Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

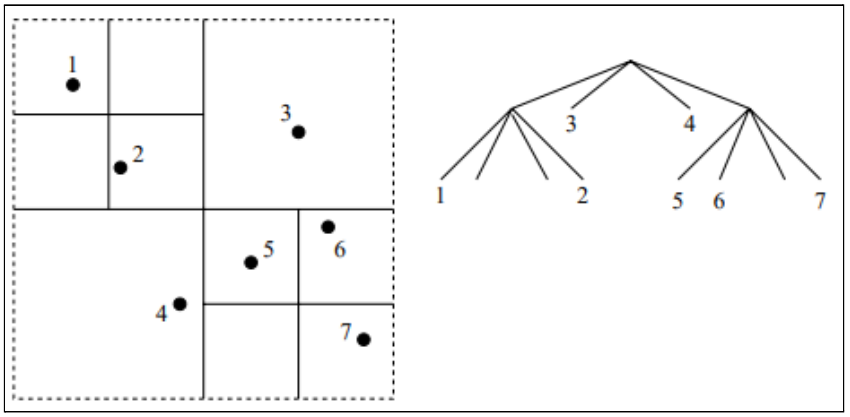
JUNAID GIRKAR
60004190057
TE COMPS A4

ASSIGNMENT 5

Q1. Explain the different data structures used in Spatial Mining.

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

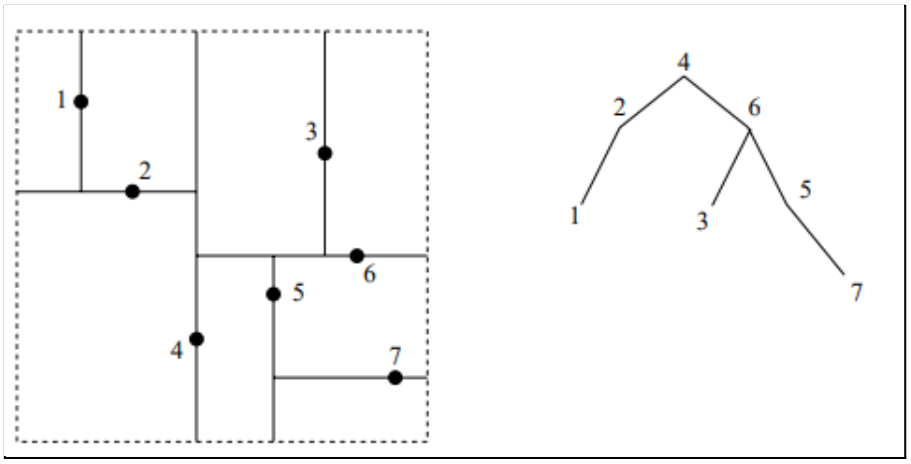
1) **QUAD TREE :** It is used to index 2-D space. Each internal node of the tree splits the space into NW, NW, SW and SE regions according to the axes. Each subspace is recursively split until there is at most 1 object inside each of them



QUAD TREE

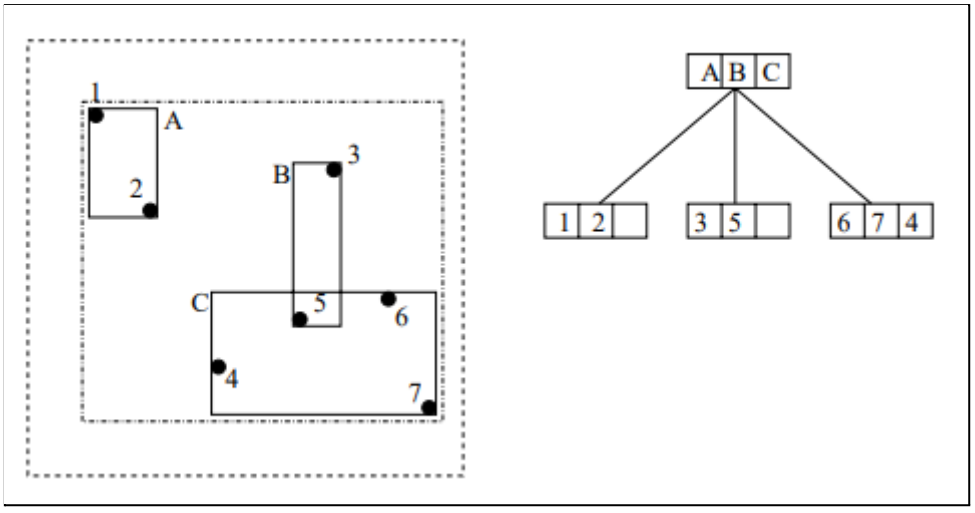
The Quad tree is not balanced as its balance depends on the data distribution and order of inserting points

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. Let L (node) be the length of path from root to the node and suppose axes are numbered 0 to k-1. At L (node), in every node the space is split according to the coordinate number



