National College of Ireland

Pre-processing Techniques for Optimizing Association Rule Mining Algorithms

MSc Research Project
Data Analytics

Saurabh Shantkumar Yeramwar
Student ID: x20131283

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Saurabh Shantkumar Yeramwar
Student ID:	x20131283
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	31/01/2022
Project Title:	Pre-processing Techniques for Optimizing Association Rule Mining Algorithms
Word Count:	8061
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Saurabh
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Pre-processing Techniques for Optimizing Association Rule Mining Algorithms

Saurabh Shantkumar Yeramwar
x20131283

Abstract

Association Rule Mining (ARM) algorithms are machine learning techniques for the discovery of relationships between variables in datasets. Market basket analysis is one of the applications of association rule mining. It generates rules that give information about how frequently an item gets picked by customers depending on the current products present in the kart. The Apriori algorithm is perhaps the most well-known association rule mining algorithm, but it is computationally inefficient. There are multiple improvements to the original Apriori which take less time and less computational power to generate rules. ARM algorithms focus on frequent rules, at the expense of less frequent sets. Infrequent sets are however important as they can give clues on rare events, such as in anomaly detection, fraud, and other interesting customer behaviour. ARM algorithms are inefficient to handle rarely occurring rules in the data, because these algorithms generate all the possible rules and then filters out the rules related to the specific item. This process takes a bit longer time as it generates all the rules. To avoid this delay pre-processing of the data can be done with the help of clustering techniques. With this approach, data will get clustered first and then feed the ARM algorithms with data from small clusters. The proposed pre-processing idea is applied before running the Apriori and FP-Growth algorithms. We show the pre-processing step helps in reduction the execution time 31.6% and 35.8%, respectively. It drastically reduces memory consumption, and time and computational power to generate infrequent rules.

1 Introduction

The effect of technology on industries has been massive in the last couple of decades. The business has grown dramatically with the help of newly developed tools, and each sector's capability got reached to the next level. Exploring new ways to grow the business is an important factor in today's competitive market. The retail sector is one of the impacted sectors in the industry. The way retail works before a few years and now has many differences. Information and technology have changed the logistics, billing and customer relationship domains and emerged new domains in the sector for growth. Data has become a new asset for companies, and data analysis is the next challenge. The amount of data is huge because of multiple reasons. Each transaction, movement, cancellation are getting recorded with this. Some unwanted data is also getting collected. There are multiple ways to use this data in today's world. Different technologies will help to understand data and pull insights into the data. These all things are helping the retail industry to grow day by day.

Online e-commerce companies have grabbed a good percentage of the retail industry, and customers expectations have reached new heights because of technology applications in the e-commerce companies. Customers can view billions of products with one click and purchase the product with some clicks. The customer even can read the reviews and understand the user feedback about the product. These all the applications can be used from home with comfort, helping companies improve customer satisfaction. Indirectly this is helping companies to earn more money and invest in the research and development of new software products. This is impacting the overall ecosystem of the retail sector, including every part of which can help to speed up the process.

Retail industries collect the data from each stakeholder. The main stakeholder is the customer. Customer behaviour is the most important part of the study in which we can understand more about the customer. Each time customers visit any store and purchase some products, the list of products with a unique billing id will be stored. A similar process will occur in the case of online purchases of products from e-commerce websites or applications. This data can help the industry to understand the possibility of picking different products or services together by the customers and explore the interesting facts. For example, if any customer is purchasing a laptop, there is the possibility of selling a laptop bag with good support from data.

Association Rule Mining Algorithms are widely used to finding the interesting relationship between the variables in a large scale dataset. Market basket analysis is the application of the association rule mining algorithms in which the relationship between the sales of the different products from the store is searched. Algorithms generally use a brute-force approach in which it calculates all the possible rules and then it filters the rules with specific support and confidence. This way can take a longer time if the dataset size is large as the number of comparisons increases drastically in case a large dataset has a greater number of unique items. There are multiple applications of the association rule mining, including genome sequence, healthcare sector and understanding the customer, which is mostly used. In the case of genome sequence and some other applications, the researchers are interested in finding the relationship in the items which are rarely present in the human body. This rear rule can help doctors and other stakeholders to understand the dependency of the critical elements.

Pre-processing the data before providing it to association rule mining can help to reduce the number of comparisons and make efficient processing. Multiple techniques are available in pre-processing, and clustering is one of them. In clustering, we analyse the data, and clustering depends on various factors and posts clustering. We will provide the business interested cluster data to association rule mining. Can pre-processing with clustering algorithms improve the efficiency of association mining algorithms over a large amount of data?

Financial fraud transactions can be found with the help of association rule mining. The study of association rule mining shows various application including fraud detection. The combination of the clustering and association rule mining will help to reduce the processing required to find the fraud transaction. If the data is getting clustered have a fraud transaction in specific cluster, then that specific cluster can be given to the association rule mining. The accuracy of clustering the fraud transaction is important aspect for this experiment as it will directly affect the outcome.

Association rule mining algorithms over a large-scale dataset can require high performing machines because of their complex computation. Apriori algorithm helps to find the rule, but the amount of memory required in the second pass is huge, and it is a bottleneck. There are multiple other algorithms that improve the Apriori algorithm, but it requires a large amount of time on a large-scale dataset. The data can be used to reduce the complexity clustering, and post clustering, the interested cluster data can be provided to the association mining algorithm. This will help to run the association rule mining algorithms in reduced time with major savings in computation. This can also help to find the interesting relationship between items on a focused subset of the data. Clustered data is easy to process by complex algorithms, and in some cases, if machine configuration is not permitting the operation because of memory constraints on full data. It is possible to run the algorithm on the same machine with any cluster of data.

2 Related Work

Association rule mining algorithms over a large-scale dataset can require high performing machines because of their complex computation. Apriori algorithm helps to find the rule, but the amount of memory required in the second pass is huge, and it is a bottleneck. There are multiple other algorithms that improve the Apriori algorithm, but on a large-scale dataset, it requires a large amount of time. The power of a machine learning algorithm can be used to optimise the problem. Applying the k-means clustering algorithm with the Apriori algorithm has given the improvement in the performance in the experiment Kanhere et al. (2021). Exploring all other possibilities combinations of clustering algorithms with multiple association rule mining algorithms can help to improve the accuracy and efficiency.

