Week 9 - Searching for information (Chapter 5)

Searching for information : perhaps the most important operation (task) on the web today.

Outline

- Types of searches
- resources available, strategies for searching
- search engines, how they work
- assessing credibility of sources

Types of web searches (queries)

1) A Voyager question : open-ended exploratory question. Topic → generally unknown to you, you are willing to be educated.

(open-ended → it is not clear when the question has been answered & exploration can stop)

↔ collect information
↔ not very specific about the info you look for

ex (for me) : How to care for a bonsai tree.

2) A Deep Thought question : open-ended but more focused & goal oriented. It might have many possible answers

(origin : Douglas Adams "The Hitchhiker's guide to the galaxy"
British writer & humorist
In the book, a computer called "Deep Thought" sets out

to learn the meaning of life.)

(ex) - whenever you wish to collect multiple hypotheses, opinions, or perspectives on an issue

3) A Joe Friday question : a specific question with an expected simple, straightforward answer.

(ex) - questions about names, dates, locations, etc...

(Origin : '50's TV show "Dragnet" where the character Joe Friday, a policeman, was famous for the line "The facts ma'am, just the facts".)

(source : W. Lehnert & R. Kopec, "Web 101 - 3rd Edition", 2008)

Types of resources available for queries

a) A subject tree (directories or topic hierarchies)

-> websites & online documents grouped by topic

-> topics : organized hierarchically

b) A clearinghouse

-> a collection of web sites & documents on a topic

-> similar to subject tree but the focus is more narrow. It may or may not provide a subject tree to assist browsing.

c) A generalized search engine
d) A specialized search engine (focused on a specific topic).

Type of question	Resource
Voyager	subject tree, clearinghouse
Deep Thought	subject tree, specialized search engine
Joe Friday	subject tree, general search engine

Subject Trees

- ex: yahoo directory (http://search.yahoo.com/dir)
  (oldest, largest, most popular subject tree)

→ list of topics organized in a tree like hierarchy:

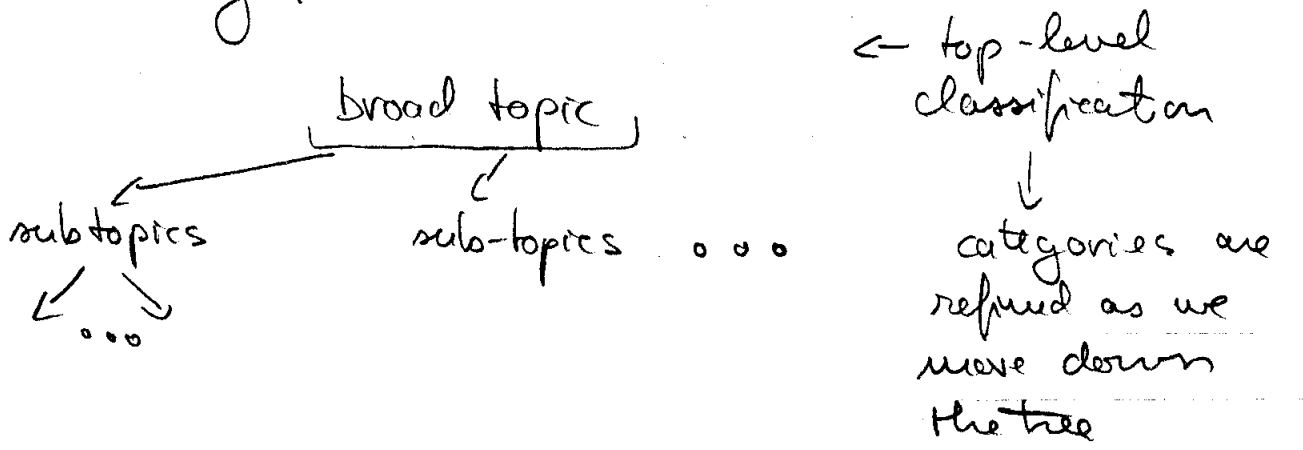


Fig. Yahoo's web resources on Britney Spears started @ Directory → Entertainment → Music → Artists → By Genre → → Rock & Pop → Britney Spears

- the information in the directory is reviewed & maintained by humans
- directory tree contains cross references (a particular topic may be relevant to several categories)
- the subject you're looking for may be listed under several categories
  - ↳ site search may be useful = search who's results (hits) are restricted to pages within the current web site
  - → category search = site search within a directory tree
    - Tips – use one keyword / search
      use simple & obvious keywords
      if several keywords are relevant, search for them one by one.

Other directory trees

- About.com
  - → resources are handpicked by experts
  - → documents can be trusted to contain high quality information
  - → less extensive tree as yahoo, good for introductory

articles & tutorials on the topics covered

- Open Directory Project (dmoz.org)
  - → focused on practical knowledge rather than academic
  - → contributed by volunteers.

