Computer 2

Introduction to Web Search Engines & Databases:

What is a search engine ? / What is a Database ?

Name_______________________________________ Date 11/20/01

Background Information: In general terms search engines on the Internet can be categorized as **Search Engines, Web Directories , White and Yellow Pages , Multi-engine searches and Specialized Sites.**

Search engines are searchable databases that appear as web pages, with built-in common, plain English query systems. **Their databases are built in two steps**.

First, **"spiders" or "robots"** roam the web, sending back information on the web pages they encounter. Alta Vista calls it's spider "Scooter" for example. Scooter is able to visit 6 million sites a day. Spiders don't "invade" your web pages. They visit web servers and politely request documents, which they then examine. They may find these documents by following links on pages they've already indexed, look at "What's New?" pages, or pay particular attention to popular web sites. Most of them allow authors to submit their sites for inclusion in the database.

Once a spider encounters a new web site, it informs the indexing software back home and indexing begins. An index can be a list of all relevant words found on a page, or perhaps just from specific location such as the title, special "meta" tags in the header, or they might read the first few paragraphs on each page. Those that index all the words on a page may give extra weight to certain words depending how often and where they appear.

Since the web changes so rapidly, spiders must revisit previously indexed sites to make sure they still exist and look for changes. The "freshness' of search results depends on how often this is done, and so a high return of 404 error messages and irrelevant sites would indicate that the engine is not revisiting sites often enough. Alta Vista appears to revisit most sites about once a month. When you use a search engine, you are submitted your query to a previously assembled index, which looks for your word or phrase and returns a list of pages that contain it. So as you search an index you are not searching the entire Web or Internet. No single engine yet contains indexes of the entire Internet or the World Wide Web.

Web directories (databases) are best used for broad, general information. They are indexed by actual people, not software programs. They categorize sites based on topics, which are either submissions for their material and as a result, their databases will not be as large. But what you do find will be more sharply focused. Although most directories do provide search engines, though their power lies in their directory structure. **Yahoo** is the best known web directory and while it uses Google’s search engine, but it's 1400 categories are well structured and easy to navigate.

One exciting development is that of **conceptual query systems**, based on artificial intelligence theory. In other words, the search engine tries to determine what you mean, not just what you say. Excite is the best known example of this technique. In concept-based searching, you enter your queries in plain English, and the search engine will return not only those documents containing your exact words, but will extend it's search and return related sites as well. It does this by calculating the frequency by which certain words appear along with the query term. It identifies relationships from the documents themselves, and learns more from each document it indexes. With these systems, the search engine will soon "learn" that heart can also mean cardiac and Porifora are marine sponges.

How do I decide which search engine to use? You'll find that loyalty runs deep when it comes to everyone's favorite search engine, and "shoot-out" reviews will vary widely in their recommendations. But to cling to one method is to fail to take advantage of the fluid, rapidly evolving state of the art. My suggestion is to periodically take the time to run some tests yourself. Everyone's an expert on something - so perform some searches on topics that are familiar to you. See if you're satisfied not only with the depth of the search but also the relevancy rankings of the sites. If you have special requirements, such as the need to explore current news sources, some sites may be better than others for your needs. Be aware that competition is fierce in this area, and all of the top contenders are constantly upgrading and refining their user interfaces and their techniques. So it pays to explore now and then, and revisit an old engine or check out the new ones that crop up.

Directions: You are to visit several of the search engines linked from the MEHS Search Engines Launch Pad. The URL is **http://www.mehs.educ.state.ak.us/mehsengines.html,** or you may link to this site from the MEHS home page. Once there examine the list of search engine links and visit several. Locate the following and be able to defend your conclusions:

(1) Name a least two "search engines"? ______________________, ___________________

(2) Name a least two "Web directories"? ______________________, _________________

(3) "Web directories" are also referred to by what other names"?

____

(4) Do you find any links to "White and/or Yellow Pages"? (name them)

______________________, ___________________

(5) Do you find links to any "Multi-engine searches", sometime referred to as metacrawlers or metasearch engines? ______________________, ____________________

(6) Did you find any links to "Specialized Sites"? ________________, _________________
... these would be links to search / data sites that are dedicated to a particular archive of information

Mark which web search tool performs the below functions:

	Search Engine	Database
Has a computer program that goes on to the Internet, finds and web sites		
Indexes (or categorizes) websites by a computer program		
Indexes (or categorizes) websites by a human		
Gives a larger view of the web		
Gives a smaller, but easier to sort through, view of the web		