Market Basket Analysis in Data Mining

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.

The adoption of market basket analysis was aided by the advent of electronic point-of-sale (POS) systems. Compared to handwritten records kept by store owners, the digital records generated by POS systems made it easier for applications to process and analyze large volumes of purchase data.

Implementation of market basket analysis requires a background in statistics and data science and some algorithmic computer programming skills. For those without the needed technical skills, commercial, off-the-shelf tools exist.

One example is the Shopping Basket Analysis tool in Microsoft Excel, which analyzes transaction data contained in a spreadsheet and performs market basket analysis. A transaction ID must relate to the items to be analyzed. The Shopping Basket Analysis tool then creates two worksheets:

- The Shopping Basket Item Groups worksheet, which lists items that are frequently purchased together,
- And the Shopping Basket Rules worksheet shows how items are related (For example, purchasers of Product A are likely to buy Product B).

How does Market Basket Analysis Work?

Market Basket Analysis is modelled on Association rule mining, i.e., the IF {}, THEN {} construct. For example, IF a customer buys bread, THEN he is likely to buy butter as well.

Association rules are usually represented as: {Bread} -> {Butter}

Some terminologies to familiarize yourself with Market Basket Analysis are:

- **Antecedent:** Items or 'itemsets' found within the data are antecedents. In simpler words, it's the IF component, written on the left-hand side. In the above example, bread is the antecedent.
- **Consequent:**A consequent is an item or set of items found in combination with the antecedent. It's the THEN component, written on the right-hand side. In the above example, butter is the consequent.

With the help of the Apriori Algorithm, we can further classify and simplify the item sets which are frequently bought by the consumer.
There are three components in APRIORI ALGORITHM:

- SUPPORT
- CONFIDENCE
- LIFT

Now take an example, suppose 5000 transactions have been made through a popular eCommerce website. Now they want to calculate the support, confidence, and lift for the two products, let's say pen and notebook for example out of 5000 transactions, 500 transactions for pen, 700 transactions for notebook, and 1000 transactions for both.

SUPPORT: It is been calculated with the number of transactions divided by the total number of transactions made,

support(pen) = transactions related to pen/total transactions

i.e support -> 500/5000=10 percent

CONFIDENCE: It is been calculated for whether the product sales are popular on individual sales or through combined sales. That is calculated with combined transactions/individual transactions.

Confidence = combine transactions/individual transactions

i.e confidence-> 1000/500=20 percent

LIFT: Lift is calculated for knowing the ratio for the sales.

Lift-> 20/10=2

When the Lift value is below 1 means the combination is not so frequently bought by consumers. But in this case, it shows that the probability of buying both the things together is high when compared to the transaction for the individual items sold.

With this, we come to an overall view of the Market Basket Analysis in Data Mining and how to calculate the sales for combination products.

Types of Market Basket Analysis

Market Basket Analysis techniques can be categorized based on how the available data is utilized. Here are the following types of market basket analysis in data mining, such as:

1. **Descriptive market basket analysis:** This type only derives insights from past data and is the most frequently used approach. The analysis here does not make any predictions but rates the association between products using statistical techniques. For those familiar with the basics of Data Analysis, this type of modelling is known as unsupervised learning.
2. **Predictive market basket analysis:** This type uses supervised learning models like classification and regression. It essentially aims to mimic the market to analyze what causes what to happen. Essentially, it considers items purchased in a sequence to determine cross-selling. For example, buying an extended warranty is more likely to follow the purchase of an iPhone. While it isn't as widely used as a descriptive MBA, it is still a very valuable tool for marketers.
3. **Differential market basket analysis:** This type of analysis is beneficial for competitor analysis. It compares purchase history between stores, between seasons, between two time periods, between different days of the week, etc., to find interesting patterns in consumer behaviour. For example, it can help determine why some users prefer to purchase the same product at the same price on Amazon vs Flipkart. The answer can be that the Amazon reseller has more warehouses and can deliver faster, or maybe something more profound like user experience.

Algorithms associated with Market Basket Analysis

In market basket analysis, association rules are used to predict the likelihood of products being purchased together. Association rules count the frequency of items that occur together, seeking to find associations that occur far more often than expected.

Algorithms that use association rules include AIS, SETM and Apriori. The Apriori algorithm is commonly cited by data scientists in research articles about market basket analysis. It identifies frequent items in the database and then evaluates their frequency as the datasets are expanded to larger sizes.

R's rules package is an open-source toolkit for association mining using the R programming language. This package supports the Apriori algorithm and other mining algorithms, including arulesNBMiner, opusminer, RKEEL and RSarules.

