# "Improved Apriori Algorithm for Reducing Execution Time of Frequent Dataset"

Macwan Kiran
*A. D. Patel Institute of Technology, Anand, Gujarat*

***Abstract*: Association rules are the main technique for data mining. Mining association rules involves a lot of memory and CPU costs .This is especially a problem in data streams since the processing time is limited to one online scan. Therefore, when to update association rules, in real time or only at needs, is another fundamental issue. A new approach is required to update discovered association rules in a database when new transactions are added to, delete from, or modified in the database. However in real time environment, data are added continuously, and therefore, if we update association rules too frequently, the cost of computation will increase drastically. The purpose of the project is to overcome the above defined problem and to present a new approach which may decrease the execution time and increasing efficiency.**

## I. INTRODUCTION

A data warehouse is a centralized repository containing comprehensive detailed and summary data that provides a complete view of customers, suppliers, business processes, and transactions, from a historical perspective with little volatility. data warehouse is defined as a subject-oriented, integrated, non-volatile , time-variant collection of data in support of management's decisions [5]. More generally, data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker, such as executive, manager, and analyst, to arrive at better and faster decisions. Data warehouses provide access to data for complex analysis, knowledge discovery, and decision-making. They support high performance demands on an organization's data and information. It provides an enormous amount of historical and static data from three tiers:

1. Relational databases

2. Multidimensional OLAP applications

3. Client analysis tools

Several types of applications such as online analytical processing (OLAP), decision-support systems (DSS) and data mining are being supported. OLAP is a term used to describe the analysis of complex data from the data warehouse.

OLAP is a software technology that allows users to easily and quickly analyse and view data from multiple points-of-view [5]. OLAP provides dynamic and multi-dimensional support to executives and managers who need to understand different aspects of the data. Activities that are supported include:

- analysing financial trends
- Creating slices of data
- finding new relationships among the data
- drilling down into sales statistics
- doing calculations through different dimensions where each category of data is considered a dimension.

DSS support an organization's leading decision makers with higher-level data for complex and critical decisions. DSS queries a data warehouse or an OLAP database for predict the impact of that decision.

Real-time Data Warehouse delivers the right information to the right people just in time. Many essential operational decisions (e.g. promotion effectiveness, customer retention, key account information) need some actual yet integrated and subject-oriented data in or near real-time [1]. However, the direct real-time operational or tactical decision support is not achieved by traditional Business Intelligence Systems. These types of analytical applications are generally completely disconnected from operational IT systems. The decisions are executed by communicating them as a command or suggestion to humans, thus always cause latency. The real-time analysis requirements demand a set of service levels like data freshness, continuous data integration, analytical environments, active decision engines, adaptive platform for the event stream processing that go beyond a traditional Business Intelligence System [2].

## II. PROBLEM STATEMENT

Traditionally Processes run on a periodic basis (weekly, daily), and bulk transfer source data to a target data source (the data warehouse). However, in today's fast moving world, where immediacy of information is becoming absolutely critical for business competitiveness and survival, historical data is no longer enough. Furthermore, with the exponential growth of raw data volumes over the past years, the previously simple process of transferring in bulk the data from source to destination, has now become unwieldy and far too time consuming. The following are the weaknesses of the current system.

**1.** This approach is inefficient when the data under the source system is not changed frequently.

**2.** It increases response time (latency), network traffic, wasteful to CPU or memory utilization.

**3.** It maximizes resource requirements and bulk transfer.

**4.** It is intrusive to the source databases.

## III. MOTIVATION

Traditional data warehouses are increasingly being challenged by demands for real-time data access, analysis of structured and unstructured data, and the need to synchronize core customer and product information across operational systems to create a single view of the enterprise [3]. These challenges are the result of new business requirements to leverage enterprise information more effectively in order to identify new opportunities and deliver new products to market faster, optimize business processes through real-time information and analytics, provide increased visibility to business performance and meet industry compliance standards for reporting. A real-time data warehouse eliminates the data availability gap. Continuous processing without delay opens up significant new opportunities for the practice of business intelligence [4].

Warehousing approaches that enable access to multiple types of data and provide real-time views of business operations to a broader audience will be required just to keep up with the competition. The challenges faced by today's organizations to leverage information effectively are given below.

- Information distributed in silos across the organization
- Volume and variety of information increasing
- Velocity of business driving real-time requirements

To overcome these many challenges, there is increased need to aggregate and analyse information dynamically [2]. Increasingly mixed workload environments and the constantly changing needs of different business constituents require more dynamic warehousing capabilities.

