**A Project on**

**ANALYZING SEQUESNTIAL PATTERNS IN CUSTOMER SUPERMARKET TRANSACTIONS**

***Submitted in partial fulfillment of the requirement for the award of the degree of***

Masters Of Computer Application

****

**Under The Supervision of Dr. Mala Saraswat Associate Professor**

Submitted By

**Shashwat Sharma – E23MCAG0085**

**Ritik Roshan – E23MCAG0012**

**Yatharth Aggarwal – E23MCAG0051**

**SCHOOL OF COMPUTER SCIENCE ENGINEERING AND TECHNOLOGY**

**BENNETT UNIVERSITY, GREATER NOIDA, U.P**

# School of Computer Science Engineering and Technology - Bennett University

**CANDIDATE’S DECLARATION**

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **“Analyzing Sequential Patterns in Customer Supermarket Transactions.”** in partial fulfillment of the requirements for the award of the **Masters of Computer Application** submitted in the School of Computing Science and Engineering of **Bennett University, Greater Noida**, is an original work carried out during the period of month, Year to Month and Year, under the supervision of **Dr. Mala Saraswat**, Associate Professor School of Computer Science Engineering and Technology, Bennett University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

**Names**

Shashwat Sharma

Ritik Roshan

Yatharth Aggarwal

**CERTIFICATE**

The Project ……… of ………………

has been held on and his/her work is recommended for the award of Master of Computer Applications.

**Signature of Examiner(s) Signature of Supervisor(s)**

**Signature of Program Chair Signature of Dean**

Date: May, 2024 Place: Greater Noida

# Abstract (250 word)

In a retail supermarket environment, understanding the connectivity patterns of shoppers in retail stores can provide better insight into customer behavior, preferences, and purchases. By analyzing the order in which products are purchased across multiple markets, retailers can identify active products and identify product synergies. The goal of this project is to create a cross-sectional model of the client's business data collection to find commonalities and extract useful patterns. This involves identifying the cross-section of products that customers purchase in a row and determining the cross-section of those purchases.

**Keywords**

Data Mining, Sequential rule mining, locally frequent itemset.

# Table of Contents

|  |  |  |
| --- | --- | --- |
| **Title** |  | **Page No.** |
| **Abstract** |  | **I** |
| **List of Table** |  | **II** |
| **List of figures** |  | **III** |
| **Chapter 1** | **Introduction**   * 1. Motivation   2. Problem statement   3. Aim and objectives of the study | **1** |
| **Chapter 2** | **Literature Survey**   * 1. A review of relevant literature and existing work related to the project.(with table) | **1** |
| **Chapter 3** | **Project Design**   * 1. Hardware/ Software Requirement   2. Proposed Methodology   3. Details on data collection methods and tools   4. Procedures and techniques used in the project | **2** |
| **Chapter 4** | **Result and Discussion**   * 1. Presentation of the findings or results of the project   2. Use of tables, charts, graphs, or other visuals for clarity. Explanation of result | **4** |
| **Chapter 5** | **Conclusion and Future Work**  **REFERENCES (at least 15)** | **7** |

# Introduction

Sequential pattern mining has gained widespread attention in many fields, including sales and marketing, due to its ability to discover patterns and relationships in changing data. In the context of consumer behaviour in retail, some studies have investigated the use of sequence mining techniques to understand consumer purchasing behaviour and develop business strategy.

Data mining is also considered an important field for database scientists. This may mean finding detailed rules from large repositories. Ad hoc data mining has recently managed to attract many people to work in this field. There are two general definitions for searching temporary data[1]. The real problem is finding regular patterns in physical data. This is called string mining. Initial research by Agrawal and Srikant (1995) introduced the Apriori algorithm for association rule mining, which laid the foundation for later research on basket trading. This algorithm identifies active objects but does not take into account the nature of the changes. To address this limitation, Han et al. (2001) specifically proposed algorithms such as SPADE (Sequential Pattern Discovery Using Equality Classes) and PrefixSpan for pattern mining.

Recent research has focused on the use of correlational models in retail environments, particularly when analysing consumer behaviour in retail stores. For example, cross-sectional approach to analyse purchasing behaviour and uncover purchasing patterns was used across multiple productcategories. The corporate structure represents the success of the business, and the organizational structure represents the relationship in the business, in the organizational mining policy of business databases, the result of mining is usually items grouped together, and these items must be from the same transaction[2]. Although the results of the mining model refer to products purchased by the same customer in a single order, these products come from different sources[2].

