, a sufficient amount of transactions needs to be collected and stored in a database [1]. Each transaction in the database consists of the items purchased in the transaction besides other information like transaction date and time, customer name, quantity, price, and other information. All what was taken in consideration is the set of items bought together in a transaction. Because the amount of these transactions' data can be very large, an efficient algorithm needs to be designed for discovering useful information from these huge transactional datasets. Mining frequent itemsets and association rules is a popular and well researched method for discovering interesting relations between variables in large databases [2]. Association rules, first introduced in 1993 [3], 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, but rather based on co-occurrence of the data items. Association rules are used to discover the relationships, and potential associations, of items or attributes among huge data [4]. These rules can be effective in uncovering unknown relationships, providing results that can be the basis of forecast and decision. They have proven to be very useful tools for an enterprise as it strives to improve its competitiveness and prosperity [4].

II. BASIC CONCEPTS

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Item

Let I= {i1, i2, i3….in} be a set of *n* binary attributes called items.

Database

Let D= (t1, t2…tm) be a set of transactions called the database.

Support

The support Supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

Confidence

The *confidence* of a rule is defined
Conf(X→Y) =Supp(X ∪ Y)/Supp(X)

Lift

The *lift* of a rule is defined as
Lift (X→Y) =Supp(X ∪ Y)/Supp(Y)× Supp(X)

Conviction

Conv**(**X→Y) = (1-Supp(Y))/1- Conf(X→Y)

III. **RELATED WORKS**

A number of association rule algorithm have been developed in the last few years [5, 6, 7, 8, 9, 10], which can be classified into two categories: (a) candidate generation/test approach such as apriori[6]. (b) Pattern growth approach [9,10]. The problem of discovering association rules was first introduced in [5] and an algorithm called AIS was proposed for mining association rules. The main drawback of the AIS Algorithm is that it makes multiple passes over the database. The Apriori algorithm from [3] is based on the Apriori principle, which says that the itemset X' containing itemset X is never large if itemset X is not large. Based on this principle, the Apriori algorithm generates a set of candidate large itemsets whose lengths are (k+1) from the large k itemsets (for k>=1) and eliminates those candidates, which contain not large subset. The DHP (Direct hashing and Pruning) Algorithm is an effective hash based algorithm for the candidate set generation. It reduced the size of candidate set by filtering any k itemset out of the hash entry does not have minimum support. The DHP Algorithm consists of three steps: First step is to get a set of large 1-itemsets and construct a hash table for 2 itemsets. The Second step generates the set of candidate itemsets Ck. The third step is the same as the second step except it does not use the hash table in determining whether to include a particular itemset into the candidate itemset. The Partition Algorithm [12] logically partitions the database D into n partitions, and requires just two database scans to mine large itemsets. The algorithm consists of two phases. In the first phase, the algorithm subdivides the database into n number overlapping partitions which can fit into main memory. The algorithm iterates n times, and during each iteration only one partition is considered. In the second Phase, the algorithm counts actual support and generates the global large itemsets. In 2000, Attila Gyenesei put forward an important method of mining association rules for market analysis [3, 13, 14, 15]. The Traditional association rule mining algorithms can only be used to data mining problems with categorical attribute. For a data mining problem with quantitative attribute, it is necessary to transform each quantitative attribute into discrete intervals. HUANG Liusheng [16] discovering a new algorithm BitMatrix, This algorithm is compared with the previously known algorithms, the Apriori and AprioriTid algorithms. LI Pingxiang [17] proposed a method explores the database to filter frequent 1-itemsets and then it obtains the candidate frequent 2-itemset, 3-itemsets up to n-itemset by evaluating their probabilities in Equation. The main task [3] of every association rule mining algorithm is to discover the sets of items that frequently appear together—the frequent itemsets. The number of database scans required for the task has been reduced from a number equal to the size of the largest itemset in Apriori . By analysing Carlos Ordonez [18], several problems come up when trying to discover association rules in a high dimensional data set. Daniel kunkle [19] focused three possible straightforward solutions. First, mine all frequent generalized itemsets, and then eliminate the non-max ones. Second, mine max frequent itemsets in the ordinary case and third choice is to dynamically browse the lattice of all generalized itemsets.

