Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Enhancement of Association Rule Mining Based Apriori Algorithm

Sachin Saxena
Department of Computer Applications, IFTM University, India

Brij Mohan
Department of Computer Science, COER, Roorkee, India

Abstract*— In the past decade, a significant amount of research was devoted to develop to analyze volumes of data mechanically in an efficient and effective way so that the users don't have to look through that massive amount of data manually for generating various association rules among them. Apriori algorithm, which is the most famous and commonly used data mining algorithm. In this manuscript an attempt has been made by proposing an improvement of the performance of the Apriori algorithm in such a way that when we will implement on a large records, it will lead to less time consuming and fast implementation for generating frequent itemsets.

Keywords*— Apriori algorithm, Association rule mining, frequent itemsets, Signature algorithm.

I. INTRODUCTION

Data mining, which is a relatively new field of computer science, is the process by which new patterns are generated in a large data set. The main goal of the data mining process is to extract information or knowledge from an existing data set and transform it into a human-understandable structure for further use [1, 2]. That information can be used in various forms such as cost cutting, increasing revenue for business orientation [1, 2]. The term data mining is somehow new, but the technology has been there for many years. In this information era, we have been collecting tremendous amounts of information because we believe that information, from the technologies such as computers, satellites and mobile Ad Hoc networking etc. leads to power and success. With the invention of data and other storage devices helps us to collect all type of data [2, 3, 5].

Unfortunately, this huge amount of data stored on different structures became overwhelming very rapidly. The initial chaos led to the creation of database management systems (DBMS) and structured database. For the large amount of data we have DBMS which very crucial assets for managing this huge data and mainly for getting effective and efficient retrieval of particular information from it. The generation of database management systems has also contributed to huge accumulation of all sorts of information. Now days, we have a lot of more information, than we can handle, from scientific data and business transactions, to satellite images, text reports and military intelligence. Information retrieval or data mining is not enough for decision-making. With huge combination of data, we have now created new requirements to help us make better managerial choices for business. These needs are automatic summarization of data, extraction of the "essence" stored information, and the patterns discovery in raw data [5, 7, 10, 15]. In This paper, the next section describes a survey of earlier association rules algorithms. Section Three and four discusses includes basics of association rule in mining, define the difficulties in apriori based algorithms. In section five is purposed work and finally paper is concluded in section six.

II. LITERATURE REVIEW

There has been an exponential growth in businesses and interest people in these businesses there has been a dramatic increase in the amount of information from Giga bits (GB's) to Tera bits (TB's).It has been figured out that the amount of information in the world doubles every 12 months and the number and size of databases are increasing even more quickly [4, 5]. The increase in use of electronic data gathering devices such as point-of-sale or remote sensing devices has added to this explosion of available data [2, 5 ,8]. The storage of data became easier and cheaper as the cost of computing power and electronic data storage devices decreased rapidly.

The organizations had concentrated so much attention on the accumulation of data; the problem was now what to do with this valuable resource. It was realized that information is at the core of business operations and that decision-makers could make use of the data stored to gain valuable insight into the business. Traditional on-line transaction processing systems are good at feeding and saving data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Data analysis can provide more knowledge about a business by going beyond the data explicitly stored, to derive important knowledge about the business. This is where Data Mining or Knowledge Discovery in Databases (KDD) has obvious gains for any enterprise. Data Mining, or KDD as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [6, 7, 8]. This includes a number of different technical approaches, such as clustering, data reduction, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies.

III. DESCRIPTION OF ASSOCIATION RULE IN MINING

In This Section, a brief description of association rule in mining, it is a technique for discovering unsuspected data dependencies and is one of the best-known data mining techniques. The basic idea is to identify from a given database, consisting of itemsets (e.g. shopping baskets), whether the occurrence of specific items, implies also the occurrence of other items with a relatively high probability. Association rules are one of the most researched areas of data mining and have recently grabbed much attention from the database community. They have been proved to be quite useful in the marketing and retail communities as well as other more diverse fields

Association rule mining (ARM) is a core technique for data mining which discovers patterns or rules among items from large database of variable-length transactions. The goal of ARM is to identify groups of items that are often occurring together. One of the most important and well researched techniques of data mining, ARM was first introduced in [11, 12, 13]. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc.

