# Unit 1 Search Engines

## Unit Objectives (单元目标)

***After you have completed this unit, you will be able to***

- identify and describe features and functions of different types of search engines.
- explain how search engines work.
- evaluate the web pages from the Internet.
- search the Internet effectively and efficiently.

Search engines are programs on the Web that help users search for files and information. There are basically three types of search engines that people use to find what they want on the Web:

Crawler-based search engines
Human-powered directories
Meta search engines

None of these search engines allow you to search the entire Web; that would be an almost impossible task. However, you can use these search engines to scour different parts of the Web, and obtain different types of information.

## Task 1 Different Types of Search Engines

*Compare features and functions of crawler-based search engines, human-powered directories and meta search engines, and determine which one to use for specific searching tasks.*

## Crawler-based search engines

Crawler-based search engines are large, robots (called crawlers or spiders) created databases of web pages that help searchers find specific information on any given subject. The crawlers or spiders visit the sites and rank them for relevance without human interception. Some of the well known crawler-based search engines are:

- Google
- AltaVisa
- All the Web (FAST Search)
- Excite
- HotBot
- Inktomi

To search for business and other information in Chinese, you can use the famous search engine *Baidu*.

**Table of features**

- built by computer robot programs;
- not organized by subject categories. All pages are ranked by a computer algorithm;
- contain full-text of the web pages they link to;
- huge and often retrieve a lot of information;
- unevaluated: They contain the good, the bad and the ugly. You must evaluate everything you find.

## Human-powered directories

Human-powered directories, also known as subject directories, are smaller and more selective than crawler-based search engines. They are based on human indexing of sites in different categories. The editors review and select sites to add to their directories on the basis of determined selection criteria. Some good examples of human-powered directories are:

- Yahoo!
- LookSmart
- Open Directory Project (DMOZ)
- MSN

**Table of features**

- built by human selection, not by computers or robot programs;
- organized into subject categories. The subjects are not standardized and vary according to

the scope of each directory;
- never contain full-text of the web pages they link to. You can only search what you can see (titles, descriptions, subject categories, etc.);
- small and specialized to large, but smaller than most search engines—huge range in size;
- often carefully evaluated and annotated.

## Meta search engines

Meta search engines, also called metacrawlers, do not crawl the web themselves to build their own listings. Instead, they query other search engines for results. The results are then blended together on one web page. Metacrawlers present the results of their searches in one of the two ways:
- **Single list.** Most metacrawlers display multiple-engine search results in a single merged list, from which duplicate entries have been removed.
- **Multiple lists.** Some metacrawlers do not collate multiple-engine search results but display them instead in separate lists as they are received from each engine. Duplicate entries may appear.

Below are some of the major meta search engines:

- Ixquick
- Metor
- Vivisomo
- Dogpile
- Webcrawler

**Table of features**
- useful in obtaining a quick overview on a subject and/or unique term;
- take advantage of differences in pages indexed by individual search engines;
- good for hard-to-find information;
- quick but not thorough, with loss in precision and complexity of searching.

# Task 2 How Search Engines Work

*Could you find out how a spider works for a crawler-based search engine?*

Search engine is actually a general term for any internet portal that has World Wide Web search capabilities. The way crawler-based search engines collect information about the website is quite different from human-powered directories. This doesn't mean one is better than the other, but it does mean that the success of each search engine is based on meeting its specific requirements. Here is a quick look at how search engines work.

## Crawler-based search engines

Crawler-based search engines "crawl" or "spider" the Web from link to link, identifying and perusing pages. Usually, crawler-based search engines are made up of three major elements: the spider, the index, and the software. Each has its own function and they together produce what we have come to trust (or distrust) on the SERPs (Search Engine Results Pages).

### ◆ The Spider

Also known as a web crawler or robot, a search engine spider is an automated program that reads web pages and follows any links to other pages within the site. This is often referred to as a site being "spidered" or "crawled".

Spiders start their journeys with a list of page URLs that have previously been added to their index (database). As it visits these pages, crawling the code and copy, it adds new pages (links) that it finds on the page to its index. The spider returns to the sites in its index on a regular basis, scanning for any changes. Web page owners may submit their URLs to search engines for "crawling" and finally inclusion in their databases.