IV. **PROPOSED WORK**

Analysis of existing algorithm

Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy to counting the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

Based on the frequency of collection of Apriori search algorithm using a layer of the iterative approach is simple and clear, not complex theoretical analysis, but also easy to realize. However, it is having following shortcomings:

Scan the database **too many times**. Apriori algorithm in the scan, each set to generates a further option; the database should conduct a comprehensive search. To generate maximum length for the frequency of collection K, K to the database scans. When the database storing a large number of data services, the limited memory capacity, and the system I/O load considerable time scanning the database will be a very long time, so efficiency is **very low**.

To overcome the above problem, we suggest a new technique known as Hash-set technique that will help to increase the efficiency of algorithm by avoiding multiple scanning the database. Our Hash Techniques based algorithm is used to reduce the frequently scanning of item set also it helps to mine interesting association rules on the basis of lift and conviction technique.

Proposed algorithm

Only 2 Steps I have implemented.

Let,

I={I1,I2,I3- - - - - -In} represents the attributes of database D.

T={T1,T2,T3- - - - - -- Tm} represents the transactions of database D.

CH={CH1,CH2- - - - - - -CHk} represents the list of candidate hashset.

MinS= Minimum support threshold value.

MinC= Minimum threshold value of confidence.

MinL = Minimum threshold value of lift.

Step 1: Creation of Corpus (Candidate hash-set).

Scan the transaction database D to find the each unique item (attribute) of each transaction and store all the unique items in hash set in such a way that each item represents the key of hash set and the value represents the list of all transaction IDs in which the item is available, these collection of all unique items is also known as corpus.

This hash set will represent the candidate hashset-1 (CH1) in which each key of hash-set will contain only 1-item. Each key is known as candidate.

Support of each candidate=size of list of IDs/n

Only step1 is responsible for scanning of database and other step will not scan the database.

Step 2: Removal of unwanted candidate from Candidate Hash-set-1 (base hash-set).

Candidate hash-set (CH1) is the base for all the other steps. Find the support of each candidate of CH1 and remove those candidate from CH1 having support less than the minimum support (MinS).

V. CONCLUSIONS

This paper discusses an enhanced version of Apriori algorithm. The aim of this paper is to improve the performance of the conventional Apriori Algorithm. The approach to attain the desired improvement is to create a more efficient new algorithm by adding new features to the Apriori Algorithm.

REFERENCES

[1] Show-Jane Yen and Arbee L.P. Chen, *A Graph-Based Approach for Discovering Various Types of Association Rules*, IEEE Transaction on Knowledge and Data Engineering, vol. 13, no. 5, September/October 2001, pp. 839 – 845.

[2] Petra Perner (Ed.), *Advances in data mining, Medical Applications, Ecommerce, Marketing, and Theoretical Aspects*, Book. 8th industrial Conference, ICDM 2008, Leipzig, Germany, July 2008, p 192.

[3] JR. Agrawal, R. Srikant,, *Fast algorithm for mining association rules in large databases*, Proceedings of 1994 International Conference on VLDB, 1994 pp. 487–499.

[4] JYuh-Jiuan Tsay, Jiunn-Yann Chiang. CBAR: *An efficient method for mining association rules. Knowledge-Based Systems* ,18 (2005) 99–105.5.

[5] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, *Dynamic Itemset Counting and Implication Rules for Market Basket Data*,, Proceedings of the ACM SIGMOD, Tucson, A AZ,May 1997, pp. 255-264.

[6] J. Han, J. Pei, and Y. Yin, *Miniing Frequent Patterns without Candidate Generation*, Proceedings of th e ACM SIGMOD, Dallas, TX, May 2000, pp. 1-12.

