Search Engines:
Technology, Society, and Business

Prof. Marti Hearst

Oct 22, 2007

The Economics of Web Search

(Recap of Prof. Varian’s lecture)

The Economics of Web Search

- Why does Prof. Varian call Google a Yenta?
- What is an organic search result?
- What does it mean to bid on a keyword?

The Economics of Web Search

- Some search engine ad statistics:
  - 99% of Google's profit is from search ads
  - 2% of ads shown are clicked on (20 out of 1000)
  - 2% of these clicks turn into a purchase (.4 out of 1000)
    - So 1 out of 2500 people who see an ad on Google buy something.
  - Advertisers pay about $2 per 1000 (?) impressions
    - =about 10 cents (?) per click
- Need to reach a *lot* of people to make money
  - Radio example … what was the lesson there?

A History Lesson

- When radio was invented, no one knew what to do with it. The killer-app was ship-to-shore communication that everyone could hear.
- Why would everyone want to hear a message sent to no one in particular?
  - Turned out people liked to tune in
- Then it wasn’t clear how it would support itself
  - Public support (BBC in Britain)
  - Tax on hardware providers
  - Donations as seen for public radio
  - Tax on vacuum tubes to produce content
  - Commercially supported (“brought to you by”)
    - People hated this.
    - Needed to scale; hampered by the curvature of the earth. So needed a dense population, because the response rates were so low. So only the big cities had radio.
    - Key breakthrough: AT&T carried the shows through the TV network after it was recorded, broadcast it at radio shows throughout the country, and this made the product support pay for itself.

Costs of Supplying Search

- There are big fixed costs to build the search engine but low marginal cost to serve the next impression.
- But user switching costs are very low.
  - People disappear from the site after a negative search result; this means people do switch when one engine fails (they may be giving up or going to some other mechanism though.)

Search Market Structure

Low switching costs for search customers +

Advertisers go where the customers are =

A market structure like this:

- Several large search engines in each language group/country,
- Highly contestable market for users
  - This means there is competitive pricing,
- No force driving everyone to the same supplier
  - As opposed to, say, computer operating systems.

Flexibility of Web-Based Businesses

- You can tinker with your product while it's being used, and you can keep changing it and seeing how people respond.
  - "kaizen" = "continuous improvement"
- This is quite difficult for traditional software products.
- Example: slideshare.net
- Example:
  - Google's first idea for monetization was to just sell an intranet appliance
    - But that wasn't all that successful.
  - Next tried advertising.
    - But having salespeople price ads doesn't scale

Early Ad Auctions

- Goto.com startup decided to use an auction model: advertisers bid on query terms asked by the user.
  - Later changed name to Overture; was later bought by Yahoo.
  - Big difference from standard web search: at first Goto *only* showed paid-for search results; no organic results.
  - Later they switched to showing both organic and sponsored results.
- Auction type: Highest bidder gets the better positions on the page; highest bidder pays what they bid.

More Sophisticated Ad Auctions

- Google changed this in two ways.
  - (1) Charge based on expected revenue (likelihood of someone clicking on the ad)
  - Why?

More Sophisticated Ad Auctions

- Google changed this in two ways.
  - (1) Charge based on expected revenue (likelihood of someone clicking on the ad)
  - Why?
    - If the search engine keeps showing ads that are irrelevant to the query, or are otherwise low-quality, people will stop looking at the ads.

More Sophisticated Ad Auctions

- Google changed this in two ways.
  - (2) Used a second-price model auction.
    - This means that if you bid highest, you pay the amount offered by the *second*-highest bidder.
  - Why?

More Sophisticated Ad Auctions

- Google changed this in two ways.
  - (2) Used a second-price model auction.
    - This means that if you bid highest, you pay the amount offered by the *second*-highest bidder.
  - Why?
    - Don't have to game your price to be one cent above the ad of your top competitor.
      - Lots of load in the system with people logging in constantly.
      - Instead, the search engine sets the price you pay to be the minimum necessary to get your position.

