Finding Improved Frequent Data Query Sets using Genetic Algorithm
Shankey Gupta1 Ms. Amrita Chaudhary2
1P.G Student
2Department of Computer Science & Engineering
Karnal, Haryana, India

Abstract— In the recent years, several methods are proposed for mining frequent data query sets, but almost all of them suffer from the problems like generating large number of candidate generation and large number of database scans. The proposed approach uses Genetic Algorithm to produce offsprings (under certain fitness test or conditions) to become frequent pattern and then these offsprings become ancestor for next query set. The proposed approach requires only one scan of the database and also this approach pruned the useless individuals or candidates in less time. Experimental results shows that the proposed approach worked better than existing approach using MATLAB programming.

Key words: Data Mining, Web Mining, Apriori Algorithm, Genetic Algorithm (GA)

I. INTRODUCTION

A. Data Mining

mining of knowledge from data. Data Mining[21][22] means extracting useful information from the huge set of data. The raw data is of no use until converted into useful information. Data Mining is defined as extracting the information from the huge set of data.

B. Web Data Mining

Web Data Mining is the application of data mining techniques to Web data. Web mining[20] is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data. The process may involve pre-processing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data mining operations.

Web content mining involves efficiently extracting useful and relevant information from different web sites and databases. Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the web pages.

C. Frequent Dataset Mining

Frequent Dataset Mining is the process of pattern mining. Frequent dataset mining generates the data sets that are frequently occurred in the system. It provides pattern of interesting data sets and helps in the user behaviour. Frequent item sets are sets of item that are frequent occurred in the datasets.

There are various traditional approaches used in mining of frequent item-sets. One of the approaches is Apriori algorithm[15].

In this approach, association rule is applied on the datasets. The support value(s) of each item in the dataset is calculated if the value satisfies the minimum threshold value (c) then that item will be added in the frequent item-set. It is a level by level approach. At each level, numbers of candidate item-sets get increased. It is a bottom-up approach where items in frequent subsets are increased by one at each level which is known as process of candidate generation. At each level, it generates the candidate item-sets of length k from the k-1 length of item-sets and after pruning of candidate item-set which is done by eliminating the infrequent item-sets. Scanning of dataset is done at each level for pruning process.

D. Association Rules

Association rule[15] is one of the most widely used data mining concepts. The goal of an association rule mining algorithm is to discover associations between data items. It is classically defined as:

Let I be a set of items i1, i2, i3, . . . in. Let T = t1, t2, . . . tm be a set of transactions. Each transaction t in T has unique id and contains a subset of items in I.

An association rule implies the following X => Y; where X is a subset of I, Y is a subset of I, and X∩Y =Ø.

Examples of association mining applications are market basket analysis, medical diagnosis and research, web site navigation analysis and home and security. An association rule is illustrated in example:

Bread → Cheese (support = 10%, confidence = 90%)

This rule says that 10% of customers bought bread and cheese together and those who bought bread, also bought cheese 90% of the time. Support and confidence are two important measurements association rule mining.

Support and confidence are defined as follows:

E. Support

The support of a rule, X →Y, is the percentage of transactions in T that contains XUY, and can be seen as an estimate of the probability P(X U Y). The rule support thus determines how frequent the rule is applicable in the transaction set T.

Let n be the number of transactions in T.
The support of the rule X →Y is computed as follows:

\[
\text{support} = \frac{\text{count}(X \cup Y)}{n} \quad (1.1)
\]

F. Confidence

The confidence of a rule, X →Y, is the percentage of transactions in T that contain X also contain Y. It can be seen as an estimate of the conditional probability, P(Y | X). It is computed as follows:
G. Genetic Algorithm (GA):
The idea with Genetic Algorithm[16][19] use power of evolution to solve optimization problems. The father of the original Genetic Algorithm was John Holland who invented it in the early 1970's. The basic techniques of the GAs follows the principles lay down by Charles Darwin "survival of the fittest".

In a genetic algorithm, a population of candidate solutions (called individuals, creatures, or phenotypes) of an optimization problem is evolved toward better solutions.

Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered. Solutions are represented in binary as strings of 0s and 1s, and other encodings are also possible.

The steps to be followed are:
1) Selection: Individual solutions are selected through a fitness-based process, where fitter solutions more likely to be selected.

2) Genetic Operators:
The next step is to generate a second generation population of solutions from those selected through a combination of genetic operators: crossover (also called recombination), and mutation.

1) Initial population/Reproduction: On the basis of minimum fitness value, the population will be selected for reproduction.

2) Crossover: Select two individuals randomly or on the basis of fitness value. Apply Crossover on these two selected individuals to produce offspring.

3) Mutation: Define probability of mutation (inverted bit of offspring) to improve fitness of individuals.

Repeat Crossover and Mutation until the complete solution has been obtained.

3) Termination:
Process is repeated until a termination condition has been reached.

GA process is a three step searching goal which is described using the flowchart:

![Flowchart of Genetic Algorithm](image)

Fig. 1: Flowchart of Genetic Algorithm
II. LITERATURE SURVEY

Naili Liu, Lei Ma[1] proposed the improved algorithm, which constructs the directed graph and generate candidate item sets by using the directed neighbor nodes set, the algorithm need traverse the directed graph only once. The algorithm verifies whether a candidate itemset is a frequent itemset by logic AND operation.

Kamika Chaudhary, Santosh Kumar Gupta[2] proposed a new method for prioritizing the web pages based on web usage and web content data. The proposed method uses Genetic Algorithm for providing good quality web pages as a result of user query. The method includes the parameters from both web usage and web content mining.

R. Vijaya Prakash, Dr. Govardhan, Dr. S.S.V.N. Sharma [3] have proposed the application of Genetic Algorithm for improvement of the generation of Frequent Item set with numeric attributes instead of binary or discrete attributes. They have claimed that this approach will be advantageous in the discovery of frequent item sets during global search with relatively less time due to greedy approach.

Xiaowei Yan et al. [4] designed a genetic algorithm based strategy for identifying association rules without specifying actual minimum support their work was based on elaborate encoding method and for fitness function they used relative confidence.

T. Sunil kumar, Dr. K. Suvarchala[5] aim is to explore the role of data mining for information extraction in web content, structure and usages mining in current web models, and the outlines the process of extracting patterns from data.

Pratima Gautam, K. R. Pardasani[8] presented an efficient version of Apriori algorithm for mining multi-level association rules in large databases to finding maximum frequent itemset at lower level of abstraction. She proposed a new, fast and an efficient algorithm (SC-BF Multilevel) with single scan of database for mining complete frequent item sets.

Abhijit Raorane, R.V.Kulkarni[11] objective is to know consumer behaviour, and to know consumer psychological condition at the time of purchase and how suitable data mining method apply to improve conventional method.

Arvind Jaiswal, Gaurav Dubey[12] proposed a different approach finding frequent item sets. Frequent item sets are generated using the Apriori association rule mining algorithm. Then genetic algorithm has been applied on the generated frequent item sets to generate the rules containing positive attributes, the negation of the attributes with the consequent part consists of single attribute and more than one attribute.

III. ALGORITHM

The proposed algorithm is described below:

Consider Sessions denoted as $S = S_1, S_2, ..., S_n$ and Queries are denoted as $Q = Q_1, Q_2, ..., Q_m$.

Let $Q_k$ is the chromosome in the population $P$, $n$ is the number of sessions and $f(i)$ is the fitness value function.

Following steps have been given in the algorithm of the proposed approach:

1. First of all binary encoding is performed to generate chromosome or individual on the sessions and queries so apply the encoding algorithm.