Nowadays, with the rapid development of information technology, especially Web services, service-oriented architecture and cloud computing, comprehensive information has been brought together and led to the creation of useful information, various methods were used to search for data[3]. Association Rule Mining (ARM) is one of the best methods. ARM-related issues, particularly parallel and distributed data search, include reducing I/O, increasing speed, and reducing communication costs[3].

For marketing policy in the retail sector, it is necessary to know some characteristics of the consumer market and understand its purchasing behaviour, the sequential pattern mining method finds similar patterns in sequences, although many studies have been done to uncover some patterns in data connectivity, link mining is the most common method[4]. It involves discovering the next occurrence in a sequence of events, Sequential rule mining (SRM) is another approach to the prediction problem, order rules say that if certain elements appear in a sequence, certain elements will disappear after a certain confidence or probability[4].

## **Literature Survey**

The problem of mining arrays is studied in . This issue was raised by Agrawal and Srikant in 1995. But in all of the above, the pattern is inside trading[5]. One thing related to this study is the problem with the rules of the mining organization [6]. Since the problem occurred in the supermarket, using data changes to study customer behaviour is also a purchasing problem. Joint rules are rules for things bought together on an exchange and are therefore a different business model, unlike the joint model. The mining policy organization includes active items to identify the mining problem, which is a common research in the mining model. In it, an algorithm called the A-priori algorithm was proposed for the extraction of active objects[7]. Apriori algorithm has been proposed for extracting active objects.

The association rule is an expression of the form X => Y, where X and Y are database objects. This example might be a large item that is frequently purchased together. Two methods have been developed to measure interdependence: support and reliability. Rules associated with high support and trust are called negative rules of the organization and act as X = > ~ Y, ~ X = > Y, ~ X = > ~ Y, where X and Y are database objects, ~. X is the rejection of database object ~Y. Negative organization rules determine the presence and absence of objects in the database and the negative effects of database objects[3], [8].

The clustering method is one of the best methods of analysing customer behaviour by dividing customers into segments. They are also used to analyse human movement. In these studies, many customer groups were created and many opinions were discussed. Classification problems or extraction of features from customer profiles are also considered as another type of customer profile research. Some researchers analyse people's movements by taking into account individual paths and use mining exploration algorithms to explain people's daily movements. Use mining techniques to investigate and report important behavioural variables in gendered narratives. Policy mining, also known as shopping cart research, is a popular way to learn about consumer products and preferences in retail[4], [9].

Interim corporate mining rights are an important extension of traditional corporate mining rights and may be of interest to prospectors. Considering the duration of the document's content, many of the rights at issue could be infeasibly removed. Ale and colleagues proposed a method to determine the life cycle of a business configuration. The lifecycle of a program is the time between the first transaction containing the program and the last transaction containing the program, and is not necessarily equal to the lifecycle of all data. Ale et al. The method suggested by. However, this algorithm cannot extract many types of patterns that can be found in physical data, for example. Calendar, cycle etc[10].

In their paper, Mazarbhuiya F focuses only on extracting patterns from active local objects based on interaction names obtained from the interactive method. Our method is a two-stage approach; In the first stage, we use an algorithm to extract all active local objects and variable names. In the second stage, the elements obtained in the first stage are fed into the algorithm discussed in this article and the connection patterns of the elements are extracted. The algorithm is step by step. In the first stage, we have all the local objects that are frequently obtained by the process, which are our 1 or 1-dimensional arrays[1], [11].

In the study, Taser P and Dogan O performed multiple experiments, Experiments 1 and 2 show that, in general, the number of detailed rules increases with the number of visits. This shows that customers choose the store not only for their specific needs, but for all their needs every time they visit. Otherwise, it can be concluded that there are no interesting rules and customers go to the supermarket only because there are certain needs. The highest order (8039) came from the fifth customer who visited the same supermarket. In addition, when the same general rules for different number of visits are examined, it is seen that customers visit some places more frequently in their next visits. The results also show that some customers may visit sites they did not visit on their first visit and make purchases on their next visit. Additionally, in experiment 3, the CMRules algorithm was successful when varying support and confidence. Apart from this, the result of Exercise 4 is that when the times obtained using SRM algorithms in the experiment are compared with each other, the results show that EMiner runs faster than our algorithms[4].

