Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Optimization Technique for Removing the Deficiencies of Current Apriori Algorithm

Anupriya, Kapil Ghai
Graphic Era Hill University, Dehradun
Uttarakhand, India

Abstract— In last few decades there is an increase in storing the information in electronic format. This massive amount of data needs analyzing. For this there is a technique, which is designed for mining the databases and is less time consuming. The purpose of this study is on existing Apriori algorithm in which databases are converted into numerical format and these images are further used in coding to create association rules. Threads will help for making pairs and if the itemsets are large enough then in that case a minimum criterion will be set to remove the pairs for next itemset generation. This technique will help to reduce the reading overheads of the database and removal of pairing, results in fast execution for making associating rules.

Keywords—Apriori algorithm; threads; association rules

I. INTRODUCTION

The current technological trends lead to massive amount of data. Most of the data is generated from banking, telecom, business transactions, scientific experiments space explorations, biology, and other on the web, especially in text, image, and other multimedia format.We have been collecting this tremendous amount of information and yet believe that this information helps lead to power and success. This mixture of information initially leads to Databases and to make sense of that data needs data mining. Database management System (DBMS) provides data sets to use for analyzing in Data mining. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security. With huge amalgamation of data only Information retrieval or data mining is not enough for decision-making. We need automatic summarization, extraction and patterns[1-6].

The term data mining is somehow new, but the technology has been there for many years. Data mining 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. That information can be used in various forms such as cost cutting or increasing revenue [7-11].

II. TYPES OD DATA FLOOD

We have been collecting massive amounts of data, from simple numerical measurements and text documents. Following is the list of a variety of information collected in digital form in databases-[12]

For example:-

A. *Scientific data:*

There are huge amounts of scientific data that need to be analysed. We can capture and store more new data at a faster rate than we can analyse the old data which is already accumulated.

For example - Europe's Very Long Baseline Interferometer (VLBI) has 16 telescopes, each of which produces **1 Gigabit/second** of astronomical data over a 25-day observation session. AT&T handles so many calls per day that it cannot store all of the data .The estimated data is 5 Exabyte (5 million terabytes) was created in 2002. Twice information was created in 2002 as in 1999 (~30% growth rate). US produces 40% of new stored data worldwide. As of 2003, according to Winter Corp. Survey, France Telecom has largest decision-support DB, ~30 TB (terabytes); AT&T was in second place with 26 TB database.

Some of the largest databases on the Web, as of 2003, include

- Alexa internet archive 500 TB
- Internet Archive ~ 300 TB
- Google, over 4 Billion pages, many, many TB

B. *Business transactions:*

Each transaction in the business industry is “memorized” for perpetuity. Such transactions consume time and can be intra-business such as exchanges, purchases, banking, stock, etc., or intra- business operations such as management of in-house wares and assets. The efficient utilization of the data in a reasonable amount of time for competitive decision-making is definitely the most crucial problem to solve for businesses that struggle to survive in a highly competitive world.

For example-Verizon Wireless is the largest wireless service provider in the United States with a customer base of 34.6 million subscribers as of 2003 .Verizon built a customer data warehouse that Identified potential attire's, Developed multiple, regional models, Targeted customers with high propensity to accept the offer.

C. Personal and Medical rule:

Right from governmental census to personnel and customer files, huge collection of information is continuously collected about groups and individuals. Companies, governments and organizations are gathering very important quantities of personal data to help them manage human resources, better understand a market.

III. TECHNIQUES AVAILABLE

As the size of data is increasing rapidly from gigabytes to terabytes and now to Exabyte, sequential data mining algorithms may or may not deliver the accurate results in required time. Hence there is an increasing interest in the research for data mining algorithms. There are many challenges associated with data mining algorithms like minimizing I/O, reducing communication and increasing processing speed [7-11].

Among the available data mining algorithms high performance is big issue. So many researches are done for sequential data mining algorithms which include Apriori, Éclat, D-CLUB and FP-growth. Our main focus is on the Apriori based algorithms by implementing it with multiple threads to remove the database scan overhead by using a file because read operations are faster in files. The previous related work done in the area of Apriori algorithm to make our algorithm more efficient and optimum. The pseudo code of our proposed algorithm explained through an example dataset.

IV. DATA MINING TASK

Data mining is about many different types of patterns, and there are correspondingly many types of data mining tasks. Some of the most popular are

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection**: finding changes
- **Estimation**: predicting a continuous value
- **Link Analysis**: finding relationships

Mining association rules is one of the most important data mining applications. Association rules are used to identify patterns among a set of items in a database. These relationships or patterns 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 rules are usually used by retail stores to analyze market basket transactions. The identified association rules can be used by the management to increase the effectiveness and gains (and reduce the cost) associated with advertising, marketing, inventory, and stock location on the floor. Association rules are also helpful in other applications such as prediction of failure in telecommunications networks by identifying what events occur before a failure.

V. APRIORI ALGORITHM

It is one of the most popular data mining algorithms. 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 itemsets are found. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified.

VI. CONTRIBUTION

1) Read the file for creation of numerical image.
   (i) Read the first transaction and fill the corresponding cell of the item with their count in the transaction.
   (ii) Do it for all the transactions.

To show this lets take an example of six transactions as:

fdbe
feb
adb
aefc
ade
acfe

Now when the code runs, the numerical image so formed will be of this sort:

0	1	0	1	1	1
0	1	0	0	1	1
1	1	0	1	0	0
1	0	1	0	1	1
1	0	0	1	1	0
1	0	1	0	1	1

For every occurrence of the item in the transaction there count is mentioned in that translation in the numerical image. The first position always shows the **‘a’** item and second position always shows the ‘**b’** item and so on.

