Intelligence Service Of Web Mining With Genetic Algorithm

Kolli Prabhakara Rao¹, G.Kalyana Chakravarthy ²

¹M.Tech Student, Dept of CSE, Sri Sai Aditya Institute of Science & Technology, Surampalem
²Asst.Professor, Dept of CSE, Sri Sai Aditya Institute of Science & Technology, Surampalem

Abstract: The Main objective of this paper is to eliminate user uninterested rules and find the optimal patterns or association rules. Association rules are important basis of describing Web user’s behavior characteristic. Traditional algorithms of Web association rule mining, based on statistics, usually pays attention to the analysis on existing data, they can't offer effective means and optimizing measure and cannot find out the latent and possible rules. In this paper we are proposing an efficient web association rule mining approach with Apriori algorithm and genetic algorithm features like cross over and mutation for generation of the optimal patterns.

I. INTRODUCTION

Data Mining is one of the fastest growing research areas for Knowledge discovery. In Knowledge discovery Association Rule Mining plays a vital role. Association Rules Mining is an important research topic among the various data mining problems. Association rules have been extensively studied in the literature for their usefulness in many application domains such as market basket analysis, recommender systems, and diagnosis decisions support, telecommunication, intrusion detection, and etc [1][2].

Mining Association rules is not full of reward until it can be utilized to improve decision-making process of an organization. When mining association rules, we adopt another minimum support threshold to mine frequent item sets. With a correlation coefficient measure and pruning strategies, the algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. The complexity and large size of rules generated after mining have motivated researchers and practitioners to optimize the rule, for analysis purpose. All the traditional association rule mining algorithms were developed to find positive associations between itemsets. Several algorithms have been developed to cope with the popular and computationally expensive task of association rule mining. With the increasing use and development of data mining techniques and tools, much work has recently focused on finding negative patterns, which can provide valuable information [3][4]. However, mining negative association rules is a difficult task, due to the fact that there are essential differences between positive and negative association rule mining.

The various application areas in which association rules can be applied for extracting useful information from the huge dataset are:

A) Market basket analysis

A typical and widely-used example of association rule mining is market basket analysis. For example, data are collected using bar-code scanners in supermarkets. Such ‘market basket’ databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together [6][7]. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns.

For example, if a supermarket database has 100,000 point-of-sale transactions, out of which 2,000 include both items A and B and 800 of these include item C, the association rule "If A and B are purchased then C is purchased on the same trip" has a support of 800 transactions (alternatively 0.8% = 800/100,000) and a confidence of 40% (=800/2,000).

One way to think of support is that it is the probability that a randomly selected transaction from the database will contain all items in the antecedent and the consequent, whereas the confidence is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent [6].

Now days every product comes with bar code. The software supporting these barcode based purchasing/ordering systems produces vast amounts of sales data, typically captured in “baskets” (records in which the items purchased by a given consumer at a given time are grouped together). This data was quickly recognized by the business world as having immense potential value in marketing. In particular, commercial organizations are interested in discovering “association rules” that identify patterns of purchases, such that the presence of one item in a basket will imply the presence of one or more additional items. This “market basket analysis” result can then be used to suggest combinations of products for special promotions or sales, devise a more effective store layout, and give insight into brand loyalty and co-branding.[8][10]
Market basket can be defined as collection of items purchased by a customer in a single transaction (e.g. supermarket, web) Association rules are used for pattern discovery, each rule has form: A->B, purchase whole wheat bread.”

Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B, i.e.

Support = Probability (A and B)
Support = (# of transactions involving A and B) / (total number of transactions).

Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A, i.e. Confidence = Probability (B if A) = P (B/A)

II. RELATED WORK

Typically association rule algorithms generate a huge number of rules, not all of which are truly interesting to a data analyst. Various methods have been proposed to prune the set of generated rules and discard irrelevant rules during the rule generation phase (Sahaaya & Malarvizhi, 2010). Balcazar (2010) presented a formal approach to determining a rule set basis and thus eliminate redundant rules. Although various pruning methods can significantly reduce the rule set size, it typically still remains huge [9][13].

An important issue in association rule mining is selecting the right interestingness measures (Huang & Xiangji, 2007). A recent study (Kannan & Bhaskaran, 2009) analyzes the distribution of rule clusters over various interestingness measures. A discussion of the limitations of association rule mining is given in a study by Webb (2011). These include the over-generation of rules, many of which are false discoveries, discovery of rules that are hard to understand, and limitations imposed by the minimal support constraint. The authors propose an approach called “filtered top-k association discover” that alleviates some of these problems. Association rule mining in the context of web usage data suffers from additional issues. Namely, web usage data is specific in the sense that it contains a large number of tightly correlated items (web resources or web pages) due to the link structure of a website. Web pages that are tightly linked together often co-occur in the same session, is why the generated set of association rules contains a high number of so-called “hard” association rules that have very high confidence, but are not truly interesting to the user. To alleviate this problem, additional rule pruning methods need to be applied (Wang, 2005; Dimitrijevic & Bosnjak 2010). A line of research deals with finding additional web usage association rules that may be missed during the classical association rule mining. For example, an approach that discovers association rules between web pages that rarely occur together, but with other pages, where they occur frequently, is taken by Kazienko (2009). This method is used to add new, so called “transitive” association rules to the typical association rule set, which brings new potentially useful information to a webmaster[11][12].

