# Analysis of Optimized Association Rule Mining Algorithm using Genetic Algorithm

Shanta Rangaswami, Assistant Professor
Shobha G., Professor and Head
Pallavi Gupta, Anusha R., Meghana H., Final Year Students
Department of Computer Science and Engineering,
R. V. College of Engineering, Bangalore

## ABSTRACT
Apriori algorithm is a classic algorithm for frequent item set mining and association rule learning over transactional databases. The algorithm determines frequent item sets, which in turn can be used to determine association rules. These rules indicate the general trends in the database. Genetic algorithm is a search heuristic that mimics the process of natural selection using a greedy approach. This heuristic is routinely used to generate useful solutions for optimization and search problems. In this paper, we apply genetic algorithm to optimize the frequent item sets generated by Apriori algorithm and identify all possible significant association rules by analyzing the working of the algorithm on real data sets.

## General Terms
Data Mining, Genetic Algorithms, Algorithms

## Keywords
Apriori, Genetic, optimization, transaction, association rule mining

## 1. INTRODUCTION
Association rule mining [1] is a classic algorithm used in data mining for learning association rules and it has several practical applications. For instance, in market basket analysis, shopping centers use association rules to place the items next to each other so that users buy more items. Using data mining techniques, the famous beer – diapers-Wal-Mart analysis showed that on Friday afternoon young American males who buy diapers also tend to buy beer. This resulted in Wal-Mart placing beer next to diapers and the beer-sales going up. This was a legendary observation as it showcased a cross - selling opportunity indicating the power of data mining.
The e-commerce giant, Amazon, uses association mining to recommend items to users based on the current item that they are browsing or buying. Another application of rule mining is seen in the Google auto - complete feature, where after a word is typed, frequently associated words that a user types after that particular word, is searched. [2]
Association analysis is applicable in all the major application domains such as bioinformatics, geo informatics, data mining, web mining, medical diagnosis and scientific data analysis. Over the years, traditional association rule mining algorithms developed were used to find a set of associations between items in transactions. These associations are referred to as positive associations. However, valuable information can be mined with the help of negative associations. [3]
The Apriori algorithm finds frequent item sets in an incremental manner starting from an item set containing a single element. In this algorithm, the output that is generated is given as an input to the genetic algorithm that typically uses an incremental approach to generate new items from old ones [4]. The three basic operations used by the genetic algorithm are - selection, crossover and mutation.

The selection operation is used to select individual items from population. This is usually done by applying fitness criteria to all the items and normalizing them. These values are arranged in descending order and accumulated from which one value is chosen. The crossover operation refers to the process of taking more than one parent solution and producing a child solution from them [5]. Finally, the mutation operation maintains genetic diversity from one generation of a population of the genetic algorithm to the next. In mutation, there is a possibility that the solution might change entirely from the previous solution. This helps prevent the population from stagnating at any local optima. Mutation occurs during evolution according to a user-definable mutation probability. Thus, applying these basic operations of genetic algorithm can result in a better solution.

## 2. DESIGN
The frequent item sets generated by Apriori algorithm are optimized using Genetic Algorithm as follows [3]:
Step 1: Begin
Step 2: Read the sample record file, which fits in the memory
Step 3: Apply Apriori algorithm on the sample data by setting the support and confidence values to generate the sets of frequent item sets
Step 4: Apply selection method of genetic algorithm to this item set collection to select two members.
Step 5: Apply crossover and mutation on selected item sets to find association rules.
Step 6: Repeat the steps 3-5 till desired number of generations is obtained.
Step 7: End

## 3. EXPERIMENTATION
### 3.1 Discussion
The following observations were made when the optimized version of Apriori algorithm was tested with a real data set. The dataset comprised of 1000 entries and 5000 entries [5]. The data set is that of bakery sales, which consists of entries in the form of a sparse vector representation:

*Receipt# followed by item #'s that are on that receipt*
This dataset [6] of a *bakery* chain has a menu of about 40 pastry items and 10 coffee drinks. It has a number of locations in West Coast states (California, Oregon, Arizona, Nevada). The dataset file describes the contents of the EXTENDED BAKERY dataset developed for CPE 466, Knowledge Discovery in Data course at Cal Poly. It contains information about one year worth of sales information for a couple of small bakery shops. The dataset contains information about the different store locations, the assortments of baked goods offered for sale and the purchases made

## 3.2 Observations

The following details were observed by applying the optimization of the routine association rule mining algorithm can be done with the help of genetic algorithm.