Clustering algorithm is used by researchers to enhance the association rule mining algorithm and this experiment is successfully implemented by researchers. Reduction in size of the input data is achieved with the help of clustering algorithm and the number of rules generating from original data as well as the reduced data are same. Moreover, the generated rules from reduced data are having more interesting relationship between the items, this experiment is carried out over the UK retail dataset Kanhere et al. (2021). This experiment contains multiple observation, most highlighting result is that the input data is reduced to 7% of total data, which will directly impact on the execution time of the algorithm. Researchers have used K-means algorithm for clustering and Apriori, FP Growth and Eclat algorithm for association rule mining. Multiple combinations of data pipelines are developed and whole process is analysed with multiple evaluation factor including execution time of the algorithm, product exclusion rate and varying support threshold for algorithm.

Anomaly detection is the crucial task in case of data with higher number of dimension as it requires complex processing and high performing hardware. Association rule mining can be used to find the infrequent rules which can be applied on dataset to find the anomaly. Credit card fraud transaction detection is the application of the stream data infrequent rule mining algorithm which helps to find the fraud transactions from high volume of the stream data. Researchers used Minimal Infrequent Pattern Pattern algorithm on the synthetic publicly available sensor dataset to find the minimal infrequent pattern mining Sweetlin Hemalatha et al. (2015). The proposed method requires the categorical Boolean data only which require to conversion of continuous. The proposed method is

tested on the on sliding datasets with various values of support. The proposed solution performs better than the classic anomaly detection algorithms and execution time is not increasing after increasing the input transactions which is comparatively lower than the anomaly detection algorithm.

Detail study on the association rule mining as well as clustering algorithms is carried out. This section gives a detailed analysis of various association rule mining algorithms specially apriori and fp growth with overview of clustering algorithms.

2.1 Association Rule Mining

A survey over various association rule mining algorithms is carried out by the researchers to understand the applications and working on it Solanki and Patel (2015). Researchers explained details about the basic association rule mining with an example of market basket analysis. Stock market data can be analysed with the help of association rule mining to understand the pattern of stock and the relationship between events before or after the sudden hike in stock price. Website logs can be provided to the association rule mining algorithms to find the frequent errors that occurred together and find the impact of the different errors on each other. The Healthcare sector got multiple devices that share data about the human body. These different datasets can be pre-processed and can correlate the connection between different events that occurred before getting infected with the disease. This can help doctors and medical workers to understand the root cause of the problem and avoid similar issues with other people. Further various algorithms are discussed by researchers, including Apriori, FP Growth and its different improved variants.

Apriori algorithm is a fundamental algorithm in association rule mining. This algorithm generates all possible combinations of rules from a dataset with a given support threshold. Apriori algorithm runs internally with multiple phases. The second phase of the algorithm requires huge memory as the frequency of all combinations of a unique itemset is getting calculated. Generally, this second phase shows a memory bottleneck, and if a machine is not able to provide the required memory slot, then this algorithm will not work as the machine is not able to allocate the required memory space, which generally takes place while processing large dataset. Researchers have implemented the modification in the second phase of the fuzzy apriori algorithm, and the proposed algorithms are working six times faster as per analysis over the data Mangalampalli and Pudi (2009). Pseudo-code for the proposed algorithm is provided in the paper, and this proposed algorithm is then run multiple times with different support thresholds. Results show that the partitioning of data and modified logic have improved the efficiency of the algorithm.

Association rule mining is the most important topic, and many researchers have done studies on it. The resultant is multiple algorithms are present in this machine learning technique, its starts from apriori and have multiple other improvements of the same. There are other multiple algorithms where it uses tree structure and has a more efficient way to find association rules. The researcher has completed a survey and used the criteria process to choose the appropriate algorithm accordingly to dataset attributes Addi et al. (2015). Overall, all the algorithms have step one in which finds the frequency of the candidate set, and then in the second step, it calculates the frequency of pair of candidate sets. As apriori have a memory management bottleneck for step two, other algorithms

try to improve it in multiple ways. Researchers have provided an overview and internal working of multiple algorithms, including Apriori, AprioriTID, FP-Growth, Partition, Dic, Eclat, Apriori-Hybrid and Relim Addi et al. (2015).

Association rule mining algorithms require high-performance computing while processing a large amount of data. In case the number of items are more than the calculations to find all necessary variables are more. In the apriori algorithm, support for each item of the dataset is get calculated with each other item. In case the total number of items is more, the amount of memory required to store and calculate these numbers is more. The high requirement of mail memory is causing the low efficiency of the algorithm. Multiple variations of Apriori, PCY and FP Growth are present, which improves algorithms complexity Caroro et al. (2018).

2.2 Clustering Algorithms

The clustering algorithm is used by researchers to enhance the association rule mining algorithm, and this experiment is successfully implemented by researchers. Reduction in size of the input data is achieved with the help of a clustering algorithm, and the number of rules generated from original data as well as the reduced data is the same. Moreover, the generated rules from reduced data have a more interesting relationship between the items. This experiment is carried out over the UK retail datasetKanhere et al. (2021). This experiment contains multiple observations. The most highlighting result is that the input data is reduced to 7% of total data, which will directly impact the execution time of the algorithm. Researchers have used the K-means algorithm for clustering and Apriori, FP Growth and Eclat algorithm for association rule mining. Multiple combinations of data pipelines are developed, and the whole process is analysed with multiple evaluation factors, including execution time of the algorithm, product exclusion rate and varying support threshold for the algorithm.

Urban scale data face multiple issues while processing for clustering algorithms due to large size. It needs complex computations to find the universal clusters in streaming data, these problems can be handled by online clustering algorithms. BIRCH is a online clustering algorithms for streaming data which optimizes the algorithms as per the data behaviours which gives better performance in case of big data. The datasets which have difficulty in managing the memory while performing clustering algorithms is given to the BIRCH algorithm and the clustering is performed with minimum computational power with good efficiency.

