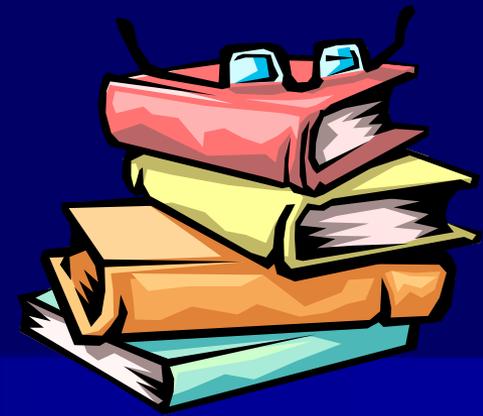# Internet Fundamentals

Lecture-13

WWW Search Engines

Outline

- Introduction
- Directories, Search Engines, Metasearch Engines
- Search Fundamentals
- Search Strategies
- How does a search engine work?
- Searching Tips
- Your site's ranking?
- Summary

Introduction

- You have probably been using search engines, but perhaps may not be as effectively as possible.

- A lot of information is available on-line, but not all of them is completely accurate.

- The web-page addresses are constantly changing, it may be only available for a short time.

Search Engine History

- In 1990, before the WWW, Alan Emtage created Archie, the first search tool for finding files on FTP sites.
- In 1993, Veronica is developed. Followed by Jughead, Wandex, …
- In 1994, Galaxy, WebCrawler, Yahoo! and Lycos debuted.
- In 1995 and afterwards, Excite, Infoseek, Alta Vista, MetaCrawler, …
- Next generation: specialized hybrids

5

Directories

- A Web Directory or Web Guide is a hierarchical representation of hyperlinks.

- The top level is typically a wide range of very general topics.

- Each topic contains hyperlinks of more specialized sub-topics.

- Very easy to use.

Hierarchical Representation

Popular Directories

- AOL anywhere – search.aol.com
- CNET Search.com – www.search.com
- Excite – www.excite.com
- E-Wild life – www.ewildlife.com
- Lycos – www.lycos.com
- Yahoo! – www.yahoo.com
- Google – www.google.com

Some figures

Search Engines

- A search engine is a computer program that does the following:
  - Allows user to submit a query that consists of a word / phase
  - Searches the database
  - Returns a list of suitable URLs which match your query.
  - Allows user to revise and resubmit.

