Search Engines

Search engines are a program that searches for and identifies items in a database that correspond to keywords or characters specified by the user, used especially for finding particular sites on the World Wide Web. Search engines utilize automated software applications (referred to as robots, bots, or spiders) that travel along the Web, following links from page to page, site to site. The information gathered by the spiders is used to create a searchable index of the Web.

Every search engine uses different complex mathematical formulas to generate search results. The results for a specific query are then displayed on the Search Engine Results Page. Search engine algorithms take the key elements of a web page, including the page title, content and keyword density, and come up with a ranking for where to place the results on the pages. Each search engine's algorithm is unique, so a top ranking on Yahoo! does not guarantee a prominent ranking on Google, and vice versa. To make things more complicated, the algorithms used by search engines are not only closely guarded secrets, they are also constantly undergoing modification and revision. This means that the criteria to best optimize a site with must be surmised through observation, as well as trial and error — and not just once, but continuously.

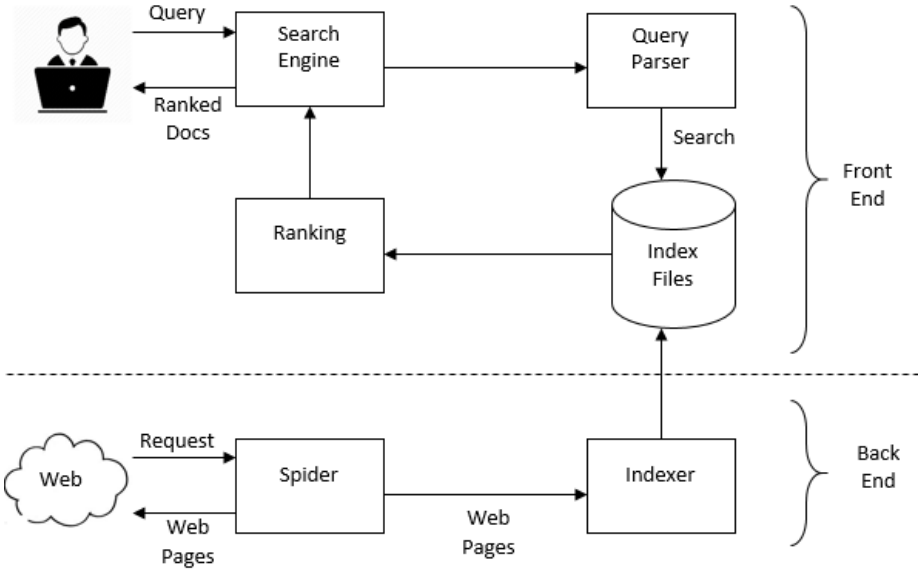

Figure 1 Architecture of a Search Engine

When the index is ready the searching can be perform through query interface, a user enters a query into a search engine (typically by using keywords). The application then parses the search request into a form that is consistent to the index. The engine examines its index and provides a

listing of best matching Web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. In this stage the results ranked, where ranking is a relationship between a set of items such that, for any two items, the first is either "ranked higher than", "ranked lower than" or "ranked equal" to the second. It is not necessarily a total order of documents because two different documents can have the same ranking. Ranking is done according to document relevancy to the query, freshness and popularity among many other factors.

Characteristics of Search Engine

i. Unedited

- Anyone can enter content Quality issues; Spam

ii. Varied information types

- Phone book, brochures, catalogs, dissertations, news reports, weather, all in one place!

iii. Different kinds of users

- Lexis-Nexis: Paying, professional searchers
- Online catalogs: Scholars searching scholarly literature
- Web: Every type of person with every type of goal

iv. Scale

- Hundreds of millions of searches/day; billions of docs

Functions of Search Engine

i. Crawling

The crawler, or web spider, is a vital software component of the search engine. It essentially sorts through the Internet to find website addresses and the contents of a website for storage in the search engine database. Crawling can scan brand new information on the Internet or it can locate older data. Crawlers have the ability to search a wide range of websites at the same time and collect large amounts of information simultaneously. This allows the search engine to find current content

on an hourly basis. The web spider crawls until it cannot find any more information within a site, such as further hyperlinks to internal or external pages.

ii. Indexing

Once the search engine has crawled the contents of the Internet, it indexes that content based on the occurrence of keyword phrases in each individual website. This allows a particular search query and subject to be found easily. Keyword phrases are the particular group of words used by an individual to search a particular topic.