2.3 Apriori and FP Growth

Internet has changed the way of life. Users across the world started using web applications for essential and non-essential activities. Business is liable to predict the next request from the user to keep the process ready and give the user a sense of personalisation which can directly affect a positive effect on advertising. This process will help businesses to increase revenue, which is the goal. This process will take place with the help of access sequence mining. Each movement of the user is getting recorded, and a series of events is getting stored as an access sequence. Most of the users have a large size of access logs as multiple requests come to web applications leads to large access sequences. Apriori algorithm requires multiple scans of the whole dataset and generates a large number of

candidate keys which leads to inefficient execution with respect to time and memory. Researchers have proposed AC-Apriori algorithm, which is based on Aho-Corasick (AC) automation which leads to AC-Apriori Yang et al. (2017). Results show that as the size of the sequence grows, the execution time is better for AC-Apriori as compared to the Apriori algorithm. If the support is minimum, the amount of computation is maximum in such cases. Also, the proposed algorithm works better. Researchers have experimented on a small scale dataset and executed processes on local machines. The proposed method needs to be implemented on a large scale dataset with clustered setup.

Web users interact with multiple websites for multiple reasons, and each time they interact, the weblog gets generated. Researchers have analysed this weblog data with reverse apriori, an improved association rule mining algorithm Puneeth and Prasad Rao (2019). This proposed algorithm takes advantage of the Fp tree and combine it with the apriori algorithm and shows the significant improvement in execution time on the same dataset. The rules generated by these web usage mining algorithms can help to detect fraudulent elements over the web by connecting the events that occurred over the web page. These logs also can help in the improvement of the application. When users are not able to use the application, they will stop using the web application. Analyses of the weblog data over events that occurred before closing the web application can help to improve the product. If there is a pattern in these events, then there is an issue in the application which can be rectified by the developer.

Apriori algorithm is the key algorithm in association rule mining which helps to find the rules with the help of the n frequent item list. This algorithm scans the whole database in each pass which leads to the consumption of a lot of computational power and time in the case of large-scale datasets. Researchers proposed an improvement in the apriori algorithm in which the one item set frequency generated by the algorithm is getting saved in the linked list, and the two-item set frequency is getting stored in hash tables Mar and Oo (2020). Once the linked list and hash table gets the required data inside them, this algorithm does not need to scan the database for every pass, which helps to reduce in computation required for the execution and the time for the same. The result shows that the amount of time and memory required for the proposed algorithm is less than the apriori algorithm over a sample dataset.

MapReduce, a parallel processing Hadoop framework, have the ability to use the power of distributed computing environment. Large scale dataset needs distributed environment for smooth processing, in which MapReduce plays an important role. In market basket analysis or any other association, rule mining algorithm, the size of data is huge as the number of transactions are more. The researcher proposed the implementation of Apriori algorithm with MapReduce where the input data is stored on Hadoop Distributed File System, also known as HDFS Chang (2015). In HDFS the stored data gets distributed over multiple nodes, which will help in parallel processing. The data replicated across the cluster which makes system fault-tolerant in case of node failure in the cluster. Researchers have tested the newly implemented algorithm from 10 GB to 30 GB of massive data to analyse the efficiency, and result shows the positive outcomes. The execution time is inversely proportional to number of nodes in the cluster as they improve parallel processing.

Intrusion detection to identify the anomaly in the data is the important task in the many tests case, including credit card fraud transaction detection and fraud loan applicant

detection. This detection has many problems, including high possibility of false detection which means the system suggest that the credit card transaction is fraud but this is authentic transaction. This false detection will trouble application users, which can impact revenue of the product. While processing high amount of data, these algorithms show the poor performance in memory and time management. Researchers have improved the process of isolation forest which helps in finding the abnormal transactions and then improvement in the FP growth an association rule mining algorithm Zhou et al. (2020). The combination of these two improved algorithms is impacting on less false alarms and fast processing of a large amount of data.

3 Methodology

CRISP DM stands for CRoss Industry Standard Process for Data Mining is the plan which includes the process for model building. Figure 1 shows the fundamental components CRSIP DM which also can be considered as data science life cycle. This is a cyclic process at any stage. If the process is digressing from main purpose because of any reasons, stake holders can go back and start process again from any step to start again. In the research over association rule mining, CRISP DM methodology is used which is helping in improving the quality of the solution.

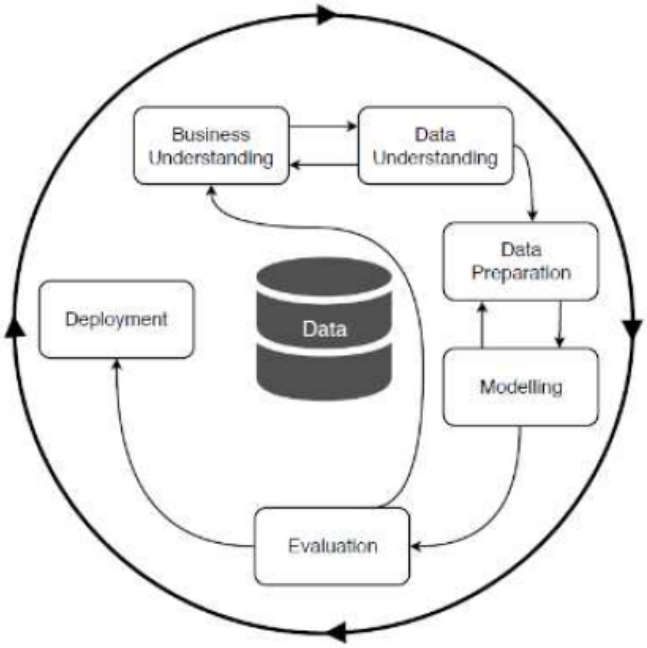

Figure 1: CRISP DM

The ethical challenges for information technology are getting evolved with arising problems in society. Researchers have emerged a methodology that can comply with the regional ethical laws with an agile framework and make product more efficient Hobbs (2021). In the competitive business environment, the improving and changing requirements for the product needs adaptive framework, agile is the framework that can help in this environment.

Agile framework works with sprint model whereas per requirement work done by developers get deployed in the production at the end of sprint. Machine learning is evolving with multiple stake holders and changing applications, integration of the agile framework with the machine learning is need of time. Researchers have studied about the machine learning and its fast-changing framework. Currently, MLOps is capturing market where the roles and responsibilities are divided among stake holders with high communication Mäkinen et al. (2021). All the proof of concepts are getting evolved with MLOps where the developed application can get deployed with more ease and the whole team works together. While working on the improvements over the association rule mining with clustering, the steps are involved in an iterative manner, whereas the process can get changed, which will directly impact the quality of the final research. This section explains step by step approach to complete research on how pre-processing will help in the association rule mining to improve efficiency.

