A Survey on FP (Growth) Tree Using Association rule Mining

Princy Shrivastava [1], Ankita Hundet[2],Babita Pathik[3], Shiv Kumar[4]

[1] M.Tech Scholar, Department of Computer Science & Engineering, Lakshmi Narain College of Technology & Excellence Bhopal (M.P), India,

[2,3] Assistant Professor, Department of CSE, Lakshmi Narain College of Technology & Excellence, Bhopal (M.P), India

[4] Professor & Head, Department of CSE, Lakshmi Narain College of Technology & Excellence, Bhopal (M.P), India

---***---

ABSTRACT: *Data mining is passed down to arranged with the data stored in the backend to extract the required information and expertise. It has number of ways for the finding data; association rule mining is the very robust data mining approach. Its main work to find out required hidden pattern from bulk storage. Authors go through with many techniques out of them frequent pattern growth is effective algorithm to extract required association rules. It examines the directory two times for handling. FP growth algorithm has some concern to generate an enormous conditional FP trees. Authors introduce a new technique which extracts all the frequent item sets without the generation of the conditional FP trees. It also catches the frequency of the usual item sets to extract the required association rules. This paper present a survey for Association rule mining.*

Keyword- Data mining, Association rules, Apriori Algorithm, FP Growth Algorithm.

1. INTRODUCTION

Data mining refers to the process of extraction or mining expertise from data storage. In this (Association) data mining suggest picking out the unknown inter-connection of the data and concludes the rules between those items. Its aim is to find out striking correlations, persistent patterns and association among set of item in the storage. It is used in all real life applications of business and industry.

Association rules illustrate how frequently items are bayed together. E.g., an association rule beer => chips (80%) tells that four out of five customer that bought beer, they bought chips too . such rules may become useful for decision regarding product pricing, promotions, store outline and many others.[3].

Data mining encompasses many different techniques and algorithms. They differ in the kinds of data that can be analyzed and the kinds of knowledge. representation used to convey the discovered knowledge. In this paper, we review these techniques briefly and highlight the interestingness concept defined to overcome the huge number of patterns induced as a result of the mining process. We also review the categorization of the interestingness measures and give numerous examples in the literature.

Extraction of Knowledge in databases (KDD) refers to the overall operation to find out useful knowledge from data, and data mining refers to a particular step in this process. KDD involves many steps including preparation of data, selecting & cleaning, data mining and proper interpretation of the results of mining.

The data mining step is the adequate of previously unknown, valid, interesting, novel, potentially useful, and understandable patterns in big storage. The knowledge that we seek to discover describes patterns in the data as opposed to knowledge about the data itself. Classification rules, association rules, clusters, sequential patterns, time series, contingency tables; summaries obtained using some hierarchical or taxonomic structure, and others. Association rule mining is done to find out association rules that satisfy the predefined minimum support and confidence from a given database.[1]. The problem of finding association rule is usually decomposed into two sub problems:

(a) Find all Frequent Item sets using Minimum Support.
(b) Find Association rules from Frequent Item sets Using Minimum Confidence

1.1 Data Mining Techniques

Knowledge or Information for decision making in a business is very poor even though data storage grows exponentially. Data mining also known as Knowledge. The Knowledge extracted allows predicting the behaviour and future behaviour .This allows the business owners to take positive, knowledge driven decisions. Data mining is applied on various industries like aerospace, education etc.

Knowledge is extracted from the historical data by applying pattern recognition, statistical and mathematical techniques those results in the knowledge in the form of facts, trends, associations, patterns, anomalies' and exceptions. There are some areas where we will apply Data Mining.
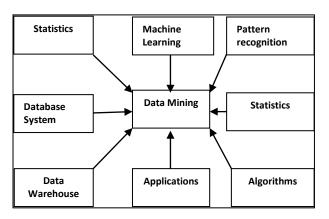
Figure 1: Uses of Data mining

1.2 Knowledge Discovery Process From Data

The KDD process is ultimately data mining methods to extract patterns from data. Each method has different aim, which decides the outcome of the KDD process entirely.
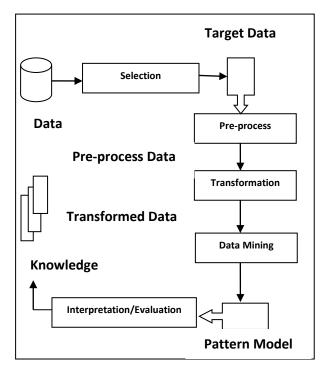

Figure 2: Process of Data Mining in KDD

Data Pre-processing: Data Pre-processing generate real time data for the mining process.

