Comparative Analysis of Frequent Pattern Mining Algorithm on Water Quality Data

[1]Ms. P Mahalakshmi*, [2]Ansh Kaul, [3]Yash Aggarwal

[1,2,3]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, India-603 203.

[1]mahalakshmi.p@ktr.srmuniv.ac.in, [2]anshkaul98@gmail.com, [3]yashbindal652@gmail.com

Abstract

Every day humans are generating very large volumes of data. In fact 2.5 Exabytes(10^{18}) are generated every day. Therefore many data mining methods were proposed in order to achieve better execution and time complexity for such humongous volumes of data. One such type of algorithms consists of extracting the most frequent going on patterns in the transactional databases. Such algorithms are called association mining algorithms. Dependency in transactions between the time and location in addition increases the complexity of the frequent object set mining task. The proposed work aims to recognize and extract the common styles from such transactional information. The dependency of water data on various factors is used to discover regularly co-occurring pollution over numerous water bodies, in different states of India. These consist of Nitrates and Coliform. Other factors which influence the quality of water are pH,Temperature,Biochemical Oxygen Demand (BOD) and Dissolved Oxygen(DO).Various strategies have been proposed to mine frequently occurring patterns quickly and accurately. However our work promotes a general hash based methodology which may be carried out to any numerical information, which also includes water quality data. This method is called hash based apriori algorithm [1]. This is a modification of the traditional apriori algorithm. Also, an assessment with respect to the FP growth algorithm is shown in terms of execution time.

Keywords: *Water quality, association mining, hash based apriori algorithm, Frequent pattern growth algorithm.*

1. Introduction

The web generates full-size volumes of records via sources which include business enterprises, social media platforms, and transactional databases to name a few. There are various data mining methods which are role-playing major positions in processing, analyzing and such facts in order to mine more useful statistics. Therefore, an important role is portrayed by data mining in a large number of situations. These consist of market basket analysis,[8] risk analysis and management and other such applications. However, the actual data-sets tend to be very complex and have a large amount of dimensions. The need for efficient mining of data cannot be ignored.

For instance, there are datasets with a large number of dimensions that generate transactional data.This kind of data consists of numeric data which must be converted to categorical values so that very strong association rules are generated. In other words association rule mining is required in such cases. We also have to determine that the association mining algorithms are efficient and take very less time in builb ding the association rules. Thus a comparative analysis is done which involves comparing the hash based Apriori algorithm[1] and the FP Growth algorithm.

There are a vast variety of pollutants in the water bodies of India. In this case we are considering the major pollutants and factors which influence the water quality of the rivers of India. These include Nitrates and Fecal Coliform. Other factors which influence the water quality are Biochemical Oxygen Demand, Dissolved Oxygen, pH and Temperature.

A correlation matrix is also generated which shows the relationship among the various variables. It also becomes clear which variables tend to influence the other variables. It also helps to isolate the variables which are unrelated as well.

Water pollutant quantities are growing day by using day. Extremely rising water pollutants degrees within the urban and rural areas are the most primary environmental challenges in the current times. Therefore the proposed algorithm is used in identifying and isolating the most pattern of pollutants. The data comprised of attributes consisting of pollutant concentrations, area and city. All of this information combined with the factors that influence water quality makes the dataset useful for mining.

2. Related Work

Association Mining consists of two major types of algorithms [2] :-Apriori and FP Growth. The Apriori algorithm generates candidate sets and determines the item-set which is frequent. It states, if the item-set is found to be more frequent, then most subset’s are considered frequent consequently. FP Growth algorithm does not use candidate item sets. Instead it will use a pattern fragmented growth in order to get the most frequent pattern present in the huge data repository. An extended tree is used to store information about such frequent patterns.

Jong Soo Park and others [3] used hash based Apriori for association mining in 1997.Here the Apriori algorithm was used. Hashing is used in order to improve the functions of Apriori algorithm. It helps to generating a smaller number of candidate

2-itemsets. In fact the order of magnitude is also reduced. The amount of disk I/O is also significantly reduced.