3.1 Business Understanding

Association rule mining algorithms are complex and find all possible rules present in the dataset. Apriori algorithm is classic algorithm in the association rule mining, which works with the help of frequency of candidate set. Apriori needs repeated scan of the whole dataset, which leads to the more execution time and complex computational. In some cases, if the system available to execute the algorithm is of basic configuration and support is less then the amount of computation is more and need of memory is more. This high memory requirement leads to out of memory exceptions for basic system configuration. The possibility of this exception increases in the case of a large-scale dataset, even though the commercial servers can compute complex problems, this leads to high amount of execution time. This research project analyses the application of clustering algorithm as pre-processing technique for dataset.

3.2 Data Understanding

The research is based on the implementation of pre-processing framework for the association rule mining, there are multiple applications of association rule mining. Market basket analysis is the one of the most used application, dataset [1] selected for this experiment is sales data of online store which is publicly available at the UCI machine learning archive. This data does not contain any personalised information about the customer. Dataset contains more then 54,000 transactions, each transaction contains the details about each product sold in the transaction with quantity, invoice number and all other details. For same invoice number there can be multiple transactions, which means to find all the product sold to single customer with same invoice number we need to group the dataset with invoice number.

3.3 Data Preparation

Association rule mining algorithm works with text data. Preparation of data is completed with downloading the data from the public repository. The rules generated by the dataset are the information stored in the data in a more understandable and business-oriented

[1] `http://archive.ics.uci.edu/ml/machine-learning-databases/00352/`

format. If we generate the rules from different association rule mining algorithms with the same parameters, then not even the number of rules generated are the same but also each rule will be same. The downloaded file from the public repository is in the xlxs format which can be directly consumed by python programme.

3.4 Data Preprocessing

The data which we are using in this research contains various fields, including the invoice number, product description, which is a string explaining the product. Quantity of products sold also provided along with the price of each unit of the product and country in which it got sold. The data needs to be in format explaining all the products purchased in a single invoice. Grouping the data on the invoice number is the first step post-cleaning of data, which can give the list of products with each invoice number. This then should get encoded with the help of one-hot encoding. This means at the end of the process it will give a single encoded dataframe which have multiple columns having only Boolean values.

3.5 Modeling

Association rule mining is a classic machine learning problem where there are multiple algorithms got evolved to improve the quality of the algorithm. Irrespective of increased research, the public libraries to implement these algorithms are very limited, and related documentation is also limited. This research use mlxtend [2] library which is open source library which implemented multiple machine learning algorithm and have basic documentation over multiple sources. Apriori and Fp Growth are the association rule mining algorithms getting implemented by using this library.

3.5.1 Apriori

Apriori algorithm is a classic association rule mining algorithm, which process data to find the candidate sets the frequency. Apriori algorithm works in multiple passes according to the size of the rule expected to generate from the process. Minimum two passes will run as rules need a minimum of two items, and to generate a rule of two items, it needs a frequency of pairs. There are multiple algorithms present that improves the Apriori algorithm, but the fundamentals of these algorithms are Apriori algorithm.

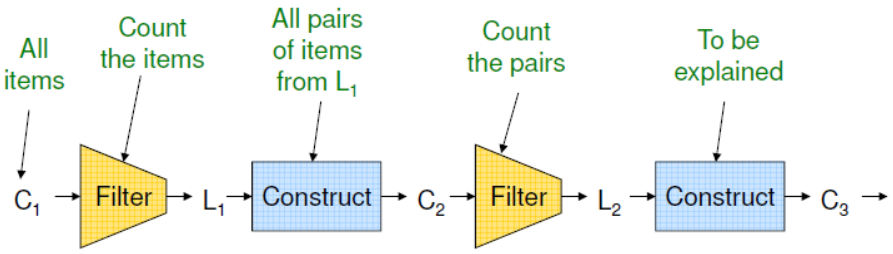

Figure 2: Apriori Algorithm

[2] http://rasbt.github.io/mlxtend/

Figure 2 shows the process getting internally triggered inside the Apriori algorithm. C1 is the candidate key which is the frequency of the singletons, with this algorithm will calculate the support of each key. In the filter part of the process, it checks the support of each key with minimum support provided to algorithm. If the support of key is more than the minimum support, then only the key is stored in L1.

L1 is the all the singletons which have minimum support, which then used to construct all the pairs with its frequency which gets stored in C2. This is step where more memory gets consumed as number of pairs are maximum. Constructed pairs need to follow two conditions, first one is the support of pair is more than the minimum support. The second condition is both elements of the pair individually should have supported more than minimum support. If any pairs satisfy these two conditions, then it will be stored in L2. This is an iterative process as per the value provided for the size of the rules.

Figure 3 shows the main memory picture after running the Apriori algorithm. In the first pass, it only calculates the frequency of unique items in the dataset. This pass needs less amount of main memory as in dataset comparatively fewer numbers of unique items are present. In the second pass, it stores the frequency of singletons as well as all frequency of combinations of the size of two. These numbers of pairs are more, and the memory required for this is more as compared to the pass one. In some cases, if local hardware on which Apriori is not able to allocate the memory required in the second pass, out of memory exception can get raised. Memory management bottleneck is present in the Apriori algorithm in pass two.

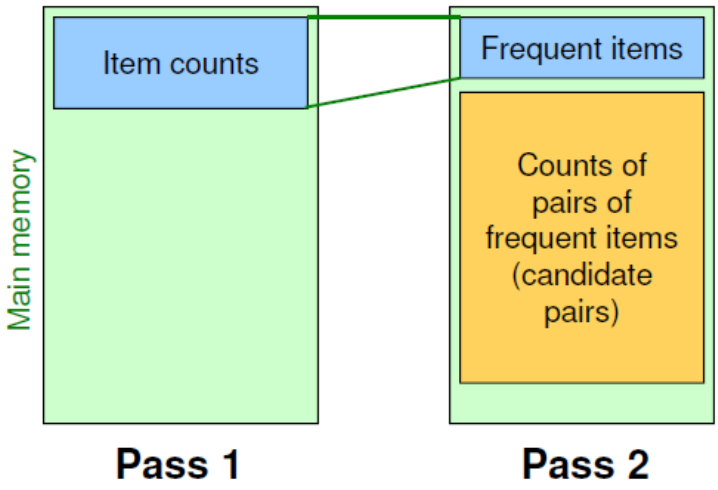

Figure 3: Apriori Memory Picture

3.5.2 FP Growth

Fp Growth is the algorithm that is more efficient than the Apriori as it is not using the approach which scans the dataset repeatedly. Fp Growth algorithm works on the fp tree which is generated at the start of the process. This Fp tree contains all the information about each item and its relation with other item with frequency of its relation. This

optimised tree development is one time activity, once the tree is generated, each node of tree shows the item and its frequency.

