National College of Ireland

Enhancement of Apriori Algorithm for Applications of Data Mining using Frequent Pattern Tree

MSc Research Project
Cloud Computing

Shivam Pandey
Student ID: 20167725

School of Computing
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shivam Pandey
Student ID:	20167725
Programme:	Cloud Computing
Year:	2022
Module:	MSc Research Project
Supervisor:	Majid Latifi
Submission Due Date:	31/01/2022
Project Title:	Enhancement of Apriori Algorithm for Applications of Data Mining using Frequent Pattern Tree
Word Count:	XXX
Page Count:	22

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:	
Date:	31st 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):	

Enhancement of Apriori Algorithm for Applications of Data Mining using Frequent Pattern Tree

Shivam Pandey
20167725

Abstract

This research work shed light on the topic of enhancement of the Apriori algorithm in the application of Data Mining using cloud database platform. In this study, a clear description about the Apriori algorithm is stated and an example based on grocery database is taken to execute the Apriori algorithm for generating the frequent item sets. Finally a comparison among Apriori algorithm and FP growth algorithm has been done on the discussion section. During the implementation we have used cloud services from AWS like EC2 which helps to create instance which is easy to mange and configure. To sore the large amount of data S3 service of AWS is used which allows data to store and retrieve from anywhere. As our main aim to reduce to time complexity or time of execution of the dataset from cloud. EMR service of AWS has been utilized which internally uses Map-Reduce which help to process the large amount of data in a very less time which saves and execution cost of the large databases. In our research we have used 4 different size of database of 1000, 4000, 15000, 25000 records.

Keywords : Apriori algorithm, Improved Apriori algorithm, Itemsets, frequent itemset, candidate itemset, Time and space complexity, Big Data, cloud computing, MapReduce

Contents

1 Introduction **3**
- 1.1 Background and Motivation . . . 3
- 1.2 Research question . . . 3
- 1.3 Aim . . . 4
- 1.4 Research Objectives . . . 4
- 1.5 Research significance . . . 4

2 Related Work **4**
- 2.1 Concept of Data mining and its related algorithms . . . 5
- 2.2 Explanation of Apriori Algorithm based on Variation . . . 5
- 2.3 Comparison between Apriori algorithm and Eclat algorithm . . . 7
- 2.4 Role of Apriori algorithm in cloud-based Supply chain management . . . 7

3 Propose Research Methodology **8**
- 3.1 Framework . . . 9
- 3.2 Research Methods . . . 9

4 Design Specification **10**

5 Implementation **10**

6 Evaluation **14**
- 6.1 Discussion . . . 17

7 Conclusion and Future Work **19**

1 Introduction

Data mining refers to the process of finding patterns, correlations, and valuable information within a large set of databases to make predictions about outcomes. It is an effective scientific tool that helps to discover hidden information to predict future trends within an organization or working environment. Meaningful trends and patterns are extracted by implementing a data mining process to explore and examine a large number of datasets. Cloud Computing is related to the delivery of data and information from hardware and software tools by the access of the internet. It consists of databases, networking systems, and data storage to perform computing operations. In the last few years, this tool has rapidly grown due to the modification of information technology. Apriori algorithm is an important and useful tool and technique of data mining that makes relationships among various variables of a dataset. Association rule represents the relationship among different variables and items of the dataset. With the help of these relationships, irregularities within the large dataset could be found easily.

An interesting pattern from the dataset related to the association rules of mining could be extracted. In the following study, utilization of the Apriori algorithm to get frequent pattern trees through a data mining process has been properly highlighted. For this purpose Data mining and its related algorithms have been explained in this dissertation. In the methodology section of the study, the workflow of the Apriori algorithm has been designed and explained. The result and analysis part of the study has expressed the coding process of the Apriori algorithm to find patterns from the data set. The discussion part of the dissertation has discussed all the vital points related to methods and processes applied by the Apriori algorithm to get the desired outcome.

1.1 Background and Motivation

Cloud computing is one of the latest and useful tools of information technology that has transformed the traditional model of business into digital form. In the last few years, it has provided services to software engineering on a large scale by fluffing its demand. It is based on the advanced switches and servers that help to optimize the data mining method to enhance the extraction of data (Azeez et al. (2019)). Large databases could be stored with the help of technology and access to transferred data can be easy in any part of the world. Due to its specific and advantageous characteristics, it has been adopted by most business organizations across the world since 2012. Cloud computing needs different types of algorithms to store and manage data from a dataset. Apriori algorithm is one of the most desirable and effective algorithms models that make valuable patterns from data sets and result in important outcomes (Jia et al. (2019)). It is a common type of algorithm adopted in the process of data mining that maximizes the data storage and transfer in a cloud-based computing environment.

1.2 Research question

- How the complexity of Apriori algorithm can be reduced using frequent pattern tree to extract frequent itemsets in Data mining?
- How does the Apriori algorithm differ from other kinds of algorithms utilized in the process of data mining using cloud database?

1.3 Aim

The main aim of this dissertation is to evaluate the improvement in the performance of Apriori Algorithm by reducing the time and space complexity used in data mining for getting frequent pattern trees using cloud database.

1.4 Research Objectives

- Study on the state of the art works in data mining and Cloud computing algorithm
- To design the framework of the Apriori algorithm with coding by the use of Cloud Computing.
- To examine Apriori algorithms based on variation and its importance.
- To compare the Apriori algorithm with other kinds of algorithms implemented in data mining.
- To evaluate the main concept of data mining and its related algorithms.

1.5 Research significance

The following research is highly important and beneficial for getting ideas about the role and benefits of Cloud Computing in business activities. After conducting this research, a brief idea about the data mining process has been evaluated to store and manage data about a given dataset from the cloud database (Calcaterra et al. (2020)). By conducting this research, the concept of the Apriori algorithm has been identified and its function to find patterns from a dataset. Designing framework of Apriori algorithm would help learners to collect, store and extract data from the whole dataset. With the help of the Apriori algorithm framework, business analysts of an organization could find valuable information and patterns from the database to predict and analyze future events in their companies. Hence, this study paper would be highly beneficial for learners and business analysts to know the importance of applying the Apriori algorithm for finding information from large databases.

This research paper is divided into multiple sections as per the research conducted on the topic. The Section 2 provides information of Apriori algorithm used in various fields and the changes done in the algorithm to conduct research in the respective fields. In Section 3, methodology has been described with specifications to conduct the research on the enhancement of Apriori Algorithm. Section 4 provides the design and implementation of the architecture and framework of the research. Section 5 provides details of the process of the implementations of the tools and languages used in the research. Evaluation and results of the output are shown in section 6. In the last section 7 of the paper conclusion and future work has been discussed.

