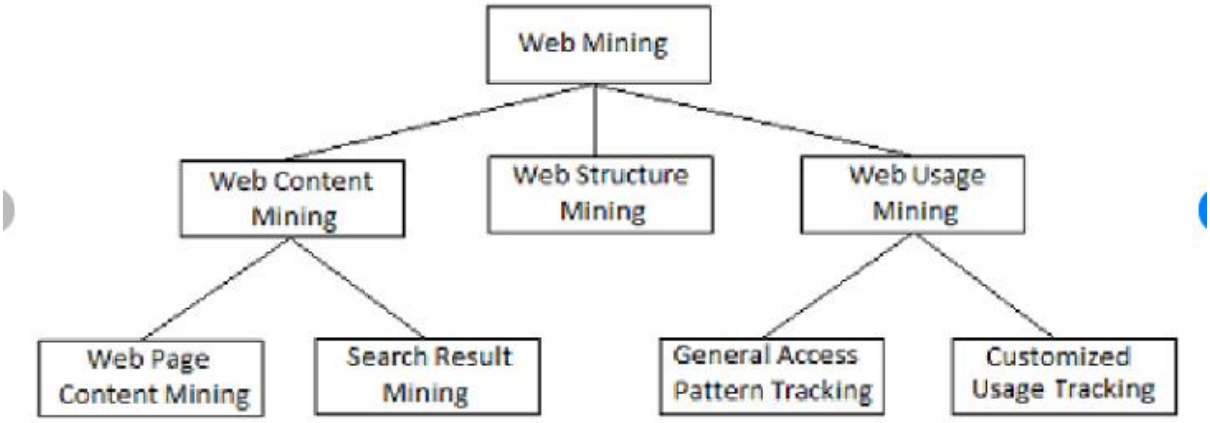
Module 5

1. Briefly describe about web mining taxonomy.



Web data mining is a sub discipline of data mining which mainly deals with web. Web data mining is divided into three different types: web structure, web content and web usage mining. All these types use different techniques, tools, approaches, algorithms for discover information from huge bulks of data over the web.

2. Distinguish between focused crawling and regular crawling.

A robot (spider/crawler) is a program that traverses the hypertext structure in the Web. The page (or set of pages) that the crawler starts with are referred to as the seed URLs. By starting at one page, all links from it are recorded and saved in a queue. These new pages are in turn searched and their links are saved. A crawler may visit a certain number of pages and then stop, build an index, and replace the existing index. Crawlers are used to facilitate the creation of indices used by search engines. Traditional crawlers usually replace the entire index or a section thereof. Because of the tremendous size of the Web, it has also been proposed that a focused crawler be used. A focused crawler visits pages related to topics of interest.

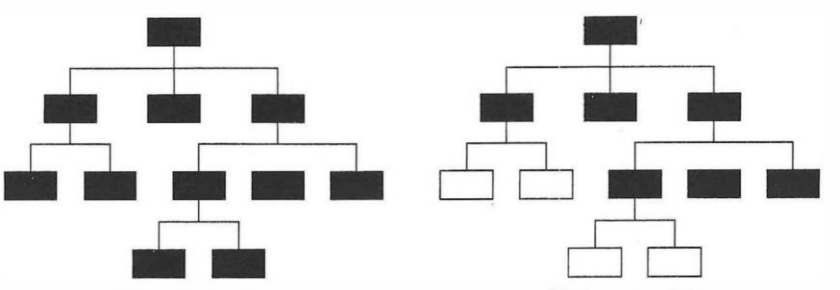


FIGURE 7.3: Focused crawling.

3. Write a CLEVER algorithm for web structure mining.

Clever, is aimed at finding both authoritative pages and hubs. The authors define an authority as the "best source" for the requested information. In addition, a hub is a page that contains links to authoritative pages. The Clever system identifies authoritative pages and hub pages by creating weights. Hyperlink-induced topic search (HITS) finds hubs and authoritative pages

The HITS technique contains two components:-

a) Based on a given set of keywords, a set of relevant pages is found.
b) Hub and authority measures are associated with these pages. Pages with the highest values are returned.

Input:

```
W        //WWW viewed as a directed graph
q        //Query
s        //Support
```

Output:

```
A        //Set of authority pages
H        //Set of hub pages
```

HITS algorithm

$R = SE(W, q)$

$B = R \cup$ (pages linked to from R) $\cup$ (pages that link to pages in R);

$G(B, L) =$ Subgraph of W induced by B;

$G(B, L^1) =$ Delete links in G within same site;

$x_p = \sum_q$ where $(q,p) \in L^1$ y_q; // Find authority weights;

$y_p = \sum_q$ where $(p,q) \in L^1$ x_q; // Find hub weights;

$A = \{p \mid p$ has one of the highest $x_p\}$;

$H = \{p \mid p$ has one of the highest $y_p\}$;

4. Describe different Text retrieval methods.

   Information retrieval (IR) is a field, which has focused on query and transaction processing of structured data, and retrieval of information from a large number of text-based documents.

   Various text retrieval methods are

   a. Document selection methods, the query is regarded as specifying constraints for selecting relevant documents. A typical method of this category is the Boolean retrieval model, in which a document is represented by a set of keywords and a user provides a Boolean expression of keywords, such as “car and repair shops,” “tea or coffee,” or “database systems but not Oracle.”
   b. Document ranking methods use the query to rank all documents in the order of relevance. For ordinary users and exploratory queries, these methods are more appropriate than document selection methods.