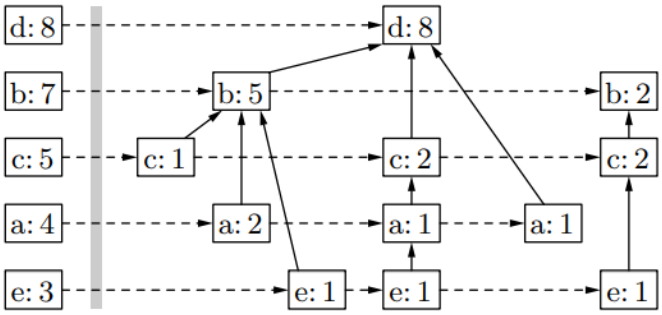

Figure 4: FP Tree

Figure 4 shows the sample Fp tree which is generated for sample transactions of small size dataset. Tree structure is generated once calculating the frequency of singletons in the detaset, the root node is the node with item with highest freq ency, then the related items are append in next levels of the tree. Each node contains the item and its relevant frequency with parent node. Simple logic helps to generate this tree. On left side it shows process to calculate items total frequency, for example b item is present 7 times in the dataset. This 7 number is generated from two branches of parent node d, one branch have 5 frequency and other have 2. Then the total 7. To generate rule, traverse the tree from leaf node towards root node with considering the frequency number.

3.5.3 PCY

Aprioir algorithm have memory bottleneck in pass two in which large number of pairs frequency get calculated. PCY algorithm resolves the memory concerns of Apriori algorithm with the help of bitmap constructed while pass one. Bitmap is datatype which holds some Boolean values, information related with bucket is frequent or not is stored in this bitmap. While scanning database for pass one other than calculating frequency of the candidate keys of the transactions it generates data to construct bitmap. In each transaction, it use all possible combinations of pair of items and then hashing will get applied to these pairs. Hash value of the pair is then mapped to a specific bucket number. At the end of the pass one, total number of pairs in each bucket get calculated. The bucket which have number of pairs more than support threshold is considered as frequent bucket.

With the help of data generated about the frequent and infrequent buckets in pass one a bitmap gets generated. Bitmap contains true value if the relevant bucket is frequent and false if the relevant bucket is infrequent. Figure 5 shows the memory picture of PCY algorithm, generated bitmap is present as part of main memory in pass two. Pass two use this bitmap to generate frequency of all the pairs in dataset. Same hashing function is applied on the all the pairs and the bucket number is found. Whether the pair is element of frequent bucket or not is find with the help of bitmap generated in pass one. In some cases the non-frequent pair get added in frequent bucket, this pair will get filtered in the pass two. Even though the number of pairs generated in pass two of PCY algorithms

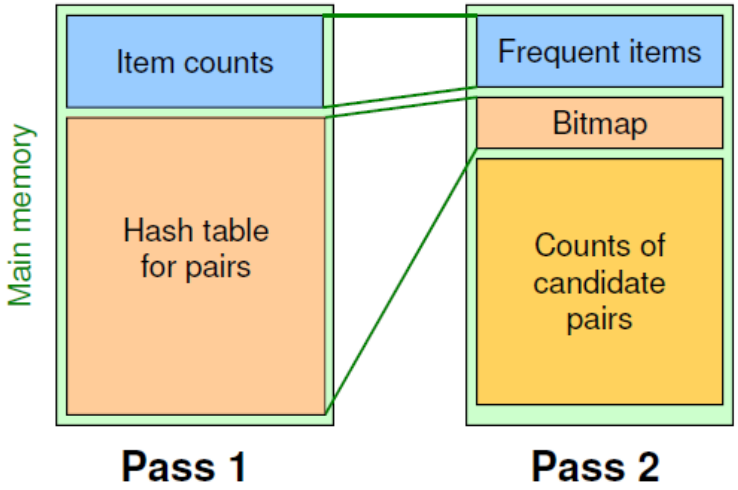

Figure 5: PCY Memory Picture

should be less the Apriori algorithm. Post pass two all the processes is same as Apriori to generated rules of size more than two.

3.5.4 K-Means

Unsupervised machine learning algorithms helps to understand and analysis the data with various ways. Clustering is one of the unsupervised clustering algorithms which helps to divide the data in various clusters. K-means is the classic clustering algorithm which work on the mean average distance between the centre of cluster and other points of the cluster. Number of cluster is decide on the basis of the variance elbow graph. Researchers have used the clustering algorithm to divide the process running over infrastructure with various parameters Sharma and Bala (2018). This is helping in the setup of infrastructure over the cloud environment to reduce cost.

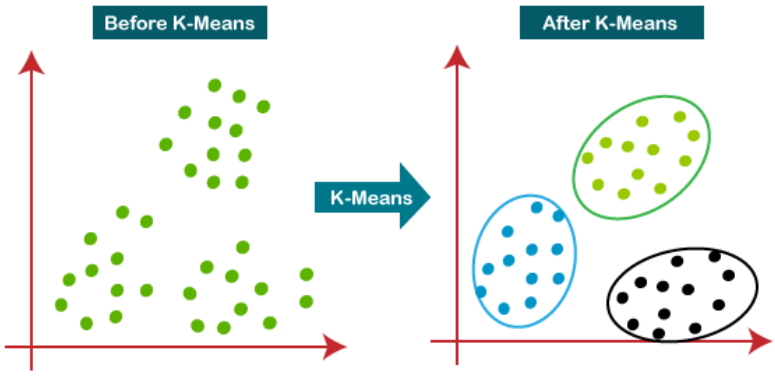

Figure 6: K-Means Clustering

Figure 6 shows that the data points across the plane before the applying the clustering algorithm. These data points are plotted with the different attributes available in the

dataset. The right hand side shows the output of a conceptual k-means algorithm with $k=3$. After applying the k-means algorithm it covers the whole data into three small clusters.