A) Optimization of Association Rules Using GA

In this section describes the GA algorithm for optimization of association rule associated .First, explanation of how GA algorithm represents the rule individually and encodes scheme and the chromosome structure (Representation of rule) shown. After that, description of genetic operators and fitness function assignment and selection criteria are listed. Finally, the algorithmic structure is given.

B) Genetic operator

Genetic Algorithm uses genetic operators to generate the offspring of the existing population. This section describes three operators of Genetic Algorithms that were used in GA algorithm: selection, crossover and mutation.

1) Selection: The selection operator chooses a chromosome in the current population according to the fitness function and copies it without changes into the new population.GA algorithm used route wheel selection where the fittest members of each generation are more chance to select.

2) Crossover: The crossover operator, according to a certain probability, produces two new chromosomes from two selected chromosomes by swapping segments of genes.GA algorithm used single-point crossover operation with probability 0.1.

3) Mutation: The mutation operator is used for maintaining diversity. During the mutation phase and according to mutation probability, 0.005 in GA algorithm, value of each gene in each selected chromosome is changed.

Figure 1: Single Point Crossover

III. PROPOSED SYSTEM

In this paper we are proposing an efficient web association rule mining approach with Apriori Algorithm and genetic algorithm features (Cross over and mutation) to find the optimal patterns or web association rules from the rules
which are generated by the traditional pattern mining approach. The entire architecture of the project includes Data Acquisition, Apriori Algorithm and genetic algorithm which includes cross over and mutation.

A) Data Acquisition

In this Module, Initially it loads the dataset, which contains different URLs with respect to the node details and session id details (Transaction details), to find the Unique urls in the transactions, here each individual url is acts as gene and combination of more than one url acts as chromosome for genetic approach.

B) Apriori Algorithm

Apriori is the one of the basic algorithm for finding the frequent itemsets or web association rules from the different data items. This algorithm uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as level wise search, where k itemsets are used to explore (k+1) itemsets. It generates the association rules

C) Crossover and Mutation

Genetic algorithm is one of the most efficient evolutionary algorithms for solving the problems like NP hard problem (i.e. if a problem can have more than one solution) by applying the process of crossover and mutation.

Crossover and mutation are two basic operators of GA. Performance of GA very depends on them. Type and implementation of operators depends on encoding and also on a problem.

Method of merging the genetic information of two individuals; if the coding is chosen properly, two good parents produces good children. This called Cross over.

In this approach paper Initially we consider the individual urls from the dataset and assigns unique id for each url and generates the pattern which is a combination of more than one url and then forwards the patterns to the apriori approach by specifying the minimum support count value. Our pattern mining algorithm generates the number of patterns until the frequent pattern condition terminated.

After the generation of the patterns, we consider the initial population as number of chromosomes (i.e combination of more than one url), we specify iteration details and we apply crossover and mutation. In cross over, there is a mutual exchange of the genes in the two parent chromosomes then new combination of chromosome or pattern generated and again checks the pattern with the minimum threshold value, accept the chromosomes which satisfies the minimum threshold value and ignore the others.

During the Mutation there is a change of place or order within the chromosome and followed by it checks the minimum threshold value, if it satisfies the minimum threshold value and ignore the others.

IV. CONCLUSION

The result of analyzing show that we can realize the optimizing and forecasting for exist and latent association rules, by utilizing GA to mine the statistic mining result, and it surely can compensates halfway results that by the no sufficient visiting amount of web users on one certain class swarm.

The final results of the proposed system shows the optimal patterns, by solving the problem of NP hard problem by eliminating the unnecessary patterns and retrieving the interesting patterns from the frequent patterns which are generated by the Apriori Algorithm approach.

References


BIOGRAPHIES

Mr.G.Kalyana Chakravarthy is working as an Asst.Professor in Sri Sai Aditya Institute of Science & Technology, Surampalem,Andhra Pradesh. He has received his M.Tech Post graduation in Vignan's Engineering. College, Guntur from JNTU Kakinada University.

Kolli PrabhakaraRao pursuing M.Tech, department of CSE, in Sri Sai Aditya Institute of Science & Technology, Surampalem. His interested area of research is data mining.