Where to submit Query?
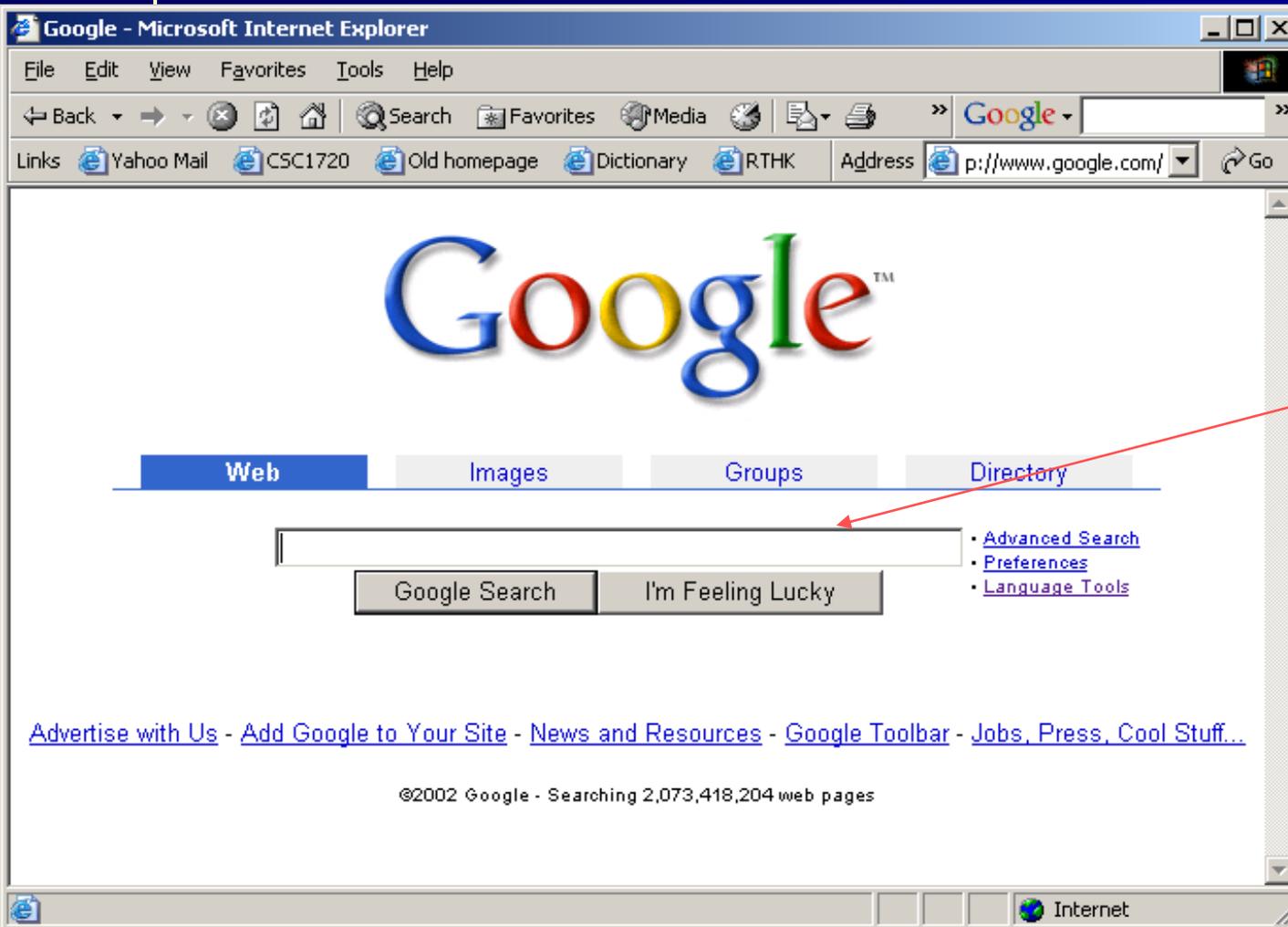
Submit your Query

Popular Search Engines

- AOL anywhere – search.aol.com
- AltaVista – altavista.digital.com
- Excite – www.excite.com
- HotBot – www.hotpot.com
- Magellan – www.mckinley.com
- Google – www.google.com

Metasearch Engines

- A metasearch or all-in-one search engine performs a search by the use of more than one other search engine to complete the search job.
- The duplicate retrievals are eliminated.
- The results are ranked according to how well they match with the query.
- Advantage:
  - A single query can access lot of search engines.
- Disadvantage:
  - A high noise-to-signal ratio, lot of matches will not be suitable for you.

Popular Metasearch Engines

- Metasearch – www.metasearch.com
- Metacrawler – www.metacrawler.com
- MetaFind – www.metafind.com
- Dogpile – www.dogpile.com

Some Figures

White Pages / Yellow Pages

- White pages allows user to lookup information about individuals.
- We can use white page to track down the telephone numbers, email address.
- People can abuse white pages
- Some people think that white pages are an invasion of their privacy.
- Yellow pages contain information about businesses.

Popular White Pages & Yellow Pages

- Bigfoot – www.bigfoot.com
- Yahoo! People Search – people.yahoo.com
- WhoWhere – www.whowhere.com

- Yahoo! Yellow Page – yp.yahoo.com
- SuperPages – www.superpages.com

Some Figures – White Pages

Some Figures – Yellow Pages

Comparison

Directory	Search Engine
A directory allows you to explore and get what you want eventually.	A search engine brings you to the exact page on the words or phrases you are looking for.
Use a directory to find cooking-related websites.	Use a search engine to find a specific recipe, by providing the name of the ingredients.
Use a directory to find travel guides in a country.	Use a search engine to find the transport trains schedule in South Africa.

Search Fundamentals

- Example: www.yahoo.com
- Header: Yahoo Logo and some advertising.
- Information bar: contains other hyperlinks.
- Search form area: consists a form which allows you to type a query.
- Directory area: a large number of categories, channels.
- Yahoo Links: Link to other yahoo sites.
- Footer: contains information about yahoo, copyright and a disclaimer.