There are multiple advantages of using K-means algorithm over small scale dataset. K-means work based on the mean distance of the datapoints from the centroid of the clusters. Main advantage of K-means algorithm is that it is relatively simple to understand and implement with the help of python libraries. As it is relatively simple algorithm it takes comparatively less processing on the dataset which can be implemented on small as well as large scale datasets. K-means algorithm clusters the data with the help of large number of centroids start points which will help to detect complex clusters. In terms of processing small scale data for association mining algorithms (ARM), it adds the overhead, over large-scale dataset the proposed approach will compensate this overhead as it avoids the complex computation of the ARM over large scale data. This will also help in case the available hardware to execute ARM is of low configuration for large dataset.

3.6 Evaluation

Evaluation of newly developed module is dependent on the amount of time it is taking to generate model and rules. Whole dataset loading into main memory and then computing all frequency of items is complex task, which can take more time. The new process gives some advantage by clustering data in different classes and then provided the class data to association rule mining algorithm should save some time. Then the rules generated should be more concentrated and should have high support. Amount of memory consumed in the process is also a parameter of evaluating the new process as the amount of memory required is more in case of Apriori algorithm.

4 Design Specification

Solution to every problem starts with designing the steps, if designing is proper implementation will get completed efficiently. Each step has its importance in the design, in this research the design is developed with two main components including clustering and association rule mining algorithm. High performing configuration is required to run association mining algorithms on large datasets. In some cases, it is not possible to complete these hardware requirements for short amount of time. Large amount of rule generation can lead to confusion with most helpful rules for large dataset. The proposed solution helps in running the association rule mining algorithm with low level of hardware configuration in case the dataset provided is large.

Design of the experiment is shown in Figure 7, the input data is first given to the clustering algorithm which divides data in multiple clusters. These clustered data should get analysed and then the selected and small size clustered data should get provided to the association rule mining algorithm. This algorithm then generates the frequent items frequency with rules required. These rules are more relevant and concentrated around similar set of items.

Figure 7 shows the overall data flow of the proposed model which will help in the improving the efficiency of the model. This will reduce the number of rules generated for same support value but these rules will be more interesting and help business.

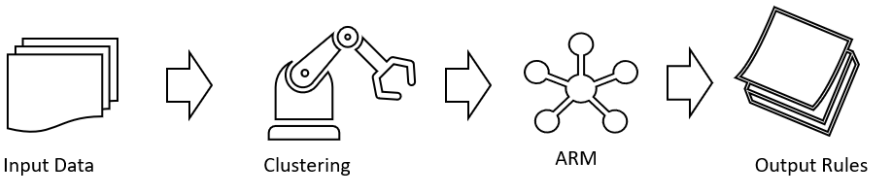

Figure 7: Design

Large dataset generates massive association rules with minimum support value. We can either increase the support value provided to the algorithm or study the rule with respect to the provided values to understand the behaviour. Example of support, confidence and lift is shown in Figure 8. Support is the ratio of the number of time rules is present in the dataset to the total number of records in the dataset. Support value close to 1 is considered as the strong support, this varies for each dataset as per the frequency of the items. If the dataset contains more equal distribution of multiple frequent items, then the support will be less. For example, if the apple and banana is present 4 times out of 10 total records the support for apple and banana is 0.4 which is 40%.

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Figure 8: Support Confidence Lift

Confidence is the amount of time the rule which is generated in the dataset is getting followed, it gives the ratio of total number of times all the elements are present in the dataset to number of times rule is getting followed. Rule can have high support which shows that the strength of rule is more but when we apply it over dataset, we can get a smaller number of transactions are following it. This is the case where the rule has high ratio of support and the confidence ratio is less. Lift is possibility of purchase of an item from transaction on other item.

5 Implementation

This section all the steps included in the implementation of project. These steps are mandatory to execute all the steps and get similar output for the reproducibility of results. Results are totally depending on the hardware as well as software as each update comes with some additional efficient approach to resolve the complex problem.

5.1 Environment Setup

Table 1 gives important information about the operating system, processor and other setup details for this project to reproduce.

Table 1: Environment Setup

	Version
Operating System	Windows 10 Home
Processor	11th Gen Intel(R) Core(TM) i5-1135G7
RAM	24.0 GB
Graphics	NVIDIA GeForce GTX 1650Ti

5.2 Model Building

Market Basket Analysis models helps to understand the data from customers are for generation of rules. While generating the rules the model needs to be provided with multiple parameters. Used these parameters to tune and check the performance of the models. Minimum support parameter gives the threshold value for every rule support, all the rules generated from the model should have at least this support. This parameter helps to reduce number of rules generated. If minimum support very less, then large number of rules get generated. To reduce these rules, we can take help of minimum support. This parameter can not be ideal for all datasets, the value is totally depending on the dataset.

Performed multiple experiments in the analysis of the model behaviour. In first experiment, changing support value to check the number of rules generated in each step. Next experiment is with the apriori algorithm, in which same dataset is given to apriori algorithm and post pre-processing with clustering data is provided to the apriori algorithm. The execution time is analysed in all the steps. Last experiment is with Fp Growth algorithm, this is two step experiments like the second experiment. Dataset is directly provided to Fp- Growth algorithm in first step. In second step, with the help of clustering data will get processed and then provided to Fp Growth algorithm. In this experiment also we will record all the execution time to analyse the models.

6 Evaluation

Understanding of internal working of association rule mining to mitigate the problem of memory and speeding up the process is the goal of this research. In this section, analysis of the results of the various experiments are present. Each experiment is carried out over same hardware and software configuration, this will help in comparison of the results. As

all other factors are same, the difference between the readings is because of the change in the approach of process.

6.1 Experiment 1

Number of rules generated from association rule mining is the important factor, the large number of rules make it difficult to understand the pattern in the items. Very less rules generated by algorithm have not all the information about the relationship of the item. Market basket analyses if the dataset is of some global store chain having thousands for products. For these datasets, medium support is preferred as it explains all about the relationship between the items.

