Learn about the operation of search engines

The Internet and the World Wide Web have hundreds of millions of websites available that carry information on a variety of topics. However, most of them are titled according to the author's liking, and are placed on servers with confusing names. when the

The Internet and the World Wide Web have hundreds of millions of websites available that carry information on a variety of topics. However, most of them are titled according to the author's liking, and are placed on servers with confusing names. When you need to know about a topic, which pages will you have to read? Most people, when wondering about this issue, will use an Internet search engine.

Search engines on the Internet are special sites on the web, designed to help people find information stored on different sites. There are many different ways in this search, but they all perform three basic tasks:

1. Search the Internet - or select pieces of information on the Internet - based on important words
2. Keep an index for the words found with the address they found
3. Allows users to search for words or phrases searched in that index.

Search engines previously kept indexes of hundreds of thousands of websites and documents, often receiving as many as one or two thousand search requests per day. Today, the leading search engine indexes hundreds of millions of pages and responds to tens of millions of requests every day. In this article we want to show you how the main tasks will be performed, and how these search engines will handle them to allow you to find the information you need on the web.

Web review

When most people talk about Internet search engines, they all consider it a World Wide Web search engine. But before the Web became the most visible part of the Internet, there were search engines to help people find information online. Programs with names like "gopher" and "Archie" kept the index of files stored on the server connected to the Internet, significantly reducing the amount of time needed to search for chapters submission and documents. In the late 80s, getting important values ??from the Internet meant knowing how to use gopher, Archie, Veronica and some other similar programs.

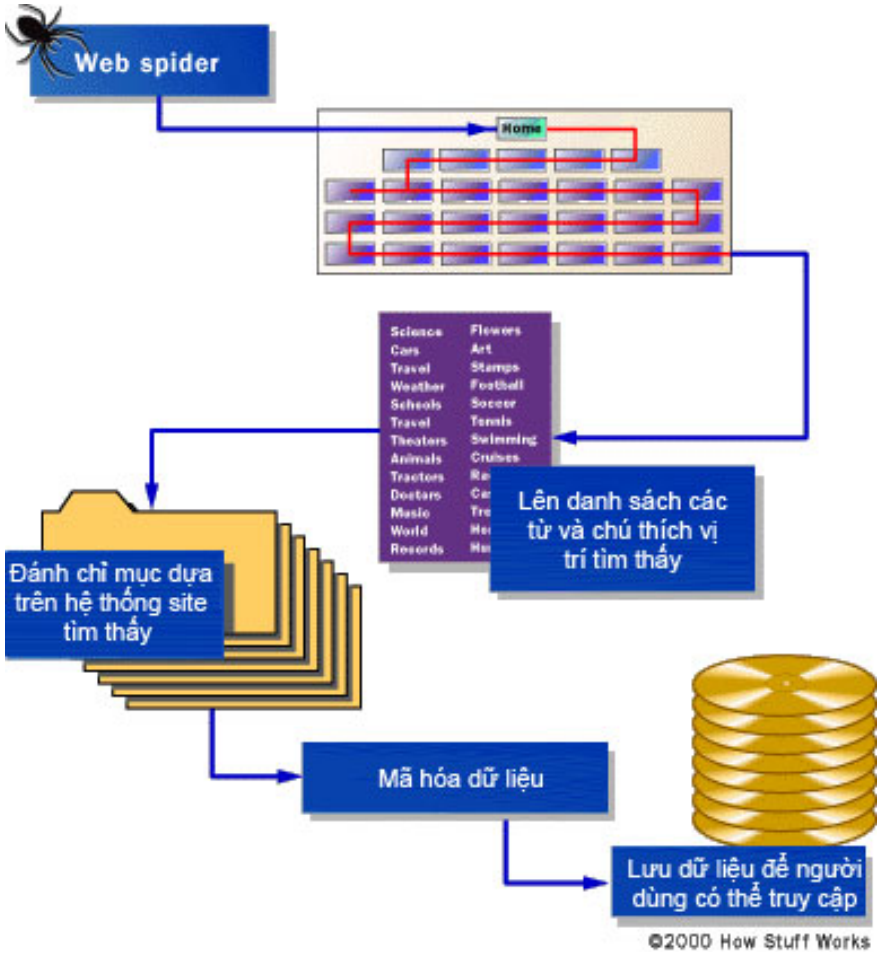
Today, most Internet users do not understand much about their search on the Web, so we will introduce this issue by focusing on the content of the website.

Start

Before the search engine can tell you where a certain file or document is stored, it must find them. To find information on hundreds of millions of existing websites, every search engine has to use special software robots, these software robots are called **spiders** , to form a name. Books of words found in websites. The process of

building a list of spider is called **Web crawling** . To create and maintain a useful search list, the search engine's spider must see the content of the web pages.

So how does spider do its work on the Web? The starting point is a list of popular servers and websites. The spider will start with a popular site, index the words on its page and follow the links found inside this site. In this way, the Spider system will quickly do its job and spread out the most widely used parts of the web.



Spider takes the content of the website and creates search keywords to allow
Online users can find the desired page.

Google.com started out as a university search engine. In the article describing how the system was built, Sergey Brin and Lawrence Page took an example of how fast their spider can work. They built the initial system to be able to use multiple spiders, usually up to 3 spiders working simultaneously. Each Spider can keep up to about 300 connections with websites every time. With its peak performance, using up to 4 Spiders, their system can find more than 100 pages per second, generating about 600KB of data per second.

Keeping the search speed fast also means building a system that can provide the information needed for Spider. The Google system previously had a dedicated server used to provide URLs for the Spider. Instead of relying on the service provider for DSN to translate the server's name into an address, Google already has their own DSN so that the slow hold takes place only in the minimum time.

