DECLARATION ON PLAGIARISM

Names:	Teenu Prathyush (21262966), Teena Sharma (21261593)
Programme:	MSc. in Computing (Data Analytics)
Module Code:	CA683
Assignment Title:	CA683 Project – Movie Recommendations using Association Rule Mining
Group No.:	11
Submission Date:	2022-04-07
Module Coordinator:	Prof. Andrew McCarren
Link for project:	https://drive.google.com/drive/folders/1qmAoFWeEhLJxkkGu-mmQnEOgT7qPCb0E?usp=sharing

I/We declare that this material, which I/we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I/We understand that plagiarism, collusion, and copying are grave and serious offenses in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion, or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines.

Name: Teena Prathyush Date: 7th April, 2022

Name: Teenu Sharma Date: 7th April, 2022

MOVIE RECOMMENDATIONS USING ASSOCIATION RULE MINING

ABSTRACT

In the current research landscape, recommendation systems are one of the most popular topics. We can use recommendation systems for recommending movies, music, television series, books, websites, and other entities. In this study, we investigated how to utilize the most appropriate recommendation system for movie retail shops. Online services such as Netflix, Prime Video, and Search Engines use recommendation systems, but movie retail shops don't typically utilize them. The Apriori algorithm, implemented by us, identifies frequent itemsets in a transactional dataset and generates relevant association rules between movies. By using the Apriori algorithm, movie retail shops can benefit and come up with different marketing strategies based on the obtained rules. In this research study, we will compare the Apriori algorithm's performance with another association rule mining algorithm called ECLAT (Equivalence Class Clustering and Bottom-Up Lattice Traversal) and also generate relevant association rules for different genres of movies using the Apriori algorithm.

Keywords: Recommendation Systems, Apriori, Association Rules, ECLAT, Marketing.

I. INTRODUCTION

Recommendation systems have gained a lot of traction in the past decade. There has been a significant amount of research conducted in this area and researchers have come up with various algorithms and improved existing algorithms for generating recommendations between items. A majority of organizations today implement large scale recommendations systems to fulfil their customer needs. Amazon, Netflix, and Facebook are a few organizations that implement recommendation systems in their products. For example, Amazon in its e-commerce website, recommends items that are related to existing items in the cart as well as previously bought together items. Facebook also utilizes recommendation systems to identify relevant connections of people that users might recognize. Netflix is also very similar to this since it takes into account the users' watch history and makes recommendations based on this [13]. A recommendation system can be broadly classified according to how it works. Content-based filtering, Collaborative filtering, and Hybrid recommendation system are some of the popular recommendation systems used in real world scenarios. By analysing the user's prior behaviour and identifying patterns, a content-based filtering system can recommend items that are similar to their past behaviour. Using collaborative filtering, you can analyse the user's ratings and previous experiences. Finally, the hybrid recommender system is a combination of the content based filtering and collaborative filtering recommendation systems [13].

In this research study, we proposed a content-based movie recommendation system. This recommendation system recommends users the most related movies based on previous purchases. The movie recommendation systems will help the user with movie recommendations, saving them time and introducing them to new movies that they might like. This paper will present a model that has been implemented using association rule mining concepts to recommend movies on the basis of different genres. The paper contains the following sections: Section II explains related work which has been conducted in the area of Association Rule Mining, Section III explains the Data Mining methodology used for this research, and section IV describes the Evaluation and Results of our research and Sections V will provide the Conclusion and Future Scope.

II. RELATED WORK

Saltz et al. [1] proposed an approach to identify the strengths and weaknesses of CRISP-DM methodology and also explained the key actions while considering the weakness of a model. They provided the benefit of utilizing CRISP-DM for data science in that it outlines a series of simple, common-sense processes for the team to follow.

