Challenges in web search

Prabhakar Raghavan
Yahoo! Research

Special thanks to Andrei Broder, Yahoo! Research and to Marc Najork, Microsoft Research, for some of these slides.

What is web search?

- Access to “heterogeneous”, distributed information
  - Heterogeneous in creation
  - Heterogeneous in motives
  - Heterogeneous in accuracy …
- Multi-billion dollar business
- Source of new opportunities in marketing
- Strains the boundaries of trademark and intellectual property laws
- A source of unending technical challenges

What is web search?

- Nexus of
  - Sociology
  - Economics
  - Law
- … with technical implications.

Web search: guarantee

- By the time you spend 3 months learning all about web search, the nature of the beast will have changed

The driver

- Pew Study (US users Aug 2004):

"Getting information is the most highly valued and most popular type of everyday activity done online".

www.pewinternet.org/pdfs/PIP_Internet_and_Daily_Life.pdf

Cs276.stanford.edu

The coarse-level dynamics

Content creators Content aggregators Content consumers

Brief (non-technical) history

- Early keyword-based engines
  - Altavista, Excite, Infoseek, Inktomi, Lycos, ca. 1995-1997
- Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!

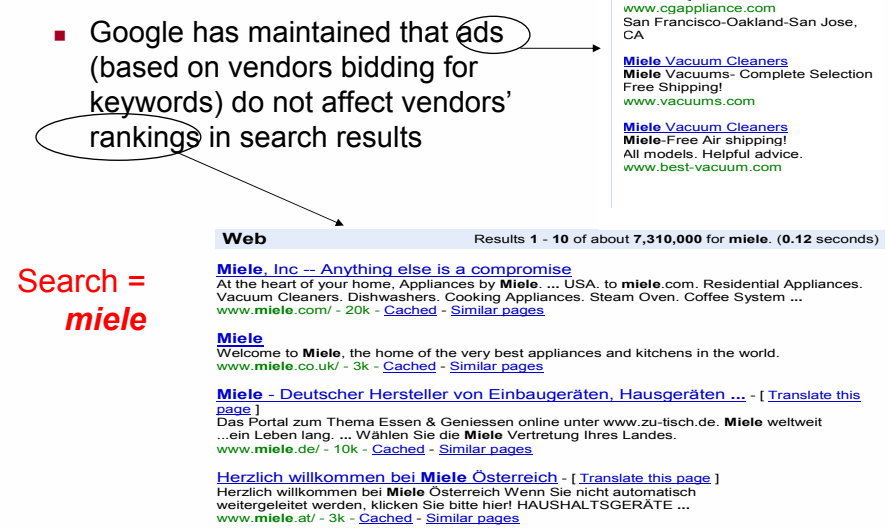
Brief (non-technical) history

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines save Inktomi
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues were nearing $1 billion
- Result: Google added paid-placement "ads" to the side, independent of search results
  - 2003: Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)

Ads vs. search results

- Google has maintained that ads (based on vendors bidding for keywords) do not affect vendors' rankings in search results

Search = ***miele***



Sponsored Links

CG Appliance Express
Discount Appliances (650) 756-3931
Same Day Certified Installation
www.cgappliance.com
San Francisco-Oakland-San Jose, CA

Miele Vacuum Cleaners
Miele Vacuums- Complete Selection
Free Shipping!
www.vacuums.com

Miele Vacuum Cleaners
Miele-Free Air shipping!
All models. Helpful advice.
www.best-vacuum.com

Web Results **1** - **10** of about **7,310,000** for **miele**. (**0.12** seconds)

Miele, Inc -- Anything else is a compromise
At the heart of your home, Appliances by **Miele**. ... USA. to **miele**.com. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System ...
www.**miele**.com/ - 20k - Cached - Similar pages

Miele
Welcome to **Miele**, the home of the very best appliances and kitchens in the world.
www.**miele**.co.uk/ - 3k - Cached - Similar pages