### **Project Design**

In our research, we adopted qualitative research involving data collection, prioritization, pattern mining, analysis, evaluation algorithm and recommendations to identify recurring patterns in the consumer market. We initially collected transaction data from major retailers and ensured it included important information such as transaction ID, product ID, and time. We then preprocess the data by removing duplicates, retaining missing values, and converting it into a suitable format (such as clustering or basket) for analysing pattern mining.

The key to our approach is the use of the Apriori algorithm, an evidence-based data mining technique, to uncover patterns in business data. Through algorithms, we aim to discover the cross-section of products that customers will combine, which will allow us to gain insight into customer behavior and improve the business basket. We also set parameters such as minimum support and minimum trust to control the discovery process and ensure effective collaboration between products.

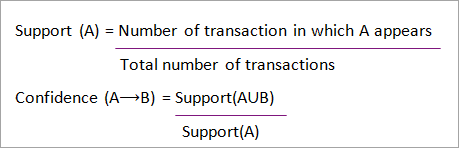
After extracting sequence patterns from the data, we continue to analyze and interpret the results to reach consensus. Our goal is to create a powerful visual system that can easily extract important information from complex models. This framework will play an important role in developing business strategies and improving the overall customer experience. We also evaluate the effectiveness of the Apriori algorithm in discovering important patterns in the data and evaluate its suitability for the immediate environment by comparing its performance with other pattern detection algorithms.

Based on the insights derived from the sequential pattern analysis, we provide actionable recommendations for supermarkets to enhance their marketing strategies and improve customer engagement. These recommendations encompass personalized marketing approaches, targeted promotions, and tailored product recommendations aimed at enhancing customer satisfaction and loyalty. Finally, we validate our findings through qualitative and quantitative analyses and document the research methodology, findings, and recommendations in a comprehensive research report, which will be presented to relevant stakeholders to facilitate informed decision-making and strategy implementation.

## **Algorithms used**

In our research, we use the Apriori algorithm to identify connectivity patterns in big business data, or connections between customers. Apriori allows us to identify connections between purchase links by setting the required minimum support and minimum trust. This allows us to gain insight into customer behaviour and improve analysis of marketing packages.

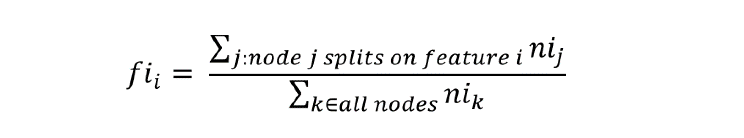
In our research, we use the Apriori algorithm to identify connectivity patterns in big business data, or connections between customers. Apriori allows us to identify connections between purchase links by setting the required minimum support and minimum trust. This allows us to gain insight into customer behaviour and improve analysis of marketing packages. The formula for the above is as follows.



## **Models Used**

In this part we give a brief description of the main models used in the research:

**Random forest:** Random Forest is a powerful learning method used to detect patterns in consumer products. Technology is particularly important for its ability to process complex data and provide accurate predictions. In the context of analyzing connected patterns of consumer transactions, random forests can provide important insights into consumer behavior and purchasing. Classification and prediction. This algorithm works by creating multiple decision trees during training and outputting the class model (classification) or mean estimate (regression) of each tree. This combination increases model robustness and reduces the risk of overfitting, providing better results. It is suitable for analysis of many transfer cases commonly encountered in supermarkets. The ability to evaluate the importance of features allows researchers to identify important factors that influence buyer behavior and thus helps make good decisions for marketing and placement. Researchers can uncover important information about consumer preferences, product relationships, and purchasing patterns. By understanding these trends, retailers can adjust marketing strategies, improve products and improve customer service, ultimately increasing sales and customers. It provides a powerful and effective way to discover valuable insights from transactional data. Its versatility, accuracy, and interpretation make it a valuable tool for researchers and businesses looking to understand and use consumer behavior for better decision making and improvement. Formula for random forest is as follows:

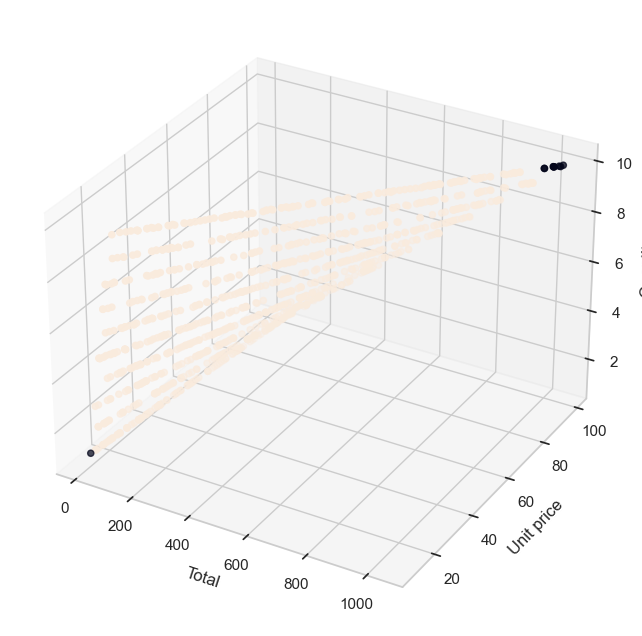


**Arima Model:** The ARIMA (Autoregressive Integrated Moving Average) model is a powerful tool often used in time series analysis, especially to predict future outcomes based on past observations. ARIMA can provide insight into consumers' real-time behavior in the context of studying consumer behavior patterns in retail stores. Patterns and patterns of behavior over time. This includes understanding seasonal changes in product purchases, checking trends in sales, and predicting future business models. Linked data lends itself well to modeling complex patterns in consumer behavior. Additionally, the ARIMA model can handle fixed-time and non-value data, which are common problems in real business data. time, repetition etc. This information can inform better business decisions, inventory management, and customer management in large enterprises. Understand customer behavior more deeply and make informed decisions to improve your business. Formula for ARIMA model is as follows:

#### **Result and Discussion**

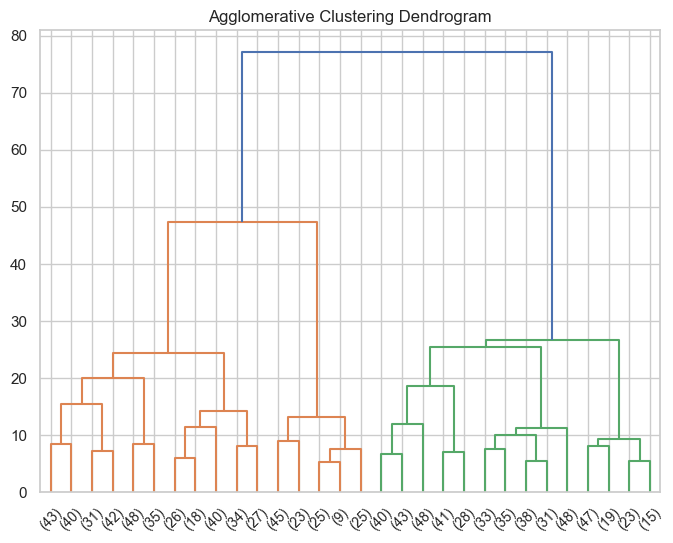
## **Data Visualization**

The graph in Figure 1. shows the results obtained using the forest separator in the Total, Unit Price and Quantity dimensions. Each point represents a change in the data set, and the main points are in different colours. By analysing these results, researchers can understand unusual purchases or trends in the consumer market. This information is useful for understanding customer preferences, detecting potential fraud or inaccuracies, and optimizing the shopping cart process to improve overall customer business strategy and customer satisfaction.



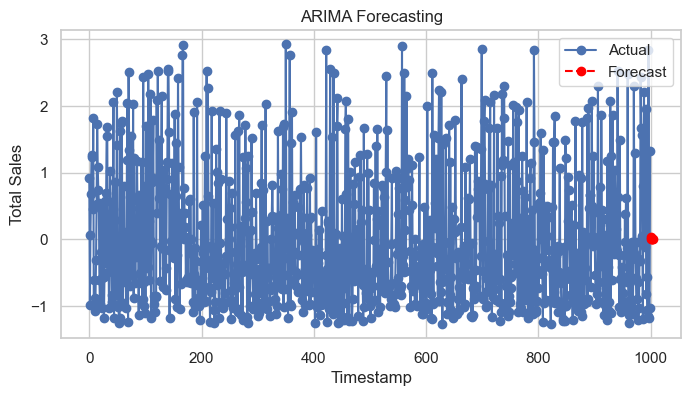
**Figure 1. Identifying Outliers in Supermarket Transaction Data**