Robu et al. [2] established a technique for identifying the most general characteristics in a transactional dataset generated by software logging users' choices. They demonstrated how to use the R opensource software to apply the Apriori and Eclat algorithms in their study. They designed a model in which two algorithms are iterated with varied minimum support thresholds to test the performance of the two algorithms and used to test the validity of common patterns. The minimum support thresholds employed in this study were defined at various levels, defining the minimum frequency of appearance of a certain item set in the dataset.

In their study, Mohapatra et al. [3] use two association rules mining algorithms - Apriori and Eclat to forecast the likelihood of subordinate products being purchased if a prime item is procured. The Apriori method finds frequent item sets by identifying the most common individual items in a database and expanding them to bigger and larger sizes, whereas the Eclat algorithm finds frequent itemsets by using a depth-first search. From their study, they have been discovered that if the rule's confidence is greater than support, the lift is positively associated, and vice versa. When confidence equals support, the lift is independent of the rule.

Feng et al. [4] have proposed the MH Apriori algorithm by improving the efficiency of the Apriori algorithm. By combining the benefits of MapReduce and HBase and forming an algorithm named MH Apriori after optimizing the original or classical Apriori algorithm. They depict that this algorithm paves the way to scan the database efficiently by saving the matching pattern. Their experimental results show that this improved MH Apriori algorithm is both efficient and scalable.

In [5] Zheng proposed methods to improve the classical Apriori algorithm by using Hash Table. At each step of execution, the standard Apriori method must scan the full transaction database, which takes up too much space and is computationally time-intensive. This paper's upgraded Hash table technology can minimize the amount of space taken by the candidate pool and the efficiency of operation is highly improved.

Yang et al. [6] proposed a new and improved - AC apriori algorithm from the traditional Apriori algorithm. This improved algorithm helps the online advertising companies to advertise based on the user's interest from their search history. To identify search histories of the user, a Contiguous serial pattern is to be used in an apriori algorithm as it helps to predict the forthcoming request from a particular user base on their search history and also it improves the design or topology based on their device from which they search.

Chang et al. [7] has proposed the apriori algorithm for performing Data Mining and Data Analytics in a cloud environment. In their work, they have mentioned that the effectiveness of the Apriori algorithm can be increased and parallel mining can be obtained if it is implemented in the Hadoop framework. They discovered and demonstrated that the MR-Apriori technique has consistent scalability as the number of nodes rises. It shows that the MR-Apriori method can successfully be implemented under the Hadoop architecture, and therefore it adapts to cloud computing applications.

Cong et al. [8] has proposed the improved apriori algorithm by identifying shortfalls of traditional association rule mining. When using the Apriori algorithm to find common itemsets, there are two distinct characteristics: First, it is a multi-iteration advanced calculating method, which means that it must start with frequent 1 and loop over the preceding frequent itemset before generating the maximum order item set; second, it is a test to locate all the frequent itemsets. The revised Apriori method not only creates the interested item set, but also excludes transactions that do not contain the items of interest, decreasing the program's running time and increasing its efficiency.

Kesarwani et al. [9] have proposed MSD-Apriori, a hybrid strategy for discovering borderline-rare components that are below but close to the minimal support level and have a high link with common items. By combining MSApriori and Dynamic Apriori, a hybrid technique is created. By mining association rules, MSApriori discovers borderline-rare item sets from weblogs, and Dynamic Apriori identifies those items among these that have a significant correlation with the frequent items.

Xu et al. [10] have proposed an MSB Apriori + algorithm. It is an advancement of the basic MSApriori algorithm. MSApriori differs from basic Apriori in various ways, and it is not as simple to interpret as the basic Apriori. They provide advantages of the MSB Apriori+ algorithm and that it might be more appropriate than MSApriori in some real-world applications as it is easier to comprehend and can be used as an efficient replacement for MSApriori.