### ◆ The Index

An index is like a giant catalog or inventory of websites containing a copy of every web page and file that the spider finds. If a web page changes, this catalog is updated with the new information. To give you an idea of the size of these indexes, the latest figure released by Google is 8 billion pages.

It sometimes takes a while for new pages or changes that the spider finds to be added to its index. Thus, a web page may have been "spidered" but not yet "indexed." Until a page is indexed—added to the index—spidered pages will not be available to those searching with the search engine.

### ◆ The Software

At the end of the day, a search engine is a software program designed to sift through billions of pages recorded in its index to find matches to a search query and rank them in an order that

it believes is most relevant.

How do search engines go about determining relevancy, when confronted with hundreds of millions of web pages to sort through? Each search engine has developed a set of rules and mathematical equations, known as an algorithm, which it uses to set the order of its rankings.

Exactly how a particular search engine's algorithm works is a closely-kept secret, but some general rules often being used to increase a website's ranking performance are clear. This is referred to as search engine optimization.

## Human-powered directories

A human-powered directory depends on humans for its listings. Website owners submit a short description to the directory for their entire sites, or editors write one for sites they review. In general, most directories link only to the home pages rather than indexing the full text of each page on the sites.

Changing the web pages has no effect on the listing. Things that are useful for improving a listing with a search engine have nothing to do with improving a listing in a directory. The only exception is that a good site, with good content, might be more likely to get reviewed for free than a poor one.

In the web's early days, it used to be that a search engine either presented crawler-based results or human-powered listings. Today, it is extremely common for both types of results to be presented. Usually, a hybrid search engine will favor one type of listings over another. For example, MSN Search is more likely to present human-powered listings from LookSmart. However, it does also present crawler-based results (as provided by Inktomi), especially for more obscure queries.

## Meta search engines

In a meta search engine, when keywords are submitted in its search box, it will transmit the search simultaneously to several individual search engines and their databases of web pages. Within a few seconds, results will get back from all the search engines queried. Meta-search engines do not own a database of Web pages; they send your search terms to the databases maintained by search engine companies. Meta search engines provide a quick way of finding out which engines are retrieving the best results for you in your search.

# Task 3 Web Page Evaluation

*Today, the Internet provides faster and more extensive ways of retrieving and sharing information than ever before. The World Wide Web can be a great place to accomplish research on many topics, because the Web opens windows to easy access and distribute information, almost anyone can publish anything on the Web. It follows that much of the information on the Web has not been for accuracy or quality. Thus, to make effective use of the information on the Web, to improve our touch and help us make our way along the line, the web pages must be evaluated. But how can it be done?*

## Reading web addresses

An important element you should look out for when evaluating a website, is the very structure of its own address, or URL (Uniform Resource Locator).

**What do you know about the sponsoring institution?**
A web address could provide a clue about an organization's integrity as well as maybe its bias. The name of the institution is usually what appears after the server (usually www) and before the domain (.com, .edu, etc.):

http://www.**phillipmorris**.com/smokingfacts.html
http://www.**communistparty**.org/viewsoncapitalism.html

Taking the first web address above as an example, let's have a look at what it all means:

- "**http**" is the Hypertext Transfer Protocol and refers to the format used to transfer and deal with information, the protocol is followed by a colon, then two slashes: http://
- "**www**" stands for World Wide Web, states the generic name of the server. It is not an essential part of the address and some websites don't use it.
- "**phillipmorris**" is the name of the institution and usually owns the server.
- "**com**" is the domain name.
- "**smokingfacts**" is the file name.
- "**html**" is the file type, stands for Hyper Text Make-up Language (that is the language the computer reads).

**What is the domain?**
It is the three or two character extension at the end of the main institution's address, which is the address before the first single slash on the right:

http://www.amazon.**com/**
http://www.epa.**gov/**
http://www.berkeley.**edu/**
http://www.rand.**org**/

A **.com** is a commercial website, meaning the institution is a corporate or small business entity.

A **.gov** indicates the sponsoring institution is a government body, which has the responsibility to provide reliable information.