With the help of the Apriori Algorithm, we can further classify and simplify the item sets that the consumer frequently buys. There are three components in APRIORI ALGORITHM:

- SUPPORT
- CONFIDENCE
- LIFT

For example, suppose 5000 transactions have been made through a popular e-Commerce website. Now they want to calculate the support, confidence, and lift for the two products. For example, let's say pen and notebook, out of 5000 transactions, 500 transactions for pen, 700 transactions for notebook, and 1000 transactions for both.

SUPPORT

It has been calculated with the number of transactions divided by the total number of transactions made,

1. Support = freq(A, B)/N

   support(pen) = transactions related to pen/total transactions

   i.e support -> 500/5000=10 percent

   CONFIDENCE

   Whether the product sales are popular on individual sales or through combined sales has been calculated. That is calculated with combined transactions/individual transactions.

   Confidence = freq (A, B)/ freq(A)

   Confidence = combine transactions/individual transactions

   i.e confidence-> 1000/500=20 percent

   LIFT

   Lift is calculated for knowing the ratio for the sales.

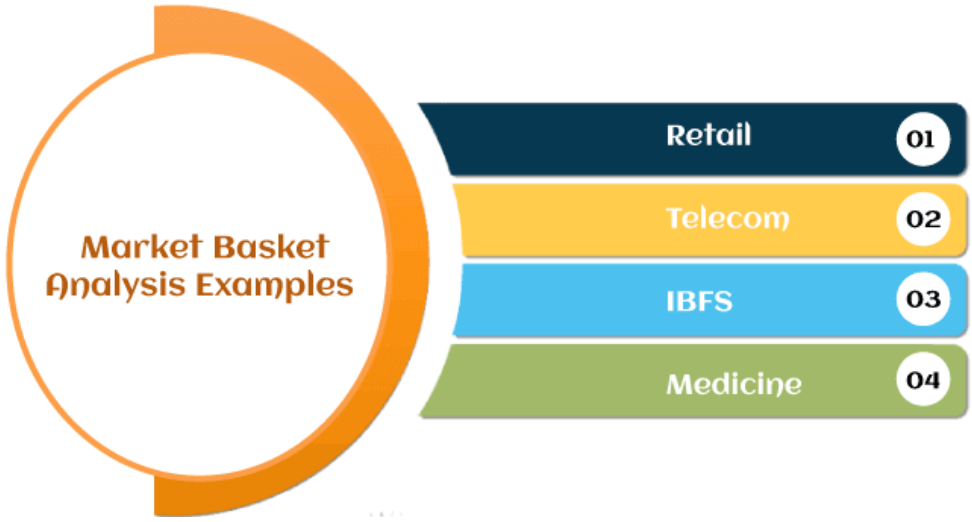
   Lift = confidence percent/ support percent

   Lift-> 20/10=2

When the Lift value is below 1, the combination is not so frequently bought by consumers. But in this case, it shows that the probability of buying both the things together is high when compared to the transaction for the individual items sold.

Examples of Market Basket Analysis

Here are the following examples that *explore Market Basket Analysis by market segment, such as:*



- **Retail:** The most well-known MBA case study is Amazon.com. Whenever you view a product on Amazon, the product page automatically recommends, "Items bought together frequently." It is perhaps the simplest and most clean example of an MBA's cross-selling techniques.
  Apart from e-commerce formats, BA is also widely applicable to the in-store retail segment. Grocery stores pay meticulous attention to product placement based and shelving optimization. For example, you are almost always likely to find shampoo and conditioner placed very close to each other at the grocery store. Walmart's infamous beer and diapers association anecdote is also an example of Market Basket Analysis.
- **Telecom:** With the ever-increasing competition in the telecom sector, companies are paying close attention to customers' services. For example, Telecom has now started to bundle TV and Internet packages apart from other discounted online services to reduce churn.

- **IBFS:** Tracing credit card history is a hugely advantageous MBA opportunity for IBFS organizations. For example, Citibank frequently employs sales personnel at large malls to lure potential customers with attractive discounts on the go. They also associate with apps like Swiggy and Zomato to show customers many offers they can avail of via purchasing through credit cards. IBFS organizations also use basket analysis to determine fraudulent claims.
- **Medicine:** Basket analysis is used to determine comorbid conditions and symptom analysis in the medical field. It can also help identify which genes or traits are hereditary and which are associated with local environmental effects.

Benefits of Market Basket Analysis

The market basket analysis data mining technique has the following benefits, such as:

- **Increasing market share:** Once a company hits peak growth, it becomes challenging to determine new ways of increasing market share. Market Basket Analysis can be used to put together demographic and gentrification data to determine the location of new stores or geo-targeted ads.
- **Behaviour analysis:** Understanding customer behaviour patterns is a primal stone in the foundations of marketing. MBA can be used anywhere from a simple catalogue design to UI/UX.
- **Optimization of in-store operations:** MBA is not only helpful in determining what goes on the shelves but also behind the store. Geographical patterns play a key role in determining the popularity or strength of certain products, and therefore, MBA has been increasingly used to optimize inventory for each store or warehouse.