## IV. OBJECTIVISE

Association rules are the main technique for data mining. Apriori algorithm is a classical algorithm of association rule mining. Apriori-based algorithms require a lot of memory and CPU costs when seeking frequent item sets. A new approach will overcome this problem by reducing execution cost. Purpose of thesis to imp lent:

1. Implement apriori algorithm

2. Implement proposed algorithm

3. Comparisons of both algorithms according to its execution time.

## V. REAL TIME WAREHOUSING

Real time warehousing is not based on the idea of accumulating a static warehouse of data that serves up data to data marts and BI tools. Instead it is based on a dynamic process that continually defines and serves information to meet known and anticipated needs. This approach distributes all types of relevant information in the right form at the right time and provides the business with dynamic end-to-end analytics.

*Design consideration for Real-time Data Warehousing*

The design of an RTDW has to consider technical aspects: scalability, high availability, frequent (i.e. just-in-time or continuously) data loading, mixed workload, etc. as well as the integration of active mechanisms which deal with the two sorts of propagation delays in Data Warehouse environments [6]:

1. Delays in capturing real world events by the operational systems, and

2. Delays in loading and integrating data into the Data Warehouse

The Business Requirements for an Active Data Warehouse

- Performance

– Within seconds

- Scalability

– Support for large data volumes, mixed workloads and concurrent users

- Availability

– 7 **X** 24 **X** 365

- Data Freshness

– Accurate, up to the minute, data

## VI. TRADITIONAL DATA WARE HOUSE

Data warehouses exist to facilitate complex, data-intensive and frequent adhoc queries. Data warehouses must provide far greater and more efficient query support than is demanded of transactional databases. The data warehouse access spreadsheet functionality includes support for state-of-the art spreadsheet applications as well as for OLAP applications programs [5].

6 The greatest benefit of a data warehouse is the ability to analyse and execute business decisions based on data from multiple sources. For example, an organization has collected valuable data and stored it in 30 databases. A data warehouse is not only a convenient way to analyse and compare data in all the databases, but it can also give historical data and perspective. Thus data warehouse is a

one-stop shop, but it is also a one-stop shop from an historical perspective as well. Using data warehouse, one can look at past trends, whether they be product sales or customers or whatever and may be do some predictions of what is going to happen in the future [5].

Also data retrieved from multiple databases is not constrained by the tables in each of those databases. A data warehouse receives application neutral data. Whatever database application is supplying the information to the data warehouse is not preconditioning the data to be presented in a way the originator of the data requires. That means, the data from the inventory system, the financial system, or the sales system is sent to the data warehouse for processing as application neutral data that is not formatted to answer only queries from an inventory database, finance database, or sales database program. If not for application-neutral data, the data warehouse would be nothing more than a collection of data marts.

## VII. PROPOSED WORK

Association rules are the main technique for data mining. Apriori algorithm is a classical algorithm of association rule mining. Lots of algorithms for mining association rules and their mutations are proposed on basis of Apriori algorithm, but traditional algorithms are not efficient. For the two bottlenecks of frequent item sets mining: the large multitude of candidate 2- item sets, the poor efficiency of counting their support. Proposed algorithm reduces one redundant pruning operations of 2 *C* . If the number of frequent 1-itemsets is n, then the number of connected candidate 2-itemsets is *n C,* while pruning operations *n C.* The proposed algorithm decreases pruning operations of candidate 2-itemsets, thereby saving time and increasing efficiency. For the bottleneck: poor efficiency of counting support, proposed algorithm optimizes subset operation, through the transaction tag to speed up support calculations. Author names and affiliations are to be centered beneath the title and printed in Times 12-point, non-boldface type. Multiple authors may be shown in a two- or three-column format, with their affiliations italicized and centered below their respective names. Include e-mail addresses if possible. Author information should be followed by two 12-point blank lines.

Algorithm Apriori is one of the oldest and most versatile algorithms of Frequent Pattern Mining (FPM). Its advantages and its moderate traverse of the search space pay off when mining very large databases. Proposed algorithm improves Apriori algorithm by the way of a decrease of pruning operations, which generates the candidate 2-itemsets by the apriori\e-gen operation. Besides, it adopts the tag-counting method to calculate support quickly. So the bottleneck is overcome [13].

