A survey of Frequent Itemsets Mining for Network Traffic

Surendra kumar chadokar, Divakar Singh, Anju Singh
BUIT-BU, Bhopal

s_chadokar@yahoo.co.in, divakar_singh@rediffmail.com,
asingh0123@rediffmail.com

ABSTRACT: *Today's digital life is based on network as the network grows the amount of data on network also increases rapidly. Due to this highly efficient algorithm is required for data mining and for accessing data from large datasets. In frequent itemsets are produced from very big or huge data sets by applying some rules or association rule mining algorithms like Apriori technique, Partition method, Pincer-Search, Incremental, Border algorithm and many more, which take larger computing time to calculate all the frequent itemsets. As the network traffic increases we need an efficient system to monitor packet analysis of network flow data. Due to this frequent itemsets mining is basic problem in field of data mining and knowledge discovery. Here in this paper a brief survey of all the techniques related to frequent item sets generation has been given.*

Key words: Genetic Algorithm, Apriori Algorithm, frequent itemsets, Association rule mining.

1. INTRODUCTION

As the network activity enhanced day by day and huge amount of data have been collected routinely from supervision in business, banking, administration, the delivery of social and health services, security, environmental protection and in politics. This type of data widely used for analysis purpose for management of the customer base activities. Traditionally, such types of data sets are very large and constantly growing and contain a large number of complex features for analysis purpose. Network traffic data enriched with detailed information of Internet usage for user, which gives information about a user, accesses a site at a time specifically. Thus, data mining on network traffic data has the problem of cooperating privacy of network users [1]. Data mining has concerned about a great deal of attention in the information industry and in society as entire in recent years, due to the wide preventability of large amounts of data and the forthcoming need for forming such data into useful information and knowledge. The information and knowledge expanded can be used for applications ranging from fraud detection, market analysis and to production control, customer retention and science exploration [2].

Frequent itemsets mining is a core component of data mining and variations of association analysis, like association-rule mining and sequential-pattern mining. Extraction of frequent itemsets is a core step in many association analysis techniques. An itemset is known as frequent if it presents in a large-enough portion of the dataset. This frequent occurrence of item is expressed in terms of the support count. Therefore, it needs complicated techniques for hiding or reforming users' private information during a data gathering process. Moreover, these techniques should not surrender the correctness of mining results [1].

The frequent itemsets are patterns or items like itemsets, substructures, or subsequences that come out in a data set frequently or rapidly. For example some common words or information that repeated frequently in a data set can be treated as frequent itemset for that data set. A subsequence, such as buying a digital camera, followed by Akash tablet and then a memory card, if it occurs regularly in a shopping database. It is known as (frequent) sequential pattern. Similarly substructure is referring to dissimilar structural forms, like sub-trees, sub-graphs or sub-lattices, which may be jointed with itemsets or subsequences. If a substructure occurs recurrently, it is called a (frequent) structured pattern. Discovery such frequent pattern plays an important role in mining relations, correlations, and many other appealing relationships along with data. Additionally, it helps in data clustering, classification, and other data mining tasks as well [2].

The process of finding out attractive and unforeseen rules from big or huge data sets is called as association rule mining. It is an *implication* or *if-then-rule* which is maintained by data. The association rule problem [3] was first invented and was broadly used in analysis in super markets, called the problem of *market-basket*. The preliminary problem was the following: given that a set of items and a huge collection of sales records, which consists of transaction date, the items that bought in the transaction, the job is to find relationships between the items contained in the different transactions [4].

The main mining algorithm based on association rule, Apriori not only predisposed the association rule mining community, but it pretentious other data mining fields as well. Genetic Algorithms are a family of computational models inspired by advancement. These algorithms are predetermine a prospective solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures as to defend significant information. Genetic algorithms are habitually inspected as function optimizer, although the ranges of problems to which genetic algorithms have been applied are quite broad. In this paper we are proposing Apriori and Genetic Algorithm (GA) based frequent itemsets mining for network traffic.