- **Campaigns and promotions:** Not only is MBA used to determine which products go together but also about which products form keystones in their product line.
- **Recommendations:** OTT platforms like Netflix and Amazon Prime benefit from MBA by understanding what kind of movies people tend to watch frequently.

Frequent Item set in Data set (Association Rule Mining)

- **INTRODUCTION:**

1. Frequent item sets, also known as association rules, are a fundamental concept in association rule mining, which is a technique used in data mining to discover relationships between items in a dataset. The goal of association rule mining is to identify relationships between items in a dataset that occur frequently together.
2. A frequent item set is a set of items that occur together frequently in a dataset. The frequency of an item set is measured by the support count, which is the number of transactions or records in the dataset that contain the item set. For example, if a dataset contains 100 transactions and the item set {milk, bread} appears in 20 of those transactions, the support count for {milk, bread} is 20.
3. Association rule mining algorithms, such as Apriori or FP-Growth, are used to find frequent item sets and generate association rules. These algorithms work by iteratively generating candidate item sets and pruning those that do not meet the minimum support threshold. Once the frequent item sets are found, association rules can be generated by using the concept of confidence, which is the ratio of the number of transactions that contain the item set and the number of transactions that contain the antecedent (left-hand side) of the rule.
4. Frequent item sets and association rules can be used for a variety of tasks such as market basket analysis, cross-selling and recommendation systems. However, it should be noted that association rule mining can generate a large number of rules, many of which may be irrelevant or uninteresting. Therefore, it is important to use appropriate measures such as lift and conviction to evaluate the interestingness of the generated rules.

Association Mining searches for frequent items in the data set. In frequent mining usually, interesting associations and correlations between item sets in transactional and relational databases are found. In short, Frequent Mining shows which items appear together in a transaction or relationship.

Need of Association Mining: Frequent mining is the generation of association rules from a Transactional Dataset. If there are 2 items X and Y purchased frequently then it's good to put them together in stores or provide some discount offer on one item on purchase of another item. This can really increase sales. For example, it is likely to find that if a customer buys **Milk** and **bread** he/she also buys **Butter**. So the association rule is **['milk]^['bread']=>['butter']**. So the seller can suggest the customer buy butter if he/she buys Milk and Bread.

Important Definitions :

- **Support :** It is one of the measures of interestingness. This tells about the usefulness and certainty of rules. **5% Support** means total 5% of transactions in the database follow the rule.

Support(A -> B) = Support_count(A ∪ B)

- **Confidence:** A confidence of 60% means that 60% of the customers who purchased a milk and bread also bought butter.

Confidence(A -> B) = Support_count(A ∪ B) / Support_count(A)

If a rule satisfies both minimum support and minimum confidence, it is a strong rule.

- **Support_count(X)**: Number of transactions in which X appears. If X is A **union** B then it is the number of transactions in which A and B both are present.
- **Maximal Itemset:** An itemset is maximal frequent if none of its supersets are frequent.
- **Closed Itemset:** An itemset is closed if none of its immediate supersets have same support count same as Itemset.
- **K- Itemset:** Itemset which contains K items is a K-itemset. So it can be said that an itemset is frequent if the corresponding support count is greater than the minimum support count.

Example On finding Frequent Itemsets – Consider the given dataset with given transactions.

TransactionId	Items
1	{A,C,D}
2	{B,C,D}
3	{A,B,C,D}
4	{B,D}
5	{A,B,C,D}

- Lets say minimum support count is 3

- Relation hold is maximal frequent => closed => frequent

1-frequent: *{A} = 3; // not closed due to {A, C} and not maximal {B} = 4; // not closed due to {B, D} and no maximal {C} = 4; // not closed due to {C, D} not maximal {D} = 5; // closed item-set since not immediate super-set has same count. Not maximal*
2-frequent: *{A, B} = 2 // not frequent because support count < minimum support count so ignore {A, C} = 3 // not closed due to {A, C, D} {A, D} = 3 // not closed due to {A, C, D} {B, C} = 3 // not closed due to {B, C, D} {B, D} = 4 // closed but not maximal due to {B, C, D} {C, D} = 4 // closed but not maximal due to {B, C, D}*
3-frequent: *{A, B, C} = 2 // ignore not frequent because support count < minimum support count {A, B, D} = 2 // ignore not frequent because support count < minimum support count {A, C, D} = 3 // maximal frequent {B, C, D} = 3 // maximal frequent*
4-frequent: *{A, B, C, D} = 2 //ignore not frequent </*

ADVANTAGES OR DISADVANTAGES:

Advantages of using frequent item sets and association rule mining include:

1. Efficient discovery of patterns: Association rule mining algorithms are efficient at discovering patterns in large datasets, making them useful for tasks such as market basket analysis and recommendation systems.
2. Easy to interpret: The results of association rule mining are easy to understand and interpret, making it possible to explain the patterns found in the data.
3. Can be used in a wide range of applications: Association rule mining can be used in a wide range of applications such as retail, finance, and healthcare, which can help to improve decision-making and increase revenue.
4. Handling large datasets: These algorithms can handle large datasets with many items and transactions, which makes them suitable for big-data scenarios.

