**Course: Advanced Bioinformatics**

**Module title: Search Engines**

**Module no. : 173**

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

How web search engines work?

A search engine maintains the following processes in near real time:

* Web crawling
* Indexing
* Searching

Web search engines get their information by web crawling from site to site. The "spider" checks for the standard filename robots.txt, addressed to it, before sending certain information back to be indexed depending on many factors, such as the titles, page content, headings, as evidenced by the standard HTML markup of the informational content, or its metadata in HTML meta tags).

Indexing means associating words and other definable tokens found on web pages to their domain names and HTML-based fields. The associations are made in a public database, made available for web search queries. A query from a user can be a single word. The index helps find information relating to the query as quickly as possible.

Some of the techniques for indexing and caching are trade secrets, whereas web crawling is a straightforward process of visiting all sites on a systematic basis.

Between visits by the spider, the cached version of page (some or all the content needed to render it) stored in the search engine working memory is quickly sent to an inquirer. If a visit is overdue, the search engine can just act as a web proxy instead. In this case the page may differ from the search terms indexed. The cached page holds the appearance of the version whose words were indexed, so a cached version of a page can be useful to the web site when the actual page has been lost, but this problem is also considered a mild form of linkrot.

High-level architecture of a standard Web crawler

Typically when a user enters a query into a search engine it is a few keywords. The index already has the names of the sites containing the keywords, and these are instantly obtained from the index. The real processing load is in generating the web pages that are the search results list: Every page in the entire list must be weighted according to information in the indexes. Then the top search result item requires the lookup, reconstruction, and markup of the snippets showing the context of the keywords matched. These are only part of the processing each search results web page requires, and further pages (next to the top) require more of this post processing.

Beyond simple keyword lookups, search engines offer their own GUI- or command-driven operators and search parameters to refine the search results. These provide the necessary controls for the user engaged in the feedback loop users create by filtering and weighting while refining the search results, given the initial pages of the first search results. For example from 2007 the Google.com search engine has allowed one to filter by date by clicking "Show search tools" in the leftmost column of the initial search results page, and then selecting the desired date range. It's also possible to weight by date because each page has a modification time. Most search engines support the use of the Boolean operators AND, OR and NOT to help end users refine the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This first form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by advertising revenue and thus some of them allow advertisers to have their listings ranked higher in search results for a fee. Search engines that do not accept money for their search results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.