Search Fundamentals

Search Terminology

- **Search Tool:** Any mean to locating information on the Internet.
- **Query:** Information typed into the form on the search engine.
- **Query syntax:** Rules for constructing a valid query.
- **Query semantics:** Rules for defining the meaning of a query.
- **Hit/Match:** A URL that the search engine returns for a specific query.
- **Relevancy score:** A value that indicates the quality of the URL (match close to the query 1 to 100).

Pattern Matching Queries

- It is also called Fuzzy Query.
- You can enter "ungrammatical sentences", "incomplete sentence fragments", "disjoint phrases", "nonsense words".
- The search engine gets a collection of keywords.
- Required keyword: Mark with "+" before the keyword.
- Prohibited keyword: Mark with "-" before the keyword.

Pattern Matching Queries

Boolean Queries

- A Boolean Query is a query that consists keywords but with logical operators (AND, OR, NOT).

- X AND Y – will return URLs that contain both X and Y.

- X OR Y – will return URLs that contain either X or Y.

- X AND NOT Y – will return URLs that contain X and do not contain Y.

- Symbol: AND - &, OR - |, NOT - !, NEAR - ~

Boolean Queries

- AND is used for narrowing a query
  - If you know that your target documents will contain a group of keywords, list them using the AND operator
- OR is used for broadening a query
  - If you can think of related words for a topic, list them using the OR operator
- NOT is used to redirect a query
  - If you find that a keyword or phrase is leading irrelevant hits, then represent it in your query as AND NOT *keyword*

Boolean Queries

Using Wildcards

- Wildcards are useful for retrieving variations of a word
- For example, art* will search for art, artwork, artist, artistry, and so forth
- An excellent way to broaden a search
- Different wildcard characters are used by different search engines
- The most common characters are: *, #, and ?

Advanced Search Options

AltaVista Advanced Search Cheat Sheet - Microsoft Internet Explorer

Advanced Search Cheat Sheet

AND	Finds documents containing all of the specified words or phrases. **Peanut AND butter** finds documents with both the word peanut and word butter.
OR	Finds documents containing at least one of the specified words or phrases. **Peanut OR butter** finds documents containing either peanut or butter. The found documents could contain both items, but not necessarily.
AND NOT	Excludes documents containing the specified word or phrase. **Peanut AND NOT butter** finds documents with peanut but not containing butter. NOT must be used with another operator, like AND. AltaVista does not accept '**peanut NOT butter**'; instead, specify **peanut AND NOT butter**.
NEAR	Finds documents containing both specified words or phrases within words of each other. **Peanut NEAR butter** would find documents peanut butter, but probably not any other kind of butter.
( )	Use parentheses to group complex Boolean phrases. For example **(peanut AND butter) AND (jelly OR jam)** finds documents with words 'peanut butter and jelly' or 'peanut butter and jam' or both.
anchor:*text*	Finds pages that contain the specified word or phrase in the text hyperlink. Finds pages that contain the specified word or phrase text of a hyperlink. **anchor:job +programming** would find pages job in a link and with the word programming in the content of the Do not put a space before or after the colon. You must repeat the keyword to search for more than one word or phrase; for example **anchor:job OR anchor:career** to find pages with anchors containing either the word job or the word career.

AltaVista Advanced Search Cheat Sheet - Microsoft Internet Explorer

applet:*class*	Finds pages that contain a specified Java applet. Use **applet:morph** to find pages using applets called morph.
domain:*domainname*	Finds pages within the specified domain. Use **domain:uk** to find pages from the United Kingdom, or use **domain:com** to find pages from commercial sites.
host:*hostname*	Finds pages on a specific computer. The search **host:www.shopping.com** would find pages on the Shopping.com computer, and **host:dilbert.unitedmedia.com** would find pages on the computer called dilbert at unitedmedia.com.
image:*filename*	Finds pages with images having a specific filename. Use **image:beaches** to find pages with images called beaches.
like:*URLtext*	Finds pages similar to or related to the specified URL. For example, **like:www.abebooks.com** finds Web sites that sell used and rare books, similar to the www.abebooks site. **like:sfpl.lib.ca.us/** finds public and university library sites. **like:http://www.indiaxs.com/** finds sites about culture on the Indian subcontinent.
link:*URLtext*	Finds pages with a link to a page with the specified URL text. Use **link:www.myway.com** to find all pages linking to myway.com.
text:*text*	Finds pages that contain the specified text in any part of the page other than an image tag, link, or URL. The search **text:graduation** would find all pages with the term graduation in them.
title:*text*	Finds pages that contain the specified word or phrase in the page title (which appears in the title bar of most browsers). The search **title:sunset** would find pages with sunset in the title.
url:*text*	Finds pages with a specific word or phrase in the URL. Use **url:garden** to find all pages on all servers that have the word *garden* anywhere in the host name, path, or filename.

