Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 8
August 2023

An Effective Association Rule Based Algorithm for Privacy Preserving Frequent Itemset Mining

Ashoktaru Pal, Dr. Ajay R. Raundale

Department of Computer Science & Engineering

Dr. A. P. J. Abdul Kalam University, Indore (M. P.) – 452010

Corresponding Author Email: pal.ashoktaru@gmail.com

ABSTRACT

Finding hidden patterns in huge data sets is a technique known as data mining or knowledge discovery. There are several data mining algorithms, each with a unique goal. It is a challenging task to extract significant and unidentified patterns from massive datasets. One of the most well-known data mining algorithms for determining the significant link between the item-sets is association rule mining, which has the capacity to uncover unanticipated data dependencies. The fundamental concept is to determine if the existence of some items strongly indicates the existence of other things in a given database of item sets (such as shopping baskets). The apriori algorithm is a popular method for extracting association rules from datasets. It involves identifying frequently occurring groups of items in transaction data and using them to generate association rules. Association rules are a descriptive data mining technique that can provide useful insights into patterns and trends in the data.In general, the traditional apriori algorithm works well for small datasets, but it may not be efficient for handling large datasets. However, by making a few modifications to the implementation of the apriori algorithm, it is possible to improve its performance for large datasets. In this work, we made some adjustments to the apriori algorithm implementation and were able to achieve better results when working with large datasets.

Keywords: Data Mining, Association Rule Mining, Confidence, Privacy-Preserving Policy, Frequent Item Sets.

1. INTRODUCTION

Privacy-preserving data mining (PPDM) is a technique used to protect data privacy while still making it useful. Data mining, which is the process of analyzing data from various angles and extracting useful information to improve profits or reduce costs, is a common analytical method used to analyze data. Data mining software allows users to categorize data, summarize relationships, and analyze it from different perspectives. The ultimate goal of data mining is to organize information from a data set in a way that can be utilized in different applications.

In this paper, we review various methods of privacy preservation and explore different privacy-preserving data mining approaches, techniques, and algorithms. Finding relevant patterns in data is a critical aspect of data analysis and mining. However, in many cases, only one organization or entity may have access to all the data, and not all parties involved may consent to open data combination due to concerns about data privacy and security.

To address this challenge and effectively mine data while upholding privacy and security standards, new systems and approaches are required. PPDM is the most effective option for balancing security and privacy requirements when processing multiparty data. Therefore, the primary objective of this work is to research and explore PPDM approaches and their real-world applicability. Additionally, a novel privacy-preserving design and implementation strategy for association rule mining is proposed.

2. BACKGROUND

2.1. PRIVACY PRESERVING DATA MINING (PPDM)

The comprehension of the many essential words that are used to explain the proposed idea of privacy-preserving association rule mining is provided in this section.

PPDM:PPDM is the name of the data mining subfield (privacy-preserving data mining). By implementing data mining techniques and algorithms in this context, the security and privacy of end user data are given priority. As a result, in order to safeguard the privacy and sensitivity of the end data owner's privacy and security, it is usually essential to utilise cryptographic techniques, noise-based techniques, and blocking-based approaches.

Data with vertical partitions:When the data is shared among several parties and has different attributes and class names, it is partitioned vertically in a database.

Horizontal data partitioning: This form of data organisation is known as horizontal partitioned data when it is shared across several parties in a particular environment with a constant number of characteristics and class labels but a variable number of data instances for each party.

Associative rule mining: Data mining approach called association rule mining establishes relationships between data attributes based on their frequency and combinations of various attribute values. These methods are employed in the construction of prediction and decision-making rules. For association rule mining, apriori and FP-tree algorithms are provided.

Multiple-party data: In some complicated situations, less attribute-based information may have an influence on the ultimate business domain decision-making process. Several data owners have chosen to combine and analyse their data in a single location in order to mine the shared judgements from the totality of the obtained data.

Private information: Numerous sets of data owners with many features. Some sensitive information about the user may be among these attributes. Credit card numbers, dates of birth, PAN card numbers, and other private information are examples of sensitive data elements that may have an influence on a person's social or financial privacy.

Mining for decisions: Decision mining or rule mining refers to the process of analyzing data or information to generate decisions based on the evaluation of available features. In this context, both decision trees and association rule mining techniques were significant. These methods can be used to extract meaningful insights and patterns from data, which can then be used to inform decision-making. Decision trees are a popular tool for visualizing decisions and the possible outcomes of a decision based on various factors. Association rule mining, on the other hand, is used to discover relationships or correlations between different variables or attributes in a dataset. By using these techniques, it is possible to identify important factors and variables that can influence decision-making and improve the overall decision-making process.

Types of privacy preservation: There are several types of privacy preservation techniques used in data mining and analysis. Some of the most commonly used techniques are:

- **Anonymization:** This technique involves the removal of personal identifiers from a dataset. This could include names, addresses, and other identifying information.
- **Generalization:** Generalization involves the aggregation of data so that individual data points are no longer identifiable. For example, ages could be grouped

into ranges (e.g., 20-30, 30-40, etc.) instead of being reported as specific values.