Lets take another example:
fdbef
feb
adbb
aefc
ade
acfe

The numerical image we get is now:

0	1	0	1	1
0	1	0	0	1
1	2	0	1	0
1	0	1	0	1
1	0	0	1	1
1	0	1	0	1

Here 2 is the count of ‘**b’** and ‘**f’** in their respective transactions.

2) And hence for every increase in the number of item their count increases. And this is how we get the numerical image. Find the minimum count.
   (i) Create an array and fill it with the corresponding global counts of the items.
   (ii) Global count calculated with the help of the numerical image.
   (iii) (Only the first time)Through a comparison find the minimum value which is equal to minimum count.

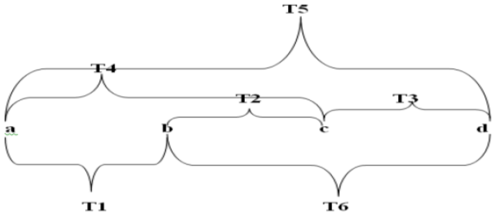
3) Pair generation
   (i) Creation of ${}^{n}_{r}C$ number of threads where **n** is the number of items and r$= 2$.
   (ii) Threads create the pairs and store them in array which then calculates their count and removes the items having count lower than the minimum count.
   (iii) Threads create the pairs and store them in array which then calculates their count and removes the items having count lower than the minimum count.

The combination formula is used here for the creation of threads because we have been taking two items at a time and creating new pairs, thus total number of combinations can be given by the above mentioned formula.

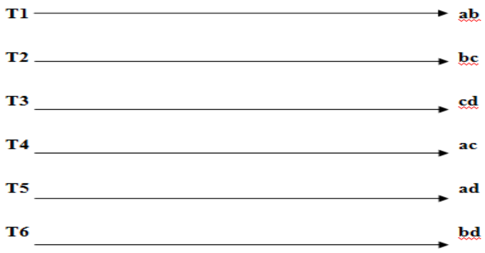
For example let us take four items:

a b c d

Now according to the formula we have to take 6 threads, namely **T1,T2,T3,T4,T5,T6.**



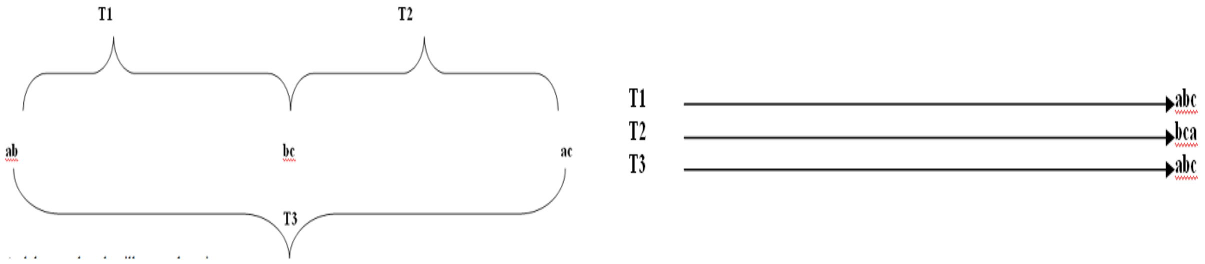
These threads now act simultaneously and provide us the pairs in form of



Now suppose **cd, ad, bd** have counts less than minimum counts so they are removed, now **ab, bc, ac** remain. The new number of threads now will be 3.

Let the threads now be **T1, T2, T3.**

And the new threads will create the pairs:



These all results in **abc** and thus if the count of **abc** is more than minimum count we get our most frequent set or else answer will be **ab, bc, ac**.

Hence, by introducing the threads with the help of java programming we have made the algorithm to create the pairs using threads. Now a very less time will be taken if there are a large number of items which need to make pairs in the algorithm. Making pairs simultaneously has greatly affected the algorithms performance thus making it faster as during the process many thousands of pairs have to be made which require a lot of time and computation, but using threads have greatly reduced the time and moreover reduced the complexity.

VII. SUMMARY

The work done by in Apriori along with finding a new way to implement it so that when it is implemented on a database , it leads to lower read through along with fast executions of the code for the frequent itemset generation. For this sole purpose we have gathered the database in a file and created a numerical image of the database which is inform of a file in our algorithm because the read operations are faster in the files, moreover threads have been implemented in the code for the pair generations so that if a large number of pairs are present, then the itemset generation will be faster and they will be removed if the minimum count is greater than their count in database. The threads and the numerical image provide an easy and effective way of increasing the speed of the code along with only one read operation of the database. The number of reads of file depends on the number of transactions, but since read operations in files are way faster than those in the database so it does not really affect the code execution

REFRENCES

[1] 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 .

[2] A Fast Apriori implementation, infomatics Laboratory, Computer and Automation Research institute, Hungarian academy of sciences

[3] Mining Large Itemsets for Association Rules ,Charu C. Aggarwal ,IBM research Lab.

[4] Mining association rules between sets of items in large databases, Rakesh Agarawal,Tomasz Imielinski*,Arun swami,IBM research lab.

[5] Fast algorithm for mining association rules,Rakesh Agarwal Ramakrishna Srikannt*,IBM research labs 650 Harry Road , San Jose , CA 95120.

[6] 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

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

[8] 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.

[9] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[10] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

[11] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. InU.M. Fayyad, et al. (eds.), Advances in Knowledge Discoveryand Data Mining, 1-35. AAAI/MIT Press, 1996.

[12] http://www.kdnuggets.com/news/2003/n19/22i.html