Disadvantages of using frequent item sets and association rule mining include:

1. Large number of generated rules: Association rule mining can generate a large number of rules, many of which may be irrelevant or uninteresting, which can make it difficult to identify the most important patterns.
2. Limited in detecting complex relationships: Association rule mining is limited in its ability to detect complex relationships between items, and it only considers the co-occurrence of items in the same transaction.
3. Can be computationally expensive: As the number of items and transactions increases, the number of candidate item sets also increases, which can make the algorithm computationally expensive.
4. Need to define the minimum support and confidence threshold: The minimum support and confidence threshold must be set before the association rule mining process, which can be difficult and requires a good understanding of the data.

Frequent Pattern Mining in Data Mining

-

Frequent pattern mining in data mining is the process of identifying patterns or associations within a dataset that occur frequently. This is typically done by analyzing large datasets to find items or sets of items that appear together frequently.
Frequent pattern extraction is an essential mission in data mining that intends to uncover repetitive patterns or itemsets in a granted dataset. It encompasses recognizing collections of components that occur together frequently in a transactional or relational database. This procedure can offer valuable perceptions into the connections and affiliations among diverse components or features within the data.
Here's an elaborate explanation of repeating arrangement prospecting:

- **Transactional and Relational Databases:**

Repeating arrangement prospecting can be applied to transactional databases, where each transaction consists of a collection of objects. For instance, in a retail dataset, each transaction may represent a customer's purchase with objects like loaf, dairy, and ovals. It can also be used with relational databases, where data is organized into multiple related tables. In this case, repeating arrangements can represent connections among different attributes or columns.

- **Support and Repeating Groupings:**

The support of a grouping is defined as the proportion of transactions in the database that contain that particular grouping. It represents the frequency or occurrence of the grouping in the dataset. Repeating groupings are collections of objects whose support is above a specified minimum support threshold. These groupings are considered interesting and are the primary focus of repeating arrangement prospecting.

- **Apriori Algorithm:**

The Apriori algorithm is one of the most well-known and widely used algorithms for repeating arrangement prospecting. It uses a breadth-first search strategy to discover repeating groupings efficiently. The algorithm works in multiple iterations. It starts by finding repeating individual objects by scanning the database once and counting the occurrence of each object. It then generates candidate groupings of size 2 by combining the repeating groupings of size 1. The support of these candidate groupings is calculated by scanning the database again. The process continues iteratively, generating candidate groupings of size k and calculating their support until no more repeating groupings can be found.

- **Support-based Pruning:**

During the Apriori algorithm's execution, aid-based pruning is used to reduce the search space and enhance efficiency. If an itemset is found to be rare (i.e., its aid is below the minimum aid threshold), then all its supersets are also assured to be rare. Therefore, these supersets are trimmed from further consideration. This trimming step significantly decreases the number of potential item sets that need to be evaluated in subsequent iterations.

- **Association Rule Mining:**

Frequent item sets can be further examined to discover association rules, which represent connections between different items. An association rule consists of an antecedent and a consequent (right-hand side), both of which are item sets. For instance, {milk, bread} => {eggs} is an association rule. Association rules are produced from frequent itemsets by considering different combinations of items and calculating measures such as aid, confidence, and lift. Aid measures the frequency of both the antecedent and the consequent appearing together, while confidence measures the conditional probability of the consequent given the antecedent. Lift indicates the strength of the association between the antecedent and the consequent, considering their individual aid.

- **Applications:**

Frequent pattern mining has various practical uses in different domains. Some examples include market basket analysis, customer behavior analysis, web mining, bioinformatics, and network traffic analysis. Market basket analysis involves analyzing customer purchase patterns to identify connections between items and enhance sales strategies. In bioinformatics, frequent pattern mining can be used to identify common patterns in DNA sequences, protein structures, or gene expressions, leading to insights in genetics and drug design. Web mining can employ frequent pattern mining to discover navigational patterns, user preferences, or collaborative filtering recommendations on the web.
Regular pattern extraction is a data extraction approach employed to spot repeating forms or itemsets in transactional or relational databases. It entails locating collections of objects that occur collectively often and possesses numerous uses in different fields. The Apriori algorithm is a well-liked technique utilized to effectively detect consistent itemsets, and association rule extraction can be carried out to obtain significant connections between objects.