A **.org** is a non-profit organization attempting to influence public opinion. Most often, the website is selling its ideas and therefore, can be more biased.

A **.edu** is coming from an educational institution (usually students take four years or more to get a degree). Many educational websites are to present factual information, however, you need to read a .edu website before the first slash to make sure.

Other domains that have been in operation and are generally accepted are:

- **.net**—Internet service provider, organizations, networks
- **.mil**—US military site
- **.info**—general use by both commercial and non-commercial sites
- **.museum**—restricted use by museums
- **.biz**—general use by business
- **.coop**—restricted use by cooperatives
- **.pro**—restricted use by certified professionals and professional entities

Country names appear as a two-letter abbreviation in the domain name. For example, .uk for United Kingdom, .fa for France, .cn for China. Because the Internet was created in the United States "US" was not originally assigned to U.S. domain names. For a complete list, go to http://www.norid.no/domenenavnbaser/domreg.html#u.

**Is it a personal web page?**

Personal web pages are usually clearly marked by a username somewhere within the address. Often the username of the person is followed by a "~" (swung line):

http://www.berkeley.edu/students/~pbailey/deathpenalty.html

The above address clearly marks that the web page is by a student, however, not all educational institutions have such a clear folder marked such as "students". (In fact "users" is more often used). But the "~" symbol is much more universal in warning you that it is a personal home page and usually not by an expert. Professors and other academics usually do not have a "~" before their usernames.

http://members.tripod.com/abortionfacts/
http://hometown.aol.com/viewsongaymarriage/
http://www.geocities.com/407689/guncontrol.html

A personal home page can also be indicated by an Internet Service Provider or web page provider. Watch out for addresses that have the words "members", any relating to a "home page" or a number used as part of the address. Sometimes an institution or company may have such an address, and sometimes graduate students or even professors (such as part-time teachers) can keep their information on such websites as well. However, most often such pages are personal web pages reflecting the opinions of a person who may or may not be an expert, or an obscure organization.

## Scanning the home pages

While not every reliable website uses the same structure, many reliable web pages tend to have a header, a body and a footer (see Figure 1.1).

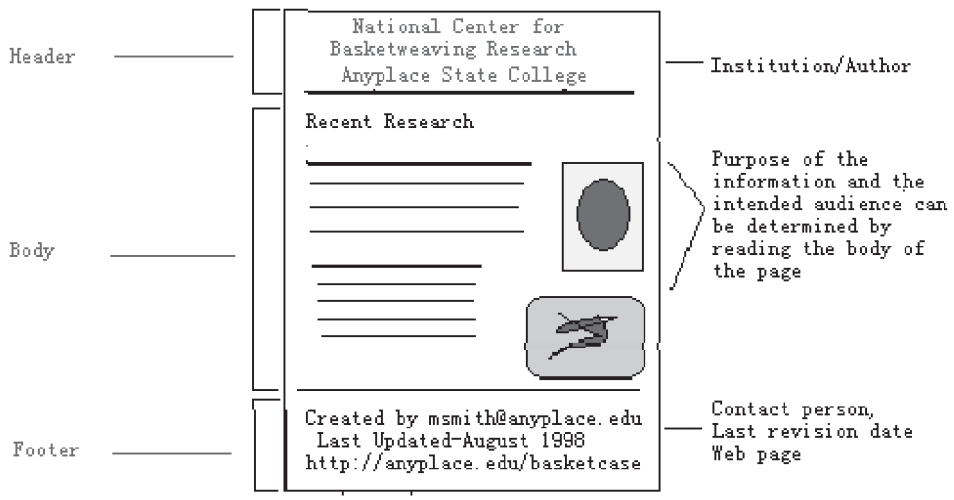


**Figure 1.1**

Within each of the following pieces you should be able to learn important vital pieces of information that determine the integrity of a website:

- **Author or contact person**—usually located in the footer
- **Link to local home page** (usually of institution or author)—usually located in the header or footer
- **Institution**—usually located in the header or footer
- **Date of creation or revision**—usually located in the footer
- **Intended audience**—determined by examining the body
- **Purpose of the information**—determined by examining the body

Sometimes depending on what page you find from a search engine, you may need to find the home page first. Find any link on the page you are on that says something such as “Home” or “Back to”. The link will usually be towards the top or the bottom of that particular page.