- **Differential Privacy:** Differential privacy involves adding noise to a dataset to prevent individual data points from being identified. This technique can help to protect the privacy of individuals while still allowing meaningful analysis of the data.

- **Cryptography:** Cryptography involves the use of encryption to protect data. This technique can be used to protect data both while it is being stored and while it is being transmitted.

- **Secure Multi-Party Computation (SMPC):** SMPC involves multiple parties jointly analyzing data without any party having access to the full dataset. This technique can be used to analyze data while still preserving privacy and confidentiality.

These techniques can be used individually or in combination to ensure that data privacy is protected during the data mining and analysis process.

Types of Attack on published data or Linkage attack: When data is published, there are several types of attacks that can be launched to re-identify individuals or link information from different sources. One of the most common types of attacks is a linkage attack, which involves combining information from multiple datasets to identify individuals or gain additional information about them. Some other types of attacks that can be launched on published data include:

- **Inference attacks:** These attacks involve using statistical analysis to infer information about individuals based on the available data. For example, an attacker might use information about the age and gender of individuals in a dataset to infer information about their income or other characteristics.

- **Re-identification attacks:** These attacks involve using information in a dataset to re-identify individuals who were thought to be anonymous. For example, an attacker might use information about the zip code, age, and gender of individuals to re-identify them.

- **Membership attacks:** These attacks involve determining whether a particular individual is a member of a given dataset. For example, an attacker might use information about a person's occupation, age, and geographic location to determine whether they are a member of a particular organization or group.

- **De-anonymization attacks:** These attacks involve using external data sources to re-identify individuals who were thought to be anonymous in a dataset. For example, an attacker might use social media data or public records to re-identify individuals in a dataset.

To protect against these types of attacks, various privacy-preserving techniques can be used, such as data anonymization, data perturbation, and access control.

2.2. CLASSIFICATION OF PPDM APPROCHES

There are six general categories that may be used to categories privacy-preserving data mining strategies.

- Heuristic Approach:In the context of privacy-preserving data mining, a heuristic method may be used in a central database setting. In this approach, the data is divided into two categories: aggregated data and unprocessed knowledge.

- Reconstruction approach:The reconstruction approach is also utilized with centralized databases, however in this instance, just one type of data—raw data—is utilized. The raw data is put via data mining processes.

- Anonymization approach:The goal of the anonymization strategy is to make each individual record invisible amid a group of records by utilizing generalization and suppression techniques. Anonymization is a method for obfuscating sensitive or private data about record holders. Even the storage of private information for analysis is acceptable.

- Randomization Approach:The randomized response technique obscures the original information by contaminating it with random information or noise, rendering it hard to tell whether or not knowledge from an individual comprises genuine expertise. As much random data or noise must be provided so that personal information about a person cannot be accessed by an untrusted party. Warner made the original suggestion to use this statistical approach.

- Perturbation approach:These techniques involve the addition of noise or randomisation to the original data to prevent the disclosure of sensitive information. Examples of such techniques include adding random values to data and swapping values between records.

- Cryptographic approach: These techniques involve the use of cryptographic algorithms to protect the privacy of data. Examples of such techniques include encryption, homomorphic encryption, and secure multi-party computation.

3. METHODOLOGY

This section has provided an explanation of the sources utilized in the association rule mining study of the chosen dataset. We developed our innovative classification technique as a starting point for data analysis using a publically available dataset for learning purposes. The transformation of data into a tidy rule method is one of the most important research topics in the field of combination of choice item sets.

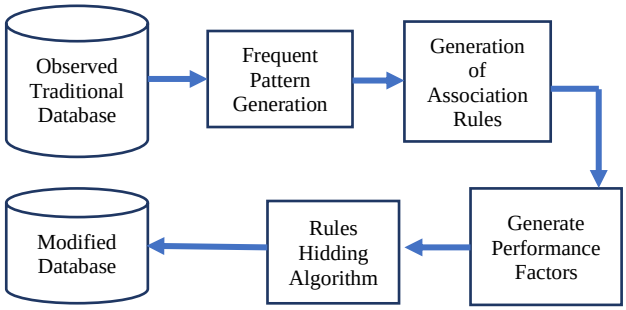

Fig. 1 Comparative study of system Architecture

3.1. DATASET

The process of collecting input data is crucial in this work, and in this case, the data refers to big store stock. The data is gathered from open sources or the Kaggle database. There were 7500 records and 20 characteristics in this collection. We applied our technique to this sizable dataset and obtained some extremely encouraging results, which we tabulated in our paper.

The dataset used in this study consists of multiple item that are choose by customer. The dataset incorporates various attributes, such as mineral water, eggs, spaghetti, chocolate,french fries, green tea, milk, etc.