There are several different algorithms used for frequent pattern mining, including:

1. Apriori algorithm: This is one of the most commonly used algorithms for frequent pattern mining. It uses a "bottom-up" approach to identify frequent itemsets and then generates association rules from those itemsets.
2. ECLAT algorithm: This algorithm uses a "depth-first search" approach to identify frequent itemsets. It is particularly efficient for datasets with a large number of items.
3. FP-growth algorithm: This algorithm uses a "compression" technique to find frequent patterns efficiently. It is particularly efficient for datasets with a large number of transactions.
4. Frequent pattern mining has many applications, such as Market Basket Analysis, Recommender Systems, Fraud Detection, and many more.

Advantages:

1. It can find useful information which is not visible in simple data browsing
2. It can find interesting association and correlation among data items

Disadvantages:

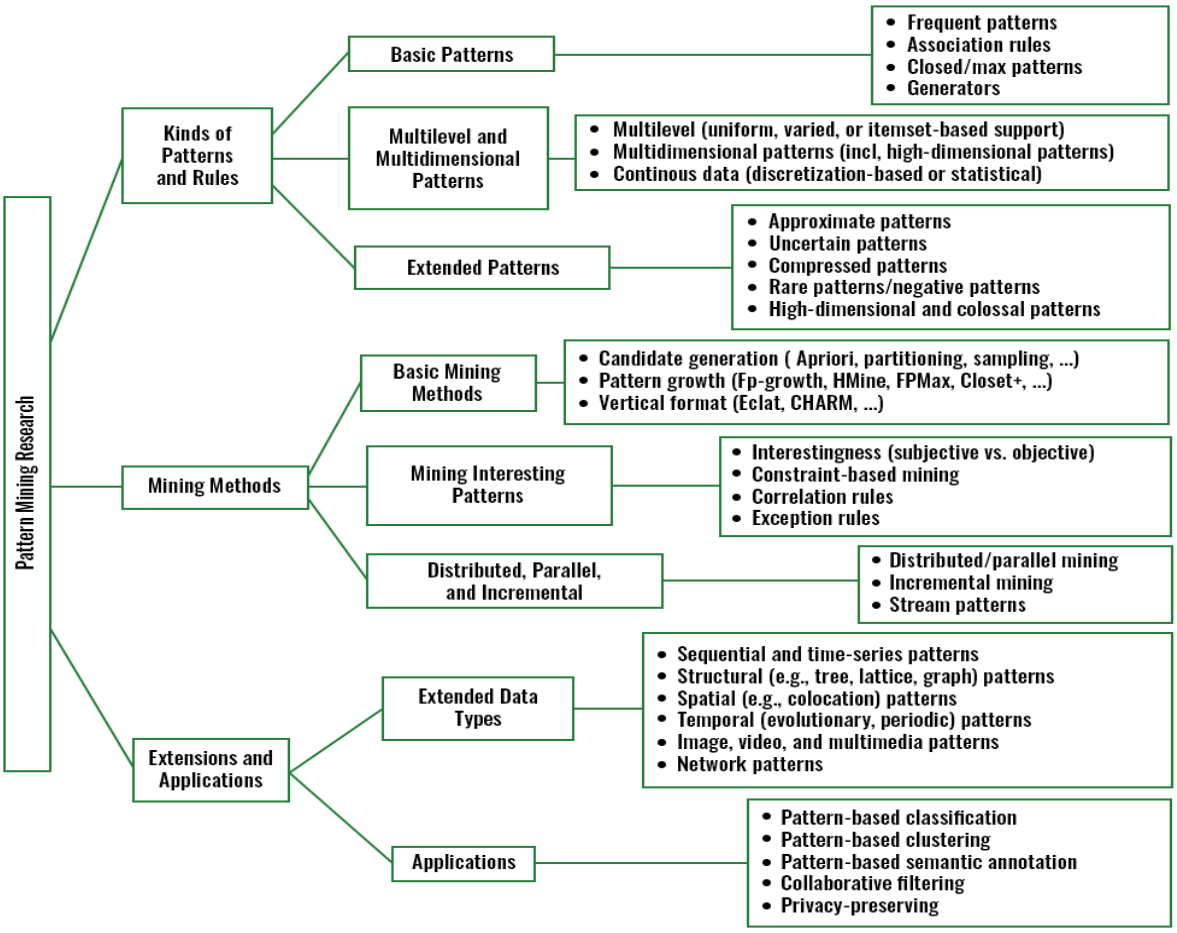
1. It can generate a large number of patterns
2. With high dimensionality, the number of patterns can be very large, making it difficult to interpret the results.

The increasing power of computer technology creates a large amount of data and storage. Databases are increasing rapidly and in this computerized world everything is shifting online and data is increasing as a new currency. Data comes in different shapes and sizes and is collected in different ways. By using data mining there are many benefits it helps us to improve the particular process and in some cases, it costs saving or revenue generation. Data mining is commonly used to search a large amount of data for patterns and trends, and not only for searching it uses the data for further processes and develops actionable processes.