Chun-Sheng et al. [11] identify the defects of the classical Apriori algorithm by determining when there will be low support, then redundant frequent itemset will be more. Local effective association rules have a higher level of confidence and a lower level of support that cannot be mined. Researchers identified that the classical Apriori algorithm fails to find local strong association rules while data mining. They proposed two algorithms as correction algorithms based on confidence and classification such as Apriori-con and Apriori classification.

Puneeth et al. [12] has constructed a reverse Apriori algorithm which is a combination of Frequent Pattern (FP) structure and Apriori algorithm. The proposed reverse apriori is more efficient than classical Apriori in terms of computational time for running and parsing the dataset in data mining. Their experimental results depict that Reverse Apriori is more efficient than Apriori algorithm in terms of computational speed, as they compared these terms between the two algorithms.

III. CRISP-DM METHODOLOGY

We have adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to carry out this project. This methodology consists of six phases that includes Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. In the following sub-sections, we are going to describe the process of conducting our project tasks under these six phases.

A. Business Understanding

In the first phase, we mainly focus on understanding the objectives and requirements of the business. It is important to determine the project requirements, business objectives at the beginning of a project. We should also assess the risks involved, and finally select the tools and technologies that will be used in our project.

From a marketing point of view, it is necessary for movie retail shops to organize their store layout and sell movies in a way that is attractive to customers. It is important for them to give discounts or benefits to customers who purchase certain products to increase their sales. From a data scientist perspective, this can be accomplished by using historical data to recommend certain products that can be used together to increase the sales.

The objectives of this business process is to determine which movies “go together” i.e., to form association rules of similar movies that are bought together. Historical data on movies that are bought together will be required to perform this task. The risks associated with this problem are – historical data of movies needs to be of good quality and consistency, implementing an algorithm that is suitable for the data involved. From a business perspective, the risks involved are – product placement, physical shelf arrangement, giving benefits, discounts and many more marketing incentives.

The main tools and technologies that are used in this project are RStudio and Microsoft Excel. We have used RStudio because it provides an easy to implement approach for using Association Rule Mining algorithms. We have also used Microsoft Excel for storing the data.

B. Data Understanding

During this phase, we mainly focus on collecting the dataset, describing the data, exploring the data and verifying the quality of data. It is essential to get an initial insight into the data, and generate hypotheses to uncover hidden information from the data. This is an important step in our data mining methodology as it will lead to producing an accurate machine learning model for our project.

We obtained our dataset from the MovieLens website. It is a non-commercial website that makes their datasets publicly available. The dataset consists of three Digital Audio Tape (.dat) files – movies, ratings, and users. Initially, we loaded these three files into Microsoft Excel for better interpretation of the data.

The 'ratings' dataset consists of 1 million movie ratings by 6040 users. There are four columns - UserID, MovieID, Rating and Timestamp. The 'movies' dataset consists of 3883 rows of movies and three columns – MovieID, Title and Genres. The 'users' dataset consists of information of 6040 users and five columns – UserID, Gender, Age, Occupation and Zip-code. A text file containing the metadata of all three datasets was also obtained from the MovieLens website for better understanding.

Upon further exploring the dataset, it was observed that all three datasets required extensive cleaning and processing before it could be used as a machine learning model. We will discuss more about this in the next phase which is Data Preparation.

C. Data Preparation

During this phase, all activities are carried out to develop the final dataset, which can then be used to create the model. Preparing a dataset means establishing it in a manner that can be used by the model. This phase typically includes five tasks – select data, clean data, construct data, integrate data and format data.

Our business requirements specify that we need to conduct our project by selecting the top 3 genres from the movies dataset. The top 3 genres was obtained using a word cloud which displays the most popular genres in the movies dataset. In our case, the top 3 genres were Action, Comedy, and Drama which is shown in Figure 1.