2 Related Work

The purpose of the research paper is to enhance the classic Apriori algorithm by minimizing the complexity of the classic algorithm. In this section, it is shown that how other researchers have used the Apriori algorithm in their respective field to get the maximum

advantage from the algorithm. The researches have proposed different methods to improve the algorithms as well the short coming of these methodologies. Our basic aim is to use the algorithm in cloud computing database and check till what extend the algorithm has improved. Apriori algorithms used in the field of data mining for the purpose of knowledge discovery. It basically tries to find the special occurring pattern in the large database. As now the data is moving on cloud so the development of the algorithm in the cloud computing can provide a significant output. In the past research the Apriori algorithms has used in many variations or the small changes with the algorithms to find association rules in their respective areas of research. Researches also compared the most used variation of the Apriori know as Eclat algorithm in some the cases where vertical approaches were required. Some of the most recent researches are group into the specific areas and show in the subsection mentions below:

2.1 Concept of Data mining and its related algorithms

The term Data mining refers to the extraction and discovery of a large set of data that is used in making relationships between product portfolio and architecture. It involves the machine learning method and statistics to predict the desired information about any event. Due to the adaptation of this technology, a sudden rise in data discovery has been seen in recent years (Abbasi and Moieni (2021)). Big data and warehouse technology of data have suddenly made a revolution in the information technology industry.

Organizations of these industries easily collected and analyzed the raw data about their customers and clients to make a proper analysis. This has performed a great toke making the proper examination of business activities at the current marketplace and making proper strategies to take appropriate action. A long-term and effective decision-making plan has been made by the organizations by getting proper data and information about a system or event.

There are different types of tools and methods that help data analysts to collect and analyze specific data. Tools like data visualization and analytics have been highly approved by most of the business and marketing analyses across the world. The process to conduct the data mining is related to machine and artificial intelligence-based technology. A specific algorithm is required to make the model of data mining that is based on a set of calculations and heuristics. Algorithms assist to analyze the types of data associated with the particular pattern. A pattern of data set clusters is formed that depend upon the types and gathered or collected data (Albahri et al. (2020)). By the design of algorithms, a mathematical and statistical model is formed that helps to predict a certain outcome.

There are various forms of algorithms adapted to design the model in data mining. This is the most desirable and useful algorithms adopted by most of the data analysis across the world are Apriori Algorithm, K-mean Algorithm, AdaBoost Algorithm, and CART Algorithm. Each of them has their specific role and performance in data mining, however, the Apriori algorithm is considered to be the best among them that possess the quality to evaluate certain association rules from a high and a large dimensional database.

2.2 Explanation of Apriori Algorithm based on Variation

Apriori Algorithm is considered to be the best and most useful algorithm used in designing data mining frameworks. It is based on association rules that make relationships between different variables connected to the database. Algorithms based on this type are highly

efficient and very different to be executed. A large amount of memory space is required to scan the data and store them for a long time safely (Leonard et al. (2019)) one of the major challenges for data analysis is to make and choose an easy form of Apriori algorithm that requires less space and scans the data quickly. This assists to enhance the versatility and robustness of the Apriori algorithm model. Improved Apriori algorithms always play a better and more effective role than the original form of the Apriori algorithm. The main reason for choosing the improved and latest version of the Apriori algorithm is its two major drawbacks during implementation (Hussain et al. (2018)). The drawbacks are a generation of a large amount of the candidate's data set and frequent scanning of the database. Due to these two reasons, an improved and latest model of the Apriori algorithm is adopted by the analysts to modify the process of data mining for their business performance and activities.

Apriori algorithm is the most desirable set of algorithms used to design the model of data mining by the analyst. Identification of items from the database could be done as easily by the application of this algorithm form. A proper and effective process could be generated that assists to make a basket analysis at the marketplace (Luna et al. (2019)). It is based on association rules that could help to learn the pattern of data collected from different sources. In the following study, some major advantages of using the Apriori algorithm have been described below.

- Apriori algorithm is most easy to understand the design of model among another learning algorithm
- Rules made for intuitive and communicating to the end issuer are suitable for them.
- An effective extension is found during the design and implementation of this kind of algorithm that helps to make an analysis of the ordering and numbering of items.
- No need to have labeled data for collection and examination that helps business analysts to use them in different ways and situations that enhance the accessibility of information (Perry et al. (2017)).
- Within this algorithm, all the rules could provide support for the system that maximizes its potential on a large scale.
- Apriori algorithm is highly important for splitting a large set of data into smaller parts and merging all of them into a single set of results.
- With the help of this algorithm, the dynamic of programming could be achieved quickly and in the most sophisticated way.
- It is highly beneficial for removing all the duplicate data and information and minimizing the length of the scanning number.
- Reduction of time can be done by the implementation of the Apriori algorithm to avail the data mining process and activity for ant organizations.
- With the application of association rules, the overall execution and performance of algorithms can be maximized for conducting data mining processes.

2.3 Comparison between Apriori algorithm and Eclat algorithm

The Equivalence Class Clustering and bottom-up Lattice Traversal algorithm known as the Eclat algorithm is one of the most popular processes used in association rules of data mining. It is a highly efficient and powerful tool that makes the execution of data faster. This kind of algorithm is highly applicable for retrieving and analyzing data from medium and small data sets (Soni et al. (2020)). During the transfer of data value from a data set it plays a vital role to support them according to the most desirable form. Like the Apriori algorithm, an Eclat algorithm also follows the mining rules to find the data from a large set. However, based on some specific criteria and conditions, they perform differently which creates differences among each other.

In Table 1 share some major differences between the Apriori algorithm and the Eclat algorithm have been shown.

Table 1: Apriori algorithm vs Eclat algorithm

Apriori algorithm	**Eclat algorithm**
Horizontal approach is followed by this kind of algorithm	A vertical approach is chosen to be followed in this algorithm.
As compared to the Éclat algorithm it has more parameters (Abbasi and Moieni (2021)).	This kind of algorithm consists of parameters as compared to the Apriori algorithm.
It is a lower process of execution than the Eclat algorithm.	It is a faster process than the Apriori algorithm.
More nodes are taken by the Apriori algorithm as compared to the Eclat algorithm for approaching targeted nodes.	As compared to the Apriori algorithm it takes a fewer number of nodes for reaching a targeted node (Wang et al. (2018)).
More than one time, the scanning of all the databases is done to generate desired output values.	Only one-time scanning of the database is done to approaches for desired value of output used for the data mining process

2.4 Role of Apriori algorithm in cloud-based Supply chain management

In the world of digitalization and information technology, large and complex data are collected from different sources. Different kinds of data like business data, cyber security data, health, and social media data are gathered. Based on the structured and unstructured form of data, various tools and techniques are implied related to machine learning and data mining. Data mining and management tools play an immensely effective role to collect and examine data about users and objects in more sophisticated ways. By the execution of the Apriori algorithm in data mining, several kinds of business activities could be made easier and autonomous (Keung et al. (2020)). In the Supply chain system of an organization, this algorithm plays an important role to perform the activities automatically and in multiple forms. RMFS[1] is one of the real-world examples of Apriori algorithm-based robotics systems that maximize the performance and efficiency of tasks and performance.

