CS 345A
Data Mining
Lecture 1

Introduction to Web Mining

What is Web Mining?

Discovering useful information from the World-Wide Web and its usage patterns

Web Mining v. Data Mining

- Structure (or lack of it)
  - Textual information and linkage structure
- Scale
  - Data generated per day is comparable to largest conventional data warehouses
- Speed
  - Often need to react to evolving usage patterns in real-time (e.g., merchandising)

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

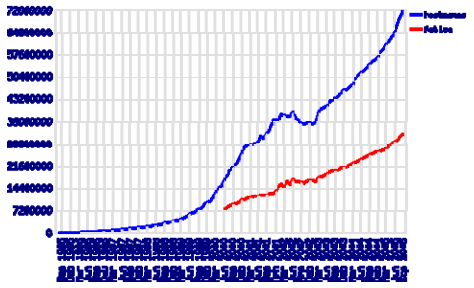
Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Size of the Web

- Number of pages
  - Technically, infinite
  - Much duplication (30-40%)
  - Best estimate of “unique” static HTML pages comes from search engine claims
    - Google = 8 billion(?), Yahoo = 20 billion
- Number of web sites
  - Netcraft survey says 72 million sites
    (http://news.netcraft.com/archives/web_server_survey.html)

Netcraft survey



http://news.netcraft.com/archives/web_server_survey.html

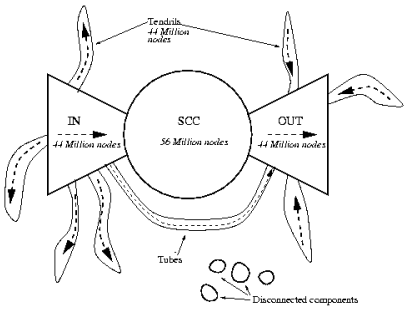
The web as a graph

- Pages = nodes, hyperlinks = edges
  - Ignore content
  - Directed graph
- High linkage
  - 8-10 links/page on average
  - Power-law degree distribution

Structure of Web graph

- Let's take a closer look at structure
  - Broder et al (2000) studied a crawl of 200M pages and other smaller crawls
  - Bow-tie structure
    - Not a "small world"

Bow-tie Structure



Source: Broder et al, 2000

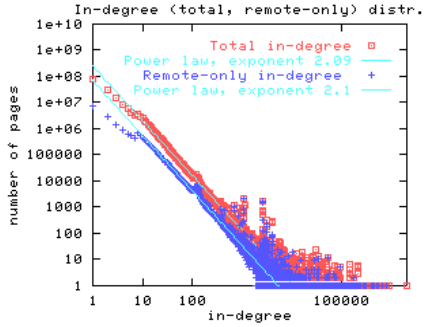
What can the graph tell us?

- Distinguish "important" pages from unimportant ones
  - Page rank
- Discover communities of related pages
  - Hubs and Authorities
- Detect web spam
  - Trust rank

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Power-law degree distribution

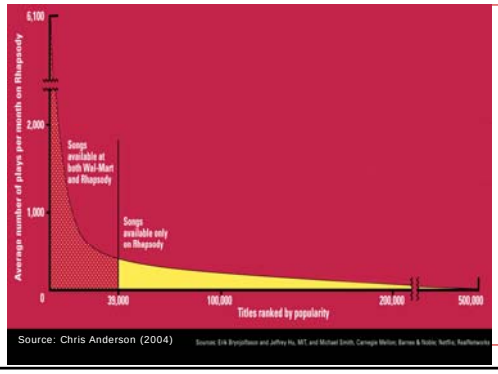


Source: Broder et al, 2000

Power-laws galore

- Structure
  - In-degrees
  - Out-degrees
  - Number of pages per site
- Usage patterns
  - Number of visitors
  - Popularity

The Long Tail



The Long Tail

- Shelf space is a scarce commodity for traditional retailers
  - Also: TV networks, movie theaters,…
- The web enables near-zero-cost dissemination of information about products
  - Action moves from Hits to Niches