[7] Qin Ding and gnanasekaranSundaraj, A*ssociation rule mining from XML data*,,Proceedings of the co onference on data mining.DMIN'06

[8] VirendrakumarShrivastava, Dr.parveenkumar and DR. K.R.pardasani, *FP-Tree and COFI Based App roach for Mining of Multiple Level Association Rules in Large database* , IJCSIS, International Journal of Computer Science and Information Security,Vol.7 No. 2,2010

[9] Irina Tudor, *Association rule mining as a data mining technique*, ,BULETINULuniversitatii Petrol-Gaze din Ploiesti,Vol.LX No1/2008,page 49-56.
[10] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data,"
[11] C. Silverstein, S. Brin, and R. Motwani, *Beyond Market Baskets: Generalizing Association Rules to Dependence Rules*, Data Mining and Knowledge Discovery, 2(1), 1998, pp 39–68
[12] Qihua Lan,Defu Zhang, Bo Wo ,*A new algorithm for frequent itemset mining based on apriori and FP-tree*, Global Congres on Intelligent System,2009.
[13] Jiawei Han, jian pei, and Yiwen Yin, *Mining frequent patterns without candidate generation*,paper id :196, SIGMOD '2000.
[14] N.Balaji Raja, Dr. G.Balakrishnan, A *Model of Algorithmic Approach to Itemsets Using Association Rules*.
[15] Zhiyong Zeng, Hui Yang, Tao Feng,*Using HMT and HASH_TREE to Optimize Apriori Algorithm*, 2011 International Conference on Business Computing and Global Informatization
[16] Hamid Mohamadlou,*A method for mining association rules in quantitative and fuzzy data*, 978-1-4244-4136-5/09, IEEE,pp.453 –458 Year of Publication 2009..
[17] Agrawal R, Srikant R. *Fast algorithms for mining association rules*, In: Proceedings of the 20th VLDB Int'l Conf., pp 487–499,1994.
[18] Agrawal R, Srikant R, *Fast algorithms for mining association rules in large database*, Technical report Fj9839, IBM Almaden Research Center, San jose, CA, jun.1994.
[19] Agrawal R, Imielinski T, Swami A. *Mining association rules between sets of items in large database,* In Proc,ACM SIGMOD, May 1993,pp.207-216.
[20] XindongWu and et al, *Top 10 algorithms in data mining*, Knowl Inf Syst ,Springer-Verlag London Limited ,pp 14:1–37,2008.
[21] Attila Gyenesei, *A Fuzzy approach for mining quantitative association rules*, Technical Report: TUCS-TR-336 Year of Publication, 2000.
[22] Zhu Ming, datamining, University of Science and Technology, china Press, Hefel, pp: 115 – 126, 2002.
[23] Balaji Raja.N and Balakrishnan.G, *Evaluation of Rule Likeness Measures for a Learning Environment,* In: Proceedings of the ICOREM Int'l Conf., pp: 1682 – 1690, 2009.
[24] Ahmed Riadh BABA-ALI, *A Novel Two Level Evolutionary Approach For Classification Rules Extraction*, IEEE Congress on Evolutionary Computation (CEC), pp 3306 – 3313, 2009.
[25] LI Pingxiang, CHEN Jiangping, BIAN Fuling, *A Developed Algorithm of Apriori Based on Association Analysis*, Geo-spatial Information Science ,Vol. 7, Issue 2, pp 108-112, june 2004.
[26] HUANG Liusheng, CHEN Huaping, WANG Xun, CHEN Guoliang, *A Fast Algorithm for Mining Association Rules*, J. Comput. Sci. & Technol., Vol. 15 No. 6, pp 619-624, Nov. 2000.
[27] Carlos Ordonez, Norberto Ezquerra, Cesar A. Santana, *Constraining and summarizing association rules in medical data*, Knowl Inf Syst, 9(3), pp 259-283, 2006.
[28] Berry M.J.A and Linoff, G.S., *Data Mining Techniques for Marketing Sales and Customer Support*, John Wiley & Sons, Inc., 1997.
[29] Brin, S., Motwani, R., Ullman, J., and Tsur, S., *Dynamic itemset counting and implication rules for market basket data*, In Proc. of the ACM-SIGMOD Int'l Conf. on the Management of Data, PP. 255-264, 1997.
[30] Daniel Kunkle, Donghui Zhang, Gene Cooperman, *Mining Frequent Generalized Itemsets and Generalized Association Rules Without Redundancy*, J. Comput. Sci. & Technol., Vol. 23(1), pp. 77-102, JAN. 2008.