The dendrogram in Figure 2. shows the results of aggregative hierarchical clustering applied to the retail store dataset. Hierarchical clustering is a technique used to group similar data into groups based on related connections. Technology helps inform the structure and relationship in information. Dendrograms visually represent the clustering process by showing the hierarchical association of clusters at each step. while the branches are represented together. The height of each leg represents the distance or difference between mixed groups. By cutting the dendrogram and focusing on a specific part of the cluster, we can gain insight into the best solution for analysis. Make it easy to discover and interpret recurring patterns in customer behaviour.



**Figure 2. Agglomerative Hierarchical Clustering Dendrogram**

The chart in figure 3 shows the results of predicting total sales using the Autoregressive Integrated Moving Average (ARIMA) model in the context of our research examining a continuous sample of users. ARIMA models are powerful tools for estimating time, including the nature of the data. Here we apply an ARIMA model to all sales data in a retail store dataset. The blue line represents total sales for the period, while the red dotted line represents forecast values ​​for the next five points in time. Indicators show future data points and predictions.



**Figure 3. ARIMA Forecasting of Total Sales**

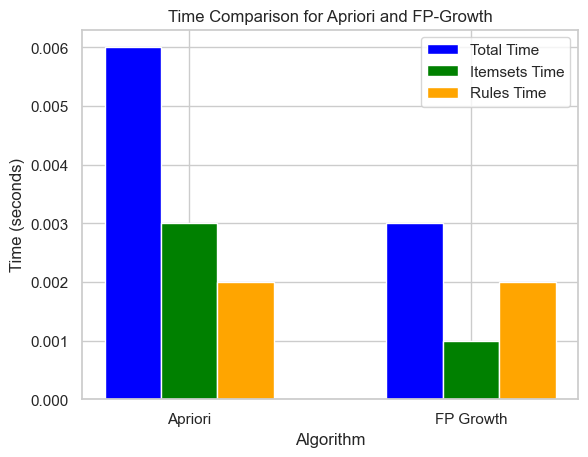
The chart in Figure 4 shows the distribution of consumers according to their purchasing behaviour during the period, within the scope of our research examining consumer behaviour in the retail market according to standards. Similarities in purchasing patterns divide consumers into different groups. Each group represents a different type of consumer with purchasing habits and preferences. Different colours represent groups from which consumers are selected based on their purchasing behaviour. Customer satisfaction in the supermarket environment.

A chart with many dots

Description automatically generated with medium confidence

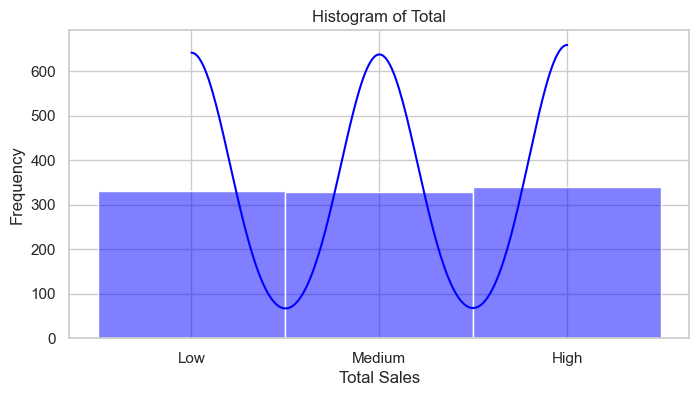
**Figure 4. Customer Segmentation Over Time**

The graph in Figure 5. shows the execution time comparison between Apriori and FP-Growth algorithms in the context of our research examining the connectivity pattern in a large consumer market. Each algorithm introduces three variables: total execution time, time spent creating objects, and time spent creating organizational rules. and business. By comparing the total execution time and the breakdown of time used to create plans and rules, we can determine which algorithm is better suited to the sequential mining model in big business data. The algorithm aims to illustrate the use of computational resources and show meaningful patterns in consumer behaviour.