[1]Remotely Managed Franking System

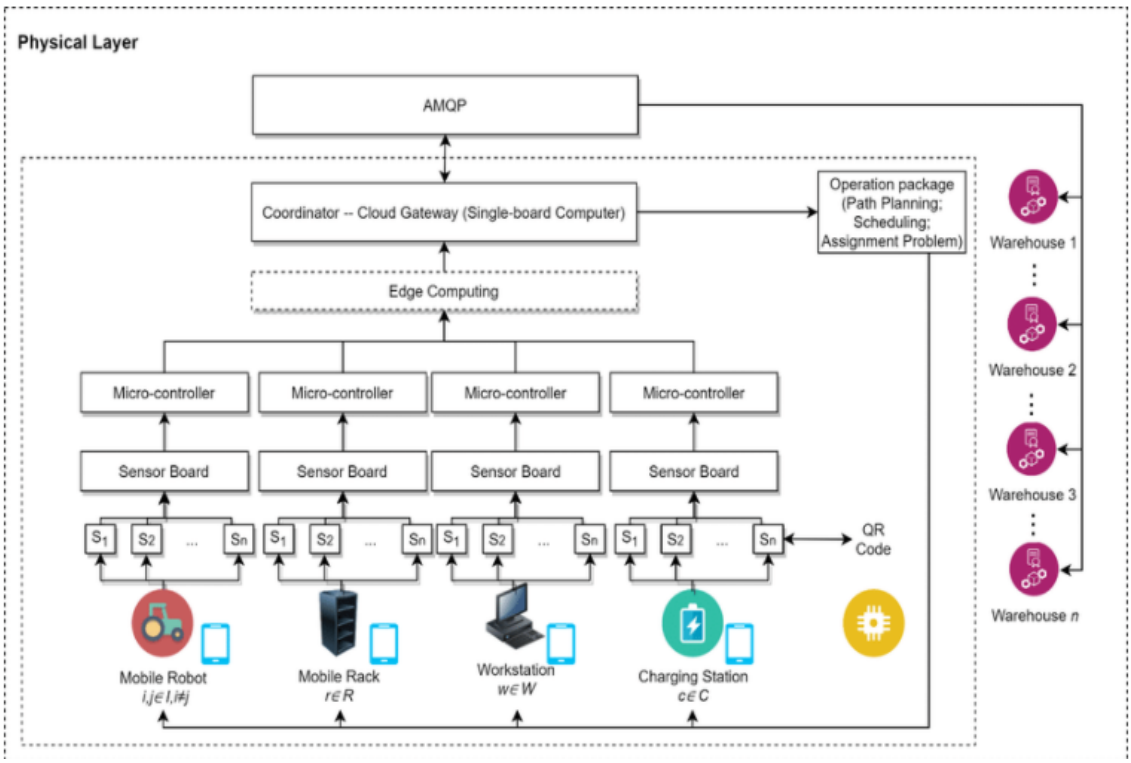

Figure 1: Physical structure of cloud-based CPS used in RMFS

By the application of this automatic system, multiple orders could be picked up to permit multiple tasks through robots. Different robots are assigned to perform specific tasks to pick and deliver the orders at a Supply chain system. It is based on cloud computing that can access the activity of robots from the desired location to perform their activity on an online platform. A proper Apriori algorithm is designed so that working schedule and methods of robots, avoidance of robot collisions and direction planning can be done from a remote location within the supply chain system of an organization.

In the Literature part of the study, all the essential points about data mining based on related algorithms have been mentioned. For this purpose, ideas about the algorithm named Apriori Algorithm, AdaBoost Algorithm, and K-mean Algorithm have been expressed. Brief information about the Apriori Algorithm and its importance on data mining has been highlighted in this part of the study. Also, the difference of comparison between two algorithms called the Apriori algorithm and Eclat algorithm has been clearly explained. A further role of using the Apriori algorithm in cloud-based software in Supply chain management has been discussed in this part.

3 Propose Research Methodology

Apriori algorithm is one of the most classical algorithms executed in the process of data mining with the help of cloud computing. Due to the growth of digitization and information technology, this tool has performed a great role in storing and transforming large sets of data from cloud-based platforms. Most of the successful business analysts from organizations of the world have adopted these scientific tools to design effective patterns and extract valuable information from large sets of databases. In the following study, the

methodology of designing and implementing the Apriori algorithm has been expressed as below.

3.1 Framework

Apriori algorithm is defined and run by the association rules used for data mining. It is a well-explored and prominent method to build relationships among different variables of a dataset. Different attributes of the database are also called items are set out for the transaction of data (Son et al. (2018)). Every transaction has a unique ID and also contains its subsets items.

In the following study, a data frame of the Apriori algorithm has been attached to explore the process of implementing this tool for getting valuable patterns from large datasets.

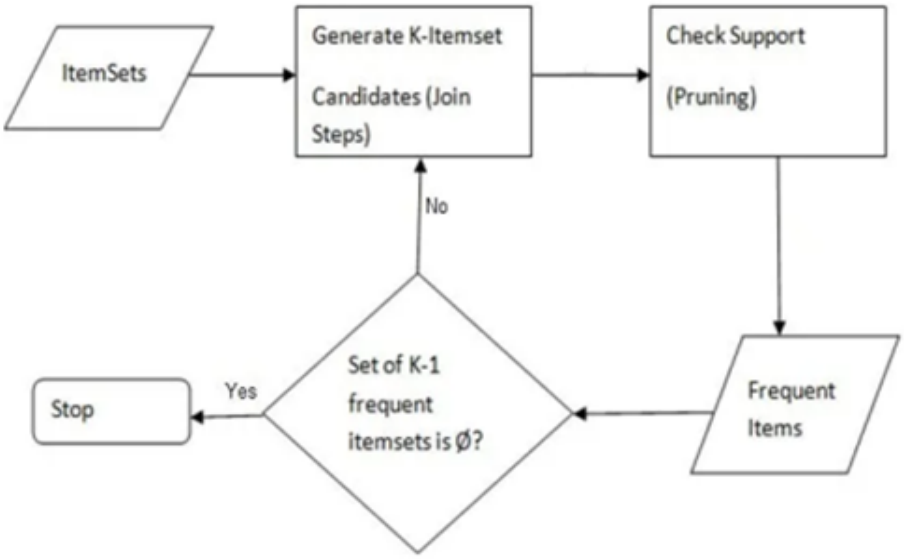

Figure 2: Apriori algorithm Framework

From the above figure, steps for approaching the Apriori algorithm by the enhancement of association rules have been identified. By over-viewing the framework it has been noticed that minimum items are required to generate k items in the given dataset. It is preceded by the self joining file by adding one attribute in the next step (Leonard et al. (2019)). After generating the k items, k+1 items are enhanced to generate frequent items in the dataset. After the generation of items, again a scan of the database is done to get frequent k items.