3.2. ASSOCIATION RULE MINING

Association rule mining is a data mining technique used to discover relationships between variables in large datasets. The technique relies on finding strong associations between variables to develop association rules. In order to establish the rules, several measurements are used in the databases. Association rules have two parts: the first antecedent section defines the "if" portion using conjunction and disjunction operators, while the second consequent section defines the "then" portion and produces the answer when combined with the antecedent component. Support and confidence values are used to classify the majority of significant associations between databases and to develop association rules by searching the volume of data for common if/then patterns. Support indicates how frequently the objects in the database appear, while confidence represents the number of times an if/then expression is found in the databases. The association rule-mining algorithm is an effective technique for discovering relationships between variables in large datasets with many variables, and is particularly useful for datasets with millions of records.

Let's start with some key concepts of Frequent Itemset Mining. Consider S1 = {I1, I2, ..., Id} as the set of all items, and S2 = {t1, t2, ..., tn} as a transaction (or market) database, where each uniquely labeled transaction ti contains a subset of items chosen from S1. An Itemset is a collection of zero or more items, and an Itemset containing p items is called a p-itemset. A transaction tj is said to contain an itemset X if X is a subset of tj. The width of a transaction is defined as the number of items it contains. The property of an itemset X is its support count σ(X), which counts the number of transactions that contain itemset X, i.e., $\sigma(X) = |\{ti|X \subseteq ti, ti \in T\}|$. The relative support of X is $supp(X) = \sigma(X)/|T|$. An itemset X is called frequent if it has $supp(X) \geq$ minsup, where minsup is the minimum support threshold provided as input.

In the KDD dataset, the number of instances is represented by n = {N1, N2, N3, …,∝}, and the number of features is represented by m = {C1, C2, C3, …, C40}. An association rule mining algorithm is proposed to determine relationships among the KDD dataset, which is represented as X Y ⇒. The association between X and Y is such that whenever X appears, Y is also likely to appear. X and Y may be single conditions or sets of conditions. X is called the rule's antecedent part, and Y is called the consequent part.

A. The concept of support in association rule mining is used to measure the frequency of occurrence of an itemset in a given dataset. In the context of KDD dataset, the support for a rule X → Y is the percentage of transactions in the dataset that contain both X and Y. If the support of a rule is greater than or equal to a user-specified minimum support threshold, then it is considered a frequent itemset. This can be mathematically represented as:

$$\text{Support}(x)=\frac{Numberoftimes X appears}{TotalNumberofRecords}$$

$$\text{Support}(x)=\frac{Numberoftimes X and Y appears together}{TotalNumberofRecords}$$

B. Specifically, the rule X ⇒ Y holds with confidence conf if conf % of the transactions in the KDD Dataset that contain X also contain Y. A rule with a confidence greater than a user-specified confidence is called a minimum confidence rule. In other words, the confidence value indicates the reliability or strength of the association between X and Y.

$$\text{Confidence}(X \rightarrow Y)=\frac{\text{Support}(xy)}{\text{Support}(x)}$$

C. To find frequent itemsets, the frequencies of all the features are calculated using the records in the dataset. Frequent 1-itemsets are generated based on the minimum support values. Then candidate 2-itemsets are generated by joining the frequent 1-itemsets to find their frequency. The minimum support value is used to generate candidate 3-itemsets, and the process continues until no new frequent itemsets can be found. This process is known as the Apriori algorithm.