Break Time – 10 minutes

Search Strategies

- You should find a search engine that meets the following conditions:
  - A user-friendly interface
  - Easy-to-understand documentation
  - Convenient to access
  - A large indexed database
  - Assigning good relevancy scores.

- Learn the syntax of this particular search engine, but not several different engines.

Search Generalization

- Too few hits?
  - Needs to generalize your search query.
- Pattern matching query: eliminate one of the more specific keywords of the query.
- Boolean query: remove the keywords with AND operator, or delete the NOT item, or use the OR operator.
- Use a directory or metasearch engine if still cannot locate the matched URL.

Search Specialization

- Too many hits?
  - Needs to specialize your search query.
- Pattern matching query: add more keywords.
- Boolean query: use AND with other keyword, or add NOT operator to excluded some unwanted pages.
- Try capitalizing proper nouns or names.
- Use a directory to locate your information.

Sample Searches

- Queries about Kayaking in Alaska
- Example: Using infoseek

Query:	No. of Hits
alaska	176,954
Alaska	176,064
+"Prince William Sound" +Alaska	778
+kayak +"Prince William Sound" +Alaska	44
+kayaking +"Prince William Sound" +Alaska	60
+kayaking +"Prince William Sound" +Alaska +rental	20

How does it work?

- **User Interface** – Allows you to type a query and displays the results.
- **Searcher** – The engine searches the database for matching your query.
- **Evaluator** – The engine assigns scores to the retrieved information.
- **Gatherer** – The component that travels the WEB, and collects information.
- **Indexer** – The engine that categorizes the data collected by the gatherer.

User Interface

- Provides a mechanism for a user to submit queries to the search engine.
- Uses forms, very user friendly.
- The user interface displays the search results in a convenient way.
- A summary of each matched page is shown.

Searcher

- It is a program that uses the search engine's database to locate the matches for a specific query.

- The database of a search engine holds extremely large indexed pages.

- A highly efficient search algorithm is necessary.

- Computer Scientists have spent years to develop the searching and sorting methods.

- You can refer to computer books.

Evaluator

- The searcher returns a set of URLs that match your query.
- Not all of the hits equally match your query.
- More references to the page, the ranking of the page will be higher.
- How the relevancy score is calculated?
  - Varies from one engine to another one.
  - The number of times of the word appears?
  - The query words appear in the title?
  - The query words appear in the META tag?

Link Popularity

reference

Gatherer

- It is a program that traverses the Web and gathers information about the Web documents.

- It runs at a short and regular intervals.

- It returns information and will be indexed to the database.

- Alternate names: Bot, Crawler, Robot, Spider and Worm.

Spiderlist

- http://www.spiderhunter.com/

Spider Hunter: Learn to write cloaking scripts and track spiders - Microsoft Internet Explorer

Address: http://www.spiderhunter.com/spiderlist/index.cfm?FuseAction=Results&Key

Spider Hunter.com

SimpleApps.net
Co-branded applications
on your own domain name

Calendar | Contacts | Briefcase | Email

Compare Scripts
Ethics
Free Scripts
Links
Non-Spiders
Pseudocode
Spider Scope
Spiders
Spider List
Spider Trap
Tools
Tutorial

Help Support this site

Donations

Search at Interrogative

PayPal

Results

Spider IP	Spider DNS Name	Search Engine	Spider User Agent
209.185.253.175	crawler.googlebot.com	Google	Googlebot/2.1 (+http://googlebot.com/bot.html)
209.185.108.147	a10.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.155	a18.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.140	a3.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.141	a4.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.143	a6.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.145	a8.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.148	a11.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.150	a13.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.152	a15.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.154	a17.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.157	c3.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.159	a19.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.161	c4.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.164	a23.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)
209.185.108.165	a24.google.com	Google	Googlebot/1.0 (googlebot(at)googlebot.com)

Internet

by C.C. Cheung 2003.