Genetic Algorithm

Genetic algorithms are one of the best ways to solve a problem for which little is known. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem. Based on efficiency of genetic algorithm it can be applicable for search the best output of a digital circuit. Genetic algorithm is used as a search algorithm, which is an efficient and cost effective. It has various applications. Genetic Algorithms are a unit of computational models enthused by progress. The algorithm encode a potential solution to a definite problem on a simple chromosome-like data structure and apply recombination operators to these structures as to protect critical information. Genetic algorithms are often viewed as function optimizer, although the range of problem to which genetic algorithms have been applied are quite wide.

It is an adaptive heuristic search based on evolutionary idea of natural selection and inheritance. Method based on population genetics. Genetic algorithm were introduced by John Holland in the early 1970s [12] .Genetic algorithm is a probabilistic search algorithm based on the mechanics of natural selection and natural genetics. Genetic algorithm is started with a set of solutions [2].

(1) Initialization: The first process decides initial genotype, namely value and genetic length. Fig.1. shows the basic steps taken by the genetic algorithm.

(2) Evaluation: The second process calculates the fitness for each individual with the target function. The evaluation depends on each problem.

(3) Termination Judgment: If the process satisfies the termination condition, the operation finishes and output the individual with the best fitness as the optimized solution.

(4) Selection: To generate the children, this process chooses parents from individuals. For example, if we assume parents the first generation, children become the second generation. The children generate the next children again. The children inherited the characteristic of the parents are generated in this way.

(5) Crossover: This process crosses individuals chosen by selection operation and generates the individuals of the next generation.

(6) Mutation: This process mutates the chromosome of new generation. The mutation is effective to escape from a local optimum solution.

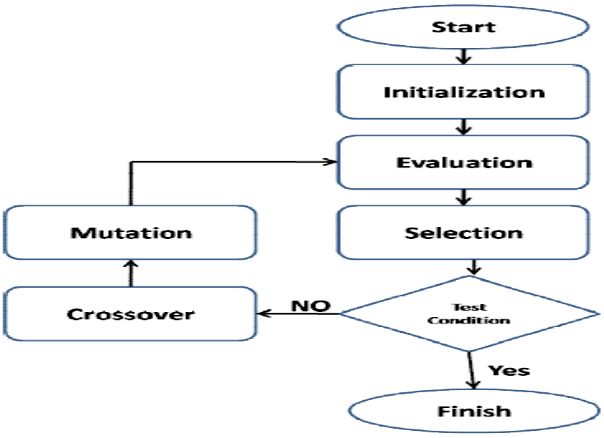

Figure 1.Flow Diagram of GA

Association Rule Mining

the problem of association rule mining is defined as: Let I1={I1,I2,I3,…….In} be a set of n binary attributes called *items*. Let D={t1,t2,t3,…..tm} be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A *rule* is defined as an implication of the form $X \rightarrow Y$ where X, YCI and $X \cap Y = \emptyset$. The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

2. RELATED WORK

Xin Li et al [1] proposed Frequent Itemsets Mining in Network Traffic Data. They think about the problem of frequent itemset mining problem in network traffic data, and propose an algorithm for mining frequent itemsets. They try to minimize the size of results and only maximal frequent itemsets are considered. To protect the privacy, intermediate mining results are encrypted using hashing method by different servers. The proposed algorithm is evaluated from the perspectives of accuracy and efficiency.

Mining of frequent itemsets using Genetic algorithm was proposed in [2]. This work carried out with logic of GA to improve the scenario of frequents itemsets data mining using association rule mining. The main benefit of using GA in frequent itemsets mining is to perform global search with less time complexity. This scheme gives better results in huge or larger data set. It is also simple and efficient.

Another frequent itemsets mining approach based on genetic algorithm for non binary dataset was proposed by G. Vjiay Bhaskar et al [4]. They present an efficient algorithm for generating significant association rules among database items. GA is used to improve the scenario and system can predict about negative attributes in generated rules. As per results obtained this scheme is simple and efficient one. The Time complexity of the algorithm is also less and suitable for non binary data sets.