3.2 Research Methods

After the designing of the framework of the Apriori algorithm, some effective and potential working methods are adopted to run the algorithm. In the following study, all the essential steps have been defined as below.

- The collection of input data is the first step to designing and running the Apriori

algorithm. In this method input data has been collected in minimum data to transit maximum numbers from a dataset (Padillo et al. (2017)).

- After collecting data, input values are converted into key values that are uploaded to the Hadoop tool.
- Input Items are generated and grouped by removing null and duplicate values.
- Similar items are combined to get the uniform whole data set having similar properties.
- By the application of the association rule, a hidden relationship among the different variables of the dataset is built that has helped for enhancing the framework of the Apriori algorithm.

The last step is Evaluation of the Enhanced Apriori Algorithm

4 Design Specification

In the research design section of designing the framework of Apriori algorithm, MapReduce programming framework to make improved Apriori algorithm. This tool is associated with the distribution of data in parallel form. It is based on two phases called mapper and reducer phases. The function of mapper pages is input applied in the pair of key-value forms. However, the reducer phase is related to taking input from the mapper phase to generate the outcome (Sevri et al. (2017)). By the application of the MapReduce programming tool, a large set of data has been distributed into parallel forms. The mapping key has enhanced the filtering of undesired input values like duplicate and null values in the input attributes. Input data has been gathered by the reducer function to generate the outcome. The whole process is run by the big data and cloud computing tools to reduce and distribute large data into parallel form. Hadoop tool has been utilized to suture the data and run software applications for data mining. By the application of this tool, an open-source platform has been built to store and compute large amounts of data from a given database. In the design of the Apriori algorithm, Hadoop has been adopted to make an open-source platform for enhancing the Apriori Algorithm executed in data mining. Also, pseudo-code has been generated to run the design Apriori algorithm based on Hadoop and MapReduce.

5 Implementation

To implement the cloud architecture the cloud services and tools used are mentioned below:

- Amazon Elastic Compute Cloud (EC2) Service: EC2 is one of the most used service of AWS. It is used to create and deploy instances with little configuration.
- Amazon Simple Storage Service (S3): AWS s3 service provide a storage on the cloud to store and retrieve the dataset. It can be access anytime from anywhere.
- Amazon Elastic MapReduce (EMR) Service: This service is used to process large amount of data. It used to lower the cost of processing data in much less time.

- Anaconda: It is an open-source software used working with data with python.
- Jupyter Notebook: It is used to run python codes.

After designing the framework and model for the Apriori algorithm, the major challenge for the researcher is to Implement the framework for getting the desired outcome. By the appropriate application of MapReduce programming and Hadoop, the design and running of programs made for the Apriori algorithm have been done.

The pseudo code for candidate generation is mentioned in the below image for Frequent set calculation:

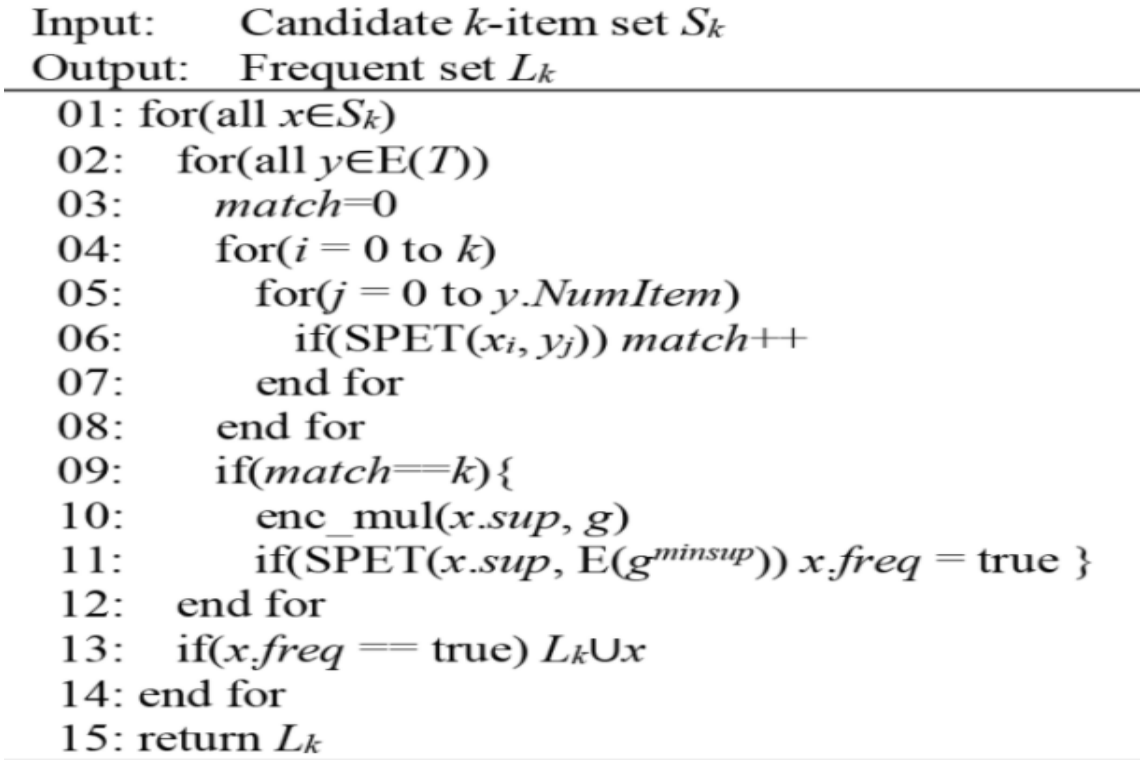

Figure 3: Frequent set calculation pseudo code

The above candidate code changes has implemented in Enhanced Apriori algorithm mentioned below:

In the following study, Appropriate steps and methods to implement the Apriori algorithm have been clearly explained.