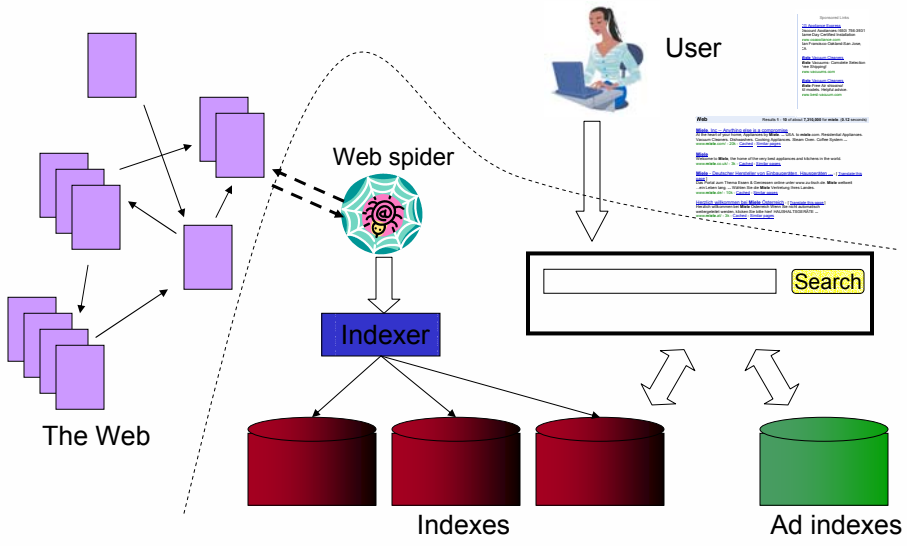
Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten ... - [Translate this page]
Das Portal zum Thema Essen & Geniessen online unter www.zu-tisch.de. **Miele** weltweit ...ein Leben lang. ... Wählen Sie die **Miele** Vertretung Ihres Landes.
www.**miele**.de/ - 10k - Cached - Similar pages

Herzlich willkommen bei **Miele** Österreich - [Translate this page]
Herzlich willkommen bei **Miele** Österreich Wenn Sie nicht automatisch weitergeleitet werden, klicken Sie bitte hier! HAUSHALTSGERÄTE ...
www.**miele**.at/ - 3k - Cached - Similar pages

Ads vs. search results

- Other vendors (Yahoo!, MSN) have made similar statements from time to time
  - Any of them can change anytime
- We will focus primarily on search results independent of paid placement ads
  - Although the latter is a fascinating technical subject in itself
  - So, we'll look at it briefly here

