UNIT-3

Data Processing

- Data Processing and Over View
- Data Cleaning
- Overview of Data Reduction Strategies
  - Histogram
  - Sampling
  - Data Cube Aggregation
- Association Rule Mining
  - Basic Concepts: Association
- Apriori Algorithm

Unit –3 Data Processing

3.1 Data Processing and Over View:

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyse this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

➢ Why Data Mining?

The simple answer to "Why is data mining required?" is that data which is the core of any business is anywhere and everywhere. Yes, it is a fact. We are living in a world where anything and everything is getting converted to data. Every click, tap, swipe, like, tweet, share, phone call, etc. generates lots and lots of data. The amount of data getting collected and stored is exploding. Just consider the case of a telecom service provider, or a banking service provider. In short, data explosion is one of the reasons that necessitate data mining.

Secondly, the technology is so advanced today that it is really easy and cheap to collect, store and retrieve large volumes of data. Data storage costs have declined dramatically which result in big data. Also, the processing power of computers is exponentially increasing. All these technological advancements help organizations in collecting, storing and retrieving large amount of data from different sources easily and quite cheaply.

Thirdly, competition necessitates the availability of information at your fingertips in the blink of an eye. Your business might be storing terabytes of data in your databases by spending lots of effort, time and money. In addition to the data available within an organization, Internet is also a great data source. But, data in its pure form might not be useful in many situations. So in today's competitive business world, there should be processes in place in order to get useful information from raw data that might help you in critical decision making and development of new strategies.

➢ What is Data Mining?

Data processing is the conversion of data into suitable for use and wished format. There is defining operating sequence by which to realize the conversion of data. The process of conversion is carried out automatically or manually. Nowadays most data is processed with the help of computer equipment. Thus, data can be converted into different forms. It

can be graphic as well as audio ones. It depends on the used software as well as d processing methods.

Processing of data is a key step of the data mining process. Raw data processing is a ti complicated task. Moreover, the results can be misleading. Therefore, it is better process data before analysis.

Data Mining is defined as extracting information from huge sets of data. In other wor we can say that data mining is the procedure of mining knowledge from data. T information or knowledge extracted so can be used for any of the following applicatio

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

3.1.1 Stages of data processing:

No matter which way data is processed any one of them requires preliminary data collection. Collected data go through such stages:

- Data collection
- Data storage
- Data sorting
- Data processing
- Data analysis
- Data reporting

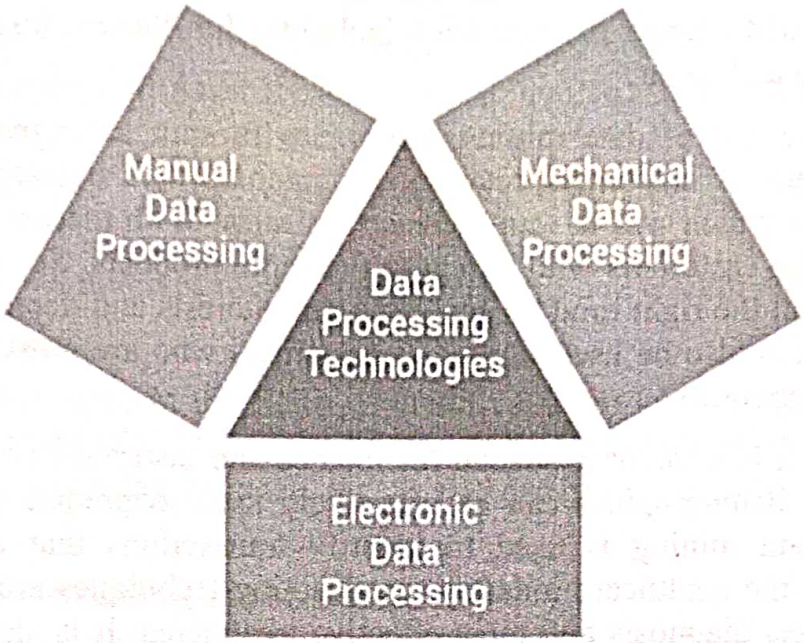
When data is gathered, there is a need to store it. The data can be stored in physical for using paper-based documents, in laptop and desktop computers or in any other da storage devices. With the rise and rapid development of such thing as Data Mining an Big Data, the process of data collection becomes more complicated and time-consumin It is necessary to carry out many operations to conduct a thorough data analysis.