1. Staring python Jupyter notebook in anaconda tool
2. Importing the required library on the jupyter notebook has been done.
3. After this process loading and putting data has been chosen (Singh et al. (2018)).
4. In the next stage of Implementation, the cleaning of extra space in the jupyter notebook has been chosen.
5. Based on the region of the transaction, splitting of input data has been done.

Input: Encrypted transaction database E(T)
Item set length k
Candidate pattern set S_k
Output: Frequency pattern set L_{k-1}

01: $L_1 = \{l_1,\ldots,l_n \mid \forall\, l \in E(T)\}$
02: $k = 2$
03: while(TRUE)
04: $E(S_k)$ = Candidate_set_generation($E(L_{k-1})$)
$=\{c_1,\ldots,c_p | c \in k$ candidate set$\}$
05: if($E(S_k)=\emptyset$) return $E(L_{k-1})$ to AU
06: $E(L_k)$= Frequent set calculation($E(T)$, $E(S_k)$)
07: k++
08: end while

Figure 4: The Enhanced Apriori Algorithm pseudo code

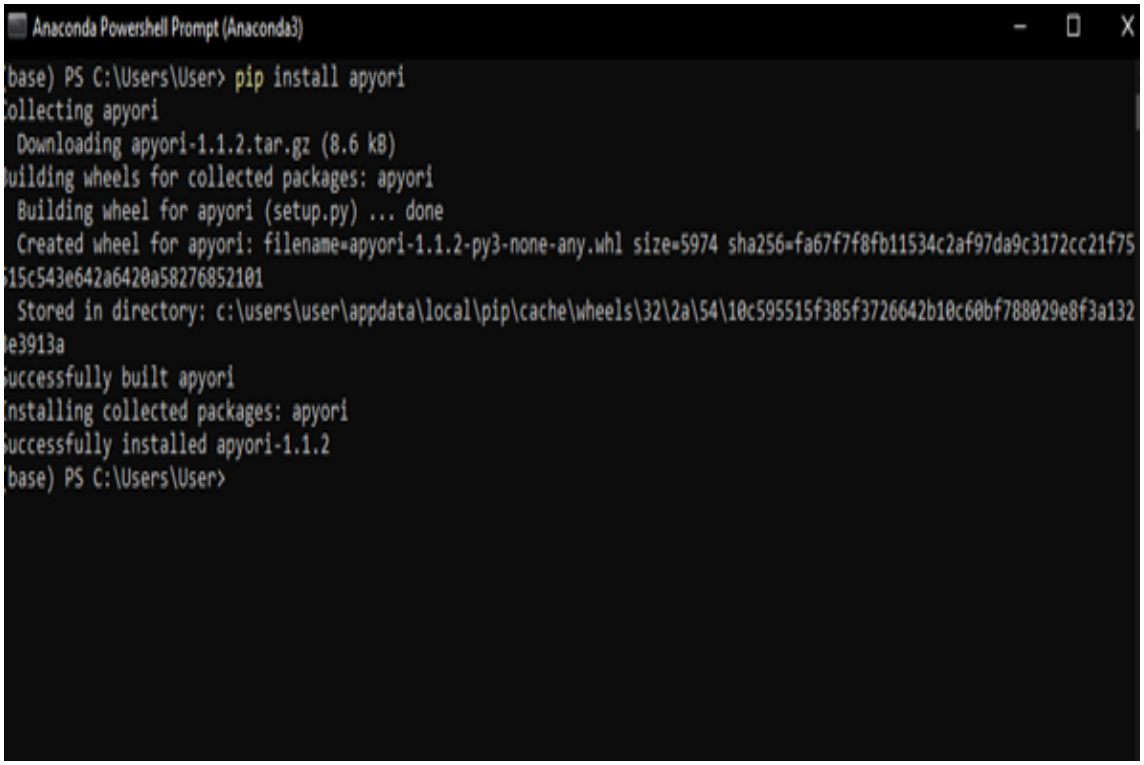

Figure 5: Anaconda console

Type *Markdown* and LaTeX: α^2

Type *Markdown* and LaTeX: α^2

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from apyori import apriori
```

Type *Markdown* and LaTeX: α^2

In [2]:
```
store_data= pd.read_csv("groceries.csv", header=None)
```

In [3]:
```
store_data.head()
```

Out[3]:

	0	1	2	3	4	5	6	7	8	9	...	23	24	25	26	27	28	29	30	31	32
0	Item(s)	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	...	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28	Item 29	Item 30	Item 31	Item 32
1	4	citrus fruit	semi-finished bread	margarine	ready soups	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 6: Data is stored and in the store_data variable from the CSV file

In [33]:
```
records = []
for i in range(1, 9836):
    records.append([str(store_data.values[i, j]) for j in range(0, 20)])
```

Figure 7: Fetching data with data values

6. Encoding of data has been performed to make suitable data.

7. Lastly, building and analysis of models have been done for the implementation of the Apriori algorithm as per the pseudo code.

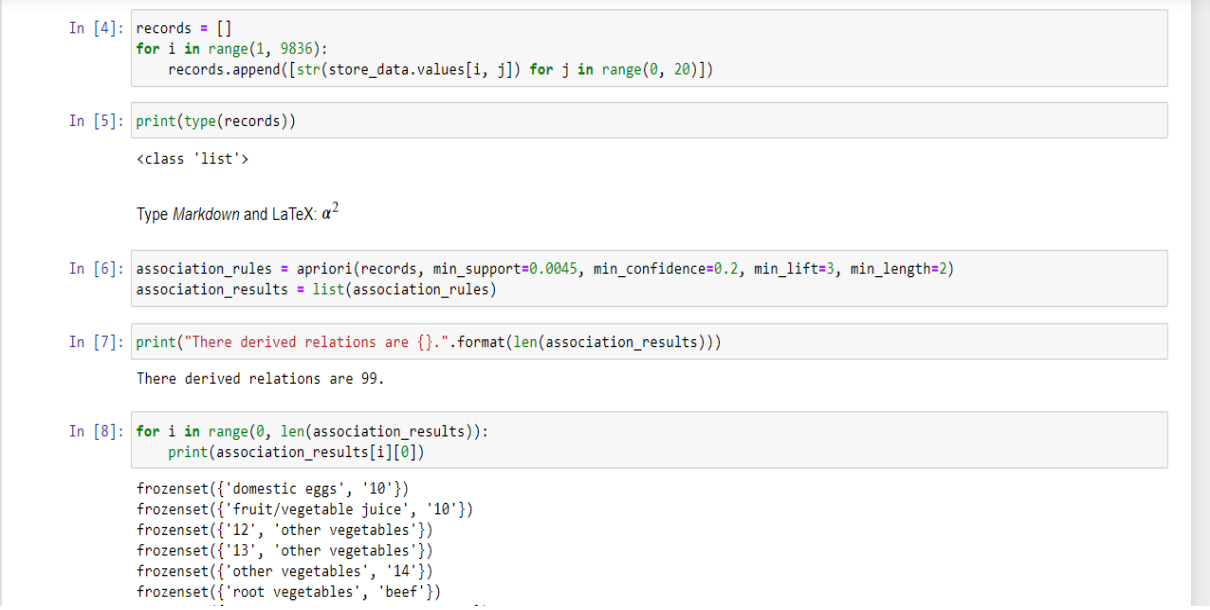

Figure 8: Appling association rules of on Enhanced algorithm

6 Evaluation

In this result and analysis section, an example dataset based on market basket data is taken from the Kaggle[2]. The name of the dataset is groceries which is a .csv file. According to the requirement, the Apriori algorithm is going to be implemented into this particular data set. Before going into the in-depth analysis of the result section, a brief overview of the method to proceed with this program is going to be described here. As the Apriori algorithm is executed in the python language, the selected platform to run the Apriori algorithm is Jupyter Notebook (Anaconda3). The methodology part to generate frequent item sets from a groceries dataset is read initially and before that, some of the pip packages need to be installed by using an anaconda console.

The packages are numpy, pandas, matplotlib.pyplot, and Apriori which are being imported executing the Apriori algorithm successfully. The available dataset is stored into a variable named stored_data in the programing (Yuan (2017)). The number of rows associated with the groceries dataset is 9836. After displaying the entire dataset into the executable platform a python function append is executed within the "for loop" to add the unique item to the mentioned dataset.

A for loop having the "range(0, len(association_results))" is going to be executed to print the frequent dataset items for the particular groceries data is defined within the CSV file.

Using this result it is visible that numerous possible association rules can be possible with this number of frequent item sets. The possible association rules are the final output

[2]https://www.kaggle.com/

