

1 → Spatial Index Structures.

R tree are hierarchical tree structures designed for indexing spatial data. They partition the space into rectangles with each node in the tree representing a boundary box that contains its child nodes.

Quad tree recursively divides a two dimensional space into quadrants until each quadrant contains a limited number of data points.

KD tree is a multidimensional binary tree that recursively partitions spaces along axes.

→ Graph Structures.

Spatial network graph represent spatial entities (nodes) and their relationships in a graph. Graph structures are valuable for modeling spatial relationships.

→ Spatial data structures

Spatial indexing is a technique that discretizes space into a grid and each cell of the grid holds a list of spatial objects that fall within that cell. This structure is useful for fast point in region queries and spatial joins.

→ Geographical Information System (GIS)

Raster data structure organizes spatial information as a grid of cell. Each cell contains attribute data, and raster structures are commonly used in GIS for representing continuous spatial phenomena such as elevation or temperature.

2 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 by how the variable rates when there is no density based clustering method.

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

The algorithm can be defined by the following steps.

- select k random points as the initial medoids.
- select random point a from L and b not in k .
- If $\sum_{i=1}^n \text{dis}(b, x_i) < \sum_{i=1}^n \text{dis}(a, x_i)$ then replace ' a ' by ' b '

→ The algorithm performs this randomized medoids search n number of times after which we arrive as a locally optimal set of medoids.

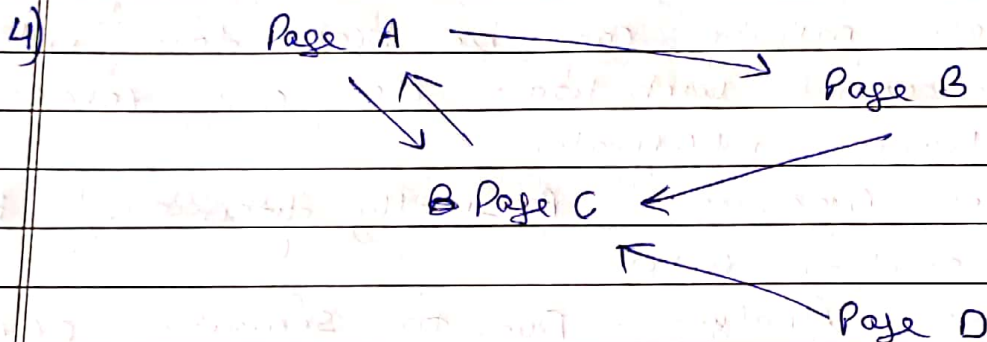
→ The process of examining the points for possible replacement is repeated till the number of replacements does not exceed the maximum number of neighbours to be examined.

3 Crawlers or spiders are programs that traverse the structure of the web. A crawler starts at some seed URL and traverse multiple links while solving the indices and storing the outgoing links in a queue. The information that they extract and store helps in improving results of the complex request in search engine.

The various types of crawlers are :-

- 1) Traditional crawlers.
- 2) Periodic crawlers.
- 3) Incremental crawlers.

These crawlers are made up of \swarrow classifier
distiller.



Page Rank is a web structure mining algorithm developed by Larry Page. It is the 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.

$$PR(A) = (1 - C) + C \sum_{u \rightarrow A} \frac{PR(u)}{d(u)}$$

where C is the damping factor & $d(u)$ is the number of outgoing links from (u)

Assuming $C = 0.85$

$$PR(A) = 0.15 + 0.85 [PR(C)] = 1.4901$$

$$PR(B) = 0.15 + 0.85 [PR(A)/2] = 0.7832$$

$$PR(C) = 0.15 + 0.85 [PR(A) + PR(D)] = 1.57$$

$$PR(D) = 0.15 + 0.85 [0]^2 = 0.15$$

- 5) web usage mining refers to the process of mining of web usage data on logs - web logs is the information of all access activities that occur on a web page is called click stream data of clicks along with information of the users. From the servers perspective, it is the information about services the site used to improve the design. The process of web page mining can be broken down into -
- 1) Preprocessing web log - clean and remove extraneous information
 - 2) Pattern Discovery - Basically discovering the association rules.
 - 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.

Eg. E-commerce sites (advertising).