Pricing Keywords

- An advertiser might want to pay different prices for different keywords
  - Example with wine sellers and the keyword “gourmet gift baskets”.
  - What about a search on “Geico”?

The Benefit of Auctions

- Bidding optimally means to bid to maximize your profit: your value per click minus the cost of getting that click.
- Example:
  - Say you sell widgets; each sale gives you a profit of $10
  - Say also that for the keyword “widget”
    - If your ad goes in the top slot, you’ll get 5 sales per 1000 views
    - If your ad goes in the second slot, you’ll get 2 sales per 1000 views
    - If your ad goes in the third slot, you’ll get one sale per 1000 views.
  - The idea of auction bidding is that you will bid the amount that makes sense for you to pay for the slot and still make a profit.
    - The advertiser has a choice of which medium they advertise in
    - Say you want to make at least $1 per item sold.
    - You’ll not pay more than $9.00 for the bottom slot
    - You’ll not pay more than $18.00 for the second slot
    - You’ll not pay more than $45.00 for the first slot
- Thus the best advertisers (those offering products people want to click on for a given query) get the most noticable slots.

Undersold and Oversold Queries

- Auction workings:
  - Minimum price is called the reserve price
  - When the auction is undersold, you pay the reserve price for the last position, and compete with other bidders for higher positions
  - When the auction is oversold, the lowest successful bidder pays the price that the first excluded bidder would have paid.
  - A lot of the revenue for the search engine comes from the need to compete for oversold pages (query terms)
  - The search engine has a relevance threshold.
    - Originally 10 clicks in a thousand. If your ad didn't meet the threshold for a query term, the ad was disabled. It's now a more sophisticated algorithm.

Combating Web Spam

(Recap of Dr. Najork’s lecture)

Web Spam

- What is a “spam” web page?
  - Why is it bad for users?
  - Why is it bad for search engines?
- What are legitimate ways to boost a web site’s organic search results rankings?

Web Spam Techniques

- “Keyword stuffing”
- “Link spam”
- “Cloaking”

Keyword stuffing

- Three variants:
  - Hand-crafted pages (ignored in this talk)
  - Completely synthetic pages
  - Assembling pages from “repurposed” content

Features identifying synthetic content

- Average word length
  - The mean word length for English prose is about 5 characters
- Word frequency distribution
  - Certain words (“the”, “a”, …) appear more often than others
- N-gram frequency distribution
  - Some words are more likely to occur next to each other than others
- Grammatical well-formedness
  - Don’t use this: natural-language parsing is expensive

Features identifying link spam

- Large number of links from low-ranked pages
- Discrepancy between number of links (peer endorsement) and number of visitors (user endorsement)
- Links mostly from affiliated pages
  - Same web site; same domain
  - Same IP address
  - Same owner (according to WHOIS record)
- Evidence that linking pages are machine-generated
- ...

Cloaking

- Cloaking: The practice of sending different content to search engines than to users
- Techniques:
  - Recognize page request is from search engine (based on “user-agent” info or IP address)
  - Make some text invisible (i.e. black on black)
  - Use CSS to hide text
  - Use JavaScript to rewrite page
  - Use “meta-refresh” to redirect user to other page
- Hard (but not impossible) for SE to detect

Detecting Web Spam

- Spam detection: A classification problem
  - Given salient features, decide whether a web page (or web site) is spam
- Can use automatic classifiers
  - Use data sets tagged by human judges to train and evaluate classifiers (this is expensive!)
- But what are the "salient features"?
  - Need to understand spamming techniques to decide on features
  - Finding the right features is "alchemy", not science
  - Spammers adapt – it's an arms race!

How well does web spam detection work?

- Overall, seems to work pretty well
- It’s an ever-escalating battle.
- Some words are more prone to spam than others
  - I recently searched for “mirror hanging hardware” and got a nasty surprise

Privacy and Search

(Recap of Chris Hoofnagle's lecture)

What is Privacy?