In continuation with this R. Vijaya Prakash et al [5] proposed similar method mining frequent itemsets for large data set using Genetic Algorithm. They implement frequent itemsets mining for numeric attributes also. Association rule mining is used to find relationship among attributes of database. This process was much time consuming and applied on discrete attributes. GA gives the facility of global search and minimum complexity. This algorithm avoids the necessity of discretizing apriori in attribute domain. They used an evolving algorithm to find the most appropriate amplitude of the intervals that be conventional a k-itemset, so that they have an elevated support value without being the intervals too extensive.

Sanat Jain and Swati Kabra [6] proposed Mining & Optimization of Association Rules Using Effective Algorithm. In this they work on association rules organization and frequent itemsets generation using positive and negative association rule mining. They proposed an apriori based algorithm to find valid positive and negative association rule in confidence framework or structure. As per result this algorithm is efficiently works for mining of positive and negative association rules in database and also optimize positive and negative association rule using genetic algorithm. This approach also reduces the search space and improved usability of mining rules that uses correlated coefficient to judge which association rule is used to mine.

A Frequent Pattern Tree Algorithm for Mining Association Rule Using Genetic Algorithm was proposed by K.Poornamala and R.Lawrance [7]. This algorithm is useful for larger datasets. In this genetic algorithm used to optimize larger data set. This algorithm also uses the advanced frequent pattern tree to mine the frequent itemset without producing conditional FP-tree. This was an efficient and less time consuming algorithm for mining of total possible frequent itemsets without producing conditional FP-tree in compressed tree structure form.

A new method for generating all positive and negative Association Rules was proposed in [8]. According to this apriori algorithm is used to generate all association rule that are hidden. They uses new name for negative rules like ANR, CNR and ACNR. They also modify correlation coefficient for improving results. Apriori Algorithm used for frequent itemset generation along with generation of positive rules, then NRGA algorithm was applied.

The multi-thread processing method with a Multi-Threaded Paralleled frequent item-set mining Algorithm (MTPA) was proposed by XuePing Zhang et al [8]. It is useful for enterprise human resources management system. It defines Hash allocation strategy for the number of threads to ensure the reasonable distribution. According to result obtained the time complexity of MTPA was superior to FP-tree algorithm. Using the multi core processing the Superiority of MTPA algorithm will enhance the efficiency of the frequent itemsets mining.

An efficient algorithm Weighted Support Frequent Itemsets (WSFI) was used to mine with normalized weight over data streams [10]. They use a new tree structure known as Weighted Support FP-Tree (WSFP-Tree) for storing crucial information in compressed form about frequent itemsets. In this approach they try to allow users to identify weight of each itemset then determine useful knowledge from data stream using one time scan using weighted support. Weighted function and weighted table are used to efficiently discover the frequent itemsets mining in single or one time scan. This algorithm was suitable for mining of frequent itemsets from stream of datasets effectively in terms of memory and run time effectiveness. It also contribute to execute frequents by generating constraint candidate itemsets effectively.

A genetic algorithm based optimized association rule mining was proposed by Anandhavalli et al [11]. In this algorithm negative occurrence of itemsets are used that are not considered in earlier algorithms like priori, incremental, border algorithm, pincer-search algorithm etc. GA algorithm is used to identify negative attributes generated by rules among more attribute in consequent part. This algorithm obtains possible optimized rules from given data set by means of genetic algorithm. The Apriori association rule mining is used to generate frequent dataset in this scheme. The positive attributes of frequent itemsets generated by genetic algorithm and subsequent part contains single attributes in negation of attributes. This algorithm gives better result for optimized rules.

3. WORKING METHODOLOGY

The work that we are implemented here is a combinatorial method of generating association rules using apriori and genetic algorithm. Here the network Data set is passed to the apriori algorithm and then the result of the frequent sets is passed to the genetic algorithm to generate less rules.