IV. ASSOSIATION RULE MINIING BASED ALGORITHMS.

A. APRIORI ALGORITHM

The Apriori-based algorithms find frequent itemsets based upon an iterative bottom-up approach to generate candidate itemsets. An Apriori algorithm generally works on databases containing transactions. The algorithm works until no more frequent item sets are found [15,16]. After frequent itemsets are obtained they are used to generate association rules having confidence larger than or equal to minimum confidence which is specified by user etc.

B. SAMPLING ALGORITHM

It is used to facilitate efficient counting of itemsets with large database. It reduces the number of database scans to one in the best case and two in the worst case. Here first any algorithm like Apriori is used to find the large item sets in the sample. The set of large itemsets is used as to find the large itemsets in the sample. The set of large itemsets is used as a set of candidates during a scan of the entire database. This results in counting of all candidates. During the second scan, additional candidates are generated and counted [2, 11, 16]. This is done to ensure that all large itemsets are found.

C. PARTITIONING ALGORITHM

PARTITION reduces the number of database scans to 2. It divides the database into small partitions such that each partition can be handled in the main memory. Since each partition can fit in the main memory, there will be no additional disk I/O for each partition after loading the partition into the main memory. In the second scan, it uses the property that a large itemsets in the whole database must be locally large in at least one partition of the database. Then the union of the local large itemsets found in each partition is used as the candidates and are counted through the whole database to find all the large itemsets.

D. PARALLEL AND DISTRIBUTED ALGORITHM

The current parallel and distributed algorithms are based on the serial algorithm Apriori. An excellent survey classifies the algorithms by load-balancing strategy, architecture and parallelism. Here we focus on the parallelism used: data parallelism and task parallelism. The two paradigms differ in whether the candidate set is distributed across the processors or not. In the data parallelism paradigm, each node counts the same set of candidates. In the task parallelism paradigm, the candidate set is partitioned and distributed across the processors, and each node counts a different set of candidates. The database, however, may or may not be partitioned in either paradigm theoretically. In practice for more efficient I/O it is usually assumed the database is partitioned and distributed across the processors.

V. DEFICIENCIES IN APRIORI BASED ALGORITHM

The Apriori like algorithms suffer from various deficiencies like too many scans of the transaction database when seeking frequent itemsets (after every iteration, the algorithm scans the whole database to find frequent itemsets), too large amount of candidate itemsets generated unnecessarily (large number of candidate itemsets are generated even though their count is less than minimum count and are then pruned after generation), the redundant generation of identical sub-itemsets and the repeated search for them in the database (item sets like ab and ba are considered to be same but still they are generated), and so on. Though the basic apriori algorithm is designed to work efficiently for large datasets, there exist a number of possible improvements:

- Transactions in the database that turn out to contain no frequent k-itemsets can be omitted in subsequent database scans.
- One can try to identify first frequent itemsets in partitions of the database. This method is based on the assumption that if an item set is not frequent in one of the partitions at least (local frequent item set) then it will also not be frequent in the whole database.
- The sampling method selects samples from the database and searches for frequent Item sets in the sampled database using a correspondingly lower threshold for the support.

VI. PURPOSED WORK

Discovery of frequent occurring subset of items, called itemsets, is the core of many data mining methods. Most of the previous studies adopt Apriori–like algorithms, which iteratively generate candidate itemsets and check their occurrence frequencies in the database. These approaches suffer from serious cost of repeated passes over the analyzed database. To address this problem, we purpose a novel method; called signature method, for reducing database activity of frequent item set discovery algorithms. The idea of signature method consists of using signature table for pruning

candidate itemsets. The proposed method requires fewer scans over the source database: The first scan creates signature, while the subsequent ones verify discovered itemsets.

A set of signature generated for each database item set. The signature of a set X is an N-bit binary number created, by means of bit-wise OR operation from the signature of all data items contained in X.