At present data is, for the most part, stored in a digital form. It allows processing da faster and converting into different formats. The user has the possibility to choose th most suitable output.

The next stage followed by data storage is sorting and filtering. At this stage, the form of stored data plays a crucial role and depends on used software. Simple data can b stored in the form of text files or tables or a mixture of both. If data is complicated an requires special handling data processing tools are used to perform tasks that are mor challenging. Data manipulation can be carried out with the help of single software or data requires analysis that is more detailed a set of software is needed to apply.

3.1.2 Technologies of data processing:

There are three data processing technologies:



1. **Manual data processing:**

It involves the processing of data by hand only. Any additional tools or devices are not applied. All of the data manipulations are carried out manually.

2. **Mechanical data processing:**

It is data processing, which entails the use of a mechanical device for work with data. In this case, ordinary electronic devices also can be used. Such devices are calculators or typewriters. Simple operations with data can be realized by means of this method.

3. **Electronic data processing:**

This one is the most progressive. It is realized by means of computers. The use of this method allows for processing an increasing amount of data and provides results that are more accurate.

➢ **The importance of data processing in data mining:**

In today's world, data has a significant bearing on researchers, institutions, commercial organizations, and each individual user. After gathering, the question arises how to store, sort, filter, analyze and present data. Here data mining comes into play. As data is often imperfect, noisy, and incompatible, it requires additional processing.

The complexity of this process is subject to the scope of data collection and the complexity of the required results. Whether this process is time-consuming depends on steps, which need to be made with the collected data and the type of the output file desired to be received. This issue becomes actual when the need for processing a big amount of data arises. Therefore, data mining is widely used nowadays.

3.1.3 Data Mining Applications:

Data mining helps businesses identify important facts, trends, patterns, relationships exceptions that are normally unnoticed or hidden. Thus, data mining techniques applied in a wide range of industries including healthcare, insurance, finance, re manufacturing and so on.

Retailers make use of data mining techniques to spot sales trends. By analysing purchase patterns of customers, retailers can come up with smarter marketing promoti and campaigns which will in turn increase the sales. With market segmentation, retail can identify the customers who purchase the same products. So, they can come up new products at the right time by analysing the interests and demographics of custome Data mining can also be used to predict customers who are most likely start purchas from your competitors.

Fraud detection is a major headache for finance and insurance companies. Studies sh that customer demographics can be effectively used to predict their fraudulent nat Nowadays, data mining is used to identify transactions that are most likely to fraudulent. In the healthcare industry, data mining techniques are mainly used for m accurate disease diagnosis and most effective treatments. It is also helpful in predict health insurance fraud, healthcare cost and length of stay (LOS) of hospitalization.

Data mining is highly useful in the following domains –

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid.

➢ **Market Analysis and Management:**

Listed below are the various fields of market where data mining is used –

- **Customer Profiling** – Data mining helps determine what kind of people buy what kind of products.
- **Identifying Customer Requirements** – Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attrac new customers.
- **Cross Market Analysis** – Data mining performs Association/correlations between product sales.
- **Target Marketing** – Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern** – Data mining helps in determining customer purchasing pattern.

- **Providing Summary Information** – Data mining provides us various multidimensional summary reports.

➢ **Corporate Analysis and Risk Management:**

Data mining is used in the following fields of the Corporate Sector –

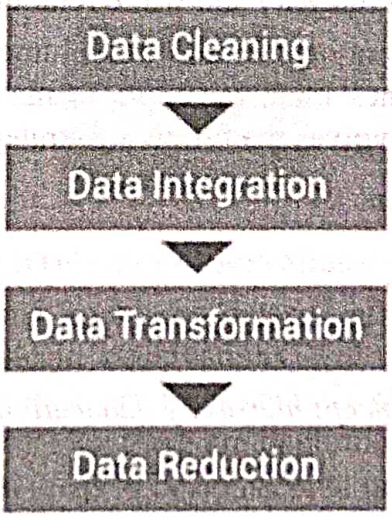
- **Finance Planning and Asset Evaluation** – It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** – It involves summarizing and comparing the resources and spending.
- **Competition** – It involves monitoring competitors and market directions.

➢ **Fraud Detection:**

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyses the patterns that deviate from expected norms.

3.1.4 Data pre-processing methods:

Results of data mining depend on the quality of source data. In order to get data of good quality, it is necessary to pre-process source data. It allows for improving efficiency and facilitating data mining. Pre-processing of data is the preparation and conversion of the original one.