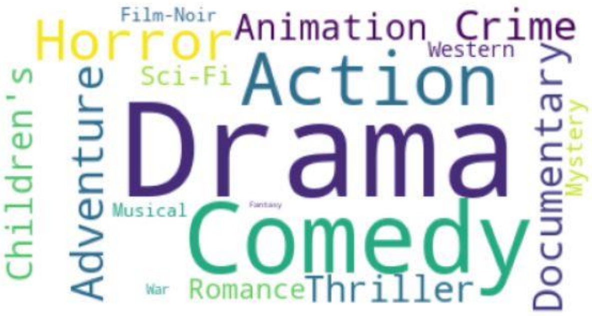

Figure 1: Word cloud of genres from the movies dataset

We initially used Microsoft Excel for cleaning and processing a part of the data. We then used RStudio for majority of cleaning and processing of data. The main tasks performed during this phase were as follows –

- Removing special characters such as punctuations that are present in the data.
- Splitting the columns, processing them and then concatenating them back together.
- Filtering the movies dataset to include only Action, Drama and Comedy movies.
- Joining the ratings dataset and the movies dataset using the inner join operation on the MovieID column present in both datasets.
- Transforming the dataset into a list of movie transactions by each user, based on the UserID column.

D. Modelling

This phase involves selecting the appropriate machine learning technique for our business needs. It also involves generating a test design where the dataset might be required to be split into training, test and validation data, however, this depends on our modelling approach. The third step in the modelling phase is to build the model and lastly it will be used for comparison against other models to assess its performance.

For our business model we need association rules between movies, therefore, we have implemented the Apriori algorithm which is used for identifying frequent itemsets in a dataset. The Apriori algorithm requires the dataset to be contained in a transactional format which we accompanied for in the Data Preparation phase of this project. The Apriori algorithm generates rules based on user-specified minimum support and minimum confidence values.

We have also implemented the Equivalent Class Clustering and Bottom-Up Lattice Traversal (ECLAT) algorithm for performance comparison purposes. The ECLAT algorithm is also an association rule mining algorithm and is an extension of the Apriori algorithm.

In the evaluation phase, we are going to evaluate both the models and compare it with each other. We are going to interpret results from the best algorithm that satisfies our business requirements.

E. Evaluation

During this phase, we are going to evaluate both the Apriori and the ECLAT algorithm by modelling them on our movie data. The ECLAT algorithm is faster when compared to the Apriori algorithm in terms of computational efficiency, however, it does not include the confidence and lift metrics that are useful in the interpretation of frequent itemsets. ECLAT algorithm only uses the support metric to filter out rules that do not meet the minimum support requirements. Also, the ECLAT algorithm does not display frequent items (movies) in a rules-based format, instead, it displays it in a set format containing items (movies) that are bought together. Therefore, as per our business requirements, the Apriori algorithm is more suitable and we are going to implement it as our modelling technique for this project.

F. Deployment

During this phase, we have used the Apriori model in R to generate association rules of 2 itemsets and 3 itemsets (movies) respectively, according to each genre. We define a prior minimum support and minimum confidence value to get the most frequent occurring itemsets from our movies data. The returned rules are sorted in order of decreasing support. We will display only the top 5 rules from each itemset. The rules are processed in such a way that duplicates are removed and are displayed in an organized way that is easy to interpret. We have performed visualization on the rules by plotting all the rules on an interactive scatter plot provided by the CRAN package - "arulesViz" with support values on the x-axis and confidence values on the y-axis. We have also visualized the top 5 rules of each genre by using a parallel coordinate's plot which connects each item in a rule using edges. During the deployment stage, we also ensured that the model is automated and can be used for any transactional dataset with only minor parametric changes according to business requirements.

IV. EVALUATION & RESULTS

In this section, we are going to evaluate and discuss about the results obtained from our analysis. Firstly, we evaluate the performance of both Apriori and ECLAT algorithm on our movies dataset.