   1. Let Sessions be $S = S_1, S_2, ..., S_n$ and queries be $Q = Q_1, Q_2, ..., Q_m$.
   2. Begin
   3. For each $Q_k \in Q$.
   4. For $i = 1$ to $n$.
   5. Then $Q_k[i] = 1$.
   6. Else $Q_k[i] = 0$.
   7. Return $Q$.
   8. End;

2. Calculate fitness value (i.e. probability support value) of each individual by applying algorithm given

   1. Let $Q_k$ be the chromosome in the population $P$, $n$ is the number of sessions and $f(i)$ is the function fitness value.
   2. Begin
   3. count = 0;
   4. For $i = 1$ to $n$.
   5. If ($Q_k[i] = 1$).
   6. count = count + 1;
   7. $f(i) = \frac{count}{n} \times 100$, //support count probability
   8. return $f(i)$;
   9. End;

Then check the fitness condition if it satisfies the fitness condition then the individual is added in the initial population(P). All the chromosome in this initial population will added in frequent 1-query sets population.

3. Apply crossover algorithm as given i.e. select the two chromosome from the population and perform arithmetic AND operation to generate new chromosome where former two chromosome will be called parents and later one called offspring.

   1. Let $Q_k$ and $Q_{k+1}$ are the two chromosomes selected from population $P$, $L$ is new offspring and $n$ is the number of sessions.
   2. Begin
   3. For $i = 1$ to $n$.
   4. if ($Q_k[i] = Q_{k+1}[i]$).
   5. $L[i] = 1$;
   6. else
   7. $L[i] = 0$;
   8. return $L$;
   9. End;
4) Inspect the mutation by checking the inversion of the bit of the chromosome in the population. Select any individual or chromosome from the population then randomly invert one of it bit value.

5) Apply fitness calculation algorithm on the offspring and then check the fitness of the offspring. If it satisfies the fitness condition then add it to the population.

6) The parent or parent of the siblings of this fit offspring will be added in the frequent query sets. Also the offspring itself is added in the frequent query sets.

7) Repeat step 3, 4 and 6; until all the frequent query sets are generated or no more offspring is added in the population.

IV. SIMULATION RESULT

This paper realized DLG algorithm, Hybrid-DLG algorithm, improved algorithm and Proposed GA algorithm, in order to verify the efficiency of the Proposed GA algorithm.

The experimental results obtained in the different support are shown in Figure 2.

![Performance Comparison of four algorithms in different minimum support.](image)

The Proposed GA algorithm has many advantages as compared to other algorithms:

1) The Proposed GA algorithm constructed candidate frequent query-sets by using the arithmetic crossover. Therefore, there is no need for the data scan during offspring generation, but improved algorithm construct the directed graph in order to generate neighbour nodes. So, Proposed GA has obvious time advantage.

2) The Proposed GA algorithm generates candidate itemsets, which is operated separately on the individual processor to perform AND operation and also simple fitness function is used to generate the population which take less time for execution than the traditional approaches. So, it reduces time overhead.

V. CONCLUSION

There are several drawbacks in finding frequent item-sets in various traditional data mining algorithms like Apriori[15], Improved Apriori algorithm[13], Partition algorithm, Hybrid-Apriori algorithm[18] etc. In proposed work we are using the Genetic Algorithm and its operators to overcome the drawbacks of traditional approaches.

From the result the conclusions are:

1) The number of data scans is reduced by using genetic approach. As it needs only one dataset scan and during scanning the datasets are converted into chromosome.

2) Candidate generation is reduced by using genetic approach. As only the offspring that satisfies the fitness condition are added in the population.

VI. FUTURE WORK

In the future, the running time can be further improved by improving the data structure (Hash Table, Tree such as bitmaps, Binary search tree) for storing the item sets.

Also, this algorithm will be further extended to work on a large data set where itemset are in a range starting from at least 1 million by avoiding the control over random number generation.

REFERENCES

[1] Naili Liu, Lei Ma, “Discovering Frequent Itemsets an Improved Algorithm of Directed Graph And Array”, 978-1-4673-5000-6/13/$31.00 ©2013 IEEE.


[7] Juan Vel´squez, Hiroshi Yasuda and Terumasa Aoki, “Combining the web content and usage mining to understand the visitor behavior in a web site” Proceedings of the Third IEEE International Conference on Data Mining (ICDM’03) 2003 IEEE.


All rights reserved by www.ijsrd.com