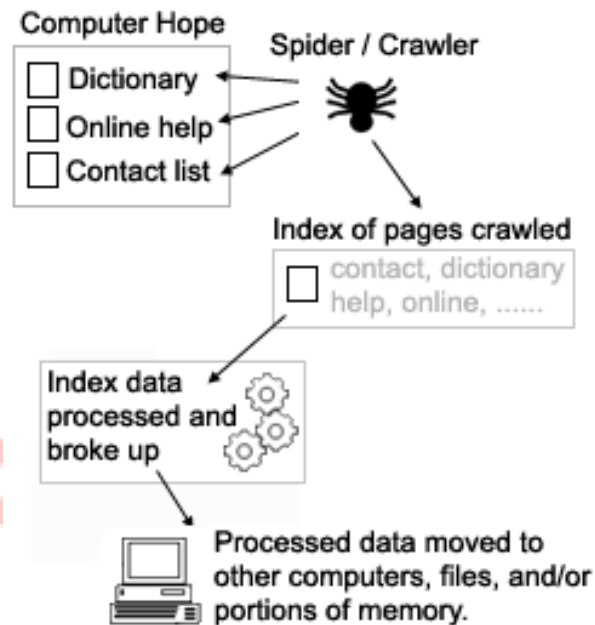## Lesson no. 2: WORKING OF SEARCH ENGINE

HOW SEARCH ENGINE WORKS

Because large search engines contain millions and sometimes billions of pages, many search engines not only just search the pages but also display the results depending upon their importance. This importance is commonly determined by using various algorithms.

Computer Hope
☐ Dictionary
☐ Online help
☐ Contact list

Spider / Crawler

Index of pages crawled
☐ contact, dictionary help, online, ...

Index data processed and broke up

Processed data moved to other computers, files, and/or portions of memory.

ComputerHope.com

As illustrated in the image on the right, the source of all search engine data is a spider or crawler, which automatically visits pages and indexes their contents.

Once a page has been crawled, the data contained within the page is processed and indexed. Often, this can involve the steps below.

a) Strip out stop words.

b) Record the remaining words in the page and the frequency they occur.

c) Record links to other pages.

d) Record information about images or other embedded media.

The data collected above is used to rank the page and is the primary method a search engine uses to determine if a page should be shown and in what order.

Finally, once the data is processed it is broken up into one or more files, moved to different computers, or loaded into memory where it can be accessed when a search is performed.

MAJOR FUNCTIONS OF A SEARCH ENGINE

a) 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.

b) 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.
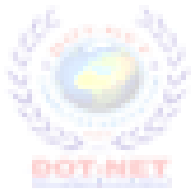
The indexing function of a search engine first excludes any unnecessary and common articles such as "the," "a" and "an." After eliminating common text, it stores the content in an organized way for quick and easy access.

c) 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.

d) 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.