IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Combining Apriori and FP Growth algorithms with Simulated Annealing for Optimized Association Rule Mining

Manoj Ganjir[1], Jharna Chopra[2]

M.E, Computer Science & Engineering Dept. Shankaracharya Group of Institutions, Bhilai (C.G.), India[1]

Assistant Professor, Computer Science & Engineering Dept. Shankaracharya Group of Institutions, Bhilai(C.G.), India[2]

Abstract: Association rule mining is the process of finding frequent patterns and associations between set of objects from information repositories. Finding optimized techniques for generating association rules from large repositories has become a major area of study. Apriori algorithm is a simple algorithm which is used for mining frequent item sets. The FP-growth algorithm on the other hand works as a solution to the problem for long frequent patterns to searching for shorter ones recursively. The output of FP-Growth is a FP Tree in the end. The current work focuses on using Simulated Annealing technique on both algorithms for optimized Association Rule Mining. The results have been also discussed in the end and analysis has also been generated.

Keywords: Apriori, FP-Growth, Simulated Annealing, Minimum Support, Minimum Confidence.

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. Data mining software is one of a number of analytical tools for analysing data. 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. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyse market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Some of the concepts of data mining are:-

A. 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 .

B.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.

C. Knowledge

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

D. Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval.

E. ARM (Association Rule Mining)

Let S= {Set1, Set2, Set3……. Setn} be a universe of Items and T= {Trans1, Trans2, Trans3……....Transn} is a set of transactions. Then expression X => Y is an association rule where X and Y are item sets and $X \cap Y = \Phi$. This rule holds support and confidence, support is a set of transactions in set T that contain both X and Y and confidence is percentage of transactions in T containing X that also contain Y. An association rule is strong if it satisfies the concepts of minimum support and minimum confidence such as support $\geqslant$ minimum_support and confidence $\geqslant$ minimum_confidence. An association rule is frequent is such that support $\geqslant$ minimum_support .

F.Minimum Support

Support is a fraction of transactions that contain an item set. Frequencies of occurring patterns are indicated by support. The probability of a randomly chosen transaction T that contain both itemsets X and Y is known as support. Mathematically it is represented as:-

Copyright to IJARCCE

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

$$P(X, Y) = \frac{\text{No\_of\_transaction containing X and Y}}{\text{Total No of transactions}}$$

G. Minimum Confidence

It measures how often items in Y appear in transactions that contain X. Strength of implication in the rule is denoted by confidence. Confidence is the probability of purchasing an itemset Y in a randomly chosen transaction T depend on the purchasing of an itemset X.

H. Market Basket Analysis

In market basket databases consist of a large no. of records and in each record all items bought by a customer on a single purchase transaction are listed. This data is used by them to adjust store layouts (placing items optimally with respect to each other), to cross-sell, to promotions, to catalog design and to identify customer segments based on buying patterns. The probability of a randomly selected transaction from the database will contain all items in the antecedent and the consequent is known as support, whereas the conditional probability of .a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent is known as confidence. This "market basket analysis" result can be used to suggest combinations of products for special promotions or sales.

II. RELATED WORK

Apriori algorithm has some limitation in spite of being very simple [1]. The major advantages of FP-Growth algorithm is that it uses compact data structure and eliminates repeated database scan FP-growth is faster than other association mining algorithms and is also faster than tree researching. According to [2] availability of determine "Which groups or sets of items are customer's likely to quality services is vital for the well-being of the economy. Market basket circles are covering all major aspects of the service analysis which may be performed on the retail data of customer transactions. According to [3] in situations with a large number of frequent patterns, long patterns, or quite low minimum support thresholds, an Apriori like algorithm may suffer from the following two nontrivial costs i.e.it is costly to handle a huge number of candidate sets and it is tedious to repeatedly scan the database and check a large set of candidates by pattern matching, which is especially true for mining long patterns. This is the inherent cost of candidate generation, no matter what implementation technique is applied.

III.METHODOLOGY

The proposed technique focuses on runtime optimization of transactional dataset by the use of Simulated Annealing technique. The optimization is performed for Apriori and FP-Growth algorithms.

A. Simulated Annealing

Simulated annealing is a method for finding a good (not necessarily perfect) solution to an optimization problem. The traveling salesman problem is a good example: the salesman is looking to visit a set of cities in the order that minimizes the total number of miles he travels. As the number of cities gets large, it becomes too computationally intensive to check every possible itinerary. An optimization algorithm searches for the best solution by generating a random initial solution and "exploring" the area nearby. If a neighboring solution is better than the current one, then it moves to it. If not, then the algorithm stays put. This technique is easy to implement.