APRIORI ALGORITHM

- Generate all frequent item sets
- All item sets with min support
- Generate all confident ARs from frequent item sets
- Downward Closure Property

GENERATE FREQUENT ITEM SETS

- Count supports of each individual item
- Create a set F with all individual items with min support
- Creates "Candidate Set" C[k] based on F[k-1].
- Check each element c in C[k] to see if it meets min support
- Return set of all frequent item sets.

GENERATE CANDIDATE SETS

- Create two sets differing only in the last element, based on some seed set
- Join those item sets into c
- Compare each subset s of c to F[k-1]- if s is not in F[k-1], delete it.
- Return final candidate set

RULE GENERATE

- Take Frequent Item Set F
- If {F[1], F[2],...F[k-1]} => {F[k]}meets some min confidence, make it a rule.
- Remove last element from antecedent, insert into consequent, check again.

Though the Apriori principle allows us to considerably reduce the search space, the technique still requires a huge computation, particularly for large databases

This research proposes an approach for finding fuzzy sets for quantitative attributes in a database by using clustering techniques and then employs techniques for mining of fuzzy Associate rules.

GENETIC ALGORITHM

Here we are implementing Genetic algorithms to reduce the frequent set for the generation of association rules.

```
Genetic Algorithm()
{
  Initialize population;
  Evaluate the initial population;
  For all population
  {
    If( Test Condition=True)
    {
      Search Element Found
    }
    Else
    {
      Apply CrossOver();
      And Mutation();
    }
  }
```

4. CONCLUSION

In this paper a survey of different techniques used for the generation of frequent item sets for the detection of network traffic has been given. The technique used for the association rules for the network traffic not only presents a whole scenario about the efficiency of the algorithms but also these techniques provides a knowledge of how can we reduce the frequent item sets for these data.

REFERENCES

[1] Xin Li, Xuefeng Zheng, Jingchun Li, Shaojie Wang "Frequent Itemsets Mining in Network Traffic Data", 2012 Fifth International Conference on Intelligent Computation Technology and Automation, pp. 394-397, 2012.

[2] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, pp. 133 – 143, October 2010.

[3] Agrawal R., Imielinski T. and Swami A. Mining Association rules between sets of items in large databases, In the Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93), pp. 207-216, Washington, USA,1993.

[4] G. Vijay Bhasker, K. Chandra Shekar, V. Lakshmi Chaitanya "Mining Frequent Itemsets for Non Binary Data Set Using Genetic Algorithm", International Journal Of Advanced Engineering Sciences And Technologies (IJAEST), ISSN: 2230-7818, Vol. 11, Issue No. 1, pp. 143 – 152, 2011.

[5] R. Vijaya Prakash, Govardhan, S.S.V.N. Sarma "Mining Frequent Itemsets from Large Data Sets using Genetic Algorithms", IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" (AIT-2011), ISSN: 0975 – 8887, Special issue No. 4, Article -7, pp. 38- 43, 2011.

[6] Sanat Jain, Swati Kabra "Mining & Optimization of Association Rules Using Effective Algorithm", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 4, pp. 281- 285, April 2012.

[7] K.Poornamala, R.Lawrance "A Frequent Pattern Tree Algorithm for Mining Association Rule Using Genetic Algorithm", International Conference on Computing and Control Engineering (ICCCE 2012), pp. 1-7, 2012.

[8] Rupesh Dewang, Jitendra Agarwal "A New Method for Generating All Positive and Negative Association Rules", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 3 No. 4, pp. 1649- 1657, Apr 2011.

[9] XuePing Zhang, YanXia Zhu, Nan Hua "Improved Paralled Algorithm for Mining Frequent Item-set Used in HRM", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 2830-2833, 2010.

[10] Younghee Kim, Wonyoung Kim and Ungmo Kim "Mining Frequent Itemsets with Normalized Weight in Continuous Data Streams", Journal of Information Processing Systems, 2092-805X, Vol.6, No.1, pp. 79 – 90, March 2010.

[11] 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 2, pp. 1 - 4, 2009.

[12] Tsoukalas, L., and Uhrig, R. Fuzzy and Neural Approaches in Engineering, Wiley, 1997.