Topic 5: Searches and Information Credibility

Eric Mlinar
CS 6
4/11/2016
Pre-Lecture

- Quizzes
  - Typically graded in a week
  - Quiz #1 grades by Friday (hopefully)
  - Grades straight to iLearn
Overview

• Searches
  • *Find something on the world wide web*

• Information Credibility
  • *Trusting what you find on the WWW*
Overview

• Searches
  • *Find something on the world wide web*

• Information Credibility
  • *Trusting what you find on the WWW*
Vocabulary

• **Query**: your question, or search argument

• **Search engine**: accesses database to return answer

• **Crawler/Spider**: programs or algorithms that search engine sends out to explore the web via links

• **PageRank**: how important a page is, given the pages that link to it
Query Types

• Informational – your standard search
• Navigational – lazy site finder
• Transactional – this is the object I want

...  

• Connectivity – (unusual)
Info Queries

• “Joe Friday” question

• “Voyager” question

• “Deep Thought” question
Info Queries

• “Joe Friday” question
  • When was the War of 1812?

• “Voyager” question

• “Deep Thought” question
Info Queries

• “Joe Friday” question
  • When was the War of 1812?

• “Voyager” question
  • What was the life of a soldier like in the War of 1812?

• “Deep Thought” question
Info Queries

• “Joe Friday” question
  • When was the War of 1812?

• “Voyager” question
  • What was the life of a soldier like in the War of 1812?

• “Deep Thought” question
  • What caused the War of 1812?
Web Crawler

• Collects text and “metadata” that we search over
  • Metadata: data which provides information about other data (e.g. library card catalogs)

• A search is not actively scanning the internet, typically just browsing an archive
  • Cached index of WWW, premade
Search Resources

• Search Engine
  • General, specialized, meta
• Web directories
  • A.k.a. subject trees
• Clearinghouse
  • Less common nowadays
Search Engines

• General Search Engine
  • Google, Yahoo (ugh), … AskJeeves, anyone?
  • DuckDuckGo if you prefer privacy
  • Most common, most used
    … heck, mostly the most useful!
  • Advanced search parameters helpful

• Specialized Search Engine
  • LegalEngine.com (for lawyers: rulings, etc.)
  • www.loc.gov (Library of Congress)
Search Engines

• Meta search engines
• *Search over search engines*
• *Useful perhaps if you aren’t sure of your usual search engine, but a dying breed*
Web Directories

• Subject trees (user-created)
• *World Wide Web Virtual Library* (without spaces)
• “Browsing” Yahoo (R.I.P.)
• *dmoz.org* – open directory project

• Usually not found by web crawlers
• Evolving with wiki-structure
Web Directories

![Dmoz Logo](image)

- **Arts**
  - Movies, Television, Music...
- **Games**
  - Video Games, RPGs, Gambling...
- **Kids and Teens**
  - Arts, School Time, Teen Life...
- **Reference**
  - Maps, Education, Libraries...
- **Shopping**
  - Clothing, Food, Gifts...
- **Business**
  - Jobs, Real Estate, Investing...
- **Health**
  - Fitness, Medicine, Alternative...
- **Regional**
  - US, Canada, UK, Europe...
- **Society**
  - People, Religion, Issues...
- **Computers**
  - Internet, Software, Hardware...
- **Home**
  - Family, Consumers, Cooking...
- **Recreation**
  - Travel, Food, Outdoors, Humor...
- **Science**
  - Biology, Psychology, Physics...
- **Sports**
  - Baseball, Soccer, Basketball...

CS 6
ERIC MLINAR
Clearinghouse

• Gathers articles on a specific topic
• Gathered by domain experts
• Similar to web directory
  • http://hrw.org/ (Human Rights Watch)
  • http://ec.wustl.edu/ (Environmental Clearinghouse)
• Clearinghouse Index, to find the right one
  • R.I.P.  http://ipl.org/
Search Guidelines

• “Successive query refinement”
• Better to rephrase and search again than to look too deeply

• Refining search:
  • Use more specific terms
  • Use phrases
  • Think about “who” behind the info
Search Directives

• Sometimes basic searches insufficient
• Search directives / parameters: add code-like behavior within search box
  • *Most common: invisible AND*
  • *Also well-known: wrapping in quotes*
Google Search

• Boolean Math (AND, OR)
• Quotes
• Wildcard (*)
• Include (+)
• Exclude (-)
• Synonyms (~)
Google Search

• Specific searches by…
  • site: riverside site:weather.com
  • related: related:cbsnews.com
  • info: info:google.com
  • intitle: intitle:whales
  • inurl: inurl:poker
  • Or you can use Advanced Search
  • www.google.com/help/operators.html
Google Tools

• Scholar (scholar.google.com)
• Local.google.com
• Video.google.com
• Images.google.com
• Books.google.com
• Blogs.google.com
The Deep Web

- Parts of the world wide web not accessed or scanned by search engines
- Only a fraction of the visible web (estimated 4%) is recovered by searches
- The rest is the “Deep Web”
The Deep Web

• Consists of:
  • *Dynamically generated pages*
  • *Pages that prohibit crawlers*
  • *Unlisted and unlinked sites*
  • *Company “cloud” storage*

• Dark Web
  • *Be anonymous*…
Overview

- Searches
  - Find something on the world wide web

- Information Credibility
  - Trusting what you find on the WWW
Questions

1. Is the information accurate?
2. Is the author credible?
3. Who is the intended audience?
4. What is the purpose of publishing this?
5. Is it current?
6. Is it superficial or deep?
Wiki-style

• Pages (content!) can be edited by users

  Sometimes users are restricted

  Most now are completely open

  Until you’re banned…

• Folksonomies: collaborative categorization, e.g. using “tag words”

  Flickr.com

  StumbleUpon.com

  digg.com
Wikipedia

• Spoiler about the discussion of Wikipedia vs. Encyclopedia Brittanica
  • Wikipedia is nearly as good
  • Slightly more errors (omissions or misleading statements)

• Make wikis better? Encourage editing
• Scientific topics are generally safe
• More subjective = more contention
Quoting the Web

• In general, bad practice
• Academic Journals
• Bibliographic databases
• Large News Outlets
• Online content still more subject to error
• Governmental sites
• Everyone should be quoting their source
The End

That's all Folks!
The Deep Web

• Consists of:
  • *Dynamically generated pages*
  • *Pages that prohibit crawlers*
  • *Unlisted and unlinked sites*
  • *Company “cloud” storage*

• Dark Web
  • *Be anonymous…*
• “Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say” – Edward Snowden