- Clearinghouses
  - → maintained by researchers (public funds) or librarians
  - → contain high quality information
  - → finding clearinghouses : clearinghouse index
    - (ex) • The Internet Public Library (ipl.org)
      - The Reference Desk (martindalecenter.com)
      - BIOTN (best info on the net)
        library.sau.edu/bestinfo
      - check University's library & public library websites ...
      - Internet Scout Project
        (www.scout.wisc.edu)
      - Netsurfer Science
        (netsurf.com/nss) → not working when I tried last time
      - Infosyssec.com (computer security)

Search Engines → computer programs to help us find info.

→ based on keyword searches: a list of web pages is returned that are relevant to the keywords given as query.

- Query: list of keywords connected by logical operators
  (ex) trails AND Nova AND Scotia

  we are interested in documents (URL's) associated with all three keywords
  - the typical query when using Google's search engine (keywords connected with AND operator)
    eg: trails Nova Scotia

  - other operators — OR
    NOT (also − minus)

    (ex) vacation (Caribbean OR Mexico)
    ↑ AND
    (we are interested in documents associated with "vacation" and either "Caribbean" or "Mexico".

- Tip → get to know the search engine... Each engine is different.

Ex → investigate google "advanced search" options

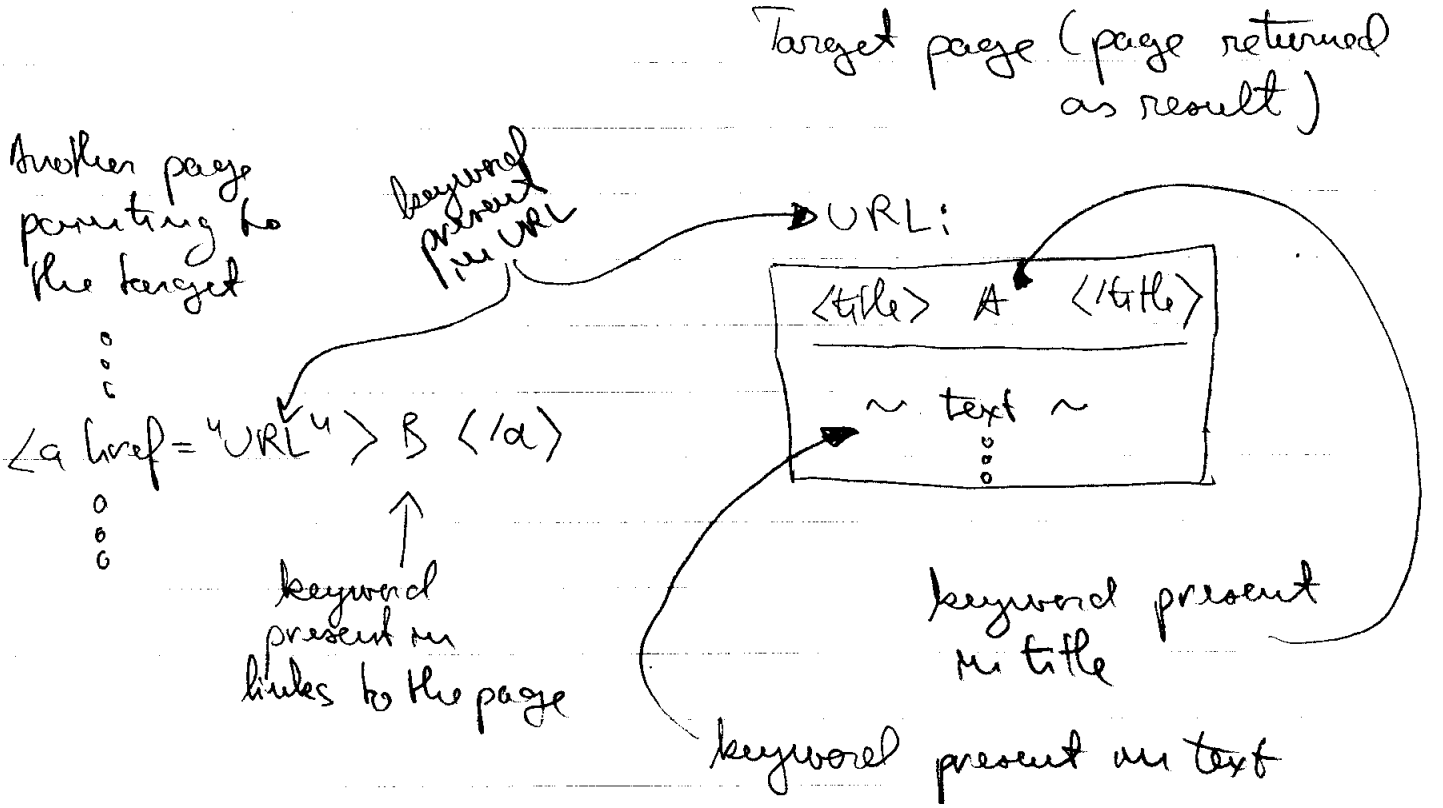
Google searches

- filters out by default common words like "is" "a" "the" etc
- exclude keywords using − (minus)
  water −mineral
- force inclusion of keywords using + sign
  Star Wars +I (otherwise capital I is excluded)
- phrase searches: "bald eagle" is interpreted as a text fragment & not as keyword search
  bald AND eagle
  vs: the text "bald as an eagle" should not be matched by the phrase search.

