A Review on Finding Improved Frequent Data Query Sets using Genetic Algorithm

Shankey Gupta[1] Amrita Chaudhary[2]
[1]P.G Student
[1,2]Department of Computer Science and Engineering
[1,2]DIET, 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. This paper is about the related work for frequent data set mining algorithms and how to overcome the drawbacks of previously constructed algorithms.

Key words: Data Mining, Web Mining, Apriori Algorithm, Genetic Algorithm

I. Introduction

A. *Data Mining:*

Mining of knowledge from data. Data Mining 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:*

is the application of data mining techniques to Web data. Web mining 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:*

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. 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 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 $i_1, i_2, i_3. . . i_n$. Let T = $t_{1,}\ t_2. . . t_{m.}$ 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 =∅.

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:

1) *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 $P(X \cup Y)$.

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:

$$\text{support} = \frac{(X\cup Y).\text{count}}{n} \qquad (1.1)$$

2) *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:

$$\text{confidence} = \frac{(X\cup Y).\text{count}}{X.\text{count}} \qquad (1.2)$$

II. Related Work

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.

Kamika Chaudhary, Santosh Kumar Gupta [2] have 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.

Jiao Yabing [13] proposed an improved algorithm of association rules, the classical Apriori algorithm. Finally, the improved algorithm is verified, the results show that the improved algorithm is reasonable and effective, can extract more value information.

Charanjeet Kaur [14] presented a survey of recent research work carried by different researchers.

III. Problem Formulation

To find out frequent datasets existing methods suffers the problem of huge candidate generation and number of database scans.

The genetic approach is proposed for frequent pattern mining in web content database using various genetic operators (Reproduction, Crossover, mutation). GA variables are represented by chromosomes.

The main objective is to reduce the large number of candidate generation and large number of database scans which are drawbacks of various previous algorithms like Apriori algorithm, Partition algorithm, Border algorithm, Pincer-search algorithm, FP-tree growth algorithm etc. There are continuous research going on finding the simple and efficient algorithm for mining frequent dataset for easy implementation and efficient results.

IV. Proposed Work

Our aim of research would be to study the existing methods and to develop a new method for finding improved data query set using genetic algorithm operators (crossover, mutation) and then check results of our method using MATLAB. We would also compare our proposed method with the existing methods.

V. Conclusion and Future Scope

Data mining field involves the study of various techniques to discover and model hidden patterns in large volumes of raw data. Various traditional approaches suffer from many drawbacks. So, need for other algorithm which has better performance than previous algorithms.

In the future, the focus is on improving the running time of traditional algorithms by improving the data structure. Also, previous algorithms can further be improved by reducing number of candidates generation and reducing the data scans.

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.

[2] Kamika Chaudhary, Santosh Kumar Gupta, “Prioritizing Web Links Based on Web Usage and Content Data”, 978-1-4799-2900-9/14/$3l.00 ©2014 IEEE.

[3] R. Vijaya Prakash, Dr. Govardhan, Dr. 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”, pp:38-43, 2011.

[4] Xiaowei Yan , Chengqi Zhang , Shichao Zhang, “Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support”, pp:3066–3076, 36, 2009.

[5] T. Sunil kumar, Dr. K. Suvarchala, “A Study: Web Data Mining Challenges and Application for Information Extraction”, IOSR Journal of Computer Engineering (IOSRJCE), ISSN: 2278-0661, ISBN: 2278-8727Volume 7, Issue 3 (Nov. - Dec. 2012), PP 24-29.

[6] Gaurav Shelke, Anuraj Jain, Shubha Dubey, “A Survey of Anomaly Detection using Frequent Item Sets”, International Journal of Computer Applications Technology and Research, Volume 2– Issue 3, 378 - 381, 2013.

[7] Juan Vel´asquez, 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.

[8] Pratima Gautam, K. R. Pardasani, “Algorithm for efficient multilevel association rule mining”, Internal

Journal on Computer Science and Enginnering, Vol. 02, N0.05, pp:1700-1704, 2010.

[9] G. Vijay Bhasker, K. Chandra Shekar, V. Lakshmi Chaitanya, “Mining Frequent Item sets for Non Binary Data Set Using Genetic Algorithm”, pp:143 – 152, 11(1), 2011.

[10] D.S Rajput, R.S Thakur, G.S Thakur, “Karnaugh Map Approach for Mining Frequent Termset from Uncertain Textual Data”, British Journal of Mathematics and Computer Science 4(3), XX–XX, 2014.

[11] Abhijit Raorane, R.V.Kulkarni, “Data Mining Techniques: A Source for Consumer Behavior Analysis”, IJDMS,Vol.3, No.3, August 2011.

[12] Arvind Jaiswal, Gaurav Dubey, “Identifying Best Association Rules and Their Optimization Using Genetic Algorithm”, International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[13] Jiao Yabing, “Research of an Improved Apriori Algorithm in Data Mining Association Rules”, International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

[14] Charanjeet Kaur, “Association Rule Mining using Apriori Algorithm: A Survey”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 2, Issue 6, June 2013.

[15] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, “Mining Association Rules between Sets of Items in Large Databses”, Proceedings of the 1993 ACM SIGMOD Conference Washigton DC, USA, May 1993.

[16] S.N. Sivanandam, S.N. Deepa, “Introduction to genetic Algorithm”.

[17] Raymond Kosala, Hendrik Blockeel, “Web Mining Research: A Survey”, S/0011033V1[cs.LG], 22 Nov 2000.