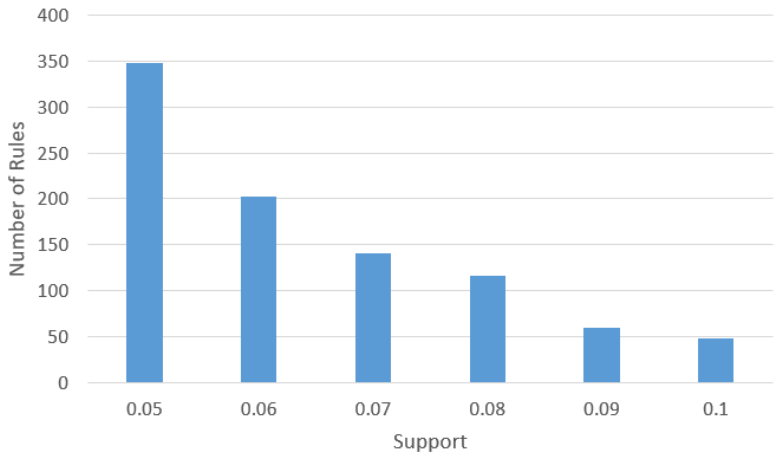

Figure 9: Support vs Number of Rules

Figure 9 shows the number of rules generated with varying support value for given dataset. Observation is that the number of rules is decreasing when the support is increasing. This means the relationship between the support and number of rules is inversely proportional to each other. All the algorithms for the same dataset with the same parameters generate equal number of rules. These different algorithms helps to improve the efficiency of the execution, execution time and computational requirements plays important role in case of large scale datasets. In large datasets if the support provided to the algorithm is very low then the very large number of rules get generated which again requires the analysis. If the support provided is very high then very less number of rules will generated. Support value for each dataset is dependent of frequency distribution of the unique items in the dataset.

6.2 Experiment 2

Apriori algorithm is applied on the part of dataset and the execution time is calculated, multiple readings are taken with varying support. Execution time of the process is noted down in which the first dataset is given to clustering then the part of clustered data is given to apriori algorithm. Figure 10 shows both execution times with respect to changing support values.

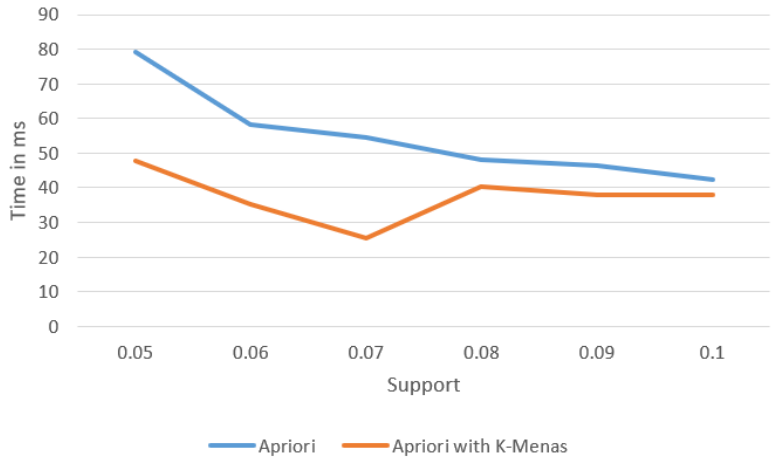

Figure 10: Apriori Execution Time

Figure 10 shows that the proposed solution in which clustering is applied on dataset as pre-processing is taking comparatively less execution time with respect to classic apriori algorithm for constant support value. Which shows that for lower support values the clustering is helping the apriori algorithm to reduce the computational overhead and complete the process in less time. It also shows that the difference between the execution time is getting decreased which signifies for higher support thresholds the proposed approach is performing like the apriori algorithm. This experiment is concentrating on the infrequent frequent items and rare rules that means the low support in which the proposed solution is performing better than apriori algorithm.

6.3 Experiment 3

Fp Growth algorithm works with the improved logic, which is more efficient than apriori. A similar experiment is completed with Fp Growth where part of the dataset is directly given to the Fp Growth algorithm, and then the execution time is measured with different support. The same part is then given to K-Means clustering algorithm and a cluster of it is then given to Fp-Growth algorithm.Figure 11 shows the graph which shows these measured execution timings for Fp-Growth and Fp- Growth with clustering.

Figure 11 shows that the proposed solution in which k-means clustering is applied on the dataset before giving it to the Fp growth algorithm is performing better than classic Fp growth algorithm with constant support. This is causing as the data is getting divided into the smaller more relevant cluster which affects the computations needed for the fp growth algorithm. It also shows that the like apriori, difference between the execution time for proposed approach and the classic Fp growth is getting reduced with increase in the support value. The research focuses on the low support values which will give rules which are having infrequent items and the infrequent rules. For this lower support the proposed solution is performing better than the classic Fp growth algorithm.

Large-scale dataset contains all datapoints which have the information related to different domain or all the domains. Post clustering this dataset, whole dataset will get divided into number of smaller parts. These clusters have datapoints which are more relevant to each other which means statistically they have relatively less mean distance

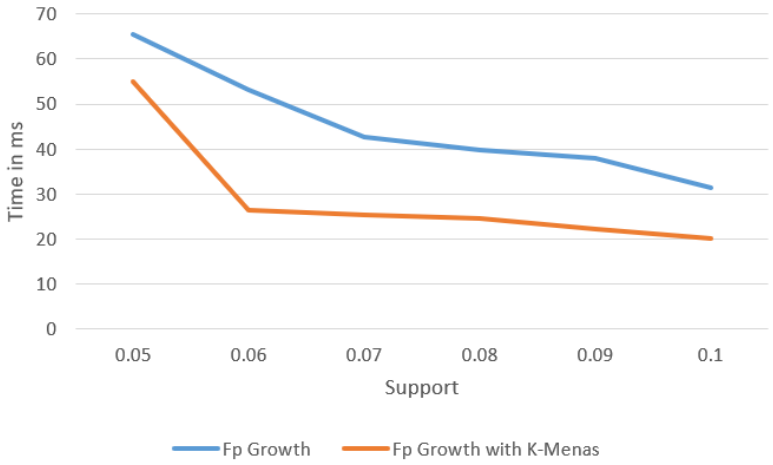

Figure 11: FP Growth Execution Time

from centroid. Logical meaning of the relevant clusters is all related datapoint of same subdomain get grouped in the small cluster. In market basket analysis if dataset contains all the products from variety subgroups. If we cluster this dataset and the if these clusters represent these subgroups, then the products within this clusters are more relevant to each other.