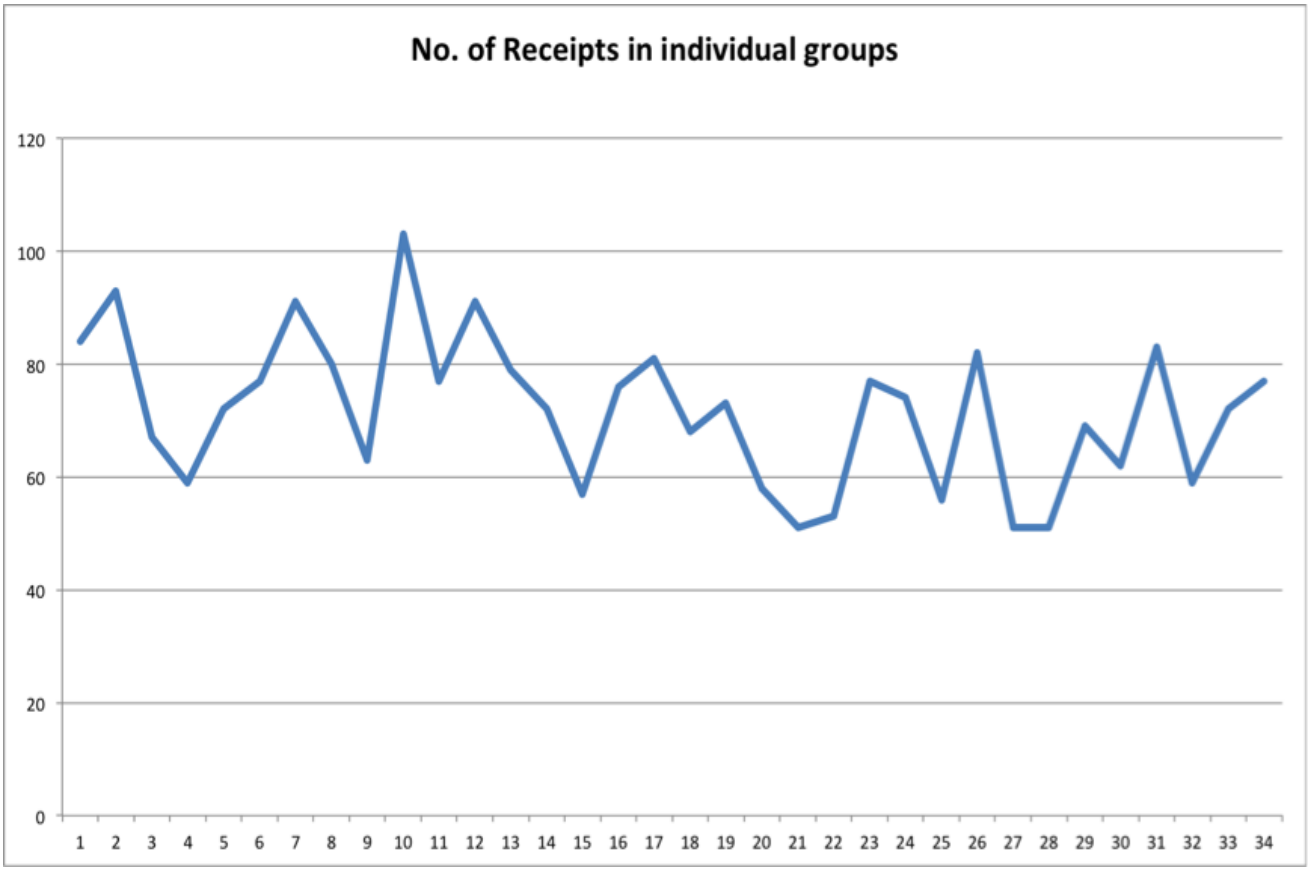


**Fig 1: Line graph representation for case 1 with 1000 transaction dataset**

Optimized Apriori Rule Mining Algorithm:

**Case 1:** ***1000 receipts dataset***
Number of groups generated: 34
Range of number of items purchased: 2 - 8
Least number of receipts in a group: 51
Highest number of receipts in a group: 103
Number of different patterns: 2097
**Case 2:** ***5000 receipts dataset***
Number of groups generated: 35
Range of number of items purchased: 2 - 35
Least number of receipts in a group: 259
Highest number of receipts in a group: 544
Number of different patterns: 10310

## 3.3 Analysis

This study shows that by applying the Apriori algorithm and genetic algorithm together, the system could be optimized with the available historical data. Such kind of information that is collected will be useful to identify trends by identifying the number of groups formed based on the items or products bought, buying pattern of the customers, frequency of buying a particular product as well as the most different pattern of purchase. The experiment that was carried out shows that This ability of the system to identify the groups of frequent item-sets, number of groups formed and also groups of different patterns based on the support value is critical in the decision making process of placing the products on the shelf of a department store to maximize profits and also helps in identifying the demand of a particular product individually and in association with other products. This knowledge discovery from the available data in the receipts is beneficial as it helps in refining the policy decisions of an organization thus maximizing gains. It also provides an insight of the growing and changing customer demands in the markets.

## 3.4 Future Enhancements

The results obtained are quite accurate since we have used genetic algorithm for optimization. This can be extended further to reduce the complexity of genetic algorithm, thereby increasing the efficiency.