The indexing function of a search engine first excludes any unnecessary and common articles such as "the," "a" and "an." After eliminating common text, it stores the content in an organized way for quick and easy access. Search engine designers develop algorithms for searching the web according to specific keywords and keyword phrases. Those algorithms match user-generated keywords and keyword phrases to content found within a particular website, using the index.

iii. Storage

Storing web content within the database of the search engine is essential for fast and easy searching. The amount of content available to the user is dependent on the amount of storage space available. Larger search engines like Google and Yahoo are able to store amounts of data ranging in the terabytes, offering a larger source of information available for the user.

iv. Results

Results are the hyperlinks to websites that show up in the search engine page when a certain keyword or phrase is queried. When you type in a search term, the crawler runs through the index and matches what you typed with other keywords. Algorithms created by the search engine designers are used to provide the most relevant data first. Each search engine has its own set of algorithms and therefore returns different results.

Parts of Search Engine

i. Spider, Crawler or robot

Web crawlers, also known as web spiders or internet bots, are programs that browse the web in an automated manner for the purpose of indexing content.

Crawlers can look at all sorts of data such as content, links on a page, broken links, sitemaps, and HTML code validation.

Search engines like Google, Bing, and Yahoo use crawlers to properly index downloaded pages so that users can find them faster and more efficiently when they are searching. Without crawlers there would be nothing to tell them that your website has new and fresh content. Sitemaps also can play a part in that process. So web crawlers, for the most part, are a good thing. However there are also issues sometimes when it comes to scheduling and load as a crawler might be constantly polling your site. And this is where a robots.txt file comes into play. This file can help control the crawl traffic and ensure that it doesn't overwhelm your server.

Web crawlers identify themselves to a web server by using the User-agent field in an HTTP request, and each crawler has their own unique identifier. Most of the time you will need to examine your web server referrer logs to view web crawler traffic.

ii. Index, Catalog or Database

Search engine indexing is the process of a search engine collecting, parses and stores data for use by the search engine. The actual search engine index is the place where all the data the search engine has collected is stored. It is the search engine index that provides the results for search queries, and pages that are stored within the search engine index that appear on the search engine results page. Without a search engine index, the search engine would take considerable amounts of time and effort each time a search query was initiated, as the search engine would have to search not only every web page or piece of data that has to do with the particular keyword used in the search query, but every other piece of information it has access to, to ensure that it is not missing something that has something to do with the particular keyword. Search engine spiders, also called search engine crawlers, are how the search engine index gets its information, as well as keeping it up to date and free of spam.

iii. Search Engine software

Search engine software is a program to traverse through index to find [age with matching keyword, phrase or any other form of input provided by user. Search engines provide an interface to a group of items that enables users to specify criteria about an item of interest and have the engine find the matching items. The criteria are referred to as a search query.

Types of Search Engine

There are basically three types of search engines: Those that are powered by robots (called crawlers; ants or spiders) and those that are powered by human submissions; and those that are a hybrid of the two.

i. Crawler Based Search Engine

Crawler-based search engines are those that use automated software agents (called crawlers) that visit a Web site, read the information on the actual site, read the site's meta tags and also follow the links that the site connects to performing indexing on all linked Web sites as well. The crawler returns all that information back to a central depository, where the data is indexed. The crawler will periodically return to the sites to check for any information that has changed. The frequency with which this happens is determined by the administrators of the search engine.

There are four basic steps, every crawler based search engines follow before displaying any sites in the search results.

- **Crawling**
- **Indexing**
- **Calculating Relevancy**

  Search engine compares the search string in the search request with the indexed pages from the database. Since it is likely that more than one page contains the search string, search engine starts calculating the relevancy of each of the pages in its index with the search string.

