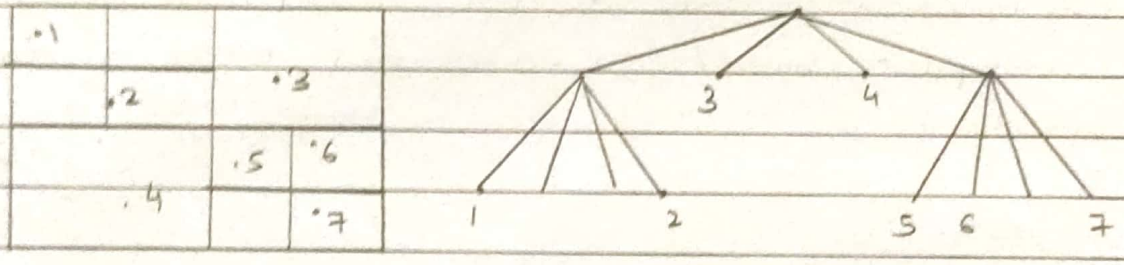
Name: Ayush Jain
SAP ID: 60004200132
Div: B2
Computer Engineering

DMW - Assignment 4

Q. 1) Explain the different data structures in Spatial mining.

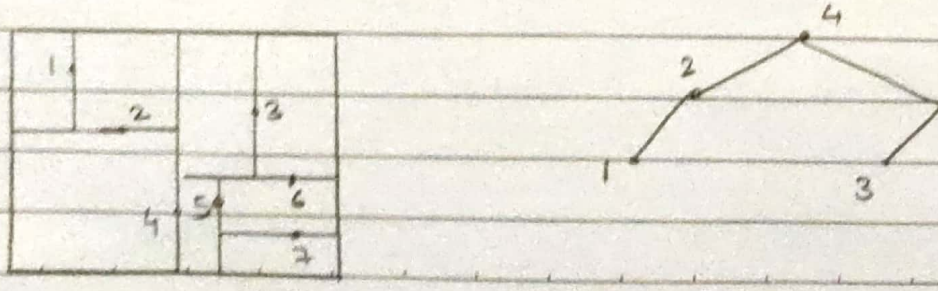
→ In spatial data mining, analyst use geo-spatial information to produce buisness intelligence or other similar results. This require specific technique and resources to get geographical data into relevant and useful format. Special data structures used in such cases are described below:

1) QUAD TREE: It is used to index 2-D space. Each internal node of the tree splits the space into NW, SW and SE regions according to the axes. Each subspace is recursively split until there is at most 1 object inside each of them.



The quad tree is not balanced as its balance depends on the data distribution and order of inserting points.

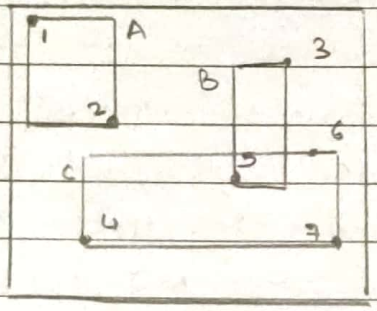
2) K-d Trees: This method uses a binary tree to split k-dimensional space. This tree splits the space into 2 subspaces according to one of the coordinates of the splitting point.

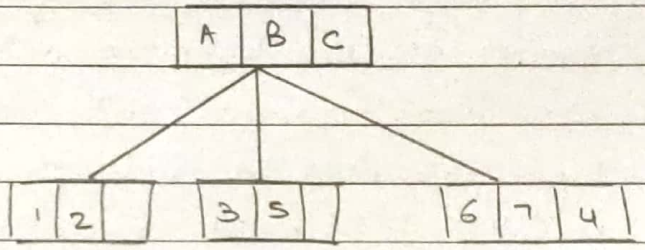


FOR EDUCATIONAL USE

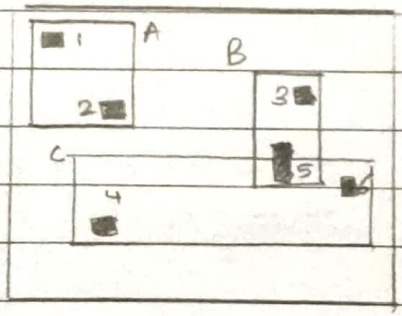
Scanned with CamScanner

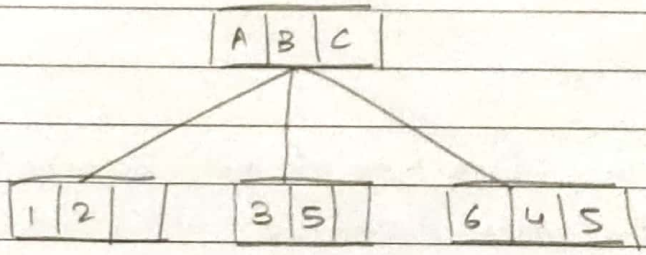
(3) R- TREE : It is a B- tree modified for spatial data. Its structure is balanced and it splits the space into rectangles that can overlap. If M is max entries in 1 node and m is minimum, then each node except root has $2 < m < M/2$ children





(4) R^+ TREE : It is an extension of R- tree. The bounding box of nodes at a level do not overlap. The increases space consumption, but the zero overlap makes it faster in practice.