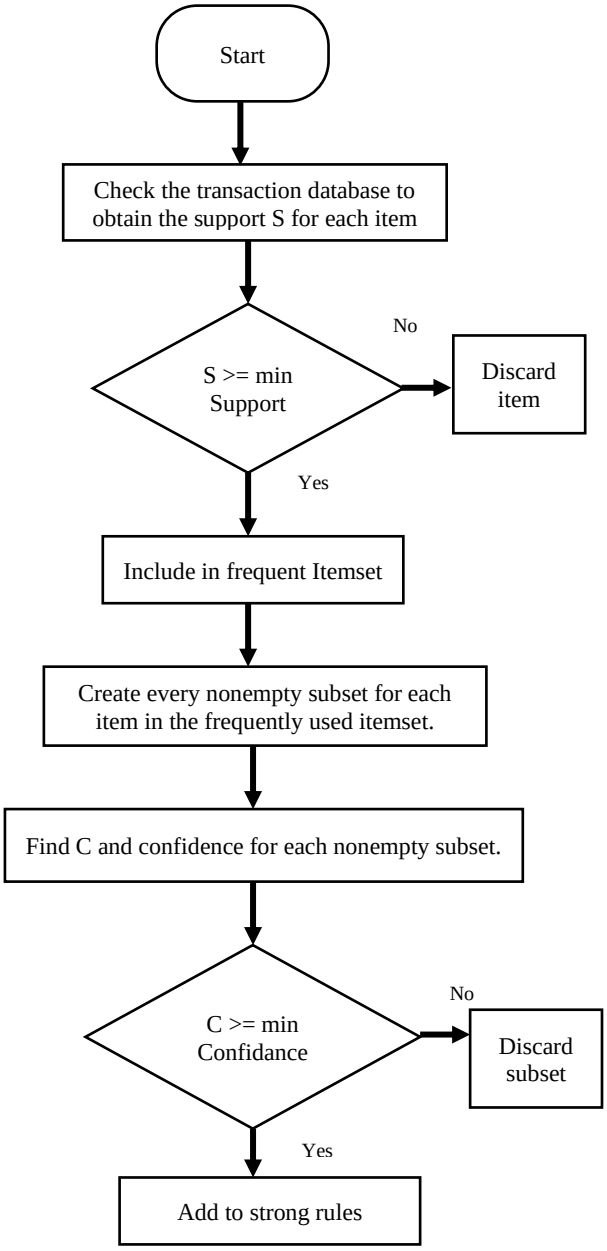

Fig. 2 System architecture of association rule mining

3.3. THE BASIC IDEA OF APRIORIALGORITHM

An technique called Apriori is used to mine frequently occurring item sets for association rules. It depends on existing data to mine frequently recurring item sets. For layer-by-layer searching, the iterative method is utilised. An algorithm for creating candidates called Apriori goes through each stage one at a time.

Apriori algorithm is two step procedures:

i) Candidate Generation.
ii) Pruning.

A candidate item-set is simply an item-set that may be frequent or uncommon, based on the user minimum support requirement. Higher level candidate item-sets (Ci) are produced by combining Li-1 or prior level frequent item-sets with one another. For this particular issue, the Apriori technique is viewed as a level-wise strategy. The Apriori algorithm's second step assists in weeding out potential item-sets whose subsets are not common. The anti-monotonic characteristic, which is the basis for this, states that any subset of a frequent item-set is also frequent. As a result, the frequent item-set and association mining procedure excludes a candidate item-set that consists of one or more a priori level infrequent item-sets.

The traditional apriori algorithm employs a support-confidence framework to assess the usefulness of association rules, but in real-world scenarios, relying solely on the support and confidence thresholds' upper and lower bounds will still produce some ineffective or even incorrect association rules. The calculation formula described in the following equations is frequently used by the apriori algorithm:

$$L = \int x^2 nx + \frac{x^{n+1}}{n+1} + C,$$

$$FV = V_A N(d_1) - e^{-rt} DN(d_2)$$

Steps InApriori

The Apriori algorithm is a popular algorithm used for frequent itemset mining and association rule learning in data mining. The basic steps involved in the Apriori algorithm are as follows:

1. Initialization: The algorithm begins by scanning the entire dataset and generating the frequent 1-itemsets based on the user-specified minimum support threshold.
2. Generating candidate itemsets: Based on the frequent 1-itemsets, the algorithm generates candidate 2-itemsets by joining pairs of frequent 1-itemsets. The candidate k-itemsets are generated by joining the frequent (k-1)-itemsets.

3. Pruning: The algorithm prunes the candidate itemsets that do not meet the minimum support threshold.
4. Repeat: Steps 2 and 3 are repeated until there are no more frequent k-itemsets.
5. Generating association rules: Once all the frequent itemsets have been generated, the algorithm generates association rules from the frequent itemsets with the user-specified minimum confidence threshold.
6. Pruning association rules: The algorithm prunes the association rules that do not meet the minimum confidence threshold.
7. Output: The algorithm outputs the frequent itemsets and the association rules that meet the user-specified minimum support and confidence thresholds.
   These steps can be adjusted and optimized based on various factors such as the size of the dataset, the minimum support and confidence thresholds, and the computational resources available.

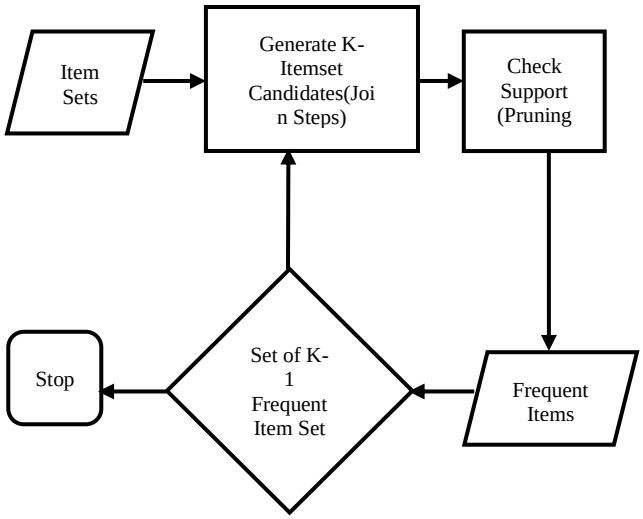

Fig. 3 Block diagram of Apriori

The Apriori Algorithm :

Here are the steps of the Apriori algorithm-

1. Generate all possible 1-itemsets in the dataset and calculate their support.
2. Prune the infrequent 1-itemsets by removing those that have a support less than the minimum support threshold.
3. Generate candidate itemsets of size k (k > 1) by joining frequent itemsets of size k-1.
4. Prune the candidate itemsets by removing those that have a subset of size k-1 that is infrequent.
5. Calculate the support of the remaining candidate itemsets.
6. Prune the infrequent itemsets by removing those that have a support less than the minimum support threshold.
7. Repeat steps 3-6 until no frequent itemsets can be generated.