J. Mennis and J.W Liu[4] had proposed to apply association mining on spatio-temporal data, in 2003. They used Geographic Information System based data preprocessing to integrate different datasets, extract relationships of spatio-temporal nature, encode them into tabular form and also classified the numeric data into the categorical data. The Multiple-level of association- rule mining are supported as well.

The FP-growth algorithm generating a large amount of conditional patterns based and FP tree for a huge dataset. This leads to a large amount of memory usage and cost. Therefore Min Chen and others [5] had proposed the parallel FP-growth algorithm , in 2009. It works independently at each node of the FP growth tree, consequently reducing the inter-node communication cost. This helps to improve memory usage, cost, efficiency and scalability of the FP-growth algorithm.

Rui Chang and Zhiyi Liu [6] had proposed an optimization algorithm, in 2011. They called it APRIORI-IMPROVE. This algorithm optimized transaction compression and 2- items generation. It used horizontal data representation and a better storage technique.In fact the performance study had showed that it worked better than the conventional Apriori algorithm.

Surbhi K.Solanki and Jalpa T. Patel[7] had tested algorithms based on association mining, in 2015. These included Apriori , FP Growth ,Dynamic FP approach,Fuzzy FP Growth and others. They had concluded that the Fuzzy FP Growth algorithm had provided the best results.

Ilham Huseyinov and others[8] had proposed to use the Eclat algorithm in order to identify buying patterns of customers, in 2017.They tested the performance of this algorithm against the conventional apriori algorithm and achieved better results.The implementation was done using the programming language, R.

Apeksha Aggarwal and Durga Toshniwal[9] had tested the Spatio Temporal Apriori algorithm on Web data. They also used direct address based hashing technique in order to access data as fast as possible. The algorithm yielded promising results.

The proposed system consists of a hash based Apriori algorithm. In order to justify that the algorithm is fast, its execution time is compared to the FP-Growth algorithm.This algorithm took lesser time to generate the candidate itemsets and the association rules. Apart from the traditional Apriori , this algorithm also consists of hashing using key-value pairs. This improves the efficiency of the algorithm to a large extent. FP Growth algorithm takes lesser memory space and lesser execution time compared to the traditional Apriori. This is why we are using hashing in order to improve the efficiency of Apriori algorithm. Thus a small number of strong association rules are generated after executing the algorithm.

3. Methodology

Hash Based Apriori Algorithm

HBA algorithm execution, utilizes the data- structures that uses a hash based table.This way it will enhance the execution time and the calculation will utilize the hash based procedure in process to minimize the quantity of item sets created in 1st pass. It was assured that the increments in item-sets that can be lessened by using hashing,and the output required will be more efficient. For a instance, while takingeach and every transaction occurred in a databaseto create the Frequent item-set,from the given candidate item sets ,we can generate the major item-set for each and every transaction, in hashmap into the bucket list of a hash table.

HBA Algorithm:

Step 1- Scan for all of the transactions to create all the possible item sets.

Step 2- Assume the Hash table be size 10.

Step 3-Every bucket must be assigned with a candidate pairs ordered by ASCII value in the item sets.

Step 4- Every bucket within the hash-table contains a count, that is increasing by one fir each item and an item set is being hashed within the bucket.

Step 5- If a count in a bucket will be equal or greater than the minimum support count,and the bit vector will be considered as 1 or in any other situation as 0.

Step 6- Candidate pairs will be hash to the location wherever the bit vector isn't set will be removed.

Step 7- Modifying transactional database will incorporate these candidate pairs.

FP Growth Algorithm

FP growth can be considered as an upgraded version of the Apriori algorithm. The frequent patterns are generated without candidate item-set generation. FP growth is used to represent the database in structure of a tree known a frequent pattern tree.