1. First, generate a random solution
2. Calculate its cost using some cost function.
3. Generate a random neighboring solution
4. Calculate the new solution's cost.
5. Compare chosen solution with new solution.
   a. If s_new < s_old: move to the new solution
   b. If s_new > s_old: move to the old solution
6. Repeat steps 3-5 above until an acceptable solution is found .

B. Apriori Algorithm

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data. Basic algorithm is as follows :-

1. Generate candidate of size n from dataset.
2. Any (n-1) itemset that is not frequent cannot be subset of frequent k-itemset.
3. Generate C_n : Candidate item set of size n.
   L_n: Frequent itemset of size n
4. Generate C_n+1 itemsets from L_n
5. Generate L_n+1 candidate with min_support
6. Return generated candidates in L_n+1.

C. FP-Growth Algorithm

The FP-growth method transforms the problem of finding long frequent patterns to searching for shorter ones recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs. When the database is large; it is sometimes unrealistic to construct a main memory based FP- tree.

Basic algorithm is as follows:-

- Scan data and find support for each item.
- Discard infrequent items.
- Sort frequent items in decreasing order based on their support
- Use this order when building the FP-Tree, so common prefixes can be shared.
- Nodes correspond to items and have a counter.
- FP-Growth reads 1 transaction at a time and maps it to a path.
- Fixed order is used, so paths can overlap when transactions share items (when items have the same prefix).
- In this case, counters are incremented
  - Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)

DOI 10.17148/IJARCCE.2015.412116

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

- The more the paths overlap, the higher the compression. FP-tree may fit in memory.
- Frequent item sets are extracted from the FP-Tree.
- Each prefix path sub-tree is processed recursively to extract the frequent item sets. Solutions are then merged.
- Divide and conquer approach

[6]

IV.RESULTS AND DISCUSSION

During experiments conducted on a set of transactions which is multidimensional in nature it was found that the results obtained after running both the algorithms without optimization and after optimization by using Simulated Annealing technique was equal. Figures 1 and 2 shows that results after and before optimization are same.

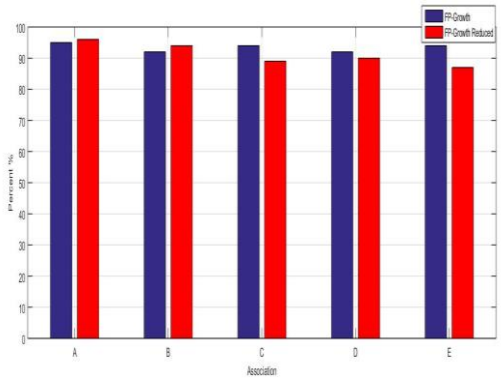

Fig 1: Percentage Analysis of equality between Association rules produced.

The above given analysis figure 1 and 2 shows the validity of association rules produced by Apriori and FP Growth algorithm with optimization. It is clearly seen that the quality of association rules produced for both the cases is same. Clearly after optimization same results are produced in same time.

V.CONCLUSION

The paper focuses on association rule mining techniques and its optimization using Simulated Annealing technique. In this paper two classical mining algorithms- Apriori algorithm and FP- Growth algorithm are discussed and the applications area of mining algorithms is also identified. After experimental analysis it is clear that Simulated Annealing greatly reduces the amount of data which is provided to the algorithms without reducing the effectiveness of results.

REFERENCES

[1] Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms, B.Santhosh Kumar and K.V.Rukmani,Int. J. of Advanced Networking and Applications,2010.

[2] Ankur Mehay, Dr. Kawaljeet Singh, and Dr. Neeraj Sharma, "AnalyzeMarket Basket Data using FP-growth and Apriori Algorithm," International Journal on Recent and Innovation Trends in Computing and Communication,2013.

[3] JIAWEI HAN, JIAN PEI, YIWEN YIN and RUNYING MAO ,"Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach",Data Mining and Knowledge Discovery, 8, 53–87, 2004

[4] Rahul Mishra and Abha choubey, "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (4) , 2012

[5] N. Naga Saranya, and M. Hemalatha, "Estimation of Evolutionary Optimization Algorithm for Association Rule using Spatial Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 51– No.3, August 2012.