Internet Search Techniques

Finding What You Want on the Web Easily, Quickly, and (sort of) Effortlessly…

Presented by: Sharon Coward
Overview

Objectives:

• Understand the Web as a repository of information.
• Explore different search tools.
• Learn to use the tools appropriately.
• Evaluate search results.
Internet Search Technologies

• Internet = network of computers
• The Web = one of the services available via the Internet; interconnected documents & other resources, linked by hyperlinks and URLs
The Web – how big is it?

• Google: 5 million terabytes = 5 trillion megabytes of data.
• Google indexes only 200 terabytes i.e. 0.004%
• 2005 – 11.5 billion pages indexable web; doubles in size every 5 yrs.
Search Tools

• Search Engines
• Meta-search Engines
• Information Gateways
• Invisible/Deep Web
What is a “Search Engine”? 

1. A (computer) program that searches documents for specified keywords and returns a list of the documents where the keywords were found.

Often used to specifically describe systems like Google and Bing that enable users to search for documents on the World Wide Web.

http://www.webopedia.com
Search Engines

• Number of pages searched can vary
• Good results depend on using proper search syntax not just the scope of the engine’s coverage
• **Good For:**
  well defined topics to search; looking for specific sites; want a large number of websites returned for topic; retrieve particular types of documents, eg. Pdf

**Not Good For:** Browsing through a subject area.
Search Engines
## Search Engine Relationship

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Web Search Results</th>
<th>Sponsored Listings</th>
<th>Directory Listings</th>
</tr>
</thead>
<tbody>
<tr>
<td>allthetext</td>
<td>WEB</td>
<td>WEB</td>
<td>N/A</td>
</tr>
<tr>
<td>altavista</td>
<td>WEB</td>
<td>WEB</td>
<td>N/A</td>
</tr>
<tr>
<td>AOL</td>
<td>Google</td>
<td>Google</td>
<td>N/A</td>
</tr>
<tr>
<td>Ask</td>
<td>Google</td>
<td>Ask</td>
<td>N/A</td>
</tr>
<tr>
<td>bing</td>
<td>Bing</td>
<td>Microsoft adCenter</td>
<td>N/A</td>
</tr>
<tr>
<td>dmoz</td>
<td>dmoz</td>
<td>N/A</td>
<td>dmoz</td>
</tr>
<tr>
<td>Go</td>
<td>WEB</td>
<td>WEB</td>
<td>N/A</td>
</tr>
<tr>
<td>Google</td>
<td>Google</td>
<td>Google</td>
<td>dmoz</td>
</tr>
<tr>
<td>HotBot</td>
<td>WEB</td>
<td>WEB</td>
<td>N/A</td>
</tr>
<tr>
<td>iwon</td>
<td>Ask</td>
<td>Ask</td>
<td>N/A</td>
</tr>
<tr>
<td>Lycos</td>
<td>WEB</td>
<td>WEB</td>
<td>N/A</td>
</tr>
<tr>
<td>Netscape</td>
<td>Google</td>
<td>Google</td>
<td>N/A</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>Yahoo!</td>
<td>Yahoo!</td>
<td>Yahoo!</td>
</tr>
</tbody>
</table>
Meta Search Engines

- Skim-search several search engines at once
- Usually reach about 10% of results of each engine they visit
- Cannot perform advanced-style searches which use engine-specific syntax
- **Good For:**
  quick search engine results overview, doing simple searches with 1 or 2 keywords; want a small # relevant results; problems finding what you want; convenient to search different content sources from one page
- **Not Good For:** comprehensive results from a complex search
Meta-Search Engines

Dogpile - [www.dogpile.com]; Metacrawler - [www.metacrawler.com]
Meta Search Engines

SurfWax – www.surfwax.com
Meta Search Engines

- Copernic - http://find.copernic.com
Information Gateways

- Subject directories, virtual libraries
- Compiled by people, not robots
- Subject categories
- More focus on sifting for relevance and quality

Good For: you have a clear topic but not unique keywords; browse for ideas

Not Good For: Quickly finding information from widely varying themes
http://dir.yahoo.com/
<table>
<thead>
<tr>
<th>Arts</th>
<th>Home</th>
<th>Science</th>
<th>Shopping</th>
<th>Society</th>
<th>Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music, TV</td>
<td>Consumers, Homeowners, Family</td>
<td>Biology, Psychology, Physics</td>
<td>Auto, Clothing, Gifts</td>
<td>Issues, People, Religion</td>
<td>Basketball, Football, Soccer</td>
</tr>
<tr>
<td>Business</td>
<td>News</td>
<td>News</td>
<td>Recreation</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>Industrial,</td>
<td>Media, Newspapers, Current Events</td>
<td>Food, Outdoors, Travel</td>
<td>Education, Libraries, Maps</td>
<td>Asia, Europe, North America</td>
<td>Asia, Europe, North America</td>
</tr>
<tr>
<td>Hardware,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Games</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roleplaying,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitness,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>World</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deutsch,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Espanol,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Francais,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italiano,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Korean,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nederlands,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polski,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Svenska</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Google Directory - [www.google.com/dirhp](http://www.google.com/dirhp)
Information Gateways