Data mining is the process of converting raw data into suitable patterns based on trends.

Data mining has different types of patterns and **frequent pattern mining** is one of them. This concept was introduced for mining transaction databases. Frequent patterns are patterns(such as items, subsequences, or substructures) that appear frequently in the database. It is an analytical process that finds frequent patterns, associations, or causal structures from databases in various databases. This process aims to find the frequently occurring item in a transaction. By frequent patterns, we can identify strongly correlated items together and we can identify similar characteristics and associations among them. By doing frequent data mining we can go further for clustering and association.

Frequent pattern mining is a major concern it plays a major role in associations and correlations and disclose an intrinsic and important property of dataset.
Frequent data mining can be done by using association rules with particular algorithms eclat and apriori algorithms. Frequent pattern mining searches for recurring relationships in a data set. It also helps to find the inheritance regularities. to make fast processing software with a user interface and used for a long time without any error.



Association Rule Mining:
It is easy to find associations in frequent patterns:

- for each frequent pattern x for each subset y c x.
- calculate the support of y-> x – y.

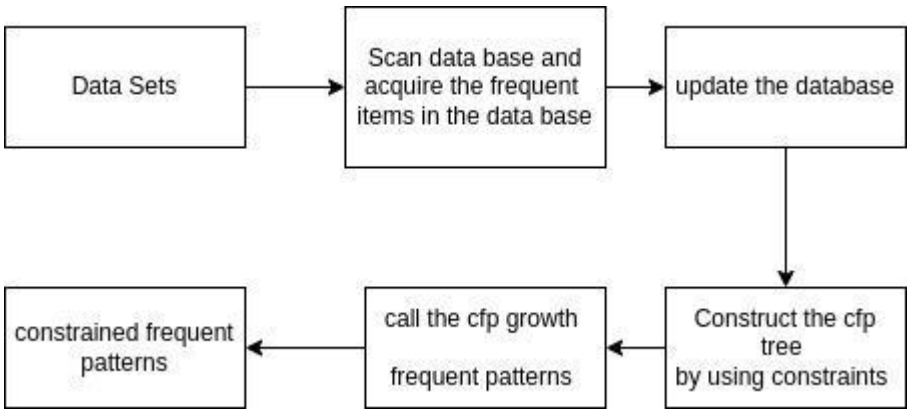
if it is greater than the threshold, keep the rule. There are two algorithms that support this lattice

1. Apriori algorithm
2. eclat algorithm

Apriori	Eclat
It performs “perfect” pruning of infrequent item sets.	It reduces memory requirements and is faster.
It requires a lot of memory(all frequent item sets are represented) and support counting takes very long for large transactions. But this is not efficient in practice.	Its storage of transaction list.

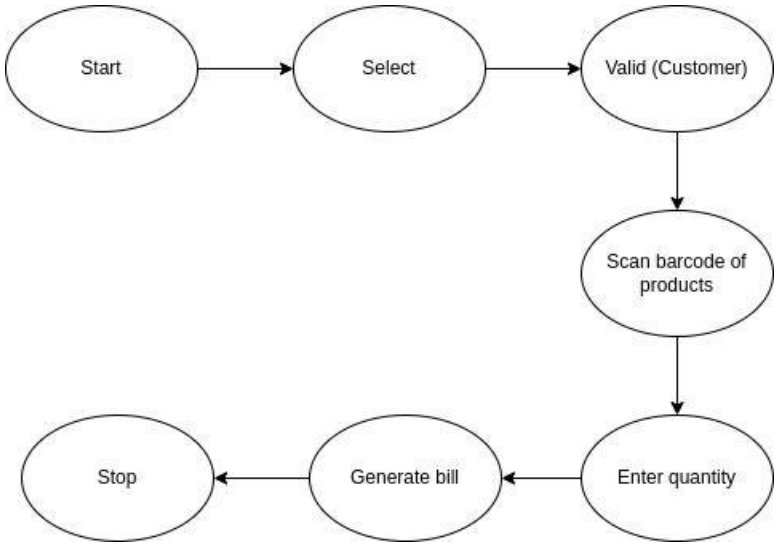
The words support and confidence support the association rule.

- **Support:** how often a given rule in a database is mined? support the transaction contains x U y
- **Confidence:** the number of times the given rule in a practice is true. The conditional probability is a transaction having x as well as y.



working principle (it is a simple point of scale application for any supermarket which has a good off-product scale)

- the product data will be entered into the database.
- the taxes and commissions are entered.
- the product will be purchased and it will be sent to the bill counter.
- the bill calculating operator will check the product with the bar code machine it will check and match the product in the database and then it will show the information of the product.
- the bill will be paid by the customer and he will receive the products.