Data pre-processing methods are provided below:

- *Data Cleaning*
- *Data Integration*
- *Data Transformation*
- *Data Reduction*

Data is not always complete. It can miss attribute values or include only aggregate ones. Moreover, data can be noisy, duplicated and inconsistent. There may be human or computer errors at data entry. Such things affect negatively the data mining process. To make the situation better, it is applied to data cleaning. This procedure allows cleaning the data through the imputation of missing values, removing outliers and reconciling inconsistencies. Due to data cleaning works, results at the output level will be more robust.

Data can be combined from different sources into one data storage. Such sources can be various databases, data cubes, and unstructured files. In order to structure data from multiple sources, data integration is used. It is realized by means of metadata (it is also called data about the data) which allows avoiding errors in the integration of data.

Data transformation is one more important procedure on a fast track to receiving final data of good quality. It presents data conversion into forms that are suitable for data mining.

Transformation of data includes normalization, noise data smoothing, aggregation, generalization. Most often, it is realized via a combination of manual and automated processing.

Data mining can be a very time-consuming process especially when there is a necessity to analyse huge volumes of data. In this case, the analysis can be unreasonable. That is where data reduction is used. It enables to analyse reduced data representation without prejudice to the source data integrity and while retaining good quality information.

In the meantime, data mining on the reduced volume of data should be performed more efficiently and the outcomes must be of the same quality as if the whole dataset is analysed. Data reduction involves the following strategies:

- *Data cube aggregation*
- *Dimension reduction*
- *Data compression*
- *Numerosity reduction*
- *Discretization and concept hierarchy generation*

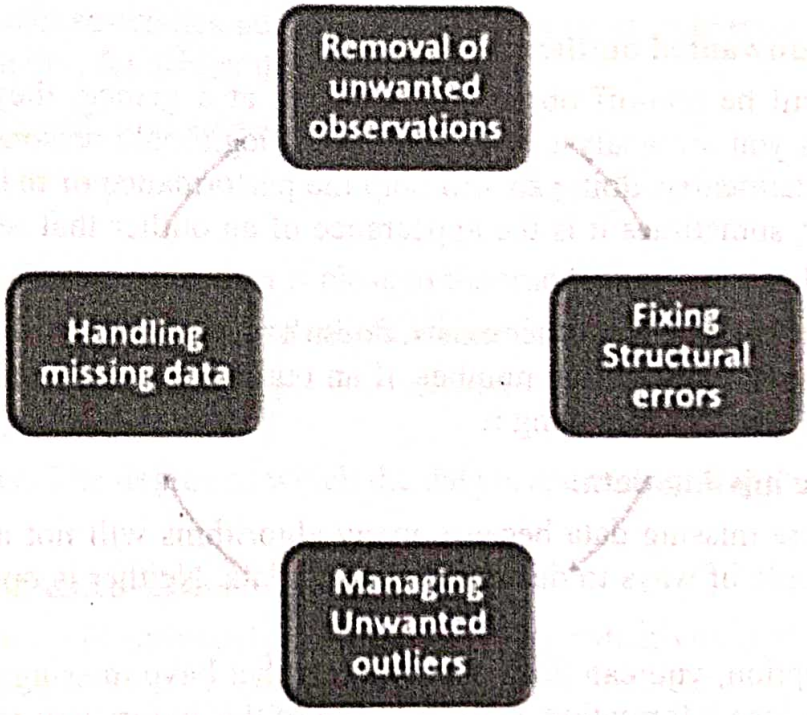
3.2 Data Cleansing:

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. Data Cleaning is one of those things that everyone does but no one really talks about. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, proper data cleaning can make or break your project. Professional data scientists usually spend a very large portion of their time on this step.

Because of the belief that, **"Better data beats fancier algorithms"**. If we have a well-cleaned dataset, we can get desired results even with a very simple algorithm, which can prove very beneficial at times. Obviously, different types of data will require different types of cleaning. However, this systematic approach can always serve as a good starting point.

3.2.1 How do you clean data?

While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to map out a framework for your organization.



Step 1: Remove duplicate or irrelevant observations:

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process.

Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze. For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more perform ant dataset.

Step 2: Fix structural errors:

The errors that arise during measurement transfer of data or other similar situations are called structural errors. Structural errors include typos in the name of features, same attribute with different name, mislabelled classes, i.e. separate classes that should really be the same or inconsistent capitalization.

