Today’s Class

- Mid term projects due
- Continue discussion of Web search
- Readings for after break

1) How Encyclopedia Britannica Was Blown To Bits
   http://www.capmag.com/article.asp?ID=807

2) The Smart Way to Sell Information, Shapiro and Varian,

3) The Recording Industry is Trying to Kill the Goose That
   Lays the Golden Egg
   http://www.bricklin.com/recordsales.htm

4) Newly-Vulnerable Markets in an Age of Pure Information
Some Recent Events

- Yahoo stopped using Google for search and replaced with own technology
  - Based on Inktomi and Overture acquisitions
- Amazon established visible R&D group focused on search
- Gates acknowledged Microsoft mistakenly did not appreciate value of search
  - Now focused on Google as challenging Web portals such as MSN
- Patent disputes between Yahoo, Google, others
Anatomy of a Search Engine

- Three main components
  - Crawler or spider
    - Finds sites and pages on the Web
    - Usually by following links from already known pages
  - Index
    - Fast access to pages based on keywords
  - Query processing
    - Displaying pages most relevant to a user request
    - Some method for ranking by relevance
      - E.g., Google’s PageRank
Web Crawlers

- Web pages are “unstructured data”
  - Text, layout, images, movies, audio
  - One key aspect of layout is hyper-links
    - Identified by particular format

- Crawlers trace chains of hyper-links in search of new pages
  - Pages that are not already in the index

- Crawlers also check for known pages that have changed
  - Rate of checking for changes can be variable based on type of content or site
Page Indexing

- Efficiently map from query terms to web pages
  - Sometimes only URL, sometimes also content
    - “Cached” copy of page
- Which terms are important for the page
  - Initial indexing methods used just the text of the page
    - Relatively easy to manipulate
  - Current methods also rely heavily on text of referring pages
    - Near link to this page
Page Relevance

- Perhaps most important aspect to user is ordering of most relevant pages
  - Initial search engines used only text information
    - On page and perhaps referring pages
  - Methods such as Google’s PageRank and Kleinberg’s hubs & authorities use link structure
    - The importance of a page depends not only on its content but in how it is linked to
  - PageRank uses popularity – pages with more links to them ranked higher and in turn their votes for other pages count more
    - Question about such popularity based measures
Constant Tuning

- Search engines constantly tune their ranking methods
  - Balancing link structure, text content, rate of change, degree of linking
    - Partly to counteract “search engine optimizers”
- Often has substantial effect on rankings
  - Google’s changes now named in style of storms
  - Recent examples
    - Rise of weblogs, which have high degree of linking, led to re-balancing
    - Incorporation of Amazon’s pages into rankings
Revenue Models for Search

- For many years was a challenge
  - Paid inclusion allows sponsors to have their pages listed in response to certain queries
    - Can lead to lowering of search quality
    - FTC investigations about un-credited paid inclusion being misleading
    - Hard for advertisers to assess value

- Google has always argued against paid inclusion advertising
  - Sought ways of having “relevant ads” appear adjacent to search results
Bidding for Terms

- Overture launched market for search terms
  - Advertisers bid a per-click fee for being listed for certain words or phrases
    - Only pay when users click on sponsored link
    - Placement ordered by bid
    - Gave advertisers "context" for their fees

- Google still sought mechanism for assessing relevance of ads
  - Focus on providing user with relevant, useful information even in ads
  - Highest bidder not necessarily relevant
AdWords

Google’s mechanism of bidding for listings
  - Placement based on product of CTR and fee
    • Items that are more relevant tend to be placed higher due to more click-throughs
      - Key to addressing relevance concern
    • Also controls frequency of appearance
  - Advertisers bid a maximum per-click fee they are willing to pay
    • Actual fee is lowest amount that is $.01 larger than next lowest ranked bidder
    • Can set daily expenditure budget as hard to predict actual costs
A Different View of Advertising

- Paying more per keyword only one component of ad visibility
  - Has to be relevant as measured by user actions (clicks)

- Initially resulted from focus on user value
  - But also valuable to advertisers in targeting their campaigns
    - Quickly learn which ad copy and which keywords are effective
    - Can select large number of keywords and several ads per keyword
      - Automatically shown proportional to CTR
Demographic is Emergent

- Ads focus on what search terms are most relevant to the product or service
  - Description and value rather than target demographic
    - Get quick feedback to validate
- Google now claims every major category of advertiser now using AdWords model
  - Last holdout was consumer packaged goods companies (article in today’s Fortune online)
- Key is continuing to attract broad audience for Google searches
Effectiveness of Sponsored Links

- Advertisers migrating from banner ads to sponsored links
  - In both cases get click through data, so indicative of higher value

- Yahoo provides interesting snapshot
  - Company sources say 60% of 4Q03 revenues from sponsored search (not broken out)
  - About 82% of revenue from advertising so vast majority now appears to be search terms

- Growth in sponsored links revenue is slowing
Localized Advertising

- Online ads can be restricted to run only in certain geographies
  - Google now offers about 200 regions in US corresponding to traditional TV markets
    - Globally currently only per country
- Based on services that attempt to determine geographic location from IP address
  - Not completely accurate
    - Some global service providers such as AOL particularly problematic
Localized Search

- Google just went from labs to beta this week
  - Searches both online phone directory and Web pages
  - Groups together
  - Can show locations on map
  - Mapquest
- Yahoo also experimenting
Localized Search

- Based on combination of search terms and a user-specified location
  - Not determined from IP address as for ad targeting
- Google labs experimented with searching only Web pages but results disappointing
  - Now combine with directory search but do add entries that are found only on Web
- Estimates (guesses) that 20-30% of current searches are local in nature
  - Not previously well served
Contextual Ad Placement

- AdSense allows any publisher of Web pages to show ads selected by Google
  - Pay a portion of the revenue to the publisher of the content
    - Same CTR*fee model as used in Google’s search term advertising
    - Competing service from Overture

- Use their analysis of the Web to determine relevant ads to display
  - Can think of in terms of Web search
  - Additional analysis of content
Deep Web

- Most of the information accessible on the Web is not available in search engines
  - Often called the Deep Web
  - Such information is contained in databases and accessible via custom search interfaces
    - E.g., airline fares, phone directories, patents, product catalogs, government records
- Sometimes this information is made available via API’s and indexed by search engines
  - E.g., Amazon and Google
Deep Web

- Often structured or semi-structured data
  - In contrast to Web pages which are relatively unstructured free-form text
- Not always sensible taken out of context
- Sometimes extremely rapidly changing
  - E.g., Airline fares and availability
- Questions about how and whether to index such data
  - Active area of R&D
  - Business cases?