3.4. GENERATION OF CANDIDATE ITEMSET AND FREQUENT ITEMSET

T1 sample of transactional data.

T ID	List of Items IDs
T1	S1,S2,S5
T2	S2, S4
T3	S2, S3
T4	S1, S2, S4
T5	S1, S3
T6	S2, S3
T7	S1, S3
T8	S1, S2, S3, S5
T9	S1, S2, S3

Scan T1 for count of each

D1

Item set	Support Count
{S1}	6
{S2}	7
{S3}	6
{S4}	2
{S5}	2

Compare candidate support count with minimum support count

K1

Item set	Support Count
{S1, S2}	4
{S1, S3}	4
{S1, S4}	1
{S1, S5}	2
{S2, S3}	4
{S2, S4}	2
{S2, S5}	2
{S3, S4}	0
{S3, S5}	1
{S4, S5}	0

Generate D2 candidate from K1

D2

Item set	Support Count
{S1}	6
{S2}	7
{S3}	6
{S4}	2
{S5}	2

Scan T1 for count of each candidate

D2

Item set
{S1, S2}
{S1, S3}
{S1, S4}
{S1, S5}
{S2, S3}
{S2, S4}
{S2, S5}
{S3, S4}
{S3, S5}
{S4, S5}

K2

Item set	Support Count
{S1, S2}	4
{S1, S3}	4
{S1, S5}	2
{S2, S3}	4
{S2, S4}	2
{S2, S5}	2

Compare candidate support count with minimum support count

3.4. LITERATURE REVIEW

Association rules in paper [1] refer to groups of pages that are visited collectively with a support value greater than a certain cutoff. On each run of the database scan, the AIS method

provided by Agrawal et al. builds candidate item sets on the fly. The presence of large item sets from the prior pass is verified in the current transaction. As a result, adding new items to existing item sets produces new item sets. This strategy is ineffective because it generates an excessive number of candidate item sets. This method generates rules with only one consecutive item, consumes more space, and performs an excessive number of passes over the whole database.

Agrawal et al. developed several Apriori algorithm variations in article [2], including Apriori, AprioriTid, and AprioriHybrid. Apriori and AprioriTid build item sets from the massive item sets found in the preceding step, disregarding the transactions. AprioriTid improves Apriori by utilising the database at the first stage. Future rounds employ the encodings from the first pass, which are considerably smaller than the database.

The predictive Apriori algorithm might provide additional rules in paper [3] that could be helpful when examining the relationship between each individual element and accident severity. Based on several hidden trends in the data, theseinsights can help the decision-makers in the department of traffic accidents take action. As a multi-objective classification problem, the swarm based algorithms to extract association rules for student performance prediction are analysed.

In article [4], the author discussed establishing association rules and identifying frequent item sets to identify sensitive products in the market basket database. Here, all of the rules from X with the same LHS are chosen, and the combined RHS of the selected rules is saved. The chosen Products will be regarded as delicate objects.

The hybrid algorithm with distortion technique, which is based on the support and confidence strategy for protecting sensitive data, was addressed in article [5]. As a result, the database owner can keep useful rules while masking sensitive rules.

The author of article [6] presented an alternative method for concealing association rules. The heuristic rule-creation method ensures data quality while preserving privacy for sensitive data by hiding as many rules as is possible at once and modifying some transactions.

The enhanced apriori technique, which mines frequent item sets without generating new candidates, is discussed in paper [7]. As a result, it uses less storage space and less often to run queries.

In publication [8], the author presented a brief summary of apriori calculation and late enhancements that were made in this area. It is assumed that the real interest in learning about various upgraded computations is to create fewer applicant sets that contain visit items in a reasonable amount of time.

3.5. RESULT

The Apriori algorithm is one of the main technique of association rule mining, which can be basically descibedasfinding the most frequent itemsets in a dataset.

We will start creating frequency plots to visualize the frequency of each item in the transactions. The height of each bar in the plot represents the frequency of the corresponding item. We can also create plots with different support levels. Support refers to the frequency of a pattern in a rule, so if we set the support level to, say, 0.1, it means that the item must occur at least 10 times in 100 transactions to be considered frequent. This is why the second plot might have more items than the first one. Alternatively, we can also specify the exact number of elements we want to include in the plot instead of setting a support level.
In our dataset the most demanded items are shown in the mentioned Fig.4.