## Checking five areas

When evaluating the site, look for the five areas:

- **Accuracy**—How reliable and error free is the information? Who is the sponsoring institution (government, university or commercial company)? How credible or well known is the sponsoring institution? Does the information consist of documented

facts or personal opinion?

- **Authority**—Is the author or source of the information identified and his/her qualifications in evidence? Does the site exhibit proper grammar, spelling, and literary composition?
- **Objectivity**—What is the site's purpose: to inform, explain, persuade or sell? Is the information presented with a minimum of personal bias?
- **Currency**—Is the content of the work up-to-date? Is the date of creation or most recent revision date clearly shown?
- **Coverage**—Is it a comprehensive coverage of the subject matter? Is the information relevant or useful for your needs?

# Task 4 Search Tips and Techniques

*It's important to give some thought to your search before you begin. Create a search strategy in your head by asking yourself some questions: Am I just beginning to amass knowledge on a fairly broad subject? Or do I have a specific objective in mind?*

*If you are more interested in broad, general information, the first place to go is Web Directory. If you are after narrow, specific information, a Web search engine is probably a better choice.*

## Basic search tips

- **Skip Stop Words**

Many search engines don't record extremely common words in order to save space or to speed up searches, they won't "stop" to look for them, and these are known as stop words. Stop words are small words such as conjunctions, prepositions and articles. Examples include: a, about, an, and, as, at, be, by, from, in, is, it, of, on, or, that, the, to, with, etc. Not all search engines recognize the same stop words. Besides, their lists can and do change frequently. If you initiate a search at a site that maintains a list of stop words and you type any of those words into your search statement, they may well continue to be ignored.

- **Use Quotation Marks ( " ")**

Use quotation marks to find words or a phrase exactly as typed. Putting quotation marks

around the words “ultrasound technologist” will find only those documents containing those two words in that exact order. Otherwise, the search results will include the words “ultrasound” and “technologist” in any order, anywhere within the document.

**Note:** Do not put quotation marks around a single word.

- **Use Plus (+) or Minus (-)**

Put a plus or a minus sign in front of words to include or exclude the words in the search results. For example, “+vegetable –tomato” will find pages containing the word vegetable and there is not the word tomato.

**Note:** Do not leave space between the plus or minus sign and the words.

- **Use Wildcards**

Wildcards searches can expand the number of matches for a particular request.

A (*) can replace one or more letters at the end of a word. This might help you search for something that can be phrased differently. For example, “nurs*” will find documents containing the words nurse, nursing and nurses.

Searching for *her* will find the words here, whether, together, gathering, and any other word that contains “her” anywhere in the word.

If a query word contains a (?**),** any character will match that position. For instance, searching for “ b?g” will find the words big, bag, beg, etc.

- **Use Appropriate Capitalization**

Capitalize proper names of people, companies or products, and remember that lower-case words will match both lower and upper case versions. For example, typing “search” will return all documents containing the words search, Search and SEARCH. Typing “Search”, however, will instruct the search engine to look only for the capitalized word.

- **Use Boolean Searches**

You can use the following Boolean operators in your search: AND, OR, NOT, etc. These operators MUST be in capital letters.

Example: (contact AND us) OR (about AND us)

## Boolean Logic

**What is Boolean Logic?**

The word Boolean is derived from the 19th-century British mathematician George Boole. Database administrators have used Boolean logic for decades. It's an information retrieval technique that allows several search words and phrases to be combined using operators or commands. Operators are instructions to the search engine such as AND, OR, NOT and NEAR. Learning to use these Boolean operators for keyword searches will help narrow your results.

**Basic Boolean Operators**

- **OR**

The OR operator combines keywords so that a document is retrieved if it contains any or all of the keywords. For example, searching for "developer OR engineer" will find all documents in which the words "developer" or "engineer" appear.

OR is the default command on most, but not all, search engines, so if you use more than one word in a search without an operator, the search engine assumes each word is linked by OR.