- Questia.com - [http://www.questia.com/](http://www.questia.com/) (full text online library >70,000 books)

- ELDIS: the Gateway to Development Information [http://www.eldis.org/](http://www.eldis.org/) (4000 sites)

- Open Learn [http://openlearn.open.ac.uk/](http://openlearn.open.ac.uk/) (open university course materials)

- Open Directory Project [http://www.dmoz.org](http://www.dmoz.org) (largest human edited directory)
Invisible/Deep Web

- 91,000 terabytes vs 167 terabytes in surface Web

- Search engines can’t access content – databases, non-text files; password protected areas; dynamic content
Invisible/Deep Web

- Dynamic content: - returned in response to a submitted query or accessed only through a form,
- Unlinked content: pages which are not linked to by other pages
- Private Web: sites that require registration and login (password-protected resources).
- Searchable - Entry pages can be found using other search tools; include term ‘database’ in search

Good For: Gathering specific kinds of data
Not Good For: Browsing through a subject area
Invisible/Deep Web
Deep Web Expanding

- Blog postings
- Comments
- Discussions and other communicative activities on social networking sites
- Bookmarks and citations stored on social bookmarking sites
Invisible/Deep Web

- www.science.gov - over 40 databases, 1,950 selected websites, 200 million pages
- http://www.deepwebtech.com/
- http://infomine.ucr.edu/
- http://www.completeplanet.com/ - 70,000+ databases
Search Strategies

1. Identify your concepts
2. Make a list of search terms/keywords for each concept e.g.
   global warming/ greenhouse effect
   greenhouse gases/ climate change
3. Specify the logical relationships among your search terms
4. Be specific – golden retriever vs dog
Search Strategies

Google

Advanced Search
Language Tools

Google Search  I'm Feeling Lucky
Advanced Search

Google Advanced Search

Find web pages that have...
all these words: 
this exact wording or phrase: 
one or more of these words: rain OR snow OR
Boolean Operators

**AND**
- Articles will contain mineral and deposit

**OR**
- Articles will contain either mineral or ore

**NOT**
- Articles will contain mineral deposit but not mining
Advanced Techniques

• Use quotation marks “…” to specify exact phrases:
  “internet marketing”

• Use the plus (+) and minus (-) sign to include and exclude words:
  “internet marketing  +facebook” vs “internet marketing  –facebook”
Advanced Search

- Wildcard (*). Searching for "looking for *" will return results that have the words "looking for dogs", "looking for cats" etc.
- Stop words ignored -“a, an, the, and”. Use a + or – symbol to force the engine to include those words in your search.
Advanced Search

• ~ Similar Words

Search for similar words, or synonyms. Searching search ~tips will return results with 'help', 'guide', 'tutorial' etc.
Advanced Search

- **Web page title**
  `allintitle:jamaica`

- **Website or domain**
  `site:barbados.org "beach hotel"`
Advanced Search

Date, usage rights, numeric range, and more

- Date: (how recent the page is) — anytime
- Usage rights: — not filtered by license
- Where your keywords show up: — anywhere in the page
- Region: — any region
- Numeric range: —

SafeSearch: Off On

(e.g. youtube.com, edu)

Caribbean Tourism Organization CARIBBEAN™
Advanced Search

• File type
define:pixel
define:“due diligence”
Advanced Search

- **Truncation**: Searches on the root of the word adding different word endings or plurals.

  *Educat* searches *educator, education, educational, educated.*

  *Colo* would find documents that contain color and colour.
Examine Results

**Authority:** Who owns the site? Credentials? .gov, .edu, .mil, .org usually more reliable than .com

**Currency:** How up-to-date is the information? Check to see if you can tell when the last time the site was updated.
**Links**

- **List of search engines:** http://thesearchenginelist.com
- **Search:** www.google.com; www.yahoo.com; www.bing.com; www.ask.com
- **Meta search:** www.dogpile.com; www.metacrawler.com; www.clusty.com; http://find.copernic.com; www.surfwax.com;
Assignment

• When was the first CTC and where?
• Find 5 job vacancies in international organizations based in the Caribbean
• Find the top sites in health, entertainment, news and trade in the Caribbean
• Find photographs of the earliest Catholic missions to 3 Caribbean countries.