Data Mining: Data Mining is application of intelligent methods that extract data patterns.

Pattern Evaluation: The patterns that are generated by the data mining are evaluated for the interest according to the target business problem.

Knowledge Presentation: Knowledge Presentation uses visualization techniques that visualize the interesting patterns and helps the user to understand and interpret the resultant patterns.

There exists many other algorithms for mining of frequent item sets viz., Apriori and Eclat, FP-Tree growth algorithm pre-processes the database only twice as follows: an initial scan of the database determines the frequencies of the items. All the uncommon items -- the items that do not appear in a minimum number of user-specified transactions -- are discarded from it as they cannot be a part of frequent item sets.

2. LITERATURE SURVEY

2.1 An optimized algorithm for association rule mining using FP tree

In this paper Author conclude that Association rule mining has a mentionable amount of practical applications, including classification, XML mining, spatial data analysis, and share market and recommendation systems. In this work, with the help of the a new improved FP tree and a new Frequent Item set mining algorithm, we are able to save a lot of memory in terms of reducing the no. of conditional pattern bases and conditional FP trees generated.[1]

The main idea of the paper is to design a technique for association rule mining which reduces the complex intermediate process of the frequent item set generation. Through this work Authors aim to design a technique for association rule mining to reduce the no. of times the transactional database is scanned, reduce the no. of conditional pattern bases generated, and remove the generation of the conditional FP trees. There are three steps involved in the proposed technique.

Step1: we scan the database one time to generate a D-tree from the database. The D-tree is basically the replica of the database under consideration.
Step2: using the D-tree and support count as an input we construct a New Improved FP tree and a node table which contains the some nodes and their frequencies.
Step3: The traditional FP growth algorithm needs to generate a large number of conditional pattern bases then calculate the conditional FP-tree, to avoid candidate generation.

2.2 Mining Association Rules Using Modified FP Growth Algorithm

In this paper Author concludes that Determining frequent objects (item sets, episodes, sequential Patterns) are one of the most important fields of data mining. It is well known that the way candidates are

defined has great effect on running time and memory need, and this is the reason for the large number of algorithms. Authors presented frequent pattern mining is expecting transpose representation to relieve current methods from the traditional bottleneck, providing scalability to massive Data sets and improving response time. Many existing algorithms are based on the two algorithms and one such is APFT, which combines the Apriori algorithm and FP-tree structure of FP-growth algorithm. To overcome the limitation of the two approaches a new method named APFT was proposed. The APFT algorithm has two steps: first it constructs an FP-tree & then second mines the frequent items using Apriori algorithm. The advantage of APFT is that it doesn't generate conditional & sub conditional patterns of the tree recursively and the results of the experiment show that it works fasts than Apriori and almost as fast as FP-growth. Author proposed to go one step further & modify the APFT to include correlated items & trim the non-correlated itemsets. This additional feature optimizes the FP-tree & removes loosely associated items from the frequent itemsets. Author call this method as APFTC method which is APFT with correlation.[2]

2.3 Market Basket Analysis using Association Rule Learning

In this paper Author concludes that a group of customers is going to purchase can be very useful for the retailers. These results could also be helpful in determining which products appeal each other so that they can be put together in a market in order to increase the sales. For the same reason, we have proposed a novel data structure, frequent pattern tree (FP-tree), for storing compressed, crucial information about frequent patterns, and developed a pattern growth method, FP-growth, for efficient mining of frequent patterns in large databases. Using the inputs of the support and confidence values, we obtain the output in the form of association rules of the item sets to be purchased thus depriving the patterns. For the same reason, we have proposed a novel data structure, frequent pattern tree (FP-tree), for storing compressed, crucial information about frequent patterns, and developed a pattern growth method, FP-growth, for efficient mining of frequent patterns in large databases.
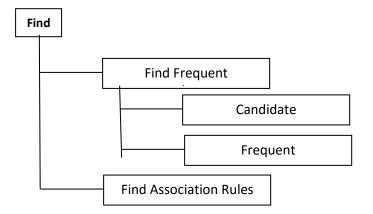