When Google's Spider views HTML pages, it notes two things:

1. Words inside the page
2. Where to find the words

Words that appear in titles, subtitles, meta tags and other relevant important sections are noted with special considerations for subsequent user searches. Google's spider has been built to be able to index every important word on a page and leave only articles like "a," "an" and "the". Other spiders use other methods.

Other methods that Spider applies often try to make Spider's system faster, allowing users to search more effectively, or both. For example, some spiders keep in touch with the words in the title, small headings and links, along with 100 commonly used words on the page and words in the first 20 lines of text.

Other systems such as AltaVista approach another direction, indexing single words on each page, including "a," "an," "the" and other 'unimportant' words. The completeness in this method is matched by other systems in the meta tag section of the Website.

Meta tags

Meta tags allow the site owner to specify keywords and concepts that will be indexed. This is one of the most useful tags, especially in many cases that the words on the page can have up to two or three meanings - Meta tags can guide the search engine to choose any of the possible meanings. is true for a certain word. However, there is still a concern about relying heavily on Meta tags because of the poor quality pages that its owners have included in these popular topics but nothing about it. To overcome this phenomenon, the Spider will correlate Meta tags with the content of the page, eliminating tags that do not match the words in the page.

All of this problem originates from the owner of this site, they want the site to be displayed in the search results of these search engines. Sometimes, the owner does not want their page to appear on the results page of a search engine or does not want Spider to access their page. (For example, a game that builds new active pages each time the page sections are displayed or followed by new links. If a Spider Web can access it and start following all the links for For the new page, the game may have a high-speed activity bug for the player and prolong the control.) To avoid this situation, the robot exclusion protocol has been developed. This protocol adds to the meta tag

section at the beginning of each page, notifying the Spider to leave its page - not indexing the words on this page or trying to follow its links.

Build index

Once the Spider has completed the task of finding information on the site (we note that this is a task that is never completed because there is always a change of the page so that means the Spider will always be in charge. Its case), the search engine must save this information in a way that is most beneficial. There are two main components involved in creating collected data to be accessible to users:

1. Information is saved with data
2. Method by which information is indexed

In the simplest case, a search engine can store only the words and URLs it finds. In fact, this will greatly limit because there will be no way to say the word has been used as important or invalid on that page, or the word has been used once or many times, or whether the page contains links to other pages that contain words. In other words, there is no way to build a ranking list to vote on the most useful pages at the top of each search result list.

To make the search results the most relevant, most search engines save many words and URLs. An engine can store the number of times a word appears on a page. It can assign 'weight' to each entry, gradually increasing the value assigned to the words when they appear near the top of the document, in sub-headings, links and in meta tags or in headings. of the page. Each commercial search engine has a different formula for assigning weight to words in its index. This is one of the reasons why a search for the same words on different search engines produces different lists, pages displayed in different order.

Regardless of the rigorous combination of additional pieces of information stored by each search engine, the data will be encrypted for storage in separate storage points.

An index has only one purpose: It allows information to be found quickly. There are many ways to build an index, but the most effective way is to build a hash table. A formula is used to add numerical values ??to each word. This formula is designed to give entries on predefined quantities of divisions. This numerical division is different from the word division in the alphabet and that is the main effect of the hash table.

In English, there are some characters that start more with other characters. For example, you will see that the 'M' part of the dictionary will be thicker than the 'X' section. This inequity means that finding a word that starts with a 'popular' character can be much longer than finding a word that starts with a less common character. The algorithm of balancing the difference, and reducing the average time to search for each entry. It also distinguishes the index from the actual entry. The hash table consists of hashed numbers along with a data pointer, which can be classified in any way that allows the most efficient storage. The combination of efficient indexing and efficient storage makes it possible for users to perform search operations quickly even when they create a complex search.

Build a search

Searching through an index requires the user to build a query and submit it through the search engine. Queries can be quite simple like a word for example. Building a more complex query needs to use Boolean operators to allow you to filter and expand as a search.

The most commonly used Boolean operators are:

1. AND - All items connected by this operator must appear in the page or documents. Some search engines replace this operator with a '+' sign.
2. OR - At least one of the items linked by 'OR' must appear in pages or documents.
3. NOT - Items or items after this operator do not appear in the page or document, some search engines replace them with a '-'.
4. FOLLOWED BY - One of the items must be followed directly by another item.
5. NEAR - One of the items must be within some specific words
6. Quotes - Words between quotation marks must be treated as a phrase, and the phrase must be found in the document or file.

Looking for the future

Searches that are defined by Boolean operators become quite normal, the engine will search for words or phrases exactly when they are entered. This can create a problem when words are entered with multiple meanings. If you are only interested in one of those meanings, you may not want to see typical pages appearing for all of its meanings. Although you can create a search by eliminating unwanted meanings, how good it is if the search engine helps you.

One of the areas of search research is conceptual search. Some of these studies involve using statistical analysis of pages that contain the words or phrases that you search for, to find other pages that may interest you. Obviously the information stored on each page is really important for concept-based search, and beyond that, processing is required for each search. Many research groups have been working hard to improve both the results and performance of these search engines. Some other groups have turned to another field of research such as natural language queries.

The idea behind the natural language query research is that you can type in a question the same way you ask people sitting next to you - no need for Boolean operators or complex query structures trash. The most natural language query site today is AskJeeves.com, which has the ability to parse queries about keywords to apply to the index of the sites it has built. However, this site only works with queries simply because the concept of a query is quite complicated.