- **AND**

Use AND to combine terms so that information is retrieved only if all terms occur in the same document. For example, "software AND engineer" will find only documents where both keywords occur.

- **NOT**

The NOT operator prevents retrieval of documents in which specified terms occur together. This operator is also helpful when a keyword has multiple meanings. For example, "engineer NOT software" will find documents in which "engineer" occurs but "software" does not.

- **NEAR**

The NEAR operator indicates that the search words you have entered must appear within a certain number of words of each other (usually between 1 and 20). For example, a search for "'sales NEAR management" would turn up results in which the two words appear close together.

- **FOLLOWED BY**

This operator means that one term must directly follow the other. ADJ, for adjacent, serves

the same function. Essentially, a search engine that will allow you to search on phrases uses the same method. For example, "endangered ADJ species" returns both "endangered species" and "species endangered".

Other Boolean operators can help you fine-tune your search results, but mastering these basics will increase your accuracy and decrease your frustration. It's also worth noting that various websites will tinker with Boolean logic to create their own versions of the search language. For example, the popular search engine Google has its own variations.

## Practice Exercises

### Exercise 1

Using different types of search engines:

1. Select one or two of the search engines listed in Task 1 of this unit and search for:
   **European Stocks**

   Now try searching for the same subject as a phrase, enclosed in quotes:
   "**European Stocks**"
   (Hint: The second search should retrieve far fewer documents than the first one.)

2. Select any one or two of the meta search engines listed in Task 1 of this unit and search for:
   **linux windows debate**
   Compare your results.

3. Select any two subject directories listed in Task 1 of this unit and follow the subject categories—do not use search box— to search for information on:
   **Depression**
   Compare your results.
   (Hint: Find the Health link and start there.)

## Exercise 2

Evaluating web resources:

1. Choose one topic area and look at two of the websites listed for it.
2. Evaluate those websites according to the criteria learned in Task 3 of this unit.
3. Write out the website's citation information.
4. Be prepared to talk about the two sites you chose, including a comparison of their strengths and weaknesses and whether you would use them as sources for a business report.

**Topic:** Nike and international labor practices
http://www.nike.com/nikebiz/nikebiz.jhtml?page=25&cat=code
http://www.saigon.com/~nike/
http://www.pbs.org/newshour/bb/business/jan-june98/nike 5-13.html

**Topic:** Teams and group decision making
http://www.csuchico.edu/sac/leaders/grpdecision.html
http://fie.engrng.pitt.edu/fie95/4c3/4c33/4c33.html
http://www.effectivemeetings.com/technology/mrtools/collaborative.asp

**Topic:** Impact of the Internet on traditional businesses
http://www.cyberlaw.com/issues.html
http://www.internetindicators.com/factfigure.html
http://www-cs-students.stanford.edu/~dwhitney/lisa/IS Final Proj.html

## Exercise 3

Search practice:

1. Select a search engine, and try searching the following keywords in the following order:

- **trade imbalance**
- **trade imbalance between China and the U.S.**
- **trade imbalance between China and the U.S. in 2005**
- **reasons behind the trade imbalance between China and the U.S. in 2005**

With each successive search, you are narrowing the topic and should be retrieving noticeably fewer results.

2. Choose one of the examples from Task 4 of this unit and try it as a search on the "advanced" page of the search engine of your choice.

(Hint: On some engines, you may have to select "Exact phrase", " Any of the words", "Must contain", etc. You may also have to indicate if it's a Boolean search.)

## Assessment for Performance

| Items | Requirements | Score assigned for the item | Your score for the item |
|---|---|---|---|
| Choosing the right search engines | Be able to use crawler-based search engines. | 10 | |
| | Be able to use human-powered directories. | 10 | |
| | Be able to use meta search engines. | 10 | |
| Web evaluation | Be able to write out the web-site's citation information. | 15 | |
| | Be able to talk about the two sites you choose, including a comparison of their strengths and weaknesses and whether you will use them as sources for a paper. | 15 | |
| Obtaining the information needed | Be able to obtain specific information for your business activity | 40 | |
| Total score | | | |