Tasks in the frequent pattern mining:

- Association
- **Cluster analysis:** frequent pattern-based clustering is well suited for high-dimensional data. by the extension of dimension the sub-space clustering occurs.
- **Data warehouse:** iceberg cube and cube gradient
- Broad applications

There are some to improve the efficiency of the tasks.

Closed Pattern:

A frequent pattern, it meets the minimum support criteria. All super patterns of a closed pattern are less frequent than the closed pattern.

Max Pattern:

It also meets the minimum support criteria(like a closed pattern). All super patterns of a max pattern are not frequent patterns. both patterns generate fewer numbers of patterns so therefore they increase the efficiency of the task.

Applications of Frequent Pattern Mining:

basket data analysis, cross-marketing, catalog design, sale campaign analysis, web log analysis, and DNA sequence analysis.

Issues of frequent pattern mining

- flexibility and reusability for creating frequent patterns
- most of the algorithms used for mining frequent item sets do not offer flexibility for reusing
- much research is needed to reduce the size of the derived patterns
- Frequent pattern mining has several applications in different areas, including:
- Market Basket Analysis: This is the process of analyzing customer purchasing patterns in order to identify items that are frequently bought together. This information can be used to optimize product placement, create targeted marketing campaigns, and make other business decisions.

- Recommender Systems: Frequent pattern mining can be used to identify patterns in user behavior and preferences in order to make personalized recommendations.
- Fraud Detection: Frequent pattern mining can be used to identify abnormal patterns of behavior that may indicate fraudulent activity.
- Network Intrusion Detection: Network administrators can use frequent pattern mining to detect patterns of network activity that may indicate a security threat.
- Medical Analysis: Frequent pattern mining can be used to identify patterns in medical data that may indicate a particular disease or condition.
- Text Mining: Frequent pattern mining can be used to identify patterns in text data, such as keywords or phrases that appear frequently together in a document.
- Web usage mining: Frequent pattern mining can be used to analyze patterns of user behavior on a website, such as which pages are visited most frequently or which links are clicked on most often.
- Gene Expression: Frequent pattern mining can be used to analyze patterns of gene expression in order to identify potential biomarkers for different diseases.

These are a few examples of the application of frequent pattern mining. The list is not exhaustive and the technique can be applied in many other areas, as well.

Conclusion:

It is impossible to give complete coverage of this topic with the limited space and our limited knowledge. Frequent pattern mining has achieved tremendous progress and claimed a good set of applications. However in-depth research is required that the field may have a long-lasting and deep impact on data mining applications.

Apriori Algorithm

Prerequisite – Frequent Item set in Data set (Association Rule Mining)

Apriori algorithm is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called *Apriori property* which helps by reducing the search space.

Apriori Property –

All non-empty subset of frequent itemset must be frequent. The key concept of Apriori algorithm is its anti-monotonicity of support measure. Apriori assumes that

All subsets of a frequent itemset must be frequent(Apriori property).

If an itemset is infrequent, all its supersets will be infrequent.

Before we start understanding the algorithm, go through some definitions which are explained in my previous post.

Consider the following dataset and we will find frequent itemsets and generate association rules for them.

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2
minimum confidence is 60%

Step-1: K=1
(I) Create a table containing support count of each item present in dataset –
Called **C1(candidate set)**

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

(II) compare candidate set item's support count with minimum support count(here min_support=2 if support_count of candidate set items is less than min_support then remove those items). This gives us itemset L1.

Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

Step-2: K=2

- Generate candidate set C2 using L1 (this is called join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common.
- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset.(Example subset of{I1, I2} are {I1}, {I2} they are frequent.Check for each itemset)
- Now find support count of these itemsets by searching in dataset.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

(II) compare candidate (C2) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L2.

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

Step-3:

- Generate candidate set C3 using L2 (join step). Condition of joining L_{k-1} and L_{k-1} is that it should have (K-2) elements in common. So here, for L2, first element should match.
  So itemset generated by joining L2 is {I1, I2, I3}{I1, I2, I5}{I1, I3, i5}{I2, I3, I4}{I2, I4, I5}{I2, I3, I5}
- Check if all subsets of these itemsets are frequent or not and if not, then remove that itemset.(Here subset of {I1, I2, I3} are {I1, I2},{I2, I3},{I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset)
- find support count of these remaining itemset by searching in dataset.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

(II) Compare candidate (C3) support count with minimum support count(here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L3.

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Step-4:

- Generate candidate set C4 using L3 (join step). Condition of joining L_{k-1} and L_{k-1} (K=4) is that, they should have (K-2) elements in common. So here, for L3, first 2 elements (items) should match.

- Check all subsets of these itemsets are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5}, which is not frequent). So no itemset in C4
- We stop here because no frequent itemsets are found further

---

Thus, we have discovered all the frequent item-sets. Now generation of strong association rule comes into picture. For that we need to calculate confidence of each rule.