	Items	Incident_count
0	mineral water	714
1	eggs	562
2	spaghetti	550
3	chocolate	529
4	french fries	521
5	green tea	430
6	milk	397
7	ground beef	281
8	frozen vegetables	276
9	pancakes	276

Min
acc
the
mos

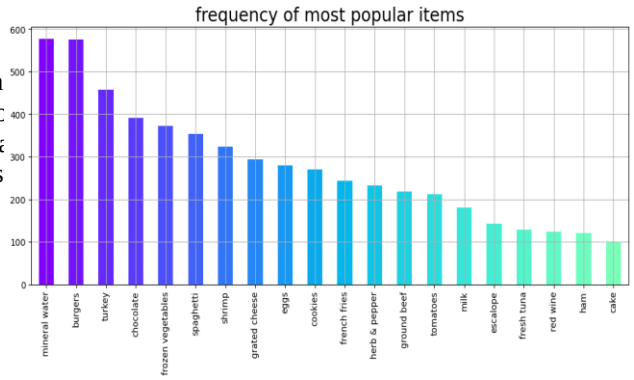

Fig. 5: Most popular items

We found support values for pair product combinations like one two and three. The below Table 1,2 and 3 shows thecombinations of itemsets with support value.

index	support	itemsets	length
1	0.238	(mineral water)	1
2	0.187333	(eggs)	1
3	0.183333	(spaghetti)	1
4	0.176333	(chocolate)	1
5	0.173667	(french fries)	1
6	0.143333	(green tea)	1
7	0.132333	(milk)	1
8	0.093667	(ground beef)	1
9	0.092	(frozen vegetables)	1
10	0.092	(pancakes)	1
…	…	…	…

Table 1: Support values for pair product combinations with

Index	Support	Itemsets	Length
1	0.053666667	frozenset({'eggs', 'mineral water'})	2
2	0.061333333	frozenset({'spaghetti', 'mineral water'})	2
3	0.056	frozenset({'chocolate', 'mineral water'})	2
4	0.032666667	frozenset({'french fries', 'mineral water'})	2
5	0.032333333	frozenset({'green tea', 'mineral water'})	2
6	0.052666667	frozenset({'milk', 'mineral water'})	2
7	0.037333333	frozenset({'ground beef', 'mineral water'})	2
9	0.037666667	frozenset({'mineral water', 'frozen vegetables'})	2
10	0.029666667	frozenset({'pancakes', 'mineral water'})	2

Table 2: Support values for pair product combinations with length 2

index	support	itemsets	length
1	0.016333333	frozenset({'spaghetti', 'eggs', 'mineral water'})	3
2	0.012	frozenset({'chocolate', 'eggs', 'mineral water'})	3
3	0.014	frozenset({'milk', 'eggs', 'mineral water'})	3
4	0.018	frozenset({'spaghetti', 'chocolate', 'mineral water'})	3
5	0.020666667	frozenset({'spaghetti', 'milk', 'mineral water'})	3
6	0.014666667	frozenset({'spaghetti', 'ground beef', 'mineral water'})	3
7	0.014	frozenset({'spaghetti', 'mineral water', 'frozen vegetables'})	3
8	0.010666667	frozenset({'spaghetti', 'pancakes', 'mineral water'})	3
9	0.011	frozenset({'spaghetti', 'shrimp', 'mineral water'})	3
10	0.011333333	frozenset({'olive oil', 'spaghetti', 'mineral water'})	3

Table3: Support values for pair product combinations with length 3

In the initial step, table 1 is created with each item being considered as a candidate 1-itemset. In table 3, candidate item sets are generated for mineral water, eggs, chocolate, and other foods using the minimum support value. As per the Apriori principle, only frequent 1-itemsets are used to generate candidate 2-itemsets in the next iteration, as all supersets of infrequent 1-itemsets must also be infrequent. Similarly, in the next iteration, only those candidate 3-itemsets are retained whose subsets are frequent.

First, we set the minimum support to 0.01 and extracted item sets of length one, resulting in several single item combinations with higher frequencies. Then, we increased the support value to 0.05 and extracted twenty-five combinations of single item sets, such as {mineral water}, {eggs}, {chocolate}, and {milk}, among others.Next, we set the minimum support to 0.01 and obtained the support value and item sets, as shown in Table 1. Then, we increased the support value to 0.05 and extracted item sets of length two, resulting in only four combinations, namely {egg, mineral water}, {spaghetti, mineral water}, {chocolate, mineral water}, and {milk, mineral water}.