narrowing search down

- searching within results
  → can be used to refine a search
  → the new query is applied only to pages returned by previous query
- limit results to − pages from a specific domain (web site)
  − pages written in a certain language
- specify where keywords appear in document
  → anywhere (no restriction)
  → in title of page
  → in the URL of the page
  → in links to the page.

About Location of keywords :



(4)

How search engines work

- Information retrieval is a branch of Computer Science dealing with finding info in large text data bases.

- search engines = fast because
  a) web documents = saved in search engine's database
  b) documents = indexed
     index, like for a book = list of words & pointers to their occurrence in the document.

- components of search engine, or crawler
  → web spider (or robot): a program that continuously downloads web pages & inspects their content → to discover new links
                                                                  ↘ to index the document
  → query processor: a program with a web interface that processes queries & looks in the indexed collection of pages constructed by the spider
    → fast operation (just lookup)
    → slow operation, but does not affect the users...

Indexing methods

- selective text : only parts deemed important from a document are scanned for keywords & indexed
  ex: title, links, headings
- full text : whole document is indexed. Still keywords found in special parts of document may have heavier weight or be more important (used for ranking)

Ranking of pages

= the order in which results are displayed to the user is "most relevant hits first"

Relevance → a search engine's guess of user intent.

Page rank strategies

- the higher the # of occurences (or the total weight) of queried keyword in a doc, the higher the relevance of that page.
  ex: query keyword in title → page more relevant.
- ranking by # of links from other pages pointing to doc. in question (popularity)
  → a page with larger # of links pointing to it from outside is considered more relevant.
- using special tags in html doc. that are not displayed by browser, but are used by

designers to tag their docs. manually.

Ranking & indexing → extremely important for search engines & have far reaching consequences economically & politically
→ check: searchenginewatch.com

Extra reads:

searchengineshowdown.com
suite101.com/reference/search_engine_reviews
wikis.ala.org → search for "Toolkit for the expert web searcher".

Other search engines:

- ASK.COM
  - -> in addition to a collection of web documents it also has a database of hand-picked queries & questions that are also searched based on keyword & returned to the user
  - -> good site for Joe Friday questions

- meta search engines
  → sites sending a query to several different search engines simultaneously

Motivation : some studies showed little overlap between pages indexed by different search engines

ex of meta search engines
- Bramboost, Dogpile, Excite, info Grid, Infonetware, ixquick, kartoo ....

Tips for efficient search

1) Think about the type of page thought for
   - home page of a person?
   - organization? compilation of resources?
2) Think about the author:
   → restrict search to a domain...
3) List terms likely to appear on page sought.
   → use boolean operators
4) Assess the results; modify the query
   → try to eliminate irrelevant pages
   → make query more general if few results
   → consider a 2 pass strategy (search within results).

invisible web (deep-web, deep net)
→ content not accessible to spiders
→ ex: • dynamically generated pages (most products bought online, etc)
• docs with restricted access
• databases, pages obtained as a result of meaningful (human) queries.
↳ to search "invisible web" means to find appropriate databases
eg: plane crash databases, toxic substances databases, etc

Assessing credibility of the information

→ Rule no 1 : do not assume too much.

- check publishers of web page
  - → use a whois client
  - → whois service returns contact info given when an organization wishes to register a domain name (with DNS servers) & assigns an IP address with it

ex: "search whois on google"

From whois info

- → address
- → contact name / tel # → try whitepages.com to verify
- → email address
  - ↳ extract domain name from e-mail & visit associated website if it exists

→ find the author of the article or info.
↳ gather information about the author (search engines)

→ check other sources to verify the info.

→ check grammar, spelling, & graphical design of webpages.

Wikipedia

- criticism → subject to erroneous entries ranging from technical errors to blatant misrepresentation of truth

(ex) Brian Chase in a prank on one of his colleagues wrote a Wikipedia article on John Seigenthaler (prominent US journalist) accusing him of involvement in assassination of JFK. Article was up for more than 100 days until finally corrected

→ Nature Journal : study comparing Encl. Britannica & Wikipedia found same # of errors in a sample of articles

Criticism of study
- small sample of articles reviewed
- bias in choosing articles
- bias in what the study defined as "error"

→ Thomas Chesney (Nottingham Univ)
- rating Wikipedia articles by 2 groups of people → experts
  → non experts

Result : articles ranked higher by experts rather than non-experts.

-obs- small sample size ...

Concl → Wikipedia : good source for infos, but avoid citing it in scholarly works before verifying the infos.