Assignment No. 2

Q.1 A) Discuss typical requirements of clustering in data mining.

→ ① Scalability :→

We need highly scalable clustering algorithms to deal with large databases.

② Ability to deal with different kinds of attributes :→

Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical and binary data.

③ Discovery of clusters with attribute shape :→

The clustering algo. should be capable of detecting clusters of arbitary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.

④ High dimensionality :→

The clustering algo. should not be able to handle low-dimensional data but also the high dimensional space.

⑤ Ability to deal with noisy data :→

Databases contain noisy, missing or erroneous data. Some algo. are sensitive to such data and may leads to poor quality clusters.

⑥ Interpretability :→

The clustering result should be interpretable, comprehensible and usable.

(Ques.18) Explain K-means algorithm.

→ ① K-means clustering is an unsupervised learning algo., which groups the unlabeled dataset into different clusters.

② It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

③ The K-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data points to its closest K-center, Those data points which are near to the particular K-center, create a cluster.

④ working :→

Step 1: Select the number K to decide the no. of clusters.

Step 2: Select random K points or centroids.

Step 3: Assign each data point to their ~~closest~~ closest centroid, which will form the predefined K clusters.

Step 4: calculate the variance and place a new centroid of each cluster.

Step 5: Repeat the third step, which means reassign each

datapoint to the new closest centroid of each cluster.

step 6: If any reassignment occurs, then go to step 4 else go to FINISH

step 7: The model is ready.

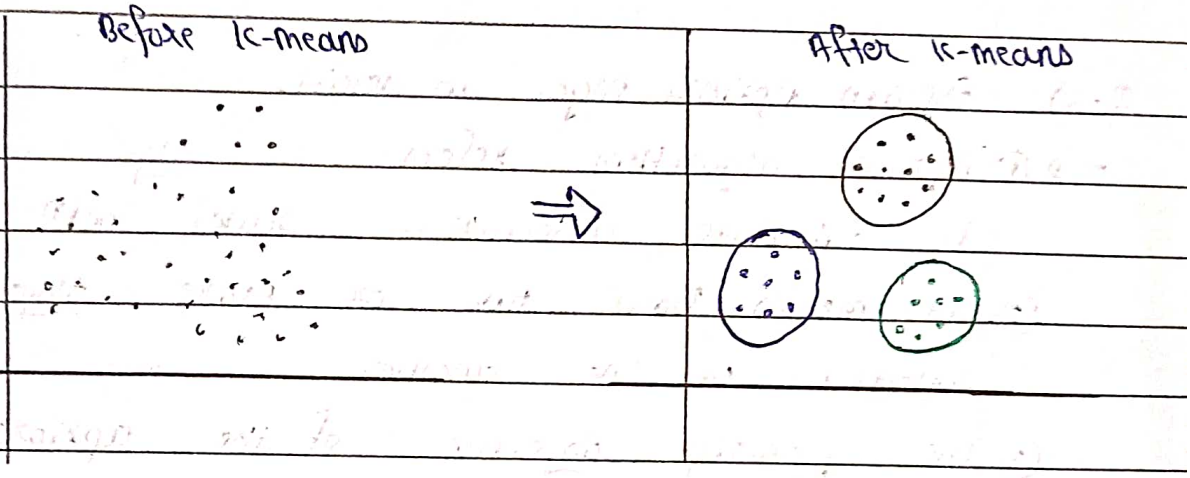

fig: clustering by K-means algo.

Ques.2) How FP growth algo. works? Explain.

→ ① The FP growth algo. is an alternative way to find frequent item sets without using candidate generations, thus improving performance.

② For so much, it uses a divide and conquer strategy.

③ The core of this method is the usage of a special data structure named frequent-pattern tree, which retain the item set / association information.

* This algo. works as follows:

- First, it compresses the input database creating

an FP tree instance to represent frequent item.

- After this, it divides the compressed db into a set of conditional db, each associated with one frequent pattern.
- Finally each such db is mined separately.

Q.3) Explain Apriori Algo. in detail.

→ (1) Apriori algorithm refers to algo. which used to calculate association rule betn objects.

(2) It means how two or more objects are related to one another.

(3) The primary objective of the apriori algo. is in to create the association rule between different objects. generally

(4) Generally the apriori algo. is operated on a db that consist of huge no. of transaction.

(5) This algo. refers to an algo. that is used in mining frequent products sets and relevant association rules.

(6) It helps the customers to buy their products with ease and increases the sales perfolmance of particular store.

(7) There are three component of Apriori algo., they are:

- Support :→

It refers to the default popularity of any product. you find the support as a quotient of the division of the no. of transaction comprising that

that product by the total number of transaction

support = Transactions relating biscuit / total transaction (Biscuit)

= 400 / 4000 = 10 percent.

- Confidence :→

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the no. of transactions that comprise both biscuit and chocolates by the total no. of transactions to get the confidence.

Confidence = Transactions relating both biscuit & chocolate / Total transactions involving biscuit

= 200 / 400

= 500 percent.

It means that 50 % of customers who bought biscuits, ~~bought~~ chocolates also.

- Lift :→

It refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equ:

Lift = (confidence (Biscuit - chocolates)) / support (Biscuit)

= 50 / 10

= 5

It means that the probability of people buying

both biscuit and chocolates together is five times more than that of purchasing the biscuit alone.

Q.4) Differentiate temporal and spatial data mining in detail.

→

	Spatial data mining	Temporal data mining
①	It requires space	It requires time.
②	It is the extraction of knowledge / spatial relationship and interesting measures that are not explicitly stored in spatial database.	It is extraction of knowledge about occurrence of an event whether they follow cyclic, Random, seasonal variations etc..
③	It deals with spatial (location, Geo-referenced) data.	It deals with implict or explict temporal content, from large quantities of data.
④	It includes finding characteristic rules, discriminant rules, association rules and evaluation rules etc..	It aims at mining new and unknown knowledge, which takes into account the temporal aspect of data.
⑤	It is the method of identifying unusual and	It deals with useful knowledge from temporal data

	un-explored data but useful models from spatial databases.	
(6)	Example:- Determining hotspots, unusual locations.	Example:- An association rule: "Any person who buys a car also buys steering lock". Temporal aspect: "Any person who buys a car also buys a steering lock after that."

Q.5) Discuss the following:

i) web content mining :→

- web content mining is referred to as text mining.
- Content mining is the browsing and mining of text images and graphs of a web pages to decide the relevance of the content to the search query.
- It can be defined as the phase of extracting essential data from standard language text.
- Some data that it can generate via text messages, files, emails, documents are written in common language text.
- Text mining can draw beneficial insights or patterns from such data.

ii) web usage mining :→

- It is used to derive useful data, information, knowledge from the weblog data and helps in identifying the user access designs for web pages.
- The management of web resources the individual is thinking about data of request of visitors of a website that are composed as web server logs.
- web usage mining can disclose relationship that were not suggested by the designer of the pages.

iii) web structure mining :→

- It is tool that can recognize the relationship bet^n web pages linked by data or direct link connection.
- This structured data is discoverable by the provision of web structure schema through db technique for web pages.
- web mining can widely be viewed as the application of adapted data mining method to the web, whereas data mining is represented as the application of algo. to find patterns on mostly structure data fixed into a knowledge discovery process.
- Structure mining uses minimize two problem that first problem is irrelevant to search outcome and second is the inability to index the large amount of data supported on the web.