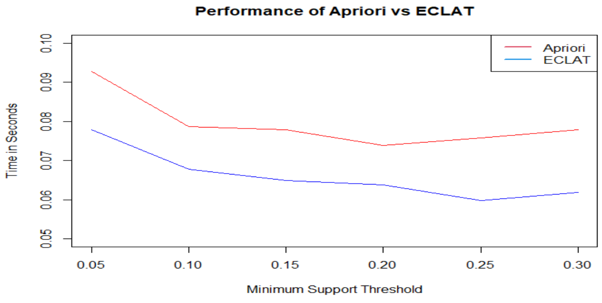

Figure 2: Performance evaluation of Apriori and ECLAT algorithms

Figure 2 shows the performance of Apriori and ECLAT with minimum support threshold values on the x-axis and time in seconds needed for association rules generation on the y-axis. It can be observed that ECLAT is slightly better in terms of performance. However, as previously stated, it returns frequent items (movies) in a set format and not a rules based format. It also does not include the confidence and lift metrics in its results. The selection of model is completely dependent on the business scenario, and in this case, since, the Apriori algorithm's computational performance is not far-off from ECLAT and due to the fact that it includes confidence and lift metrics, we have utilized the Apriori algorithm for performing further evaluation of the model and obtaining results.

The Apriori model is trained using the built-in 'apriori' function of the 'arules' library in R. The function takes in data and a list of parameters such as support, confidence, minimum length and maximum length for the generation of a rule. We have parameterised our approach with different values of support and confidence for each genre of movies. We have also parameterised our approach to include length of only two and three itemsets.

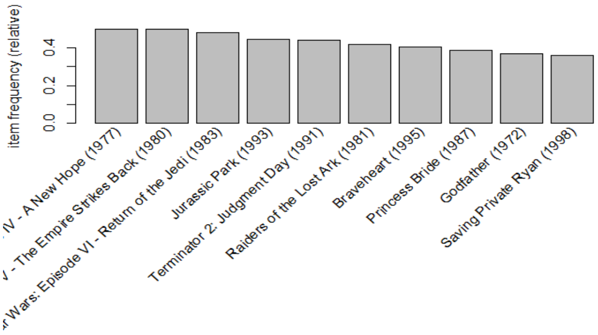

Figure 3: Item Frequency Plot of top 10 comedy movies

For the 'action' genre of movies, we identify the support values across different items by plotting an item frequency chart of the top 10 frequently occurring action movies. From figure 3, it can be observed that most of the movies have a support greater than 0.2. We have used this as the minimum support value for generating the rules.

After choosing a support value of 0.2, we chose a confidence value of 0.7 based on our business requirements. Firstly, we will return all the rules containing two items, one on the antecedent side and one on the consequent side. The rules are displayed on an interactive scatter plot, figure 4, implemented using the 'arulesViz' package where we can inspect specific rules by clicking on the visualisation. The scatter plot contains the support on the x-axis and confidence on the y-axis with the lift attribute of the rule also being displayed as a measure of intensity i.e., the higher is the lift, darker is the colour of the rule and vice versa.

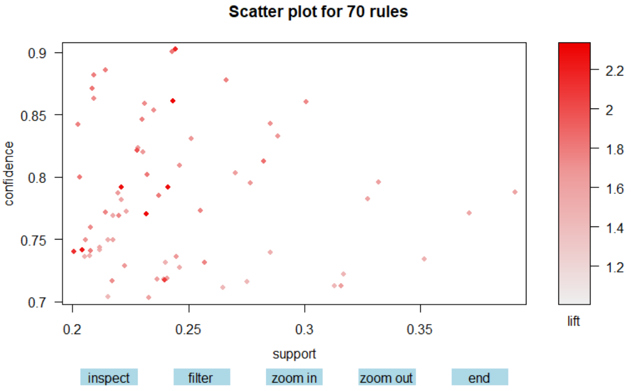

Figure 4: Scatter plot of two-item rule action movies with minimum support = 0.2 and minimum confidence = 0.7

Similarly, figure 5 shows a scatter plot of three-item rule action movies with minimum support = 0.2 and minimum confidence = 0.7 with lift attribute of the rule also being displayed.