The tree structure is used to maintain the association within the item-set. The database are fragmented by 1-frequent item called “Pattern Fragment”. The item-set in these fragmented patterns will be analyzed. So in short, searching for a frequent item-set will be reduced.

FP tree is a tree look-like structure, it was created with the help of initial item-set withthe dataset. The requirement of FP tree was to mine most number of frequent patterns. Each node on the FP tree represented the item in the item-set.The root-node will be shown as null, and the lower-nodes will represnt the item-set. The association with the node to the lower-nodes which is a item-set with the other item-sets are maintained during the creation of a tree.

FP Growth Algorithm

Step 1-Reducing the number order of frequent-item. Items which are in similar frequency, and order was provided in the alphabetical order.

Step 2- Constructing FP tree in provided dataset is important.

Step 3- The frequent pattern tree generated is used in constructing the conditional FP-tree for the item-sets.

Step 4: Mine the frequent patterns.

4. Implementation

Data and Tools

The dataset used is a file with .csv format. It has been obtained from Kaggle. This dataset consists of various water quality parameters which have been discussed in detail in the previous section.

For generating the association rules, the graphs as well as the correlation matrix, Python Development Environment 3.8.0 is used along with the Spyder 3.2.3 tool in the Anaconda Navigator. All our tests have been performed in Windows 10 installed with Intel core, i7 CPU and 2.40 Gigahertz processor and 8Gigabyte memory.

Performance Evaluation

The evaluation in the performance and execution of the HBA algorithm is done on basis of its execution time. The execution time consists of the sum of the time taken in order to generating the candidate item-sets and the time it took to generate association- rules. This is compared with the execution time in FP-Growth algorithm. Apriori algorithm includes hashing lesser candidate item-sets are generated thereby improving the algorithm’s efficiency.

Correlation Matrix

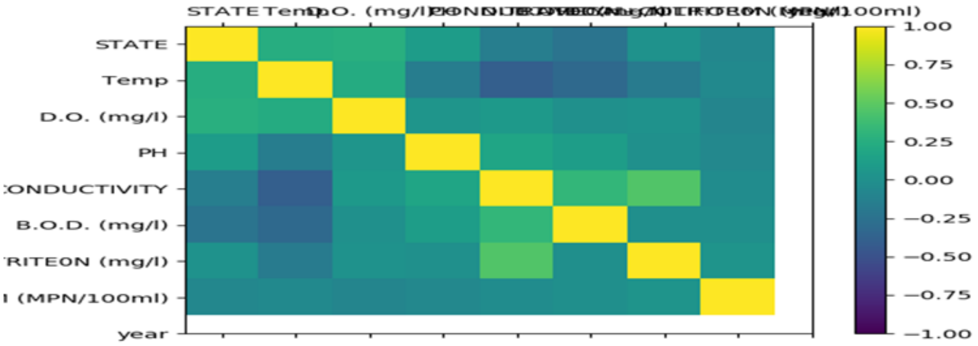

Figure 1

The correlation matrix [Figure 1] shows the relationship between the various attributes of the dataset. If the value is equal to 1 then it represents the relation between an attribute and itself. An Ifthe value range between 0 and 1 then it means that an attribute is directly proportional to the other. If it ranges between -1 and 0 then the attribute may be inversely proportional to the other. If the value is near 0 then it means that the attributes are somewhat unrelated and the second attribute's value does not influence the value of the first.