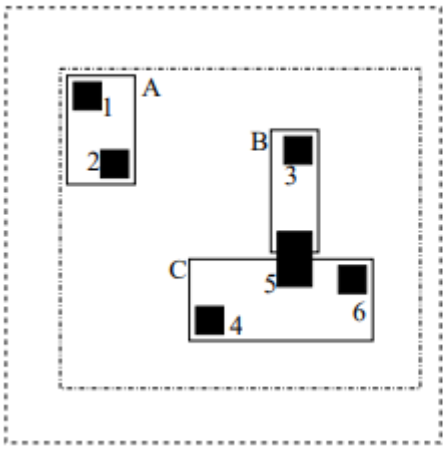
k - d TREE

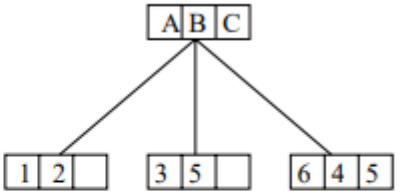
3) **R-TREE :** It is a B-tree modified for spatial data. Its structure is balanced and it splits the space into rectangles that can overlap. If M is the max entries in 1 node and m is minimum, then each node except root has 2 < m < M/2 children



R - TREE

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





Q2. How is spatial clustering different from regular clustering technique? Explain the CLARANS algorithm.

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

Spatial cluster analysis is carried out on raw variables - rates when there is no apriori hypothesis regarding the process and is a density based clustering ,ethod.

On the other hand, clustering techniques such as k-means and k-medoid are centroid based, and are sensitive to outliers.

CLARANS stands for **C**lustering **L**arge **A**pplication based on **Ran**domized **S**earch, 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 medioids. It maintains a balance between computational cost and influence of sampling on cluster formation.

The algorithm can be defined as the following steps:-

- Select 'k' random points as the initial medioids.
- Select random point 'a' 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 medoid search 'n' number of times, after which we arrive at a locally optimal set of medoids

- 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 (parameter)

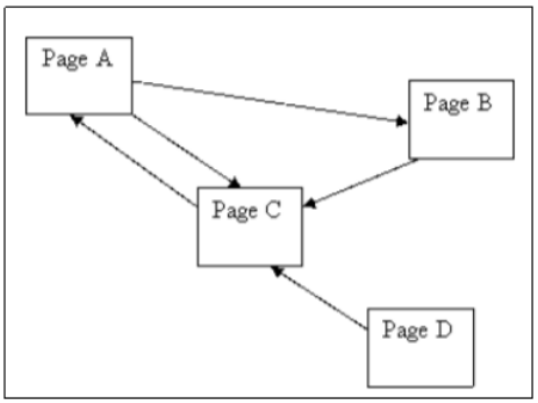
Q3. What are Crawlers? Explain the different types of crawlers.

ANS. Crawlers or Spiders are programs that traverse the structures of the web. A crawler starts at some seed URL and traverses multiple links while saving the indices and storing the outgoing links in a queue. The information that they extract and store helps in improving results of complex requests in search engines.

The various types of crawlers are :-

1) Traditional Crawler : Visits the entire web and replaces the index entirely
2) Periodic Crawler : Visits a portion of the web and updates a subset of the index
3) Incremental Crawler : Only visits links from a page if the page is determined to be relevant by a classifier. These crawlers are made up of :-
    a) **Classifier:** To determine relevance based on a specified topic
    b) **Distiller:** To identify hub pages that contain links to other relevant pages.

Q4. Explain the PageRank algorithm in web structure mining. Calculate the PageRank for the following graph

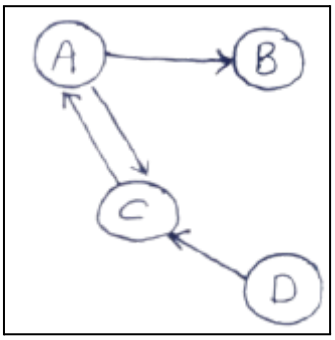


ANS. Page Rank is a web structure mining algorithm developed by Larry Page. It is a way of measuring the importance of a website by counting the number and quality of links coming into the website. The underlying assumption is that a page is only as important as the pages that link to it. The formula 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 (u) 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 equations, we get:

PR (A) = 1.4901
PR (B) = 0.7832
PR (C) = 1.5765
PR (D) = 0.15

Q5. Explain the web usage mining process in brief
ANS. Web usage mining refers to the process of mining of web usage data or logs. Web logs is the information of all access activities that occur on a web page and is called click stream data.

Click Stream data, from the client's perspective, is it's sequence of clicks along with information of the user. From the servers perspective, it is information about services the site used to improve design.

The process of web usage 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 (1 session)
2) **Pattern Discover**
    - Pattern is sequence of page visits in a session
    - The process of pattern discovery is similar to discovering association rules
    - Count the patterns that occur in sessions
3) **Pattern Analysis**

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

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