Indexer

- It organizes the data by creating a set of keys or an index.
- Indexes need to be rebuilt frequently.
- E.g. Libraries – Author, Title, ISBN, etc…
- In order to ensure the returned URL is not out of date.
- The search engine is very complex and needs to break down into different components.

Case Study - AltaVista

- Sending out Crawlers (robot programs) that capture information from the web and bring them back.
- The main crawler – "Scooter" simultaneously send out HTTP requests like blind users on the Web.
- Store all these information to the indexing engine.
- Scooter's cousins help to remove "dead" links.
- A typical day, Scooter will visit over 10 million pages.
- Web pages with no links referencing will never be found.
- You can also submit your URL to AltaVista.

Case Study - AltaVista

- **METAtags** – special keywords embedded in the headers of the webpage.
- **Full-text index** – Every word on every page is also included during searching.
- AltaVista is using Full-text indexing.

METAtag Example
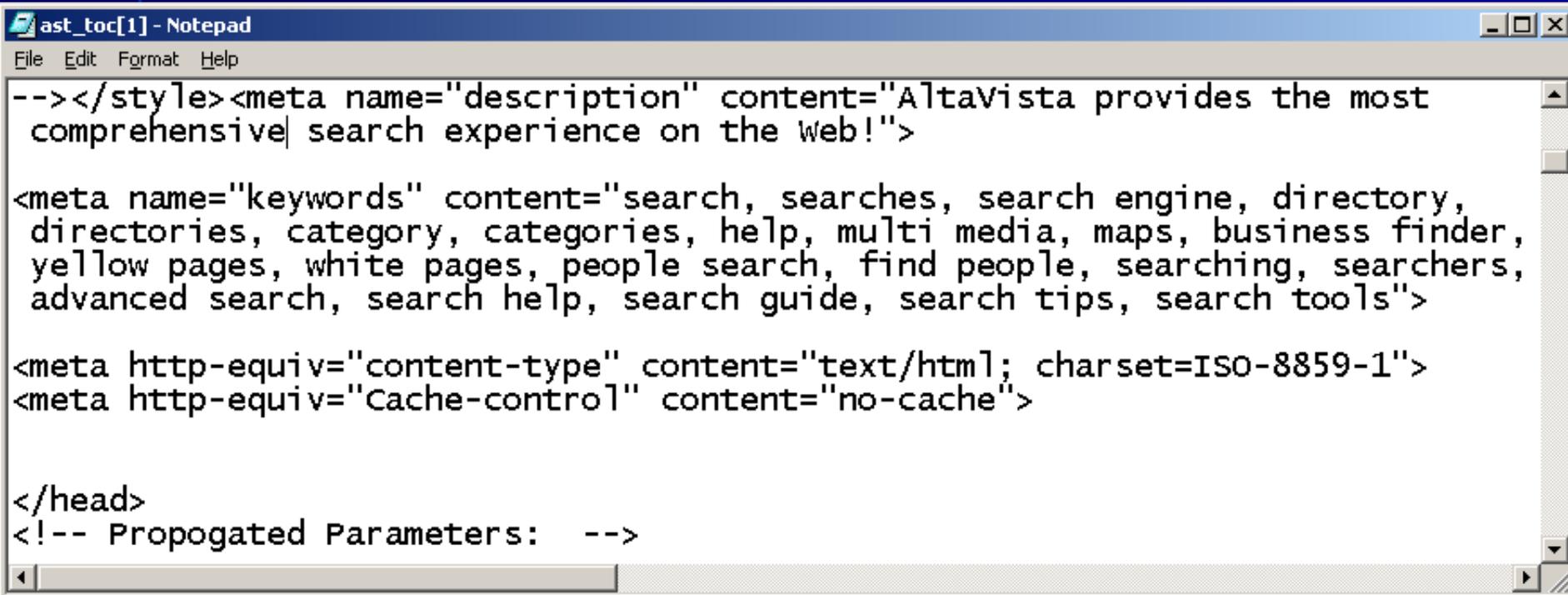

```
ast_toc[1] - Notepad
File   Edit   Format   Help

--></style><meta name="description" content="AltaVista provides the most
 comprehensive| search experience on the Web!">

<meta name="keywords" content="search, searches, search engine, directory,
 directories, category, categories, help, multi media, maps, business finder,
 yellow pages, white pages, people search, find people, searching, searchers,
 advanced search, search help, search guide, search tips, search tools">

<meta http-equiv="content-type" content="text/html; charset=ISO-8859-1">
<meta http-equiv="Cache-control" content="no-cache">


</head>
<!-- Propogated Parameters:  -->
```