6.4 Experiment 4

The last experiment is the performance check experiment, where all the data is given to the apriori algorithm first with a lower support value. This lower support value and large dataset combination require a large computation and more memory. When we run this setup, it tries to allocate a large memory chunk in the main memory, and when it fails to allocate, execution stops without of memory exception. The same dataset, when given to clustering algorithm, I execute properly, and then a cluster of the full dataset is then given to apriori algorithm. This will execute properly, and rules get generated. If in the available memory space k-means clustering is not running, then we can use the online clustering algorithm to run clustering. BIRCH (balanced iterative reducing and clustering using hierarchies) is an online clustering algorithm that requires less memory, and it is a time-efficient algorithm. This shows that the proposed solution helps in case the available memory is insufficient to run an association mining algorithm on a large dataset.

6.5 Discussion

Experiments come with some results and observations. Experiment 1 shows the relation between the support and the number of rules getting generated by the algorithm. The number of rules generated by the association rule mining algorithm increase if the support is low. Support parameters can help in concentration the high or low support rules. In the case of a large dataset, this can be used in managing the number of rules getting generated.

Experiment 2 shows the execution times for the apriori algorithm with and without

the clustering algorithm. The execution time can be seen reduced as the dataset size is shortened with the help of a clustering algorithm. The concentrated rules get generated with the more interesting fact in less amount of time and in comparatively less amount of report. Experiment 3 shows similar execution times for Fp Growth algorithm with and without clustering as the pre-processing algorithm. In this experiment also the proposed algorithm is more time-efficient with any size of the dataset. In the last experiment, a large amount of dataset is given to the apriori algorithm with low support value, which starts the high-volume computations which require a high amount of internal memory. In the first case, execution end with out of memory error, but when we apply the proposed solution to the same datasets, the small cluster executes smoothly. This shows the application of the proposed solution where the hardware configuration is not able to run the algorithm on a large scale dataset.

7 Conclusion and Future Work

Research on association rule mining is interesting as multiple algorithms present their pros and cons. Apriori algorithm is a classic association rule mining algorithm, which is the base of all algorithms. All the other algorithms have the fundamental structure of the apriori algorithm. Research of these algorithms is possible as multiple papers were published. Still, implementing these algorithms is a crucial task as there are very few python libraries that have implementations of these algorithms. Some algorithms do not have any implementations in any python libraries, making the study and implementations of the algorithms more complex. As the basic execution of these algorithms is a time-consuming task, the improvements in these algorithms are more complex tasks. In this research, apriori and fp growth algorithms are implemented, the input to these algorithms is controlled in various ways. In experiments, data is directly given to the association rule mining algorithms and then post clustering is also given to the same algorithms with the same parameters. Apriori shows an average 31.6% reduction in execution time with the help of pre-processing of the data, and the Fp Growth shows 35.8% reduction in execution time. In both cases, the execution times are impacted positively, and the time required to execute the association rule mining algorithm is reduced.

Further study in this research field is interesting as many algorithms are being implemented to improve memory and execution time performance. Multiple clustering algorithms can be implemented to divide input data into various factors and then be given to association rule mining algorithms. Accurate implementation of some association mining algorithms needs to be completed. Python open-source libraries need to be published, which will help the science community easily understand and use these algorithms. Distributed computing processes a large amount of data efficiently, and implementing the proposed solution over Spark framework will increase the processing speed, and the large size of data can get processed. The use of cloud infrastructure can help analyse algorithms performance on large scale datasets with low support.

ACKNOWLEDGEMENT

I would like to express my gratitude to Dr Giovani Estrada for his support in crucial part of research and implementation for association rule mining algorithms. His

supervision and discussions helped me in designing quality solutions and I was able to finish my work on time under his guidance.

References

Addi, A.-M., Tarik, A. and Fatima, G. (2015). Comparative survey of association rule mining algorithms based on multiple-criteria decision analysis approach, *2015 3rd International Conference on Control, Engineering Information Technology (CEIT)*, pp. 1–6.

Caroro, R. A., Sison, A. M. and Medina, R. P. (2018). An enhanced frequent pattern-growth algorithm with dual pruning using modified anti-monotone support, *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, pp. 1–5.

Chang, X.-Z. (2015). Mapreduce-apriori algorithm under cloud computing environment, *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, Vol. 2, pp. 637–641.

Hobbs, R. (2021). Integrating ethically align design into agile and crisp-dm, *SoutheastCon 2021*, pp. 1–8.

Kanhere, S., Sahni, A., Stynes, P. and Pathak, P. (2021). Clustering based approach to enhance association rule mining, *2021 28th Conference of Open Innovations Association (FRUCT)*, pp. 142–150.

Mangalampalli, A. and Pudi, V. (2009). Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets, *2009 IEEE International Conference on Fuzzy Systems*, pp. 1163–1168.

Mar, Z. and Oo, K. K. (2020). An improvement of apriori mining algorithm using linked list based hash table, *2020 International Conference on Advanced Information Technologies (ICAIT)*, pp. 165–169.

Mäkinen, S., Skogström, H., Laaksonen, E. and Mikkonen, T. (2021). Who needs mlops: What data scientists seek to accomplish and how can mlops help?, *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, pp. 109–112.

Puneeth and Prasad Rao, K. (2019). A comparative study on apriori and reverse apriori in generation of frequent item set, *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pp. 337–341.

Sharma, V. and Bala, M. (2018). A credits based scheduling algorithm with k-means clustering, *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pp. 82–86.

Solanki, S. K. and Patel, J. T. (2015). A survey on association rule mining, *2015 Fifth International Conference on Advanced Computing Communication Technologies*, pp. 212–216.

Sweetlin Hemalatha, C., Vaidehi, V. and Lakshmi, R. (2015). Minimal infrequent pattern based approach for mining outliers in data streams, *Expert Systems with Applications* **42**(4): 1998–2012.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417414006149*

Yang, J., Huang, H. and Jin, X. (2017). Mining web access sequence with improved apriori algorithm, *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, Vol. 1, pp. 780–784.

Zhou, Y., Cui, J. and Liu, Q. (2020). Research and improvement of intrusion detection based on isolated forest and fp-growth, *2020 IEEE 8th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 160–164.