Sundaram
FOR EDUCATIONAL USE

Scanned with CamScanner

Q.7) How is spatial clustering different from regular clustering technique? Explain the CLARANS algorithm.

→ a) Spatial clusters can be described as a geographically bound group of occurrences of sufficient size and concentration to be unlikely to have occured by chance.

b) Spatial cluster analysis is carried out on raw variables - notes when there is no apriori hypothesis regarding the process and is a density based clustering method.

c) CLARANS stands for Clustering Large Application based on Randomized Search, and is a partitioning method used in clustering. It is an extension of k-medoid that uses random samples of input data and computes the best medoids.

d) The algorithm can be defined as following steps:

- Select 'k' random points as the initial medoids.
- Select random point 'b' from 'k' and 'b' not in 'k'.
- If $\sum_{Vx} dist(b,x) < \sum_{Vx} dist(a,x)$ then replace a by b.
- The algorithm performs this randomized search 'n' number of times, after which we arrive at a locally optimal set of medoids.

e) The process of examining the points for possible replacement is repeated till the number of replacements does not exceed the maximum number of neighbours to be examined.

FOR EDUCATIONAL USE

Scanned with CamScanner

Q. 3) What are Crawlers? Explain different types of Crawlers?

→ 1) Crawlers or spiders are programs that traverse the structures of the web.

2) A Crawler starts at some speed URL and traverse multiple links while saving the indices and storing the outgoing links in the queue.

3) The information that they extracts and store helps in improving results of complex requests in search engines.

4) The various types of crawlers are:

(a) Traditional Crawler: Visits the entire web and replaces the index entirely.

(b) Periodic Crawlers: visits a portion of the web and updates a subset of index.

(c) Incremental Crawlers: only visits links from a page if the page is determined to be relevant by a classifier. These crawlers are made up of:

(a) Classifiers: To determine relevance based on a specific topic.

(b) Distiller: To identify hub pages that contain links to other relevant pages.

Q. 4) Explain the Page Rank Algorithm in web structure mining.

→ Page Rank is a web structure mining algorithm developed by Larry Page. It is way of measuring the importance of a website by counting the number of and quality of links coming into the website. The underlying assumption is that a page is only as important as the pages that link to it.

Sundaram
FOR EDUCATIONAL USE

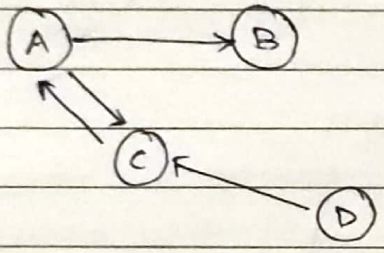
Scanned with CamScanner

The formula for Page Rank of a Page 'A' can be given as:

$$PR(A) = (1-c) + c \sum_{\forall v \to A} \frac{PR(u)}{d(u)}$$

where 'c' is the damping factor and $d(v)$ is the number of outgoing links from 'u'.

Given Graph:



Assuming $c = 0.85$ (damping factor)

$PR(A) = 0.15 + 0.85\,[PR(C)]$

$PR(B) = 0.15 + 0.85\,[PR(A)/2]$

$PR(C) = 0.15 + 0.85\,[PR(A) + PR(D)]$

$PR(D) = 0.15 + 0.85\,(0)^2$

On solving the above simultaneous equation we get

$PR(A) = 1.4901$

$PR(B) = 0.7832$

$PR(C) = 1.5765$

$PR(D) = 0.15$

Sundaram
FOR EDUCATIONAL USE

Scanned with CamScanner

Q. 5) Explain the web usuage mining process in brief.

→ 1) Web usuage mining refers to the process of mining of web usuage data or logs. Web logs is the information of all access activities that occur on a web page and is called click stream data.

2) Click stream data, from the client's perspective, is its sequence of clicks along with information of user.

3) The process of web usuage mining can be broken down into:

① Preprocessing Web log:
- Clean and remove extraneous information.
- Sessionize data or split into multiple sets of pages visited within a logical timeframe.

② Pattern Discover
- Patter is sequence of page visits in a session.
- Count the pattern that occurs in sessions.

③ Pattern Analysis

Due to security, Privacy and legal issues, we also replace any identifiable attributes in the logs with unique values during the cleaning phase.

The applications of web usuage mining include - Personalization, Improvement of site's web structure, aid in caching, improved design and improved effectiveness of e-commerce sites.

Sundaram
FOR EDUCATIONAL USE

Scanned with CamScanner