Execution Time

The execution duration of Hash based Apriori algorithm will be compared with the execution time of the FP-Growth algorithm in Figure 2.The whole dataset which consists of 1000 sets of values and 10 attributes is considered. The final graph suggests that the hash based Apriori algorithm takes considerably less time for completion compared to the FP Growth algorithm. In fact after testing multiple times the hash based Apriori has a lower execution time compared to the FP Growth algorithm.(Figure 2)

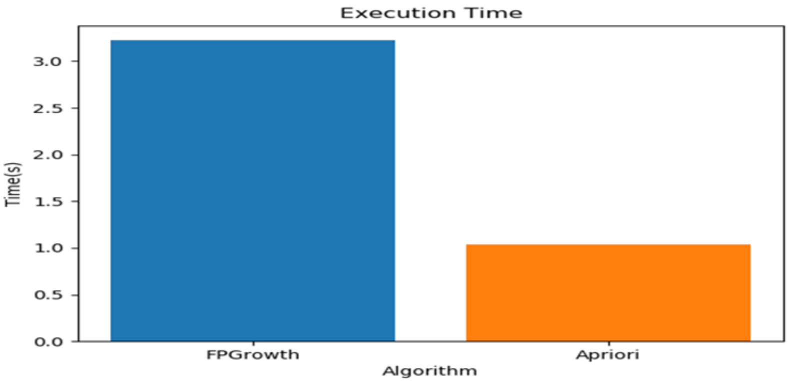

Figure 2

5. Conclusion

The Analysis of graphs shows that HBA algorithm is that the best. The execution time is shriveled when the support threshold gets enhanced. The quickest algorithms in the given data set was HBA algorithm which is followed by fp-growth algorithm. Fp growth algorithm took longer as compared to alternative algorithms for the similar dataset.The scope of the project is incredibly wide due to frequent item sets square measure helpful for applying numerous data mining technique like clustering or classification, also the association rule mining and many other techniques. There are many various techniques of fp mining that can be utilized as in many diferent ways for generating frequent itemsets. During working on project, we tend to get opportunities to grasp broad scope of applications in the field in daily life scenario. We've used python and Django platforms in the implementation of algorithms. So, the results generated may vary with the programming languages or methodologies.

References

[1] Apeksha Aggarwal, (Member, IEEE) and Durga Toshniwal(Member, IEEE),"Frequent Pattern Mining On Time And Location Aware Air Quality Data",IEEE Access,Jul.2019,Vol-7

[2] Mrs. M.Kavitha, Ms.S.T.Tamil Selvi,Department of Computer Science Tiruppur Kumaran College for Women Tiruppur, Tamil Nadu,India,"Comparative Study on Apriori Algorithm and FP Growth Algorithm With Pros and Cons" International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 4, Jul - Aug 2016

[3] Jong Soo Park, Member , IEEE ,Ming-Syan Chen,Senior Member , IEEE , and Philip S. Yu, Fellow , IEEE," Using a Hash-Based Method with Transaction Trimming for Mining Association Rules", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 9, NO. 5, SEPTEMBER/OCTOBER 1997.

[4] J. Mennis and J.W. Liu ,Department of Geography, University of Colorado,"Mining Association Rules in Spatio-Temporal Data",Jan 2003.

[5] Min Chen, XueDong Gao School of Economics and Management University of Science and Technology Services (China) Company Limited, "An Efficient Parallel FP-Growth Algorithm",IEEE, 2009.

[6]Rui Chang and Zhiyi Liu,School of Information&Engineering, Changzhou Institute of Technology," An Improved Apriori Algorithm" , 2011 International Conference on Electronics and Optoelectronics (ICEOE 2011).

[7] Surbhi K. Solanki and Jalpa T.Patel,Department of Information Technology Shri S'ad Vidya Mandal Institute of Technology Bharuch, India," A Survey on Association Rule Mining", 2015 Fifth International Conference on Advanced Computing & Communication Technologies.

[8] Ilham Huseyinov and Utku Can Aytac ,"Identification of Association Rules in Buying Patterns of Customers based on Modified Apriori and Eclat Algorithms Using R Programming Language",IEEE,2017.

[9] Apeksha Aggarwal, (Member, IEEE) and Durga Toshniwal(Member, IEEE),"Spatio-Temporal Frequent Itemset Mining on Web Data",2018 IEEE International Conference on Data Mining Workshops (ICDMW)