- For example, the model will treat America and America as different classes or values though they represent the same value or red, yellow and red-yellow as different classes or attributes, though one class can be included in other two classes. So, these are some structural errors that make our model inefficient and gives poor quality results.

Step 3: Filter unwanted outliers:

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analysing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on.

Remember: just because an outlier exists, doesn't mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data:

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

1. As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
2. As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
3. As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Validate and QA:

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?

- If not, is that because of a data quality issue?

False conclusions because of incorrect or "dirty" data can inform poor business strategy and decision-making. False conclusions can lead to an embarrassing moment in a reporting meeting when you realize your data doesn't stand up to scrutiny.

Before you get there, it is important to create a culture of quality data in your organization. To do this, you should document the tools you might use to create this culture and what data quality means to you.

3.2.2 Components of quality data:

Determining the quality of data requires an examination of its characteristics, then weighing those characteristics according to what is most important to your organization and the application(s) for which they will be used.

➢ **characteristics of quality data:**

1. **Validity.** The degree to which your data conforms to defined business rules or constraints.
2. **Accuracy.** Ensure your data is close to the true values.
3. **Completeness.** The degree to which all required data is known.
4. **Consistency.** Ensure your data is consistent within the same dataset and/or across multiple data sets.
5. **Uniformity.** The degree to which the data is specified using the same unit of measure.

3.2.3 Benefits of data cleaning:

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

- Removal of errors when multiple sources of data are at play.
- Fewer errors make for happier clients and less-frustrated employees.
- Ability to map the different functions and what your data is intended to do.
- Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
- Using tools for data cleaning will make for more efficient business practices and quicker decision-making.

3.3 Overview of Data Reduction Strategies:

3.3.1 Data reduction strategies in data mining:

Data reduction strategies applied on huge data set. Complex data and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible.

Data reduction techniques can be applied to obtain a reduces data should be more efficient yet produce the same analytical results.

Strategies for data reduction include the following-

1. Data Cube Aggregation: This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months. They involve you in the annual sales, rather than the quarterly average, so we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

2. Dimension reduction:

Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates out dated or redundant features.

- **Step-wise Forward Selection –**
  The selection begins with an empty set of attributes later on we decide best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.
- **Step-wise Backward Selection –**
  This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.
- **Combination of forwarding and Backward Selection –**
  It allows us to remove the worst and select best attributes, saving time and making the process faster.

3. Data Compression:

The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding & run-length Encoding). We can divide it into two types based on their compression techniques.

- **Lossless Compression –**

  Encoding techniques (Run Length Encoding) allows a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.
- **Lossy Compression –**

  Methods such as Discrete Wavelet transform technique, PCA (principal component analysis) are examples of this compression. For e.g., JPEG image format is a lossy compression, but we can find the meaning equivalent to the original the image. In lossy-data compression, the decompressed data may differ to the original data but useful enough to retrieve information from them.

4. Numerosity Reduction:

In this reduction technique the actual data is replaced with mathematical models or smaller representation of the data instead of actual data, it is important to only store the model parameter. Or non-parametric method such as clustering, histogram, sampling. For More Information on Numerosity Reduction Visit the link below:

5. Discretization & Concept Hierarchy Operation:

Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals. This means that mining results are shown in a concise and easily understandable way.

- **Top-down discretization –**

  If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat of this method up to the end, then the process is known as top-down discretization also known as splitting.

- **Bottom-up discretization –**

  If you first consider the entire constant values as split-points, some are discarded through a combination of the neighbourhood values in the interval, that process is called bottom-up discretization.

➢ **Concept Hierarchies:**

It reduces the data size by collecting and then replacing the low-level concepts (such as 43 for age) to high-level concepts (categorical variables such as middle age or Senior).

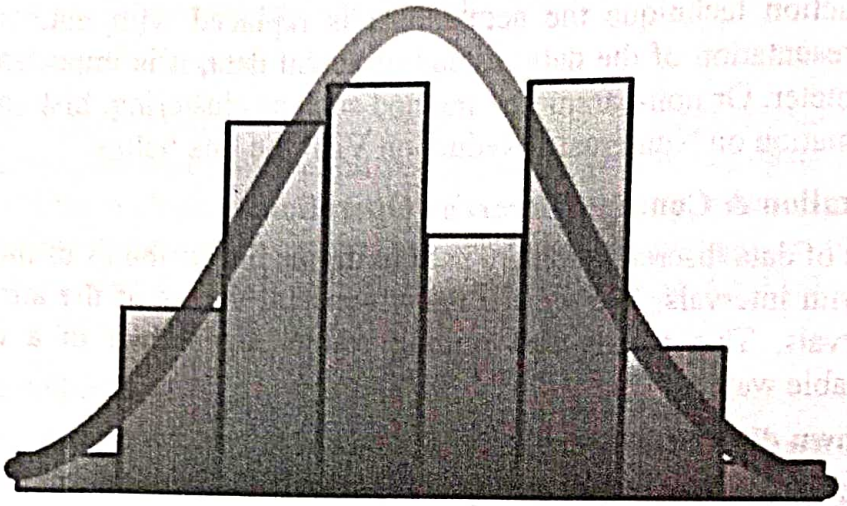
For numeric data following techniques can be followed:

- **Binning –**
  Binning is the process of changing numerical variables into categorical counterparts. The number of categorical counterparts depends on the number of bins specified by the user.

- **Histogram analysis –**
  Like the process of binning, the histogram is used to partition the value for the attribute X, into disjoint ranges called brackets. There are several partitioning rules:

  1. **Equal Frequency partitioning:** Partitioning the values based on their number of occurrences in the data set.
  2. **Equal Width Partioning:** Partioning the values in a fixed gap based on the number of bins i.e. a set of values ranging from 0-20.
  3. **Clustering:** Grouping the similar data together.

3.3.2 Histogram:



A histogram is a type of graph that is widely used in mathematics, especially in statistics. The histogram represents the frequency of occurrence of specific phenomena which lie within a specific range of values, which are arranged in consecutive and fixed intervals. The frequency of the data occurrence is represented by a bar; hence it looks very much like a bar graph.

Histogram is a Graphical Representation of a Frequency table that are taught in school level. But, they are more than just frequency table visualization.

A histogram must have some defined "interval" based on which the data is to divided or mined.

In many cases the size of the interval is unknown. For example in the problems of astrophysics. In case, have data of celestial body moving in the space , it is likely that we do not know when this event will occur again or in simple terms , we do not know the time interval . In such cases algorithms are used to identify the "interval size" and this is done with help of clustering and data mining methods.

In many cases the histogram is just a representation of some signal or image having some channels. A histogram of image can be thought of the frequencies of Red, Blue, Color Channels and similarity for Gray scales: gray intensity as 256 intervals.

To understand in a better manner, you can check your Quora account statistics. It is also showing a histogram…The frequency (views) per day (interval).

3.3.3 Data Cube Aggregation:

When data is grouped or combined in multidimensional matrices called Data Cubes. The data cube method has a few alternative names or a few variants, such as "Multidimensional databases," "materialized views," and "OLAP (On-Line Analytical Processing)."

The general idea of this approach is to materialize certain expensive computations that are frequently inquired.

For example, a relation with the schema sales (part, supplier, customer, and sale-price) can be materialized into a set of eight views as shown in fig, where psc indicates a view consisting of aggregate function value (such as total-sales) computed by grouping three attributes part, supplier, and customer, p indicates a view composed of the corresponding aggregate function values calculated by grouping part alone, etc.

A data cube is created from a subset of attributes in the database. Specific attributes are chosen to be measure attributes, i.e., the attributes whose values are of interest. Another attributes are selected as dimensions or functional attributes. The measure attributes are aggregated according to the dimensions.

For example, XYZ may create a sales data warehouse to keep records of the store's sales for the dimensions time, item, branch, and location. These dimensions enable the store to keep track of things like monthly sales of items, and the branches and locations at which the items were sold. Each dimension may have a table identify with it, known as a dimensional table, which describes the dimensions. For example, a dimension table for items may contain the attributes item name, brand, and type.

Data cube method is an interesting technique with many applications. Data cubes could be sparse in many cases because not every cell in each dimension may have corresponding data in the database.

Techniques should be developed to handle sparse cubes efficiently. If a query contains constants at even lower levels than those provided in a data cube, it is not clear how to make the best use of the recomputed results stored in the data cube.

The model view data in the form of a data cube. OLAP tools are based on the multidimensional data model. Data cubes usually model n-dimensional data.

3.4 Association Rule Mining:

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a item set occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Association rules are "if-then" statements that help to show the probability of relationships between data items, within large data sets in various types of databases.

Association rule mining has a number of applications and is widely used to help discover sales correlations in transactional data or in medical data sets.