## VII. ENHANCEMENT OF SUBSET PROCEDURE

***Procedure subset (, t) k C***

**1. for** all candidates do k sC□
**2.** if t contains s
**3.** subset = subset + □ □ s
**4. end for**

***Improved Procedure subset ( k C , t)***

***1. for*** *all candidates s∈Ck do*
***2.*** *If (first item of s* ≥ min *) (last item of s* ≤max *)*
***3.*** *subset = subset* + {*s*}
***4. end for***

Its running time is , this is a very large number. Considering k = 3, the transaction t (x2, x6, x7, x9), the item sets s (x4, x6, x10). Refinement of this sentence: "if t contain s during a judge are the following conditions: if (t contains x4) (t contains x6) (t contains x10) then….." Obviously, t does not contain s, but still have to compare one by one. □ □ | | k D The proposed algorithm improves the subset operation through adding a tag column on the transaction database, of which values are the first and the last numbers of item sets. If (first item of x > = min) (last item of x <= max), then t contains s. Otherwise t does not contain s.

The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5 x 11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

## IX. RESULT ANALYSIS

In order to study the performance, the computer used to run the experiments had Intel(R) Pentium D CPU 2.80 GHz processor and 512 MB of memory and 512 MB of memory. The operating system used was Microsoft Windows XP Professional. Programs are coded in C#.NET on the platform of Visual Studio 2008.The performance of the Apriori algorithm is largely depends on the internal characteristic of the datasets and the number of records. To make the time measurements more reliable, no other application was running on the machine while the experiments were running.

Algorithm has been tested on the three different datasets described in the previous chapter. The performance measure is the execution time (seconds) of the algorithms on the datasets. The minimum confidence is set to 50%. In result Analysis comparison of existing apriori algorithm and

improve apriori algorithm based on execution time of date set.

## X. CONCLUSION AND FUTURE WORK

The proposed algorithm for mining association rule, decreases pruning operations of candidate 2-itemsets, thereby saving time and increase efficiency. It optimizes subset operation, through the transaction tag to speed up support calculations. The experimental results obtained from tests show that proposed system outperforms original one efficiently.

The current mining methods require users to define one or more parameters before their execution; however, most of them do not mention how users can adjust these parameters online while they are running. It is not feasible for users to wait until a mining algorithm to stop before they can reset the parameters. This is because it may take a long time for the algorithm to finish due to the continuous arrival and huge amount of data. For further improvement, we may consider either let the users adjust online or let the mining algorithm auto-adjust most of the key parameters in association rule mining, such as support, confidence and error rate.

## REFERENCE


[1] The dynamic warehousing infrastructure: Establishing a foundation to meet new information requirements by IBM

[2] Conflicting, unintegrated historical data to actionable insight – An introduction to dynamic warehousing from IBM

[3] Advances in Data Warehouse Performance – White Paper by Winter Corporation

[4] Sheila A. Abaya "Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation" *Volume 3, Issue 7, July-2012 International Journal of Soft Computing and Engineering (IJSCE)*

[5] Badri Patel, Vijay K Chaudhari, Rajneesh K Karan, YK Rana "Optimization of Association Rule Mining Apriori Algorithm Using ACO " *Sep. 2008, Volume 5, No.9*

[6] HAN Feng, ZHANG Shu-mao, DU Ying-shuang "The analysis and improvement of Apriori algorithm"

[7] Ajay Acharya , Shweta Modi and Vivek Badhe "An Algorithm for Finding Frequent Itemset based on Lattice Approach for Lower Cardinality Dataset" *International Journal of Mathematical Archive 1(1), Oct.-2010, 16-19*

[8] Jacky W.W. Wan Gillian Dobbie "Mining Association Rules from XML Data using XQuery"

[9] Pradeep Chouksey, Juhi Singh, R.S. Thakurm, *R.C. Jain "*Frequent Pattern Mining using Candidate Generation approach with Single Scan of Database" Symposium on Progress in Information & Communication Technology 2009

[10] Roberto J. Bayardo Jr."Efficiently Mining Long Patterns from Databases" IBM Almaden Research Center

[11] Neelamadhab Padhy1 and Rasmita Panigrahi "Data Mining: A prediction Technique for the workers in the PR Department of Orissa (Block and Panchayat)" *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.5, October 2012*

[12] G.SenthilKumar, S.Baskar,M. Rajendran "ONLINE MESSAGE CATEGORIZATION USING APRIORI ALGORITHM" *International Journal of Computer Trends and Technology- May to June Issue 2011*

[13] Mamta Dhanda, 2Sonali Guglani, 3Gaurav Gupta "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes", *IJCST Vol. 2, Issue 3, September 2011*

[14] D. Gunaseelan, P. Uma " An Improved Frequent Pattern Algorithm for Mining Association Rules" *Volume 2 No. 5, May 2012 ISSN 2223-4985 International Journal of Information and Communication Technology Research ©2012 ICT Journal*.

[15] By Zheng, Z., Kohavi, R., and Mason, L. Real World Performance of Association Rule Algorithms

[16] By THO, M.N Zero-Latency Data Warehousing: the State-of-the-art and experimental implementation approaches