Case Study - AltaVista

- Limit a search to a domain
- E.g. searching "edu" domain
- +domain:edu +"molecular biophysics"
- The above query would only search for molecular biophysics at educational institutions.
- Here is a <u>list</u> of Top-level Internet Domains

Searching Tips

- ## Be natural
  - Is cell phone harmful?
  - Ask the search engine : "Cell phone" AND harmful

- ## Capitalize
  - Always use lowercase
  - star will search "Star, STAR, stAr, …"
  - Type "Star" unless you really want to search "Star".

Searching Tips

- Use uncommon keywords
  - The more specific results will return to you.
  - Think a valid and uncommon keyword.
- Require words
  - Add a "+" before the keyword.
  - It will be in every match.
- Exclude words
  - Use "-" before the keyword.
  - In what situation should we use?

Searching Tips

- Correct Spelling
  - Beware of the differences between English and American spellings (Color, Colour) → (color OR colour)
- Stop words
  - Ignore the most common words "the, is, …"
  - "searching the web" and the search engine will ignore "the web".
  - Add more relevant keyword.

Searching Tips

- Use wildcards
  - Use "*" in some search engines.
  - "funk*" → funk, funky, funkiest, …
- Solve dead links
  - If the search engine returns http://www.hit.com/a/b/c.html which is a dead link.
  - You can try http://www.hit.com/a/b/
  - Or http://www.hit/com/a/ …

Factors affect your site's Ranking

- Keyword prominence
- Keyword frequency
- Keyword weight
- Keyword proximity
- Keyword placement
- Click popularity & Stickiness

Keyword Prominence

- How early in a web site do the keywords first appear?
  - The first element in HTML is the title tag
  - What happen if your title is:
    - This is my homepage
    - Welcome to my company's homepage
- Include the keywords in head, Meta tag, early in the body, …

Keyword Frequency

- Search engine may determines your site's popularity by checking how frequently the keyword or phrase appears on the page.

- What is the problem if you put too many same keywords into one single page?

Keyword Weight

- It is also called keyword density
- Measure by comparing the number of keywords appearing on the web page with the total number of words on the page.
- In most case, we try not to exceed a keyword weight of 3 to 10 percent.

Keyword Density

reference

reference

Keyword Proximity

- The placement of keywords on a web page in relation to each other is measured in "Keyword Proximity".

- "Home loans" will outrank a citation about "home mortgage loans".

- E.g.
  - *Smith Brothers Inc has been selling **puppy food** for over 50 years.*
  - *Smith Brothers Inc has been selling **food** for your **puppies** for over 50 years.*

Keyword Placement

- Search engines favor web sites that contain keywords in:
    - The title tag
    - The keyword META tag
    - The headline tag <H1> …
    - The first 25 words of body
    - Hyperlinks
    - Image <ALT> tags
    - Text near the end of the document

Click popularity & Stickiness

- **Click popularity** is a measure of the number of clicks received by each site in a search engine's results page.

- **Stickiness** is a measure of the amount of time a user spends at a site. It's calculated according to the time that elapses between each of the user's clicks on the search engine's results page.