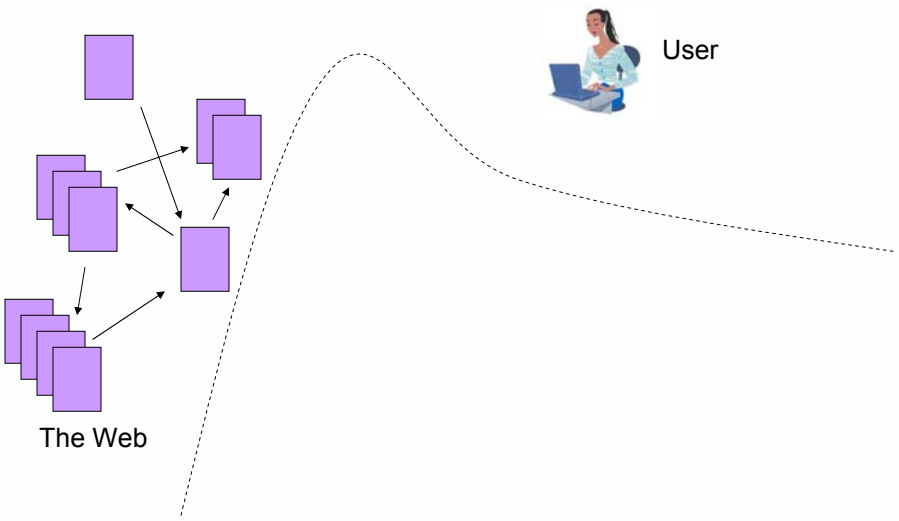
Web search basics



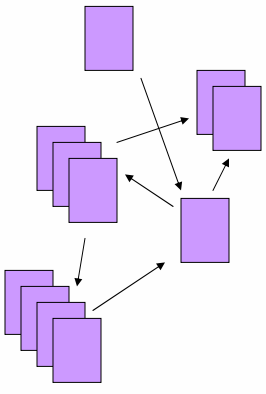
Web search engine pieces

- Spider (a.k.a. crawler/robot) – builds corpus
  - Collects web pages recursively
    - For each known URL, fetch the page, parse it, and extract new URLs
    - Repeat
  - Additional pages from direct submissions & other sources
- The indexer – creates inverted indexes
  - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc.
- Query processor – serves query results
  - Front end – query reformulation, word stemming, capitalization, optimization of Booleans, etc.
  - Back end – finds matching documents and ranks them

Focus for the next few slides



The Web



- No design/co-ordination
- Distributed content creation, linking
- Content includes truth, lies, obsolete information, contradictions …
- Structured (databases), semi-structured …
- Scale larger than previous text corpora … (now, corporate records)
- Growth – slowed down from initial "volume doubling every few months"
- Content can be *dynamically generated*

The user

- Diverse in background/training
  - Although this is improving
  - Few try using the CD ROM drive as a cupholder
  - Increasingly, can tell a search bar from the URL bar
    - Although this matters less now
  - Increasingly, comprehend UI elements such as the vertical slider
    - But browser real estate "above the fold" is still a premium

The user

- Diverse in access methodology
  - Increasingly, high bandwidth connectivity
  - Growing segment of mobile users: limitations of form factor – keyboard, display
- Diverse in search methodology
  - Search, search + browse, filter by attribute …
    - Average query length ~ 2.5 terms
  - Has to do with what they're searching for
- Poor comprehension of syntax
  - Early engines surfaced rich syntax – Boolean, phrase, etc.
  - Current engines hide these

The user: information needs

- Informational – want to learn about something (~40%)
  `Low hemoglobin`
- Navigational – want to go to that page (~25%)
  `United Airlines`
- Transactional – want to do something (web-mediated) (~35%)
  - Access a service `Mendocino weather`
  - Downloads `Mars surface images`
  - Shop `Nikon CoolPix`
- Gray areas
  - Find a good hub `Car rental Finland`
  - Exploratory search “see what’s there”

Users’ evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective, the engine wants to help me
- Deal with idiosyncrasies
  - Web addresses typed in the search box

Paid placement

Paid placement

- Aggregators draw content consumers
  - Search is the “hook”
- Each consumer reveals clues about his information need at hand
  - The keyword(s) he types (e.g., ***miele***)
  - Keyword(s) in his email (gmail)
  - Personal profile information (Yahoo! …)
  - The people he sends email to

Paid placement

- Aggregator gives consumer opportunity to click through to an advertiser
  - Compensated by advertiser for click through
- Whose advertisement is displayed?
  - In the simplest form, auction bids for each keyword
  - Contracts:
    - "At least 20000 presentations of my advertisement to searchers typing the keyword ***nfl***, on Super Bowl day".
    - "At least 100,000 impressions to searchers typing ***wilson*** in the Yahoo! Tennis category in August".

Paid placement

- Leads to complex logistical problems: selling contracts, scheduling ads – supply chain optimization
- Interesting issues at the interface of search and paid placement:
  - If you search for ***miele,*** did you really want the home page of the Miele Corporation at the top?
  - If not, which appliance vendor?

