Web Search Basics

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:
2. Raymond J. Mooney’s teaching materials
The World Wide Web (Web)

- Created in 1989 by Tim Berners-Lee at CERN (in Switzerland)
- An environment of accessing to interlinked and hypertext documents via the Internet
  - **Client-server design** for transfer text, images, videos, and other multimedia, encoded with **html** (hypertext markup language), via a protocol (**http**, hypertext transfer protocol)
    - The client side is usually a browser, a GUI environment, sending an http request to a web server (by specifying a URL, universal resource locator)
    - Asynchronous communication

```
http://www.ntnu.edu.tw/information/contact.html
```

**Domain**
Web Characteristics

- **No Control**: democratization of creation and linking (publishing). Content includes truth, lies, obsolete information, contradictions.
- **Distributed Data**: Documents spread over millions of different web servers...
- **Heterogeneity**: Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (databases)...
- ** Variety of Languages**: The types of languages used are more than 100.
- **Large Volume**: Scale much larger than previous text corpora (slowed down from initial “volume doubling every few months” but still expanding).
- **Volatile Data**: content can be dynamically generated and removed.
Rapid Proliferation of Web Content

- Total Web Sites Across All Domains August 1995 - November 2007 (http://news.netcraft.com)

- A large fraction of growth in sites has come from the increasing number of blogging sites (in particular at Live Spaces, Blogger and MySpace) in the recent past
Internet Users by Languages

• End of 2004, total 801.4 millions

- English, 35.90%
- Chinese, 13.20%
- Japanese, 8.30%
- German, 6.80%
- Spanish, 6.70%
- French, 4.40%
- Korean, 3.80%
- Italian, 3.60%
- Portuguese, 2.90%
- Dutch, 1.70%
- All Others, 12.70%
Access to Web Content

• Full-text index search engines
  – E.g., Google, Altavista, Excite, Infoseek, etc.
  – Keyword search supported by inverted indexes and ranking mechanisms

• Manual hierarchical taxonomies (directories) populated with web pages in categories (i.e., portal sites)
  – E.g., Yahoo!, Yam, etc.
  – Human editors assemble a large hierarchically structured directory of web pages
  – Users browse through trees of category labels
Growth of Web Pages Indexed

Assuming 20KB per page, 1 billion pages is about 20 terabytes of data.

- This slide is adopted from Raymond J. Mooney’s teaching materials

SearchEngineWatch Link to Note from Jan 2004

Google Inktomi AllTheWeb Teoma Altavista
Rate of Change for Web Pages

- Fetterly et al. study (2002): several views of data, 150 million pages over 11 weekly crawls
  - Bucketed into 85 groups by extent of change
Frequency of Using Search Engines

How often do you use search engines on the Internet?

- Four or more times each day: 21.2%
- At least once every day: 35.1%
- Several times each week: 22.7%
- At least once each week: 10.3%
- Several times each month: 5.5%
- Less frequently: 3.9%
- Never: 1.2%

http://www.iprospect.com
User Query Needs (1/4)

• User query roughly fall into three categories
  – Informational – want to learn about something
    • E.g., “Taroko”
  – Navigational – want to go to that page
    • E.g., “China Airlines”
  – Transactional – want to do something (web-mediated)
    • Purchasing a product, downloading a file or making a reservation

Discern which of these categories a query falls into can be challenging!
User Query Needs (2/4)

• Ill-defined queries
  – Short
    • 2001: 2.54 terms avg, 80% < 3 words
    • 1998: 2.35 terms avg, 88% < 3 words
  – Imprecise terms
  – Suboptimal syntax
  – Low effort

• Specific behavior
  – 85% look over one result screen only (mostly above the fold)
  – 78% of queries are not modified (one query/session)

• Wide variance in
  – Needs
  – Expectations
  – Knowledge
  – Bandwidth
User Query Needs (3/4)

- Query Distribution

- Power law: few popular broad queries, many rare specific queries
User Query Needs (4/4)

- How far do people look for results?

“When you perform a search on a search engine and don’t find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)”

- After reviewing the first few entries: 12%
- After reviewing the first page: 16%
- After reviewing the first 2 pages: 27%
- After reviewing the first 3 pages: 25%
- After reviewing more than 3 pages: 20%

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)
Web Search Engines (1/2)

• Goal
  – Return both high-relevance and high-quality (i.e., valuable) pages
    • Given the heterogeneity of the Web and the ill-formed queries

• Architecture: main constituents
  – Crawler
  – Indexer
  – Query Server

IR – Berlin Chen 14
Web Search Engines (2/2)

• Crawler
  – Collect pages from the Web
  – Done by distributed crawlers
    • URL server sends lists of URL to be fetched by crawlers
    • Store server compresses and stores pages (full HTML texts) into a repository
      – Duplicate content detection

• Indexer
  – Process the retrieved pages/documents and represent them in efficient search data structures (inverted files)

• Query server
  – Accept the query from the user and return the result pages by consulting the search data structures
Hyperlink and Anchor Text (1/2)

• Web as a Directed Graph - Two intuitions
  – Hyperlinks from a web page as a form of conferral of authority
    • I.e., A hyperlink between pages denotes author perceived relevance (quality signal)

  – The anchor (text) of the hyperlink describes the target page (textual context)
    • A short summary of the target page

  <a href="http://www.acm.org./jacm/">Journal of the ACM</a>
Hyperlink and Anchor Text (2/2)

- When indexing a document $D$, include anchor text from links pointing to $D$

  - Armonk, NY-based computer giant IBM announced today
  - A derogatory anchor text
  - The evil empire for computer industry
  - www.ibm.com
  - Big Blue today announced record profits for the quarter

Joe’s computer hardware links
Compaq
HP
IBM
PageRank Algorithm

- Proposed by Page and Brin, 1998
- Notations
  - A page $A$ has pages $T_1 \ldots T_n$ which point to it (citations)
  - $d$ range from 0~1, a damping factor (Google sets to be 0.85)
  - $C(T)$: Number of links going out of page $T$

- PageRank of a page $A$

$$PR(A) = (1 - d) + d \left( \frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right)$$

  - PageRank of each page is randomly assigned at the initial iteration and its value tends to be saturated through iterations

- A page with a high PageRank value
  - Many pages pointing to it
  - Or, there are some pages that point to it and have high PageRank values
Business Models for Web Search (1/3)

• Advertisers pay for banner ads (advertisements) on the site that do not depend on a user’s query
  – **CPM**: Cost Per Mille (thousand impressions). Pay for each ad display
  – **CPC**: Cost Per Click. Pay only when user clicks on ad
  – **CTR**: Click Through Rate. Fraction of ad impressions that result in clicks throughs. **CPC = CPM / (CTR * 1000)**
  – **CPA**: Cost Per Action (Acquisition). Pay only when user actually makes a purchase on target site

• Advertisers bid for “keywords”. Ads for highest bidders displayed when user query contains a purchased keyword
  – **PPC**: Pay Per Click. CPC for bid word ads (e.g. Google AdWords)

• This slide is adopted from Raymond J. Mooney’s teaching materials
Business Models for Web Search (2/3)

- Paid banner ads (news portal)
Business Models for Web Search (3/3)

- Bid keywords (search engine)
Final Project Description

- Reference site: http://140.122.185.33/IR-Demo/

- Contact TA for details of Corpus and Internet/Web Application Programs
- Project Due: 23 Jan. 2009