- Reference: http://www.directhit.com/

Submit your site to search engines

- Google – 5 pages/day, Excite – 25 pages/week
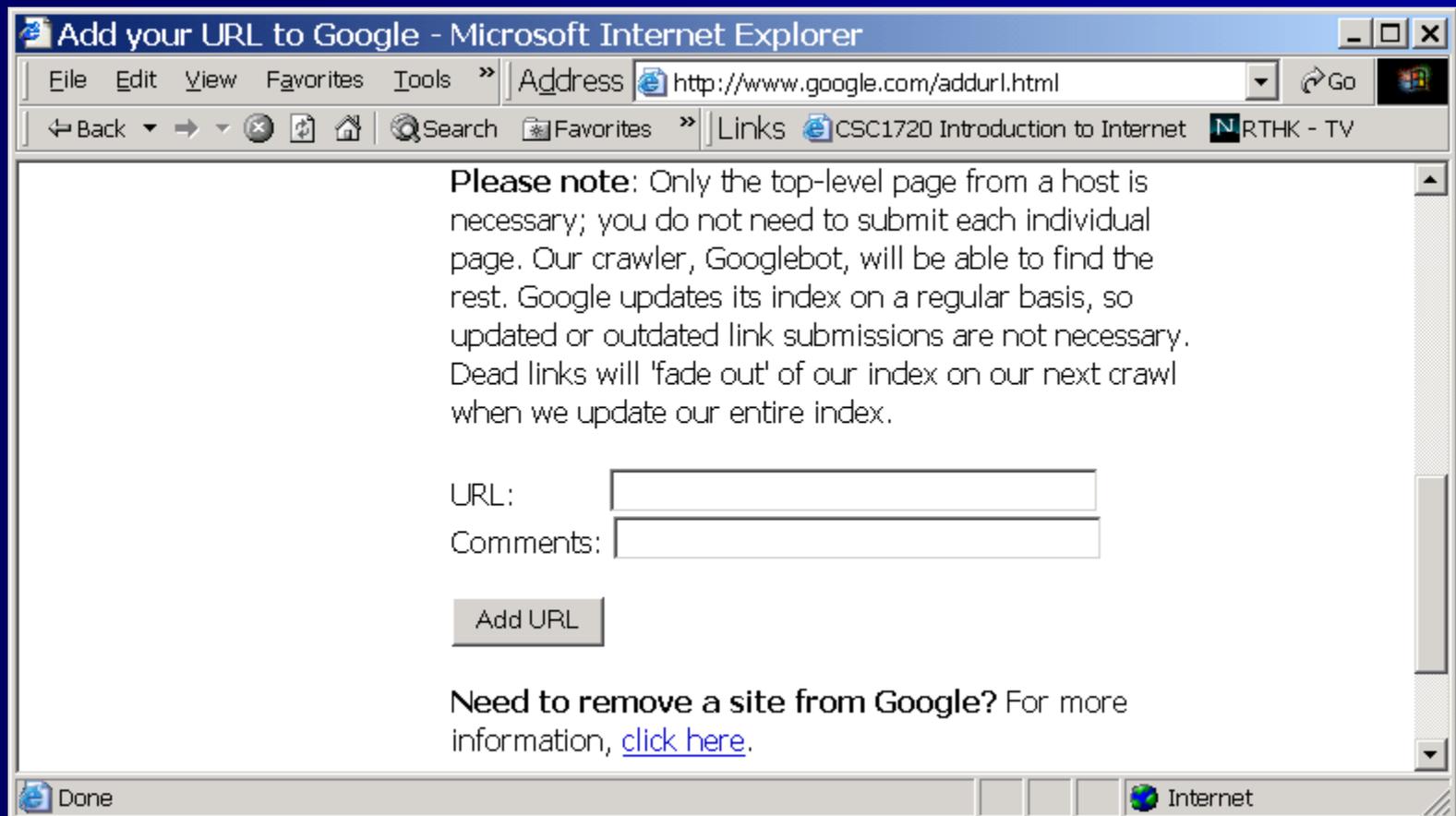
Summary

- Use different resources to find/search different kinds of information.

- Use successive query refinement to achieve effective search queries.

- Think carefully for the keywords typed in the search engine.

- Use Boolean queries when you need combinations of keywords.

- Think carefully when you create your own homepage, can it be easily indexed by search engines?

References

- searchenginewatch.com
- Information retrieval
- Search Engine Positioning – Fredrick Marckini (Wordware Publishing Inc.)

- The End.
- Thank you for your patience!