- “Peoples’ concern about privacy is an inch deep and a mile wide.”
    - -- Deirdre Mulligan, UC Berkeley Samuelson Center for Law & Technology
- Individuals have their own conceptions of privacy.
    - Differs according to the setting.
    - You know it when you lose it.

Slide adapted from Deirdre Mulligan’s

Conceptions of Privacy

- “the right to be let alone.” Samuel Warren and Louis Brandeis. “The Right to Privacy,” *Harvard Law Review,* 1890
- “the right of the individual to decide for himself, with only extraordinary exceptions in the interest of society, when and on what terms his acts should be revealed to the general public.” Alan Westin *Privacy and Freedom,* 1967
- fairness and control over personal information, anonymity, and confidentiality. Berman and Mulligan “Privacy in the Digital Age” Nova Law Review 1999.

Slide adapted from Deirdre Mulligan’s

Public vs. Private

- May someone look at you in the street?
- May someone standing on the street look at you through your kitchen window?
- May they tell someone else about what they saw?
- May they take a photo of what they see?
- What about publishing that photo?
- What if it is a photo of a crime, or an important event?
- What about modifying the photo digitally?

Public vs. Private

- May someone look over your shoulder at what you're reading?
- May they blog publicly about what they saw?
- May they listen in on your cell phone conversation?
- May they blog about it?
- What about tapping your land line?
- What about phone companies recording when and where you called?
- What about cell phone companies tracking and recording your movements?

Fair Information Practices

- Recently outlined by the Federal Trade Commission
  - Fair Information Practices (1998): http://www.ftc.gov/reports/privacy3/
  - Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress (May 2000) http://www.ftc.gov/reports/privacy2000/privacy2000.pdf
- These principles include:
  - Notice
  - Choice
  - Access
  - Security
- Still not part of one big law; manifested instead in individual state and federal laws.
  - http://www.privacyrights.org/ar/fairinfo.htm

A Right vs. Protection from Harm

- APEC model focuses on shifting from the notion of privacy as a human right to the notion of containing information to protect individuals from harm.
- What does this mean?

How Do Search Engines Break the Model?

- They don’t collect data.
- But they mediate access to content.
  - They are somewhat like library checkout records, which were protected until the Patriot Act.

Web Search Privacy Concerns

- What happened with the AOL search logs?
- Is web search the most potentially harmful privacy problem?

Difficult Information Needs

(Search tips for hard queries)

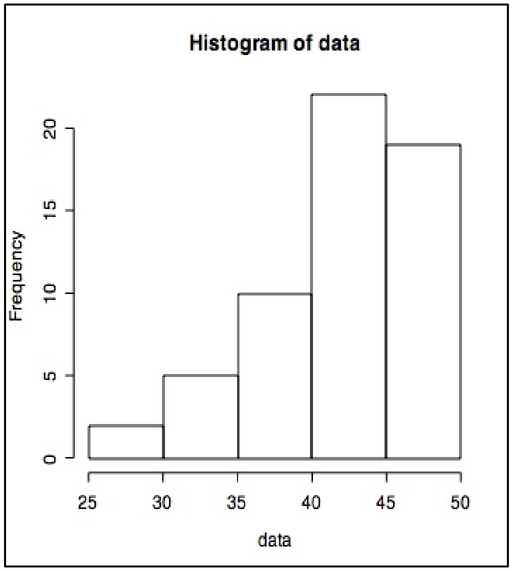
Keyword search tips

- There are many books and websites that give searching tips; here are a few common ones:
  - Use unusual terms and proper names
  - Put most important terms first
  - Use phrases when possible
  - Make use of slang, industry jargon, local vernacular, acronyms
  - Be aware of country spellings and common misspellings
  - Frame your search like an answer or question

Slide adapted from Lew & Davis

Assignments

- I'll have the new assignment posted by 5pm today.
- Here is a chart for the scores of the first assignment (out of 50 points).



Next Week

- Multimedia search
- (no reading)