```
In [37]: for i in range(0, len(association_results)):
             print(association_results[i][0])
```

```
frozenset({'10', 'domestic eggs'})
frozenset({'10', 'fruit/vegetable juice'})
frozenset({'other vegetables', '12'})
frozenset({'other vegetables', '13'})
frozenset({'other vegetables', '14'})
frozenset({'beef', 'root vegetables'})
frozenset({'whipped/sour cream', 'berries'})
frozenset({'bottled beer', 'liquor'})
frozenset({'bottled beer', 'red/blush wine'})
frozenset({'sugar', 'flour'})
frozenset({'herbs', 'root vegetables'})
frozenset({'sliced cheese', 'sausage'})
frozenset({'10', 'domestic eggs', 'nan'})
frozenset({'10', 'nan', 'fruit/vegetable juice'})
frozenset({'11', 'other vegetables', 'whole milk'})
frozenset({'nan', 'other vegetables', '12'})
frozenset({'other vegetables', 'whole milk', '12'})
frozenset({'nan', 'other vegetables', '13'})
frozenset({'nan', 'other vegetables', '14'})
frozenset({'beef', 'root vegetables', 'other vegetables'})
frozenset({'beef', 'root vegetables', 'rolls/buns'})
frozenset({'beef', 'root vegetables', 'whole milk'})
frozenset({'beef', 'root vegetables', 'yogurt'})
frozenset({'nan', 'whipped/sour cream', 'berries'})
frozenset({'nan', 'bottled beer', 'liquor'})
frozenset({'nan', 'bottled beer', 'red/blush wine'})
```

Figure 9: Frequent Item sets (frozenset) output

of the Apriori algorithm as this algorithm states that it is a procedure of identifying the frequent items individually from a database and extracting rules from frequent items. This is a generalized view for applying an Apriori algorithm into a particular data of a database. However, due to complexity and irrelevancy, not every database is suitable for applying an Apriori algorithm (Ayofe et al. (2019)). This is visible that the Apriori algorithm is mostly used for market-basket analysis and for that reason groceries data are best suited for getting a proper result in terms of generating association rules from these datasets.

```
In [9]: for item in association_results:
            # first index of the inner list
            # Contains base item and add item
            pair = item[0]
            items = [x for x in pair]
            print("Rule: " + items[0] + " -> " + items[1])

            # second index of the inner list
            print("Support: " + str(item[1]))

            # third index of the list located at 0th
            # of the third index of the inner list

            print("Confidence: " + str(item[2][0][2]))
            print("Lift: " + str(item[2][0][3]))
            print("====================================")

Rule: domestic eggs -> 10
Support: 0.005083884087442806
Confidence: 0.2032520325203252
Lift: 3.208641636978167
====================================
Rule: fruit/vegetable juice -> 10
Support: 0.005388917132689374
Confidence: 0.2154471544715447
Lift: 3.001307031483912
====================================
Rule: 12 -> other vegetables
```

Figure 10: Code for Association Rule Generation output

The above picture is going to visualize a set of output data which are final association rules for existing frequent item sets. A for loop is again going to run to print the values of rules, support of that particular frequent item sets, the confidence of the frequent item sets, and lift value of the particular data points.

In a discussion of the analysis of the above outcome, it is quite obvious that analysis of big data using an Apriori algorithm is much more effective in terms of time and