  There are various algorithms to calculate relevancy. Each of these algorithms has different relative weights for common factors like keyword density, links, or meta tags. That is why different search engines give different search results pages for the same search string. It is a known fact that all major search engines periodically change their algorithms. If you want to keep your site at the top, you also need to adapt your pages to the latest changes. This is one reason to devote permanent efforts to SEO, if you like to be at the top.

- **Retrieving Results**

The last step in search engines' activity is retrieving the results. Basically, it is simply displaying them in the browser in an order. Search engines sort the endless pages of search results in the order of most relevant to the least relevant sites.

Examples: Most of the popular search engines are crawler based search engines and use the above technology to display search results. Example of crawler based search engines are: Google, Bing, Yahoo!, Baidu, Yandex

Besides these popular search engines there are many other crawler based search engines available like DuckDuckGo, AOL and Ask.

ii. Human-powered search engines

Human-powered search engines rely on humans to submit information that is subsequently indexed and catalogued. Only information that is submitted is put into the index. Below is how the indexing in human powered directories work:

- Site owner submits a short description of the site to the directory along with category it is to be listed.
- Submitted site is then manually reviewed and added in the appropriate category or rejected for listing.
- Keywords entered in a search box will be matched with the description of the sites. This means the changes made to the content of a web pages are not taken into consideration as it is only the description that matters.
- A good site with good content is more likely to be reviewed for free compared to a site with poor content.

Yahoo! Directory and DMOZ were perfect examples of human powered directories. Unfortunately, automated search engines like Google, wiped out all those human powered directory style search engines out of the web.

In both cases, when you query a search engine to locate information, you're actually searching through the index that the search engine has created —you are not actually searching the Web. These indices are giant databases of information that is collected and stored and subsequently searched. This explains why sometimes a search on a commercial search engine, such as Yahoo!

or Google, will return results that are, in fact, dead links. Since the search results are based on the index, if the index hasn't been updated since a Web page became invalid the search engine treats the page as still an active link even though it no longer is. It will remain that way until the index is updated.

So why will the same search on different search engines produce different results? Part of the answer to that question is because not all indices are going to be exactly the same. It depends on what the spiders find or what the humans submitted. But more important, not every search engine uses the same algorithm to search through the indices. The algorithm is what the search engines use to determine the relevance of the information in the index to what the user is searching for.

One of the elements that a search engine algorithm scans for is the frequency and location of keywords on a Web page. Those with higher frequency are typically considered more relevant.

Another common element that algorithms analyze is the way that pages link to other pages in the Web. By analyzing how pages link to each other, an engine can both determine what a page is about (if the keywords of the linked pages are similar to the keywords on the original page) and whether that page is considered "important" and deserving of a boost in ranking.

iii. Hybrid Search Engines

Hybrid Search Engines use both crawler based and manual indexing for listing the sites in search results. Most of the crawler based search engines like Google basically uses crawlers as a primary mechanism and human powered directories as secondary mechanism. For example, Google may take the description of a webpage from human powered directories and show in the search results. As human powered directories are disappearing, hybrid types are becoming more and more crawler based search engines.

But still there are manual filtering of search result happens to remove the copied and spam sites. When a site is being identified for spam activities, the website owner needs to take corrective action and resubmit the site to search engines. The experts do manual review of the submitted site before including it again in the search results. In this manner though the crawlers control the processes, the control is manual to monitor and show the search results naturally.

iv. Other Types of Search Engines

Besides the above three major types, search engines can be classified into many other categories depending upon the usage. Below are some of the examples:

Search engines have different types of bots for exclusively displaying images, videos, news, products and local listings. For example, Google News page can be used to search only news from different newspapers.

Some of the search engines like Dogpile collects meta information of the pages from other search engines and directories to display in the search results. This type of search engines are called metasearch engines.

Page Ranking

- - - Refer Class Notes - - -

References

[1] S. M. Hussain, H. Al-Bahadili and S. Al-Saab, "A web search engine model based on index-query bit-level compression.," in *ACM International Conference Proceeding Series.*, 2010.