3.4.1 Use cases for association rules:

In data science, association rules are used to find correlations and co-occurrences between data sets. They are ideally used to explain patterns in data from seemingly independent information repositories, such as relational databases and transactional databases. The act of using association rules is sometimes referred to as "association rule mining" or "mining associations."

Below are a few real-world use cases for association rules:

- **Medicine.** Doctors can use association rules to help diagnose patients. There are many variables to consider when making a diagnosis, as many diseases share symptoms. By using association rules and machine learning-fuelled data analysis, doctors can determine the conditional probability of a given illness by comparing symptom relationships in the data from past cases. As new diagnoses get made, the machine learning model can adapt the rules to reflect the updated data.
- **Retail.** Retailers can collect data about purchasing patterns, recording purchase data as item barcodes are scanned by point-of-sale systems. Machine learning models can look for co-occurrence in this data to determine which products are most likely to be purchased together. The retailer can then adjust marketing and sales strategy to take advantage of this information.
- **User experience (UX) design.** Developers can collect data on how consumers use a website they create. They can then use associations in the data to optimize the website user interface -- by analysing where users tend to click and what maximizes the chance that they engage with a call to action, for example.
- **Entertainment.** Services like Netflix and Spotify can use association rules to fuel their content recommendation engines. Machine learning models analyse past user behaviour data for frequent patterns, develop association rules and use those rules to recommend content that a user is likely to engage with, or organize content in a way that is likely to put the most interesting content for a given user first.

⇨ How association rules work:

Association rule mining, at a basic level, involves the use of machine learning models to analyze data for patterns, or co-occurrences, in a database. It identifies frequent if-then associations, which themselves are the *association rules*.

An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent.

Association rules are created by searching data for frequent if-then patterns and using the criteria *support* and *confidence* to identify the most important relationships. Support is an indication of how frequently the items appear in the data. Confidence indicates the number of times the if-then statements are found true. A third metric, called *lift*, can be used to compare confidence with expected confidence, or how many times an if-then statement is expected to be found true.

Association rules are calculated from *itemsets*, which are made up of two or more items. If rules are built from analyzing all the possible itemsets, there could be so many rules that the rules hold little meaning. With that, association rules are typically created from rules well-represented in data.

⇨ **Measures of the effectiveness of association rules:**

The strength of a given association rule is measured by two main parameters: support and confidence. Support refers to how often a given rule appears in the database being mined. Confidence refers to the amount of times a given rule turns out to be true in practice. A rule may show a strong correlation in a data set because it appears very often but may occur far less when applied. This would be a case of high support, but low confidence.

Conversely, a rule might not particularly stand out in a data set, but continued analysis shows that it occurs very frequently. This would be a case of high confidence and low support. Using these measures helps analysts separate causation from correlation, and allows them to properly value a given rule.

A third value parameter, known as the lift value, is the ratio of confidence to support. If the lift value is a negative value, then there is a negative correlation between datapoints. If the value is positive, there is a positive correlation, and if the ratio equals 1, then there is no correlation.

⇨ **Association rule algorithms:**

Popular algorithms that use association rules include AIS, SETM, Apriori and variations of the latter.

With the AIS algorithm, itemsets are generated and counted as it scans the data. In transaction data, the AIS algorithm determines which large itemsets contained a transaction, and new candidate itemsets are created by extending the large itemsets with other items in the transaction data.

The SETM algorithm also generates candidate itemsets as it scans a database, but this algorithm accounts for the itemsets at the end of its scan. New candidate itemsets are generated the same way as with the AIS algorithm, but the transaction ID of the generating transaction is saved with the candidate itemset in a sequential data structure. At the end of the pass, the support count of candidate itemsets is created by aggregating the sequential structure. The downside of both the AIS and SETM algorithms is that each

one can generate and count many small candidate itemsets, according to published materials from Dr. Saed Sayad, author of *Real Time Data Mining*.

With the Apriori algorithm, candidate itemsets are generated using only the large itemsets of the previous pass. The large itemset of the previous pass is joined with itself to generate all itemsets with a size that's larger by one. Each generated itemset with a subset that is not large is then deleted. The remaining itemsets are the candidates. The Apriori algorithm considers any subset of a frequent itemset to also be a frequent itemset. With this approach, the algorithm reduces the number of candidates being considered by only exploring the itemsets whose support count is greater than the minimum support count, according to Sayad.

⇨ **Uses of association rules in data mining:**

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in customer analytics, market basket analysis, product clustering, catalog design and store layout.

Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

⇨ **Examples of association rules in data mining:**

A classic example of association rule mining refers to a relationship between diapers and beers. The example, which seems to be fictional, claims that men who go to a store to buy diapers are also likely to buy beer. Data that would point to that might look like this:

A supermarket has 200,000 customer transactions. About 4,000 transactions, or about 2% of the total number of transactions, include the purchase of diapers. About 5,500 transactions (2.75%) include the purchase of beer. Of those, about 3,500 transactions, 1.75%, include both the purchase of diapers and beer. Based on the percentages, that large number should be much lower. However, the fact that about 87.5% of diaper purchases include the purchase of beer indicates a link between diapers and beer.

3.5 Apriori Algorithm:

Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.

Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.

The rapid rise of e-commerce apps has increased the accumulation of data. To forecast outcomes, data mining, also known as KDD (Knowledge Discovery in Databases), is used to detect irregularities, linkages, trends and patterns in data.

An algorithm known as Apriori is a common one in data mining. It's used to identify the most frequently occurring elements and meaningful associations in a dataset. As an example, products brought in by consumers to a shop may all be used as inputs in this system.

An effective Market Basket Analysis is critical since it allows consumers to purchase their products with more convenience, resulting in a rise in market sales. Furthermore, it has been applied in healthcare to aid in the identification of harmful medication responses. A clustering algorithm is generated that identifies which combinations of drugs and patient factors are associated with adverse drug reactions.

⇨ What is Apriori Algorithm?

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Using breadth-first search and a Hash tree structure, apriori counts candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

➢ Apriori Property:

An essential feature known as the Apriori property is utilized to boost the effectiveness of level-wise production of frequent itemsets. This property helps by minimizing the search area, which in turn serves to maximize the productivity of level-wise creation of frequent patterns.

⇨ How to organize your data for the Apriori algorithm?

Let's start at the beginning: you have a data set in which customers are buying multiple products. Your goal is to find out which combinations of products are frequently bought together.

You need to organize the data in such a way that you have a set of products on each line. Each of those sets contains products that were bought in the same transaction.

The most basic solution would be to loop through all the transactions and inside the transactions loop through all the combinations of products and count them. Unfortunately this is going to take way too much time, so we need something better.

Two scientists Agrawal and Srikant were the first to propose a solution to this in their 1994 paper called Fast Algorithms for Mining Association Rules. Their first solution is the famous Apriori algorithm.

⇨ How Does the Apriori Algorithm Work?

The Apriori algorithm operates on a straightforward premise. When the support value of an item set exceeds a certain threshold, it is considered a frequent item set. Take into account the following steps. To begin, set the support criterion, meaning that only those things that have more than the support criterion are considered relevant.

Step 1: Create a list of all the elements that appear in every transaction and create a frequency table.

Step 2: Set the minimum level of support. Only those elements whose support exceeds or equals the threshold support are significant.

Step 3: All potential pairings of important elements must be made, bearing in mind that AB and BA are interchangeable.

Step 4: Tally the number of times each pair appears in a transaction.

Step 5: Only those sets of data that meet the criterion of support are significant.

Step 6: Now, suppose you want to find a set of three things that may be bought together. A rule, known as self-join, is needed to build a three-item set. The item pairings OP, OB, PB, and PM state that two combinations with the same initial letter are sought from these sets.

OPB is the result of OP and OB.

PBM is the result of PB and PM.

Step 7: When the threshold criterion is applied again, you'll get the significant itemset.

⇨ Steps for Apriori Algorithm:

The Apriori algorithm has the following steps:

Step 1: Determine the level of transactional database support and establish the minimal degree of assistance and dependability.

Step 2: Take all of the transaction's supports that are greater than the standard or chosen support value.

Step 3: Look for all rules with greater precision than the cutoff or baseline standard, in these subgroups.

Step 4: It is best to arrange the rules in ascending order of strength.

Advantages:

⇨ What are the advantages of the apriori algorithm?

The advantages of apriori are as follows:

- ✓ This is the most simple and easy-to-understand algorithm among association rule learning algorithms.
- ✓ The resulting rules are intuitive and easy to communicate to an end user.
- ✓ It doesn't require labeled data as it is fully unsupervised; as a result, you can use it in many different situations because unlabeled data is often more accessible.
- ✓ Many extensions were proposed for different use cases based on this implementation—for example, there are association learning algorithms that take into account the ordering of items, their number, and associated timestamps.
- ✓ The algorithm is exhaustive, so it finds all the rules with the specified support and confidence.