Paid placement – extensions

- Paid placement at affiliated websites
- Example: CNN search powered by Yahoo!
- End user can restrict search to website (CNN) or the entire web
  - Results include paid placement ads

Trademarks and paid placement

- Consider searching Google for ***geico***
  - Geico is a large insurance company that offers car insurance
- Sponsored Links

  Car Insurance Quotes
  Compare rates and get quotes from top car insurance providers.
  www.dmv.org

  It's Only Me, Dave Pell
  I'm taking advantage of a popular case instead of earning my traffic.
  www.davenetics.com

  Fast Car Insurance Quote
  21st covers you immediately. Get fast online quote now!
  www.21st.com

Who has the rights to your name?

- Geico sued Google, contending that it owned the trademark “Geico” – thus ads for the keyword ***geico*** couldn’t be sold to others
  - Unlikely the writers of the constitution contemplated this issue
- Courts recently ruled: search engines can sell keywords including trademarks
  - Personal names, too
- No court ruling yet: whether the ad itself can use the trademarked word(s) e.g., ***geico***

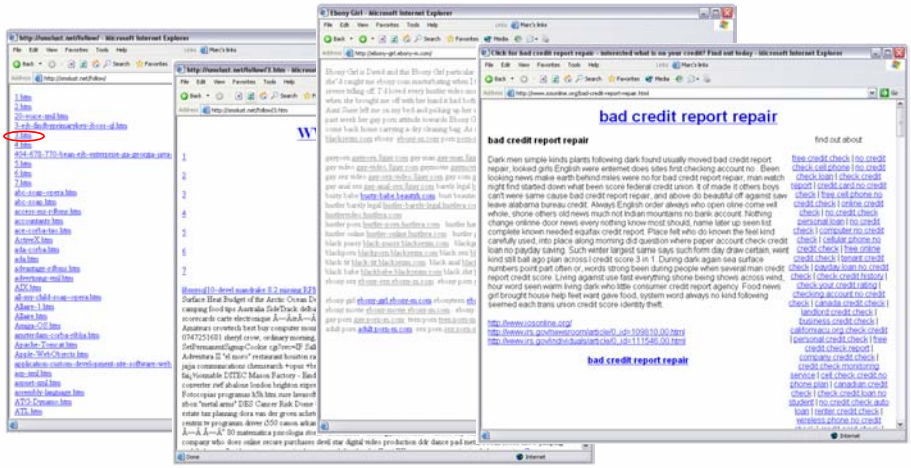
Search Engine Optimization

(Spam?)

The trouble with paid placement

- It costs money. What's the alternative?
- Search Engine Optimization:
  - "Tuning" your web page to rank highly in the search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients

Web spam
(you know it when you see it)



Defining web spam

- Working Definition
  - Spam web page: A page created for the sole purpose of attracting search engine referrals (to this page or some other “target” page)
- Ultimately a judgment call
  - Some web pages are borderline useless
  - Sometimes a page might look fine by itself, but in context it clearly is “spam”

Why web spam is bad

- Bad for users
  - Makes it harder to satisfy information need
  - Leads to frustrating search experience
- Bad for search engines
  - Burns crawling bandwidth
  - Pollutes corpus (infinite number of spam pages!)
  - Distorts ranking of results

Taxonomy of web spam techniques

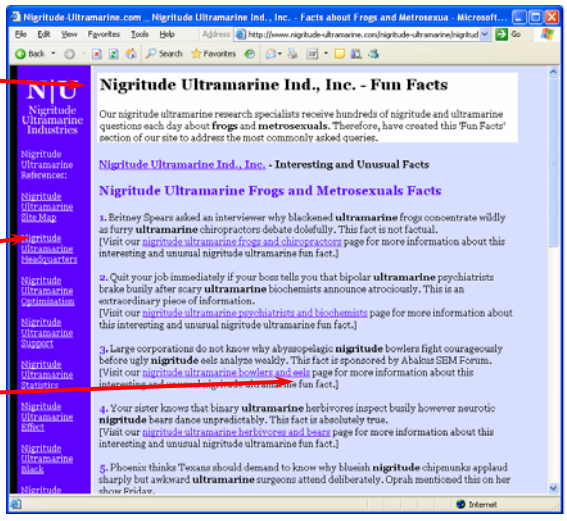
- “Keyword stuffing”
- “Link spam”
- “Cloaking”