Confidence –
A confidence of 60% means that 60% of the customers, who purchased milk and bread also bought butter.

Confidence(A->B)=Support_count(A∪B)/Support_count(A)

So here, by taking an example of any frequent itemset, we will show the rule generation.
Itemset {I1, I2, I3} //from L3
SO rules can be
[I1^I2]=>[I3] //confidence = sup(I1^I2^I3)/sup(I1^I2) = 2/4*100=50%
[I1^I3]=>[I2] //confidence = sup(I1^I2^I3)/sup(I1^I3) = 2/4*100=50%
[I2^I3]=>[I1] //confidence = sup(I1^I2^I3)/sup(I2^I3) = 2/4*100=50%
[I1]=>[I2^I3] //confidence = sup(I1^I2^I3)/sup(I1) = 2/6*100=33%
[I2]=>[I1^I3] //confidence = sup(I1^I2^I3)/sup(I2) = 2/7*100=28%
[I3]=>[I1^I2] //confidence = sup(I1^I2^I3)/sup(I3) = 2/6*100=33%

So if minimum confidence is 50%, then first 3 rules can be considered as strong association rules.

Limitations of Apriori Algorithm
Apriori Algorithm can be slow. The main limitation is time required to hold a vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets i.e. it is not an efficient approach for large number of datasets. For example, if there are 10^4 from frequent 1- itemsets, it need to generate more than 10^7 candidates into 2-length which in turn they will be tested and accumulate. Furthermore, to detect frequent pattern in size 100 i.e. v1, v2… v100, it have to generate 2^100 candidate itemsets that yield on costly and wasting of time of candidate generation. So, it will check for many sets from candidate itemsets, also it will scan database many times repeatedly for finding candidate itemsets. Apriori will be very low and inefficiency when memory capacity is limited with large number of transactions.

Association Rule Generation :

Given the **mininum threshold confidence**, Generating association rules by going through all possible combinations of frequent item sets and pruning the rules according to confidence criterion.

Following are the steps for strong Association Rule Generation:

- Generate all nonempty subsets for each frequent itemset
- For every nonempty subset S of Itemset I , output of the rule:
    - S --> (I - S)
    - **If** support_count (I) / support_count (S) > = minimum confidence threshold **then** rule is a **strong Association Rule.**

For Example

Consider the below transaction:

TID	Items Bought
1	{ A, C }
2	{ A, B, C, E}
3	{ A, D }
4	{ A, B, C, E }
5	{ A, B, C, D, E }

Given that minimum threshold support = 60% (support count = 3) and minimum threshold confidence = 80% (Confidence =

We have already generated the frequent item sets for this example

Item set	Support Count	Support
{A, B, C }	3	60%

- Generate all nonempty subsets for each frequent itemset
    - For Itemset - { A, B, C } , all non empty subsets are {A,B}, {B,C}, {A,C}, {A}, {B}, {C}

- For every nonempty subset S of Itemset I , output of the rule:
    - S --> (I - S)
        - {A,B} -> {C}
        - {B,C} -> {A}
        - {A,C} -> {C}
        - {A} -> {B,C}
        - {B} -> {A,C}
        - {C} -> {A,B}
    - **If** support_count (I) / support_count (S) > = minimum confidence threshold **then** rule is a **strong Association Rule.**
        - {A,B} -> {C}, Confidence = 3/3 * 100 = 100% - **Yes**, it is a strong association rules
        - {B,C} -> {A}, Confidence = 3/3 * 100 = 100% - **Yes**, it is a strong association rules
        - {A,C} -> {C}, Confidence = 3/4 * 100 = 80% - **Yes**, it is a strong association rules
        - {A} -> {B,C}, Confidence = 3/5 * 100 = 60% - **No**, it is not a strong association rules
        - {B} -> {A,C}, Confidence = 3/3 * 100 = 100% - **Yes**, it is a strong association rules
        - {C} -> {A,B}, Confidence = 3/4 * 100 = 80% - **Yes**, it is a strong association rules

Correlation Analysis in Data Mining

Correlation analysis is a statistical method used to measure the strength of the linear relationship between two variables and compute their association. Correlation analysis calculates the level of change in one variable due to the change in the other. A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.

Researchers use correlation analysis to analyze quantitative data collected through research methods like surveys and live polls for market research. They try to identify relationships, patterns, significant connections, and trends between two variables or datasets. There is a positive correlation between two variables when an increase in one variable leads to an increase in the other. On the other hand, a negative correlation means that when one variable increases, the other decreases and vice-versa.

Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of the relationship, the correlation coefficient's value varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables.

As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The coefficient sign indicates the direction of the relationship; a + sign indicates a positive relationship, and a - sign indicates a negative relationship.