The signature has the following property. For any two set X and Y, we have $X \subseteq Y$ if:

Signature (X) AND Signature (Y)=Signature (X)

Where AND is the bit-wise AND operator. The property is not reversible in general (when we find that the above formula evaluates to TRUE we still have to verify the result traditionally).

Algorithm:

Scan D to generate signature S and to find L_1;

For (K=2; $L_{K-1} \neq 0$; K++) **do Begin**

C_k = **apriori_gen** (L_{k-1});

Begin

Forall transactions $t \in D$ **do begin**

C_t = subset(C_K ,t);

Forall candidate $c \in C_t$ **do** c.count++;

End

End

L_k = { $c \in C_k$ | c.count $\geq$ minsup };

End

Answer= $\cup_K L_K$;

Scan D to verify the answer;

Apriori_gen()

The apriori_gen() function works in two steps:

1. Join step
2. Prune step

First in the join step, large itemsets from L_{K-1} are joined with other large itemsets from L_{K-1} in the following SQL- like manner:

Insert into C_k

Select p.item1, p.item2,……,p.item$_{K-1}$,q.item$_{K-1}$

From L_{K-1} p , L_{K-1} q

Where p.item1 = q.item$_1$

And p.item2, q.item$_2$

And p.item3, q.item$_3$

.

.

.

And p.item$_{K-1}$, q.item$_{K-1}$

Next, in the prune step, each item set $c \in C_K$ such that some (K-1) – subset of c is not in L_{K-1} is deleted:

Forall itemsets $c \in C_K$ **do**

Forall (K-1)-subsets s of c **do**

If ($s \notin L_{K-1}$) then delete c from C_K;

The set of candidate K-itemsets C_K is then returned as a result of the function apriori_gen().

VII. CONCLUSIONS

Performance study shows Signature method is efficient than the Apriori algorithm for mining in terms of time as it reduces the number of and Database scans, and also the Database pruning is efficient than the Apriori algorithm for mining in terms of time, as it prunes the database in each iteration. The Signature method is roughly two times faster than the Apriori algorithm, as Signature method does not require more than one database scan, and Database pruning method prunes the database in each iteration. So the Database Pruning algorithm is slower in the earlier stages but faster in the later stages.

REFERENCES

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[3] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, Knowledge Discovery in Databases: An Overview. In G. Piatetsky-Shapiro et al. (eds.), Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
[4] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann,2000.
[5] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM,39:58-64,1996.
[6] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
[7] G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
[8] An Effective Hash-Base Algorithm for Mining Association Rules. Jong Soo Park,* Ming Chen and Philip S . Yu, IBM Thomas J. Watson Research Center ,New York 10598
[9] An Hash – Mine algorithm for discovery of frequent itemsets, Marek Wojciechowski , Maciej Zakrzewicz Institute of computer science, ul. Piotrowo 3a Poland.
[10] An Efficient Algorithm for mining Association rules in Large databases , Ashok Savasere,Edward Omiecinski, Shamkant Navathe ,college of computing , Georgia Institute of technology, Atlanta , GA 30332 .
[11] A Fast Apriori implementation, infomatics Laboratory, Computer and Automation Research institute, Hungarian academy of sciences.
[12] Mining Large Itemsets for Association Rules ,Charu C. Aggarwal ,IBM research Lab.
[13] Mining association rules between sets of items in large databases, Rakesh Agarawal,Tomasz Imielinski*,Arun swami,IBM research lab.
[14] Fast algorithm for mining association rules,Rakesh Agarwal Ramakrishna Srikannt*,IBM research labs 650 Harry Road , San Jose , CA 95120.
[15] J Han, Y. Cai , and N,Cercone . Knowledge Discovery in database: An attribute – oriented approach. Proceeding of the 18 th International Conference on very large data bases, Page 547-559, august 1992
[16] Charu C. Aggarwal, and Philip S. Yu, A New Framework for Itemset Generation, Principles of Database Systems (PODS) 1998, Seattle, WA.