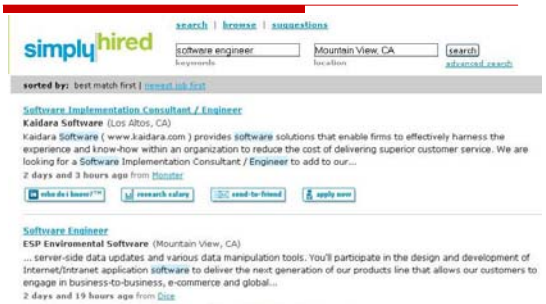
The Long Tail

- More choice necessitates better filters
  - Recommendation engines (e.g., Amazon)
  - How Into Thin Air made Touching the Void a bestseller
- In fact, page rank can be seen as a long tail filter
  - Tapping into the Wisdom of Crowds

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Extracting Structured Data



http://www.simplyhired.com

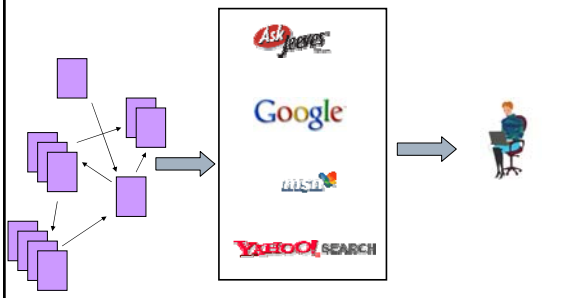
Extracting structured data

http://www.fatlens.com

Web Mining topics

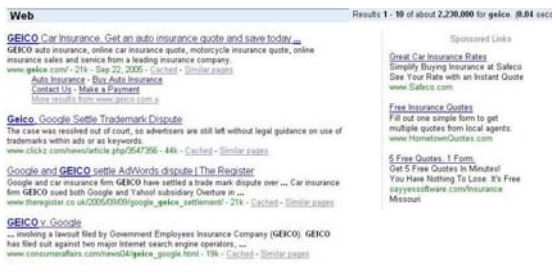
- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Searching the Web



The Web | Content aggregators | Content consumers

Ads vs. search results



Ads vs. search results

- Search advertising is the revenue model
  - Multi-billion-dollar industry
  - Advertisers pay for clicks on their ads
- Interesting problems
  - What ads to show for a search?
  - If I'm an advertiser, which search terms should I bid on and how much to bid?

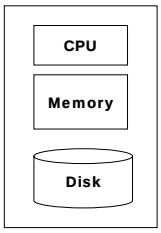
Sidebar: What's in a name?

- Geico sued Google, contending that it owned the trademark "Geico"
  - Thus, ads for the keyword geico couldn't be sold to others
- Court Ruling: search engines can sell keywords including trademarks
- No court ruling yet: whether the ad itself can use the trademarked word(s)

Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

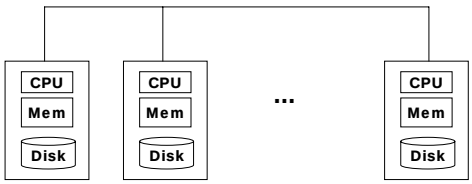
Systems architecture



Machine Learning, Statistics

"Classical" Data Mining

Very Large-Scale Data Mining



Cluster of commodity nodes

Systems Issues

- Web data sets can be very large
  - Tens to hundreds of terabytes
- Cannot mine on a single server!
  - Need large farms of servers
- How to organize hardware/software to mine multi-terabye data sets
  - Without breaking the bank!

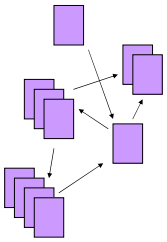
Web Mining topics

- Web graph analysis
- Power Laws and The Long Tail
- Structured data extraction
- Web advertising
- Systems Issues

Web Mining Project

- Lots of interesting project ideas
  - If you can't think of one please come discuss with us
- Data and Infrastructure
  - Webbase data (older Stanford web crawl)
  - Recent web crawl and server courtesy of Kosmix

The World-Wide Web



Our modern-day Library of Alexandria