CS276B

Text Retrieval and Mining

Winter 2005

Lecture 1

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 get up to speed on web search during this quarter, the nature of the beast will have changed

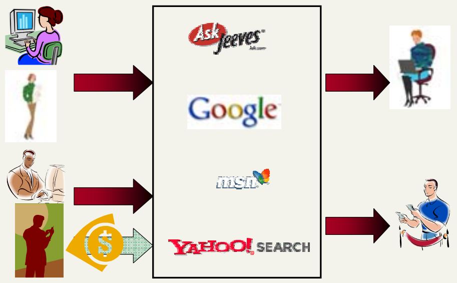
The driver

- 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

The coarse-level dynamics



Content creators Content aggregators Content consumers

Brief (non-technical) history

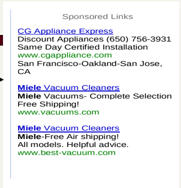
- Early keyword-based engines
  - Altavista, Excite, Infoseek, Inktomi, Lycos, ca. 1995-1997
- Paid placement ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!

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



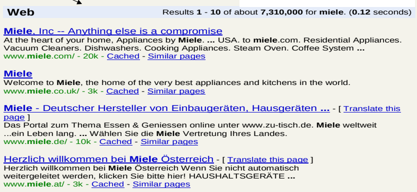
Sponsored Links

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

Search = ***miele***



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

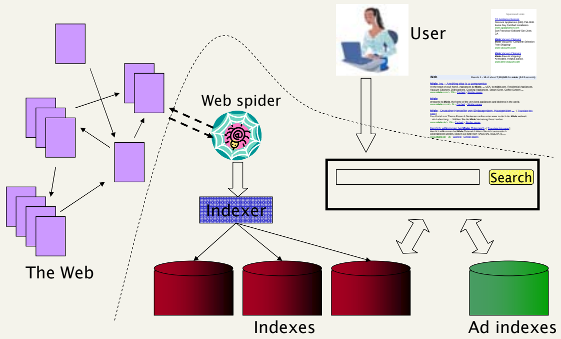
Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten **...** - [Translate this page]
Das Portal zum Thema Essen & Geniessen online unter www.zu-tisch.de. **Miele** weltweit ...ein Leben lang. **...** Wählen Sie die **Miele** Vertretung Ihres Landes.
www.**miele**.de/ - 10k - Cached - Similar pages

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
  - Deeper, related ideas in Lecture 4 (Recommendation systems)

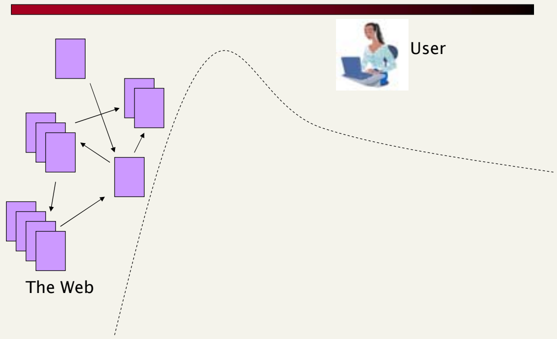
Web search basics



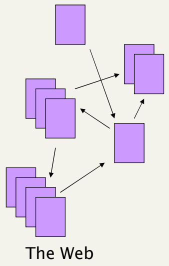
Web search engine pieces

- 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



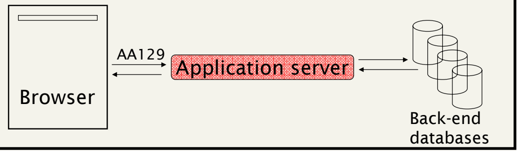
The Web



- 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 Web: Dynamic content

- A page without a static html version
  - E.g., current status of flight AA129
  - Current availability of rooms at a hotel
- Usually, assembled at the time of a request from a browser
  - Typically, URL has a ‘?’ character in it



Dynamic content

- Most dynamic content is ignored by web spiders
  - Many reasons including malicious spider traps
- Some dynamic content (news stories from subscriptions) are sometimes delivered as dynamic content
  - Application-specific spidering
- Spiders most commonly view web pages just as Lynx (a text browser) would

The web: size

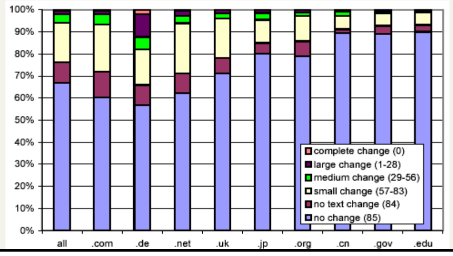
- What is being measured?
  - Number of hosts
  - Number of (static) html pages
    - Volume of data
- Number of hosts – netcraft survey
  - http://news.netcraft.com/archives/web_server_survey.html
  - Gives monthly report on how many web servers are out there
- Number of pages – numerous estimates
  - More to follow later in this course
  - For a Web engine: how big its index is