```
Rule: 13 -> other vegetables
Support: 0.004778851042196238
Confidence: 0.6025641025641024
Lift: 3.1141450072085903
====================================
Rule: other vegetables -> 14
Support: 0.00498220640569395
Confidence: 0.6363636363636364
Lift: 3.288826255195146
====================================
Rule: root vegetables -> beef
Support: 0.017386883579054397
Confidence: 0.3313953488372093
Lift: 3.0403668431100312
====================================
Rule: berries -> whipped/sour cream
Support: 0.009049313675648195
Confidence: 0.27217125382262997
Lift: 3.796885505454703
====================================
Rule: liquor -> bottled beer
Support: 0.004677173360447382
Confidence: 0.4220183486238532
Lift: 5.280598547984218
====================================
Rule: red/blush wine -> bottled beer
Support: 0.004880528723945094
Confidence: 0.253968253968254
Lift: 3.1778343228724912
====================================
```

Figure 11: Output for Enhanced Apriori Algorithm

space complexity of a program. This algorithm is an efficient way to extract frequently purchased data from any supermarket rather than applying the Brute Force Algorithm into these Market Basket data points. This algorithm not only shows an error-free result of association rules but also applies this method to generate a speed-up evaluation of mining the rules from the item sets (Alshawi (2016)). Apriori algorithm maintains two principles, this algorithm uses two steps join and prune method in order to control the space of search. This approach is an iterative one to compute the frequent items from a database.

6.1 Discussion

After reviewing the whole study, it has been observed that enhancement of the Apriori Algorithm applied in data mining using frequent pattern trees has been properly highlighted. As we compare the time time by the old apriori algorithm show highlighted with the blue colour vs the new algorthm shown with pink colour observe that with differnt 4 dataset at 1000, 4000,15000, 20000 datasets records and shown that at small number of records the Improved algo shows better result ar compare to large dataset.

In the whole section of the study, a brief idea about the Apriori algorithm has been explained. By the overview of the above paragraph, it has been noticed that Cloud Computing has transformed the data storage and transfer process. Due to the sudden rise of information technology, this tool has got a chance to grow and develop on a large scale (Sireesha et al. (2018)). Data mining is a scientific process that is enhanced by Cloud Computing to extract valuable information and patterns from large databases. By the execution of this tool, meaningful patterns and trends are collected to make proper

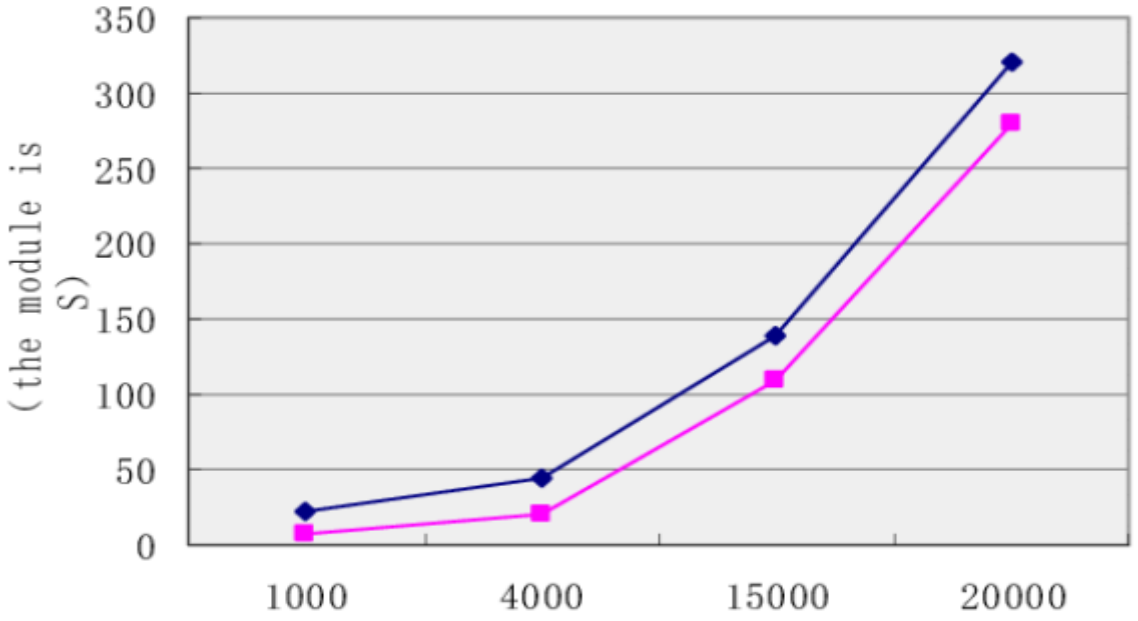

Figure 12: Apriori algorithm comparison chart before and after

analyses and predictions about future events.

Apriori algorithm is one of the most classical algorithm tools that is utilized in the process of data mining. Implementation of this tool plays a vital role in making relationships among different variables of a large dataset (Zeebaree et al. (2020)). With the help of Cloud Computing running an Apriori algorithm performed remotely to access particular information. There are different types of algorithms that help to extract information and data from a dataset. However, the researcher has chosen the Apriori algorithm to collect and store valuable data. Apriori algorithm framework that has been attached in the methodology section of the study, has explained the design model of this algorithm. This framework is run by the association rules in which relationships among different variables are built within a database.

In this rule, items are generated frequently to get support for each item for combining similar values. In the research method, the whole process of running and executing the Apriori algorithm flow chart has been clearly explained. For this purpose, both MapReduce programming and Hadoop tools have been adopted to distribute the input data into parallel forms (Zhou (2020)). These tools and methods have performed a great role in removing undesired and null values from input items within the algorithm. An open-source platform has been built by Hadoop to store and examine large data from a given database. The implementation section of the study has explained the methods for utilizing the framework and programming model to get the required outcome. All the steps mentioned in this section are essential to run and execute programming steps for getting desired Apriori algorithm used for data mining.

In this discussion section, a few pros and cons of the Apriori algorithm and some of the advantages of this algorithm over the other existing algorithm for generating frequent

itemsets is going to be discussed in this particular research work (Isong et al. (2018)).

Some of the pros of this algorithm are included as:

- The Apriori algorithm is a classic algorithm among all the other methods therefore it is easy to implement and simple to understand for beginners.
- The final outcome of this algorithm is intuitive and the generated rules are easily understandable for the users.
- This algorithm follows a completely unsupervised technique hence it does not require any labeled data as the target dataset to apply Apriori. Unlabelled data are more accessible than labeled data therefore this method works more efficiently for various datasets.

Cons

- Apriori algorithm can run nicely for small datasets but in case of large datasets, it gives erroneous results, and sometimes it takes more time to evaluates association rules which are not correct or efficient. Hence, this algorithm is not showing better result for large databases as it creates erroneous results.

The most important aspect of frequent item sets is that it takes less memory allocation and computing time (Bu (2018)). Both Apriori and frequent pattern growth algorithms is the most primary algorithm for extracting frequent item sets.

- Apriori algorithm generates frequent items by applying pairing of the item sets in a way that single item sets, double item sets, triple item sets, and so on. Whereas the FP growth algorithm creates frequent pattern trees for generating frequent items.
- Apriori algorithm follows breadth-first search method and FP growth algorithm follows depth-first search method.
- Apriori algorithm generates candidates which follow to extend frequent subsets one time in iteration. In case of the FP growth algorithm, this creates a conditional FP-Tree using each item from the dataset. Hence, in conclusion, both algorithms are competent to work for market-basket datasets.

7 Conclusion and Future Work