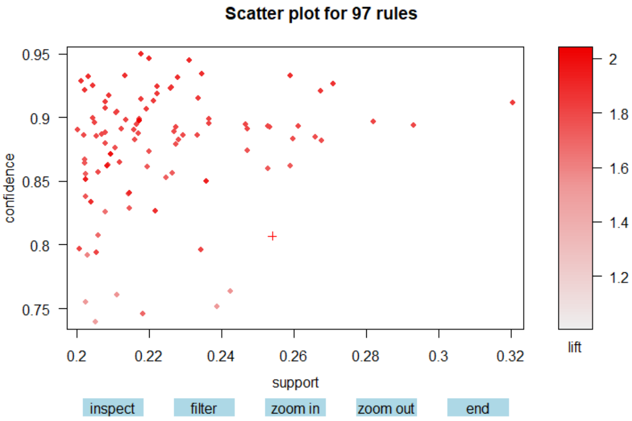

Figure 5: Scatter plot of three-item rule action movies with minimum support = 0.2 and minimum confidence = 0.7

In the final part of the results section, we will display the top 5 rules of two item rules for action movies using a parallel coordinates plot. The x-axis denotes the position in a rule i.e., first movie, second movie etc. In this case, '1' is the antecedent and 'rhs' is the consequent. The movies are displayed on the y-axis. An arrow is used to indicate that the arrowhead points to the consequent movie. Finally, the width of the arrow denotes the support and intensity of colour denotes the confidence. We observed that the rule of - Star Wars: Episode IV - A New Hope (1977) -> Star Wars: Episode V - The Empire Strikes Back (1980) had the highest support and highest confidence. Upon inspecting this rule, we inferred this rule had a support value of 0.39, confidence value of 0.78, and a lift value of 1.58.

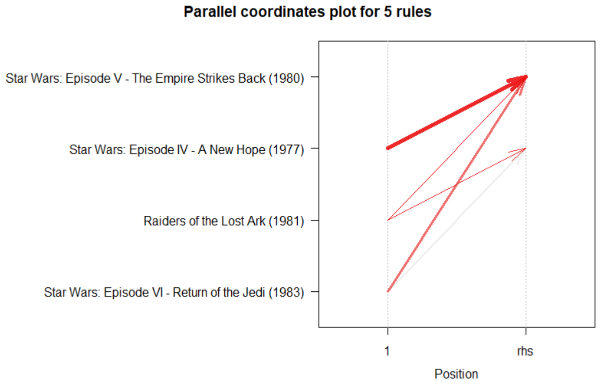

Figure 6: Parallel coordinates plot of two-item rule action movies

Similarly, in Figure 7, we will display the top 5 rules of three item rules for action movies using the parallel coordinates plot. In this case, '1' and '2' indicate the first movie and second movie of the antecedent respectively and 'rhs' represents the consequent. We observed that the rule of - Star Wars: Episode IV - A New Hope (1977), Star Wars: Episode VI - Return of the Jedi (1983)} => Star Wars: Episode V - The Empire Strikes Back (1980) had the highest support and highest confidence. Upon inspecting this rule, we inferred that this rule had a support value of 0.32, confidence value of 0.91, and a lift of 1.83.

Our approach for visualizing the top 5 rules for both two item and three item rules for each genre is a direct requirement of the business problem and it has a big impact on the DVD sales. It is possible to modify the parameters to get different rules of the same quality. This approach helps the decision makers to apply efficient marketing strategies based on these rules in order to increase the sales.

