CS 312
Internet Concepts

Web Applications:
Internet Search and Digital Preservation

Dr. Michele Weigle
Department of Computer Science
Old Dominion University
mweigle@cs.odu.edu

http://www.cs.odu.edu/~mweigle/CS312-F11/

Outline

- Generic Searching
- How Google Works
- Digital Preservation
  - » Internet Archive
  - » Memento
  - » Archive Facebook

Generic Search Engines

- Some very popular ones
  - » Google
  - » Bing
  - » Yahoo!
- Some common characteristics
  - » Huge set of results
  - » Extremely prompt retrieval
  - » Organized retrieval results

How do they achieve such amazing results?

How Do Search Engines Work?

- Prep work
  - » *crawling* – an automated way of browsing the web for pages
  - » archiving, analyzing, organizing, indexing
- Retrieval
  - » query matching
  - » ranking search results
  - » displaying search results

Search Query Syntax

- The space between keywords is interpreted as AND for some search engines, but as OR for others.

- Query: *I'm interested in cancer in adults.*
  - » Boolean logic: AND
  - » Search query: +cancer +adults

- Query: *I'm interested in radiation, but not nuclear.*
  - » Boolean logic: NOT
  - » Search query: radiation -nuclear

Outline

- Generic Searching

- How Google Works

- Digital Preservation
  - » Internet Archive
  - » Memento
  - » Archive Facebook

Google's Three Main Components

- Crawling
- Indexing
- Page Weighting

http://ppcblog.com/how-google-works/

Google's Crawler, the Googlebot

- Consists of many computers requesting and fetching pages
  - » Googlebot can request thousands of different pages simultaneously
- When Googlebot fetches a page, it adds all of the links in the page to a queue for subsequent crawling (*deep crawling*)
  - » Googlebot can quickly build a list of links that can cover broad reaches of the web
- Google continuously recrawls popular frequently changing web pages at a rate roughly proportional to how often the pages change (*fresh crawls*)

Submitting URLs to Googlebot

- http://www.google.com/addurl.html
- Rejects URLs that Google suspects are trying to deceive users
  - » including hidden text or links on a page
  - » stuffing a page with irrelevant words
  - » cloaking (aka bait and switch)
  - » using sneaky redirects
  - » creating doorways, domains, or sub-domains with substantially similar content
  - » sending automated queries to Google
  - » linking to bad neighbors.

Google's Indexer

- Parses webpages discovered by the Googlebot
- Converts them into a set of hits (word occurrences).
  - » For each document in which the word appears, indexer stores the position of the word, font size, capitalization
- Creates an *inverted index* that is indexed by word pointing to documents

Forward Index

Document	Words
Document 1	the,cow,says,moo
Document 2	the,cat,and,the,hat
Document 3	the,dish,ran,away,with,the,spoon

Inverted Index

Word	Documents
the	Document 1, Document 3, Document 4, Document 5
cow	Document 2, Document 3, Document 4
says	Document 5
moo	Document 7

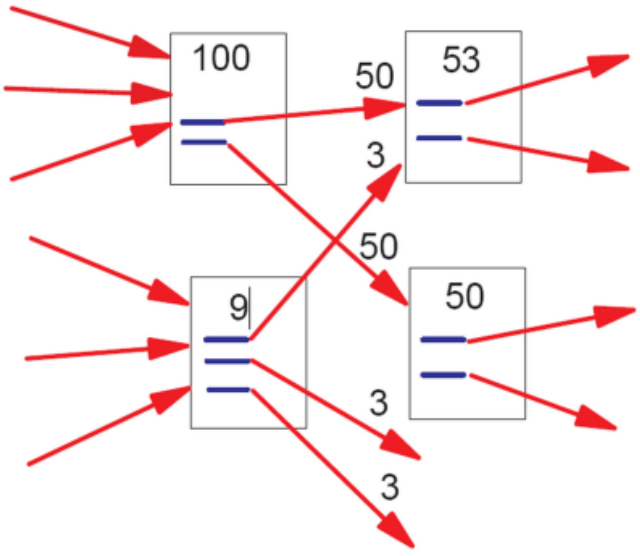
Google's Page Weighter, *PageRank*

- Assigns each webpage a relevancy score
- A webpage's PageRank depends upon
  - » frequency and location of keywords within the page
  - » how long the webpage has existed
  - » number of other webpages that link to the page
- Most important factor is number of pages that link to the page
  - » essentially, PageRank is a "vote", by all the other pages on the Web, about how important a page is

http://computer.howstuffworks.com/google1.htm

PageRank Example

PageRank: Number and importance of links pointing to you



http://www.mattcutts.com/blog/seo-for-bloggers/ Time 6:20-8:30

Manipulating PageRank / Webspam

- Google bomb
  - » is created if many sites link to the page using the same anchor text (increases the page rank)
  - » ex: In 1999, a search for "more evil than Satan himself" resulted in the Microsoft homepage

- spamdexing
  - » deliberately modifying HTML pages to increase the chance of their being placed close to the beginning of search engine results

Manipulating PageRank / Webspam

- link doping
  - » embedding a large number of gratuitous hyperlinks on a website, in exchange for reciprocal links

- Google Jacking (page hijacking)
  - » creating a rogue copy of a popular website which shows contents similar to the original to a web crawler, but redirects web surfers to unrelated or malicious websites.

How to Improve Your PageRank

- Search Engine Optimization (SEO)
- Straight from Google
  - » http://www.mattcutts.com/blog/seo-for-bloggers/
  - » keyword tool (time 17:04-18:18)
    - ❖ http://adwords.google.com/select/KeywordToolExternal
  - » be relevant, but don't overdo it! (time 23:11-25:24)

Outline

- Generic Searching
- How Google Works
- Digital Preservation
  - » Internet Archive
  - » Memento
  - » Archive Facebook

Digital Preservation

- Much of the record of our lives is digital.
  - » How do we keep these records?
  - » How do we "walk down memory lane"?
- Backup strategies
- Recording and archiving the web
- Accessing past versions of webpages

Web Science and Digital Libraries Research Group at ODU (http://ws-dl.blogspot.com/)

Internet Archive

- Internet Archive
  - » non-profit that was founded to build an Internet library, with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format
  - » http://www.archive.org
- Internet Archive's Wayback Machine
  - » allows you to browse through 85 billion web pages archived from 1996 to a few months ago
  - » http://www.archive.org/web/web.php

Memento

- Time Travel for the Web
  - » http://www.mementoweb.org/
- MementoFox – Firefox add-on
  - » https://addons.mozilla.org/en-US/firefox/addon/100298/

Demo

- Uses Internet Archive and others to present a webpage as it existed at a certain date/time

Web Science and Digital Libraries Research Group at ODU (http://ws-dl.blogspot.com/)

Archive Facebook

- Grab a stand-alone archive of your Facebook account
- Archive everything *you* want instead of what *Facebook* wants
- Preserves FB look-and-feel

- http://bit.ly/archivefb

http://www.slideshare.net/matkelly01/ndiippndsa-2011-archive-facebook

Archive Facebook

Content Dump versus WYSIWYG

Outline

- Generic Searching
- How Google Works
- Digital Preservation
  - » Internet Archive
  - » Memento
  - » Archive Facebook