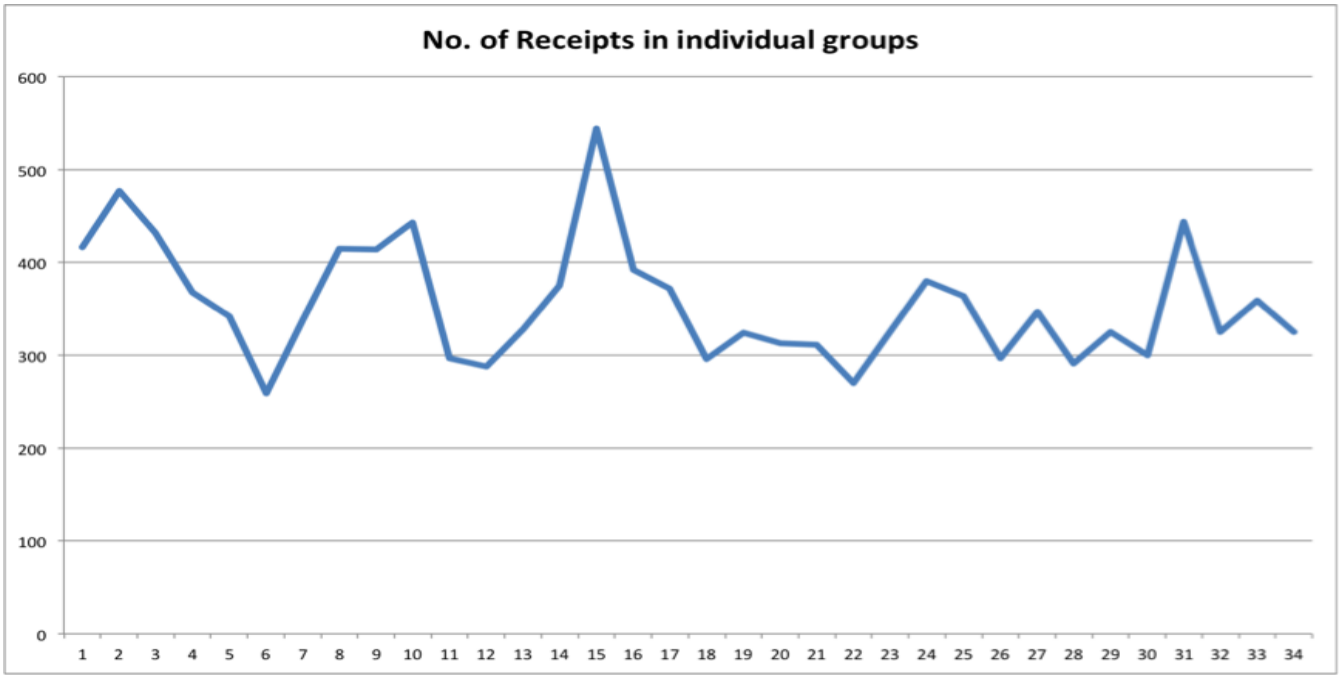


**Fig 2: Line graph for 5000 transactions dataset**

In conclusion, suppose there are records of large number of transactions at a shopping center (market basket data), data mining algorithms allow us to observe patterns associated with shopping. Learning association rules basically means finding the items that are purchased together more frequently than others.

From the example that is discussed, following interpretations were made:

- With larger data sets, the density of the groups began increasing.
- The number of groups formed remained almost the same even when the number of transactions in the data set increased tremendously.

## 4. ACKNOWLEDGMENTS

This paper was a result of the contribution from the following authors.

**Ms. Shanta Rangaswamy**, Assistant Professor, Department of C.S.E., R.V. College of Engineering, Bangalore, is pursuing her PhD from Kuvempu University. Her research areas of interest are Autonomic computing, Data mining, Machine learning techniques, Performance Evaluation of systems,Cryptography and Steganography, and System Modeling and Simulation.

**Dr. Shobha G.,** Professor and Head, Department of C.S.E., R. V. College of Engineering is associated with the college, since 1995. She has received her Masters degree from BITS, Pilani and Ph.D (CSE) from Mangalore University. Her research areas of interest are Database Management Systems, Data mining, Data warehousing, Business analytics, Image Processing and Information and Network Security.

Pallavi Gupta is a graduate student of the Department of Computer Science, R.V.C.E., Bangalore, India. Her technical interests include Data Mining, Mobile and Cloud computing, Linux internals, UX and UI. Being a technology enthusiast, she has been involved in a wide range of hobby projects and she was among one of the few students who participated in MIT Design Innovation India Workshop 2013 organized by MIT Media Labs, Massachusetts at PES Institute of Technology, Bangalore.

Anusha R. is a graduate student of the Department of Computer Science, RVCE, Bangalore, India. Her academic interests include Data mining, Algorithms, Web designing, Computer Networks and Android app development.

Meghana H. is a graduate student of the Department of Computer Science, R.V.C.E., Bangalore, India. Her academic interests include Data Mining, building Android apps. A former Google Student Ambassador, where she represented college, she is also tech savvy.

## 5. REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. In Buneman, P., and Jajodia, S., (eds.). Proceedings of ACM SIGMOD Conference on Management of Data, 1993 (SIGMOD'93), 207- 216

[2] Article titled "How Amazon is leveraging Big Data", http://www.bigdata-startups.com/BigData-startup/amazon-leveraging-big-data/

[3] Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K., Optimized association rule mining using genetic algorithm, Advances in Information

Mining, ISSN: 0975-3265, Volume 1, Issue 19 September 2011

[4] Mitchell, Melanie (1996). *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press. ISBN 9780585030944

[5] David Beasley et al., "An Overview of Genetic Algorithms: Part 1, Fundamentals", University Computing, 1993

[6] Dataset from https://wiki.csc.calpoly.edu/datasets/attachment/wiki/apriori/apriori.zip

[7] Pang-Ning-Tan, Vipin Kumar, Michael Steinbach, (2007) *Introduction to Data Mining*