index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	Antecedents length	Consequents length
399	frozenset({'soup'})	frozenset({'milk', 'mineral water'})	0.055	0.052667	0.01	0.1818	3.4522	0.007103	1.157	1	2
398	frozenset({'milk', 'mineral water'})	frozenset({'soup'})	0.0526667	0.055	0.01	0.189	3.4522	0.007103	1.1664	2	1
252	frozenset({'herb & pepper'})	frozenset({'ground beef'})	0.050334	0.093666	0.01532	0.30463	3.2523	0.010618	1.3033	1	1
400	frozenset({'milk'})	frozenset({'soup', 'mineral water'})	0.13233333	0.02633334	0.01	0.0755	2.869	0.00651	1.053	1	2
397	frozenset({'soup', 'mineral water'})	frozenset({'milk'})	0.02633334	0.132333	0.01	0.37974	2.8696	0.00651	1.398	2	1
387	frozenset({'ground beef'})	frozenset({'milk', 'mineral water'})	0.093666	0.052667	0.014	0.14946	2.8379	0.00906	1.1138	1	2
386	frozenset({'milk', 'mineral water'})	frozenset({'ground beef'})	0.0526667	0.09366	0.014	0.2658	2.8379	0.00906	1.234	2	1
…	…	…	…	…	…	…	…	…	…	…	…
388	frozenset({'milk'})	frozenset({'ground beef', 'mineral water'})	0.1323333	0.0373336	0.014	0.1057	2.8337	0.009059	1.0765	1	2
385	frozenset({'ground beef', 'mineral water'})	frozenset({'milk'})	0.0373336	0.13233	0.014	0.375	2.8337	0.00905	1.388	2	1
351	frozenset({'tomatoes'})	frozenset({'spaghetti', 'mineral water'})	0.068667	0.06133	0.01167	0.1699	2.7701	0.00745	1.1307	1	2
435	frozenset({'burgers'})	frozenset({'french fries', 'eggs'})	0.0856667	0.042665	0.01	0.11673	2.7358	0.00634	1.0838	1	2
434	frozenset({'french fries', 'eggs'})	frozenset({'burgers'})	0.04265	0.085667	0.01	0.234375	2.7358	0.00634	1.194	2	1
259	frozenset({'frozen vegetables'})	frozenset({'shrimp'})	0.092	0.077	0.0193	0.21014	2.7292	0.01224	1.168	1	1

Table 4 : Defining metric and its threshold

index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	Antecedents length	Consequents length
384	frozenset({'ground beef', 'milk'})	frozenset({'mineral water'})	0.024	0.238	0.014	0.5834	2.4509	0.008	1.8288	2	1
396	frozenset({'soup', 'milk'})	frozenset({'mineral water'})	0.0173	0.238	0.01	0.5767	2.4240	0.005	1.8010	2	1
391	frozenset({'milk', 'frozen vegetables'})	frozenset({'mineral water'})	0.0234	0.238	0.0126	0.567	2.3830	0.007	1.7604	2	1
348	frozenset({'tomatoes', 'spaghetti'})	frozenset({'mineral water'})	0.024	0.238	0.011	0.4861	2.0424	0.005	1.4828	2	1
…	…	…	…	…	…	…	…	…	…	…	…
458	frozenset({'milk', 'frozen vegetables'})	frozenset({'spaghetti'})	0.0234	0.183	0.0106	0.477	2.6051	0.0065	1.5633	2	1
312	frozenset({'spaghetti', 'milk'})	frozenset({'mineral water'})	0.0435	0.238	0.020	0.476	2.0038	0.0103	1.4567	2	1
432	frozenset({'burgers', 'french fries'})	frozenset({'eggs'})	0.0213	0.1832	0.01	0.468	2.5022	0.0060	1.5297	2	1
343	frozenset({'spaghetti', 'olive oil'})	frozenset({'mineral water'})	0.0243	0.238	0.0113	0.4657	1.9569	0.0055	1.4263	2	1

Table 5 : Sort values based on confidence

Customers who purchased eggs and ground beef are anticipated to purchase mineral water with a chance of 58.33%, according to Table 4 and 5 above (confidence). Scores for lift and conviction also lend credence to that theory. To boost sales, it would be preferable to retain them nearby. As it is the most widely used product in the sample, mineral water predominates the association findings. In order to gather more information, it is recommended to create a confidence table without the mineral water.

4. CONCLUSION AND FUTURE SCOPE

In this article, we demonstrate how to use the Apriori method to extract locally common patterns from a dataset. Several data mining approaches have been employed in the past to analyse patterns. Apriori, especially for transactional databases, is best for locating locally frequent objects. Many applications, including market basket analysis, communications, network analysis, financial services, etc., may be made use of this approach. Only tiny datasets are useful for this strategy. To create candidate sets for larger datasets and find frequently occurring objects, a lot of time and work is required. But, by modifying the algorithm in our study, we were able to produce positive findings on a sizable dataset.

REFERENCE

[1] Langfang, L., Qingxian, P., Xiuqin, Q., (2010). New Application of Association Rules in Teaching Evaluation System, International Conference on Computer and Information Application.

[2] Luo, F., Qiu, Q., (2012). The Study on the Application of Data Mining Based on Association Rules, International Conference on Communication Systems and Network Technologies.