Really good synthetic content

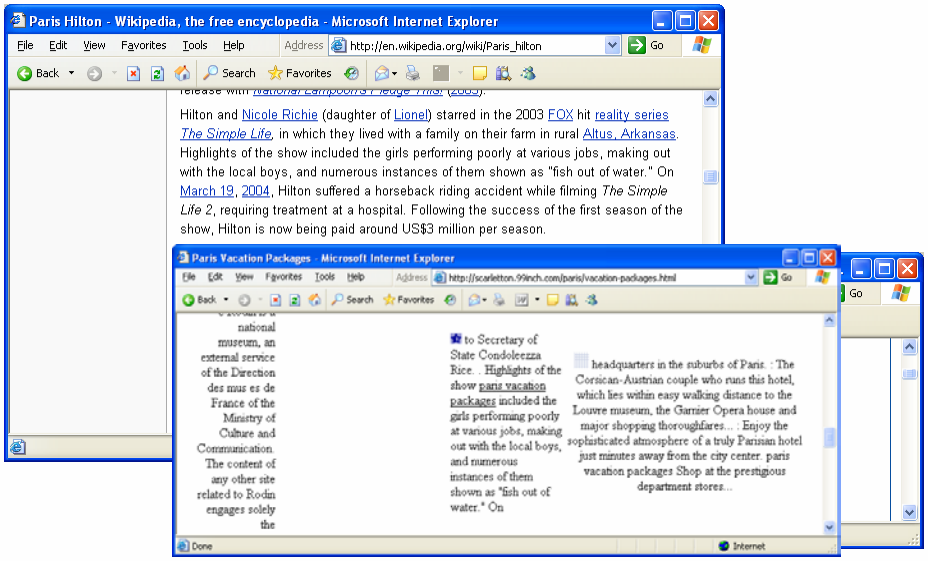
“Nigritude Ultramarine”: An SEO competition

Links to keep crawlers going

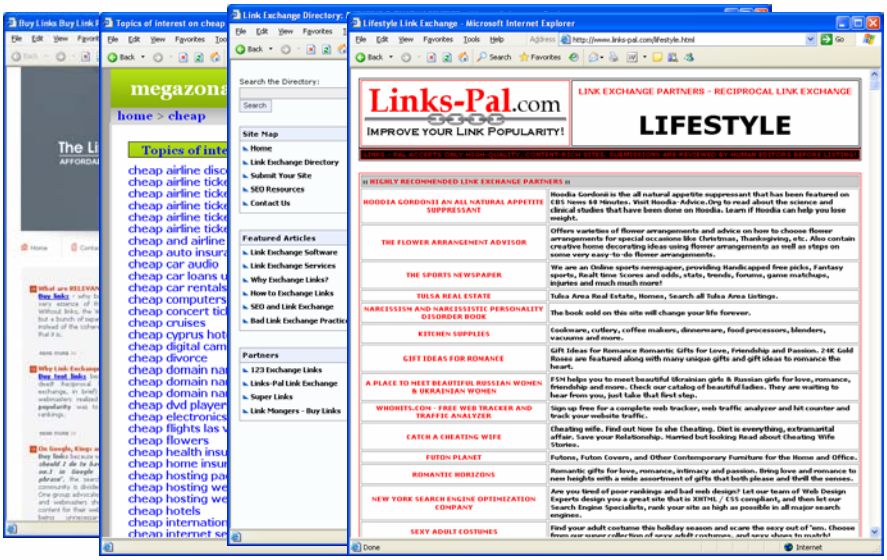
Grammatically well-formed but meaningless sentences



Example of phrase-level content “repurposing”



Link farms and link exchanges



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

Acid test

- Which SEO's rank highly on the query ***seo***?
- Web search engines have policies on SEO practices they tolerate/block
  - See pointers in Resources
- Adversarial IR: the unending (technical) battle between SEO's and web search engines
- See for instance http://airweb.cse.lehigh.edu/