Disadvantages:

One of the biggest limitations of the Apriori Algorithm is that it is slow. This is so because of the bare decided by the:

- ☒ A large number of itemsets in the Apriori algorithm dataset.
- ☒ Low minimum support in the data set for the Apriori algorithm.
- ☒ The time needed to hold a large number of candidate-sets with many frequent itemsets.
- ☒ Thus it is inefficient when used with large volumes of datasets.

As an example, if we assume there is a frequent-1 itemset with 10^4 from the set. The Apriori algorithm code needs to generate greater than 10^7 candidates with a 2-length which will then be tested and collected as an accumulation. To detect a size frequent pattern of size 100 (having v1, v2... v100) the algorithm generates 2^100 possible itemsets or candidates which is an example of an application of the Apriori algorithm.

Hence, the yield costs escalate and a lot of time wasted in candidate generation aka time complexity of the Apriori algorithm. Also, in its attempts to an improved the Apriori algorithm to check the many candidate itemsets obtained from the many sets, it scans the database many times using expensive resources. This in turn impacts the algorithm when the system memory is insufficient and there are a large number of frequent transactions. That's why the algorithm becomes inefficient and slow with large databases.

Methods To Improve Apriori Efficiency:

Many methods are available for improving the efficiency of the algorithm.

1. **Hash-Based Technique:** This method uses a hash-based structure called a hash table for generating the k-itemsets and its corresponding count. It uses a hash function for generating the table.
2. **Transaction Reduction:** This method reduces the number of transactions scanning in iterations. The transactions which do not contain frequent items are marked or removed.
3. **Partitioning:** This method requires only two database scans to mine the frequent itemsets. It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database.
4. **Sampling:** This method picks a random sample S from Database D and then searches for frequent itemset in S. It may be possible to lose a global frequent itemset. This can be reduced by lowering the min_sup.
5. **Dynamic Itemset Counting:** This technique can add new candidate itemsets at any marked start point of the database during the scanning of the database.

➢ Applications of Apriori Algorithm:

Some fields where Apriori is used:

1. **In Education Field:** Extracting association rules in data mining of admitted students through characteristics and specialties.
2. **In the Medical field:** For example Analysis of the patient's database.
3. **In Forestry:** Analysis of probability and intensity of forest fire with the forest fire data.
4. Apriori is used by many companies like Amazon in the **Recommender System** and by Google for the auto-complete feature.

➢ Conclusion:

Apriori algorithm is an efficient algorithm that scans the database only once. It reduces the size of the itemsets in the database considerably providing a good performance. Thus, data mining helps consumers and industries better in the decision-making process.

☺ ☺ ☺ ☺ ☺

Exercises

❖ Answer the following Questions in brief.

1) Why Data Mining?
2) What is Data Mining?
3) Explain Stages of data processing.
4) Write a note on Technologies of data processing.
5) Discuss of Data Mining Applications.
6) Write a note on Data pre-processing methods.
7) How do you clean data?
8) What is Components of quality data?
9) Write a note on Benefits of data cleaning.
10) Discuss of Data reduction strategies in data mining.
11) What is Association Rule Mining?
12) How association rules work?
13) Uses of association rules in data mining
14) Explain Apriori Algorithm.
15) How to organize your data for the Apriori algorithm?
16) How Does the Apriori Algorithm Work?
17) What are the advantages of the apriori algorithm?
18) What are the disadvantages of the apriori algorithm?

❖ Indicate whether the following statements are true or false

1) Data processing is the conversion of data into suitable for use and wished format. (_____)

2) There are two data processing technologies. (_____)

3) Data mining is also used in the fields of credit card services and telecommunication to detect frauds. (_____)

4) There are three Data pre-processing methods. (_____)

5) Data transformation is one more important procedure on a fast track to receiving final data of good quality. (_____)

6) Data mining can be a very time-consuming process especially when there is a necessity to analyse huge volumes of data. (_____)

7) Data cleaning is one of the important parts of machine learning. (_____)

8) Data reduction techniques can be applied to obtain a reduces data should be more efficient yet produce the same analytical results. (_____)

9) A histogram is a type of graph that is widely used in mathematics, especially in statistics. (_____)

10) apriori algorithm is not simple and easy-to-understand algorithm among association rule learning algorithms. (_____)

Answer:

1. True	2. False	3. True	4. False	5. True
6. True	7. True	8. True	9. True	10. False

☺ ☺ ☺ ☺ ☺