[3] Roghayeh, S., Mohammad,Saniee A., Jan, Z., (2015). Association Rule Discovery for Student Performance Prediction Using Metaheuristic Algorithms.

[4] Choudhary, S., Upadhyay, A., (2016). Hiding Sensitive Data Item Using Association Rule Mining, International Journal of Engineering Sciences & Management.

[5] Kalariya, D.C., Shah, V., Vala, J., (2015). Association Rule Hiding based on Heuristic Approach by Deleting Item at R.H.S side of Sensitive Rule, International Journal of Computere Application.

[6] Madhave, P., Maneadn, M., Patil, S., (2013). Data mining using Association rule based on APPIORI algorithm and improved approach with illustration, International Journal of Latest Trends in Engineering and Technology.

[7] Shridhar, M., Parmar, M., (2017). Survey on Association Rule Mining and Its Approaches, International Journal of Computer Science and Engineering.

[8] Oliveira, Stanley., R. M., Za, ïaneOsmar, R., (2002). Privacy Preserving Frequent Itemset Mining, IEEE International Conference on Data Min- ing Workshop on Privacy.

[9] Anand, S., Vibha, O., (2010). Implementation of Cryptography for Privacy Preserving Data Mining, International Journal of Database Management Systems.

[10] Jing, L., (2022). Association Rule Mining Algorithm in College Students' Quality Evaluation System, Hindawi Journal of Electrical and Computer Engineering.

[11] Poovammal,E.,Ponnavaikko,M., (2009). Task Independent Privacy Preserving Data Mining on Medical Dataset, International Conference on Advances in Computing Control and Telecommunication Technologies, IEEE.

[12] Selvamani, D.,Selvi,V., (2019). Association Rule Mining for Intrusion Detection System: A Survey, Asian Journal of Engineering and Applied Technology.

[13] Naresh, P., Suguna,R.,(2019). Implementation of Improved Association Rule Mining Algorithms for Fast Mining with Efficient Tree Structures on Large Datasets, International Journal of Engineering and Advanced Technology.

[14] Ciriani,V., De Capitani di Vimercati,S.,Foresti,S.,Samarati,P., (2008) K-Anonymous DataMining : A Survey, Spinger.

[15] Dr. Kamakshi,P., (2014). A Survey on Privacy Issues and Privacy Preservation in Spatial Data Mining, International Conference on Circuit Power and Computing Technologies, IEEE.

[16] Justin, Z., Stan, M., (2006). A Crypto-Based Approach to Privacy-Preserving Collaborative Data Mining, Sixth IEEE International Conference on Data Mining - Workshops, IEEE.

[17] Shynu,P. G., Md. Shayan, H., ChiranjiLal, Chowdhary., (2020). A Fuzzy based Data Perturbation Technique for Privacy Preserved Data Mining, International Conference on Emerging Trends in Information Technology and Engineering, IEEE.

[18] Weijia, Y., (2008). Knowledge Reserving in Privacy Preserving Data Mining, Second International Symposium on Intelligent Information Technology Application, IEEE.

[19] Supriyamenon, M., Dr. Rajarajeswari,P.,(2017). A Review on Association Rule Mining Techniques with Respect to their Privacy Preserving Capabilities", International Journal of Applied Engineering Research.

[20] Anushree, R., Rio, G.L., D'Souza, (2018). Survey on Anonymization of Privacy Preserving Data Publishing, International Journal of Scientific Development and Research.

[21] Varsha, P., Namrata, T., (2019). A Study of Privacy Preserving Data Mining and Techniques, International Research Journal of Engineering and Technology.

[22] ie Wang, Jun, Z., (2007). Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization, IEEE.

[23] Ricrdo, M.s, Joao, P. V., (2017). Privacy-Preserving Data Mining: Methods, Metrics, and Applications,IEEE.

[24] Wang, Y., Le, J., Huang, D., (2010). A Method for Privacy Preserving Mining of Association Rules based on Web Usage Mining, International Conference on Web Information Systems and Mining.

[25] Saurabh, K., (2012). Privacy Preserving Classification of heterogeneous Partition Data through ID3 Technique, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).

[26] Suzan, W., (2014). Review and Comparison of Associative Classification Data Mining Approaches, World Academy of Science, Engineering and Technology International Journal of Industrial and Manufacturing Engineering.

[27] Javheri, S. B., Kulkarni,U. V.,(2018). A Survey on Privacy Preserving Machine Learning Techniques for Distributed Data Mining", International Journal of Computer Sciences and Engineering.

[28] Ruogu, K., Laura, D., Nathaniel, F., Sara, K., (2015). My Data Just Goes Everywhere: User Mental Models of the Internet and Implications for Privacy and Security, Symposium on Usable Privacy and Security (SOUPS).

[29] Sjors, O., Marco, S.,Remko, H., (2015). Towards decision analytics in product portfolio management, Decision Analytics.

[30] Solanki,R.,(1998). Principle of Data Mining", McGraw-Hill Publication