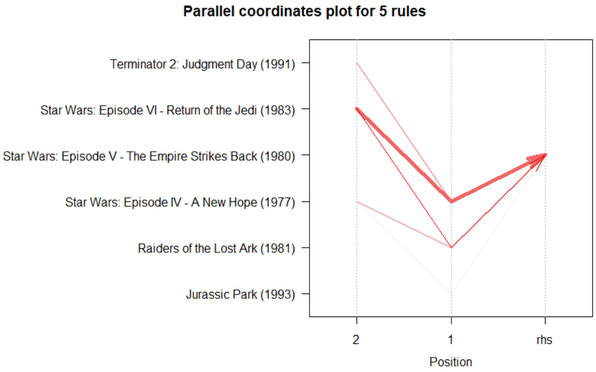

Figure 7: Parallel coordinates plot of three-item rule action movies

V. CONCLUSION AND FUTURE SCOPE

Association Rule Mining has a big impact on a wide variety of businesses especially when the data available is of good quality. In our project, we successfully implemented the Apriori algorithm to find the two item and three item rules of movies from three genres (Action, Comedy, and Drama). The results obtained were significant and has a positive impact on the business problem. The business can use these rules from the Apriori algorithm to improve their marketing strategies, in order to increase sales of movies.

Due to time constraints we had to implement the traditional Apriori algorithm for our business problem. The traditional Apriori algorithm is limited in the sense that all items in the dataset assume the same minimum support. However, in some real world scenarios, this is not the case, as different items will have different support values i.e., some items appear more frequently than others in a transactional dataset. In the future, we propose the MSApriori algorithm as a solution to this problem, which assigns minimum support values to each item in the dataset. This research study has been significant in gaining an understanding of Association Rule Mining and how it is implemented in real world scenarios.

REFERENCES

[1] J. S. Saltz, “CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps”, *2021 IEEE International Conference on Big Data (Big Data),* pp. 2337-2344, 2021.

[2] V. Robu and V. D. dos Santos, “Mining Frequent Patterns in Data Using Apriori and Eclat: A Comparison of the Algorithm Performance and Association Rule Generation”, *2019 6th International Conference on Systems and Informatics (ICSAI),* pp. 1478-1481, 2019.

[3] D. Mohapatra, J. Tripathy, K. K. Mohanty, and D. S. K. Nayak, “Interpretation of Optimized Hyper Parameters in Associative Rule Learning using Eclat and Apriori”, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC),* pp. 879-882, 2021.

[4] D. Feng, L. Zhu, and L. Zhang, “Research on improved Apriori algorithm based on MapReduce and HBase”, *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC),* pp. 887-891, 2016.

[5] L. Zheng, “Research on E-Commerce Potential Client Mining Applied to Apriori Association Rule Algorithm”, *2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS),* pp. 667-670, 2020.

[6] J. Yang, H. Huang and X. Jin, “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),* pp. 780-784, 2017.

[7] X. Chang, “Mapreduce-Apriori algorithm under cloud computing environment”, *2015 International Conference on Machine Learning and Cybernetics (ICMLC),* pp. 637-641, 2015.

[8] Y. Cong, “Research on Data Association Rules Mining Method Based on Improved Apriori Algorithm”, *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE),* pp. 373-376, 2020.

[9] S. Kesarwani, A. Goel and N. Sardana, “MSD-Apriori: Discovering borderline-rare items using association mining”, *2017 Tenth International Conference on Contemporary Computing (IC3),* 2018.

[10] T. Xu and X. Dong, “Mining frequent patterns with multiple minimum supports using basic Apriori”, *2013 Ninth International Conference on Natural Computation (ICNC),* pp. 957-961, 2013.

[11] Z. Chun-Sheng and L. Yan, “Extension of local association rules mining algorithm based on apriori algorithm”, *2014 IEEE 5th International Conference on Software Engineering and Service Science,* pp. 340-343, 2014.

[12] R. P. Puneeth and K. P. Rao, “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, 2019.

[13] SRS Reddy, Sravani Nalluri, Subramanyam Kunisetti, S. Ashok and B. Venkatesh, “Content-Based Movie Recommendation System”, 2019 Part of the Smart Innovation, Systems and Technologies book series (SIST, volume 105).