Figure 3: Generation of Association

3. PROBLEM IDENTIFICATION

An ample of research work is going in field of association rule mining and various authors have proposed different algorithms in this field. Still there exist many problem and demand in this field which need to be solved in order to get complete advantage of Association Rule technique. A graph based clustering with affinity propagation. That find a new way to clustering document based more on the keywords they contain document based clustering techniques mostly depend on the keywords. The work is modifying the FP-mining algorithm to find the frequent sub graph with clustering affinity propagation in graph. Data mining is the current focus of research since last decade due to enormous amount of data and information in modern day. Association is the hot topic among various data mining technique. Traditional FP method performs well but generates redundant trees resulting efficiency degrades.[1] To achieve better efficiency in association mining, positive and negative rules generation helps out. Same concept has been applied in the proposed method. Results shows that proposed method perform well and handles very large size of data set. Aim of Authors to improve the Apriori-Growth algorithm such that it determines the frequent patterns for a large database with efficient computation result.

4. CONCLUSION

In this paper we introduced Enhanced-FP, which does its work without any complex data structure or prefix tree. Its main strength is its simplicity. There is no need to re-representation of transactions. By comparing this frequent item set mining algorithms Apriori and FP-growth and Enhanced-FP, the strength of Enhanced-FP is analyzed. Enhanced-FP clearly outperforms Apriori and FP-Growth. It is faster than Apriori and FP-Growth and is not expensive like FP-tree. Its Transactional database is memory resident. Association is the growing topic among various data mining technique. In this article

Authors proposed a hybrid approach to deal with large size data. Proposed system is the enhancement of frequent pattern (FP) technique of association with positive and negative integration on it. Traditional FP method performs well but generates redundant trees resulting efficiency degrades. To achieve better efficiency in association mining, positive and negative rules generation helps out. Same concept has been applied in the proposed method.

5. REFERENCES

[1]Meera Narvekar, Shafaque Fatma Syed "An optimized algorithm for association mining using FP tree" international conference on advanced computing technologies and application (ICACTA-2015)

[2]V.Ramya, M.Ramakrishnan "Mining Association Rules Using Modified FPGrowth Algorithm" international journal for research in emerging science and technology, E-ISSN: 2349-7610

[3] Rakesh Agrawal Ramakrishnan Shrikant, "Fast Algorithms for Mining Association Rules",IBM Almaden Research Center.

[4] Honglie Yua, Jun Wenb, Hongmei Wangc ,Li Jun de "An Improved Apriori Algorithm Based On the Boolean Matrix and Hadoop", SciVerse Science Direct

[5] Mary C. Burton, Joseph B. Walther "The Value of Web Log Data in Use-Based Design and Testing" JCMC 6(3) April 2001.

[6]M.S.Chen, J.Han, and P.S. Yu. Data Mining: An Overview from Database Perspective. IEEE Yrans, on Knowledge and DataEng, 1996, Vol.8, Vol.8(No.6): 866-833.

[7] Chen Wenwei. Data warehouse and data mining tutorial [M]. Beijing: Tsinghua University Press. 2006

[8] Hand David, Mannila Heikki,Smyth Padhraic. Principles of Data Mining[M].Beijing:China Machine Press,2002

[9] J.Hipp, U.Guntzer, and G.Nakaeizadeh. Algorithms for association rule mining – a general survey and comparison. ACM SIGKDD Explorations, 2(1):58-64, June 2000.

[10] R.Agrawal and R.Srikant. Fast algorithms for mining association rules. In Proc. 1994 Int. Conf. Very Large Data Bases,pages 487-499, Santiago, Chile, September 1994.

[11] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. In: Proc. of the l993ACM on Management of Data, Washington, D.C, May 1993.