After reviewing the whole study, it has been observed that all the essential points about the enhancing Apriori Algorithm for applications of data mining for getting frequent pattern trees have been properly highlighted.

In Result and Analysis part, coding of the Apriori algorithm with the help of appropriate tools has been mentioned. A file of coding has been attached to express the required coding to run the desired Apriori algorithm to apply data mining processes. Also, an explanation of all the phases of coding that have been done on the Jupyter notebook platform has been discussed in this part of the research paper. The discussion section of this research paper has explained all the above-mentioned parts of the study that has been defined. All the essential points of the entire part of the dissertation have been briefly discussed in this portion of the study paper. Therefore, this research paper

would be highly beneficial for the learners and business analysts to get proper information about the Apriori algorithm for extracting information and patterns from large data-sets. As we have seen the Enhanced Apriori algorithm provides better result in small number of records but once the size of records increases it almost give the same result as old algorithm so in future this enhancing algo for high number of records can be considered.

References

Abbasi, S. and Moieni, A. (2021). Bloomeclat: Efficient eclat algorithm based on bloom filter, *Journal of Algorithms and Computation* **53**(1): 197–208.
URL: *https://jac.ut.ac.ir/article*$_8$1890*.html*

Albahri, A. S., Hamid, R. A., k. Alwan, J., Al-qays, Z. T., Zaidan, A. A., Zaidan, B. B., Albahri, A. O., AlAmoodi, A. H., Khlaf, J. M., Almahdi, E. M., Thabet, E., Hadi, S. M., Mohammed, K. I., Alsalem, M. A., Al-Obaidi, J. R. and Madhloom, H. T. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (covid-19): A systematic review.

Alshawi, A. (2016). Applying data mining techniques to improve information security in the cloud: A single cache system approach, *Scientific Programming* **2016**: 1–5.

Ayofe, N., Ayemobola, T. J., Misra, S., Maskeliūnas, R. and Damaševičius, A. R. (2019). Network intrusion detection with a hashing based apriori algorithm using hadoop mapreduce, *Computers* **8**(4).
URL: *https://www.mdpi.com/2073-431X/8/4/86*

Azeez, N. A., Ayemobola, T. J., Misra, S., Maskeliūnas, R. and Damaševičius, R. (2019). Network intrusion detection with a hashing based apriori algorithm using hadoop mapreduce, *Computers* **8**.

Bu, F. (2018). A data mining framework for massive RFID data based on apriori algorithm, *Journal of Physics: Conference Series* **1087**: 022020.
URL: *https://doi.org/10.1088/1742-6596/1087/2/022020*

Calcaterra, C., Carmenini, A., Marotta, A., Bucci, U. and Cassioli, D. (2020). Maxhadoop: An efficient scalable emulation tool to test sdn protocols in emulated hadoop environments, *Journal of Network and Systems Management* **28**.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M. and Ribata, N. (2018). Educational data mining and analysis of students' academic performance using weka, *Indonesian Journal of Electrical Engineering and Computer Science* **9**.

Isong, B., Ntshabele, K., Moemi, T., Dladlu, N. and Gasela, N. (2018). Efficient 3-tier hybrid encryption model for improved data security in the cloud landscape, *PONTE International Scientific Researchs Journal* **74**.

Jia, L., Xiang, L. and Liu, X. (2019). An improved eclat algorithm based on tissue-like p system with active membranes, *Processes* **7**: 555.

Keung, K., Lee, C., Ji, P. and Ng, K. K. (2020). Cloud-based cyber-physical robotic mobile fulfillment systems: A case study of collision avoidance, *IEEE Access* **8**: 89318–89336.

Leonard, L., Miles, B., Heidari, B., Lin, L., Castronova, A. M., Minsker, B., Lee, J., Scaife, C. and Band, L. E. (2019). Development of a participatory green infrastructure design, visualization and evaluation system in a cloud supported jupyter notebook computing environment, *Environmental Modelling and Software* **111**.

Luna, J. M., Fournier-Viger, P. and Ventura, S. (2019). Frequent itemset mining: A 25 years review.

Padillo, F., Luna, J. M., Herrera, F. and Ventura, S. (2017). Mining association rules on big data through mapreduce genetic programming, *Integrated Computer-Aided Engineering* **25**.

Perry, D. C., Brown, J. A., Possin, K. L., Datta, S., Trujillo, A., Radke, A., Karydas, A., Kornak, J., Sias, A. C., Rabinovici, G. D., Gorno-Tempini, M. L., Boxer, A. L., May, M. D., Rankin, K. P., Sturm, V. E., Lee, S. E., Matthews, B. R., Kao, A. W., Vossel, K. A., Tartaglia, M. C., Miller, Z. A., Seo, S. W., Sidhu, M., Gaus, S. E., Nana, A. L., Vargas, J. N. S., Hwang, J. H. L., Ossenkoppele, R., Brown, A. B., Huang, E. J., Coppola, G., Rosen, H. J., Geschwind, D., Trojanowski, J. Q., Grinberg, L. T., Kramer, J. H., Miller, B. L. and Seeley, W. W. (2017). Clinicopathological correlations in behavioural variant frontotemporal dementia, *Brain* **140**.

Sevri, M., Karacan, H. and Akcayol, M. (2017). Crime analysis based on association rules using apriori algorithm, *International Journal of Information and Electronics Engineering* **7**: 99–102.

Singh, S., Garg, R. and Mishra, P. K. (2018). Performance optimization of mapreduce-based apriori algorithm on hadoop cluster, *Computers and Electrical Engineering* **67**.

Sireesha, M., Vemuru, S. and TirumalaRao, S. N. (2018). Coalesce based binary table: An enhanced algorithm for mining frequent patterns, *International Journal of Engineering and Technology(UAE)* **7**.

Son, L. H., Chiclana, F., Kumar, R., Mittal, M., Khari, M., Chatterjee, J. M. and Baik, S. W. (2018). Arm–amo: An efficient association rule mining algorithm based on animal migration optimization, *Knowledge-Based Systems* **154**.

Soni, A., Saxena, A. and Bajaj, P. (2020). A methodological approach for mining the user requirements using apriori algorithm, *Journal of Cases on Information Technology* **22**.

Wang, F., Li, K., Duić, N., Mi, Z., Hodge, B. M., Shafie-khah, M. and Catalão, J. P. (2018). Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns, *Energy Conversion and Management* **171**.

Yuan, X. (2017). An improved apriori algorithm for mining association rules, *AIP Conference Proceedings* **1820**(1): 080005.
URL: *https://aip.scitation.org/doi/abs/10.1063/1.4977361*

Zeebaree, S., Shukur, H., Haji, L., Zebari, R., Jacksi, K. and Abass, S. (2020). Characteristics and analysis of hadoop distributed systems, *Technology Reports of Kansai University* **62**: 1555–1564.

Zhou, Y. (2020). Design and implementation of book recommendation management system based on improved apriori algorithm, *Intelligent Information Management* **12**.