**Figure 5. Time Comparison for Apriori and FP-Growth**

The histogram in Figure 6 describes the distribution of all products sold within our research, which analyzes patterns in the customer's store. The x-axis represents total sales, and the y-axis represents the frequency or number of each sale. The histogram is divided into boxes, and each box represents a certain percentage of total sales. By analyzing the shape and distribution of the histogram, we can detect anomalies, anomalies or patterns in overall sales, which are highly relevant to cost efficiency, inventory management and revenue optimization.



**Figure 6. Histogram of Total Sales Distribution**

##### **Conclusion and Future Work**

**Conclusion**

In our research examining consumer behaviour patterns in retail, we use data analysis and visualization techniques to gain insight into consumer behaviour, algorithm performance and sales patterns. By segmenting our customers over time, we identify different customers with specific purchasing patterns, enabling marketing plans and personalized recommendations. We use time comparison graphs to compare the results of common policy search algorithms, which helps in choosing an algorithm for pattern search. Additionally, histograms of total sales distribution provide insight into sales patterns and trends, informing decisions regarding cost effectiveness and inventory management. Overall, our research enables supermarkets and retailers to improve their marketing strategies, operational processes and customer satisfaction.

# Future Work

The potential for future research identifying connectivity patterns in the consumer marketplace offers many ways to understand and use retail. First, search data mining methods beyond traditional methods such as Apriori and FP-Growth pave the way for extracting complex patterns from large data sets. By exploring deep learning or hybrid models, such as the combination of machine learning and graphics, researchers can improve the accuracy and efficiency of models to gain a deeper understanding of consumer behaviour and business dynamics.

Additionally, predictive analytics for demand forecasting has great potential to optimize inventory and supply chain management. By using historical market data and identifying recurring patterns, retailers can predict future product demand, improve product levels and reduce stock or inventory. The best approach to demand forecasting allows retailers to simplify their operations, reduce costs and increase overall profitability.

In summary, the future of pattern analysis in the consumer market contains many opportunities, including professional data mining, integration of external information, instant recommendations, forecasting needs for prediction, and ethical decision-making. By examining these factors, researchers and professionals can understand our understanding of the retail market as a trend and create opportunities to improve merchandising, customer engagement, and marketing.

### **References**

[1] F. A. Mazarbhuiya, “Mining Sequential Patterns from Super Market Datasets,” *International Journal of Computer Trends and Technology*, vol. 30, no. 4, pp. 206–212, Dec. 2015, doi: 10.14445/22312803/IJCTT-V30P136.

[2] Q. Zhao and S. S. Bhowmick, “Sequential pattern mining: A survey,” *ITechnical Report CAIS Nayang Technological University Singapore*, no. 2003118, 2003, doi: 10.1016/j.cie.2008.11.006.

[3] V. Rastogi and V. K. Khare, “Apriori Based: Mining Positive and Negative Frequent Sequential Patterns.”

[4] P. Y. Taser and O. Dogan, “Sequential Rule Mining for Analysis of Customer Movements in Different Visits,” in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2021.

[5] R. Agrawal and R. Srikant, “Mining Sequential Patterns.”

[6] R. Agrawal, T. Imieliński, and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases,” *ACM SIGMOD Record*, vol. 22, no. 2, 1993, doi: 10.1145/170036.170072.

[7] M. J. Zaki, “Efficient Enumeration of Frequent Sequences,” in *International Conference on Information and Knowledge Management, Proceedings*, 1998. doi: 10.1145/288627.288643.

[8] H. Yun, D. Ha, B. Hwang, and K. H. Ryu, “Mining association rules on significant rare data using relative support,” *Journal of Systems and Software*, vol. 67, no. 3, 2003, doi: 10.1016/S0164-1212(02)00128-0.

[9] O. Dogan, C. Fernandez-Llatas, and B. Oztaysi, “Process mining application for analysis of customer’s different visits in a shopping mall,” in *Advances in Intelligent Systems and Computing*, 2020. doi: 10.1007/978-3-030-23756-1\_20.

[10] J. M. Ale and G. H. Rossi, “An approach to discovering temporal association rules,” in *Proceedings of the ACM Symposium on Applied Computing*, 2000. doi: 10.1145/335603.335770.

[11] A. Kakoti Mahanta, F. A. Mazarbhuiya, and H. K. Baruah, “Finding locally and periodically frequent sets and periodic association rules,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005. doi: 10.1007/11590316\_91.