The web: evolution

- All of these numbers keep changing
- Relatively few scientific studies of the evolution of the web
  - http://research.microsoft.com/research/sv/sv-pubs/p97-fetterly/p97-fetterly.pdf
- Sometimes possible to extrapolate from small samples
  - http://www.vldb.org/conf/2001/P069.pdf

Static pages: rate of change

- Fetterly et al. study: several views of data, 150 million pages over 11 weekly crawls
  - Bucketed into 85 groups by extent of change



Diversity

- Languages/Encodings
  - Hundreds (thousands ?) of languages, W3C encodings: 55 (Jul01) [W3C01]
  - Google (mid 2001): English: 53%, JGCFSKRIP: 30%
- Document & query topic

Popular Query Topics (from 1 million Google queries, Apr 2000)

Arts	14.6%	Arts: Music	6.1%
Computers	13.8%	Regional: North America	5.3%
Regional	10.3%	Adult: Image Galleries	4.4%
Society	8.7%	Computers: Software	3.4%
Adult	8%	Computers: Internet	3.2%
Recreation	7.3%	Business: Industries	2.3%
Business	7.2%	Regional: Europe	1.8%
…	…	…	…

Other characteristics

- Significant duplication
  - Syntactic – 30%-40% (near) duplicates [Brod97, Shiv99b]
  - Semantic – ???
- High linkage
  - More than 8 links/page in the average
- Complex graph topology
  - Not a small world; bow-tie structure [Brod00]
- Spam
  - 100s of millions of pages
- More on these later

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

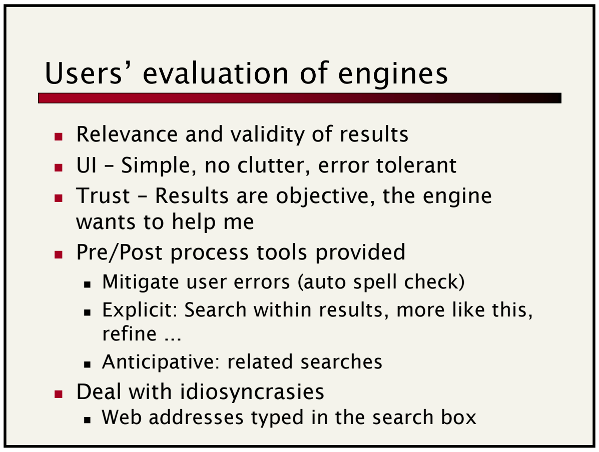
The user

- 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

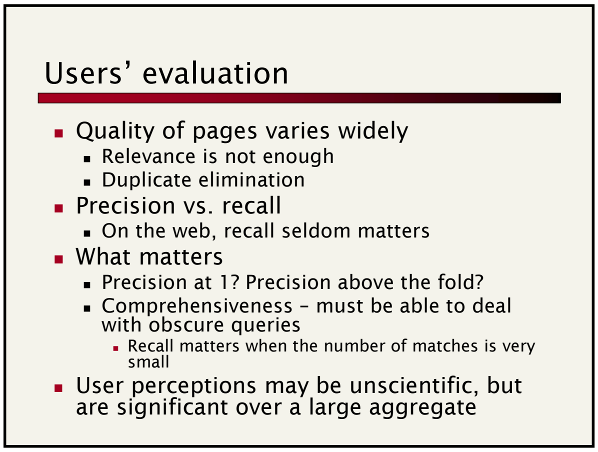
- Informational – want to learn about something (~40%) **Low hemoglobin**
- Navigational – want to go to that page (~25%) **United Airlines**
- Transactional – want to do something (web-mediated) (~35%)
  - Access a service **Mendocino weather**
  - Downloads **Mars surface images**
  - Shop **Nikon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

Courtesy Andrei Broder, IBM



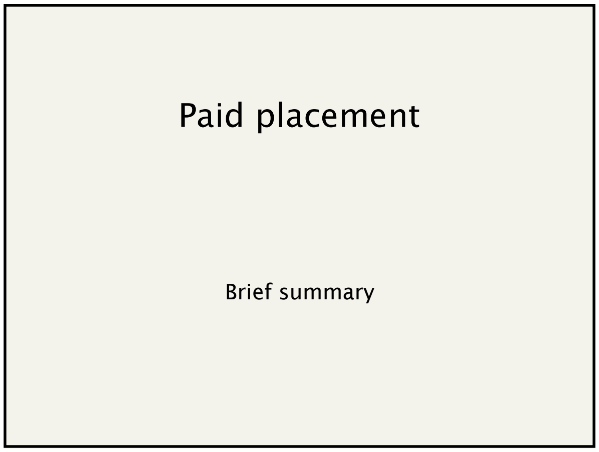
Users' evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective, the engine wants to help me
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check)
  - Explicit: Search within results, more like this, refine ...
  - Anticipative: related searches
- Deal with idiosyncrasies
  - Web addresses typed in the search box



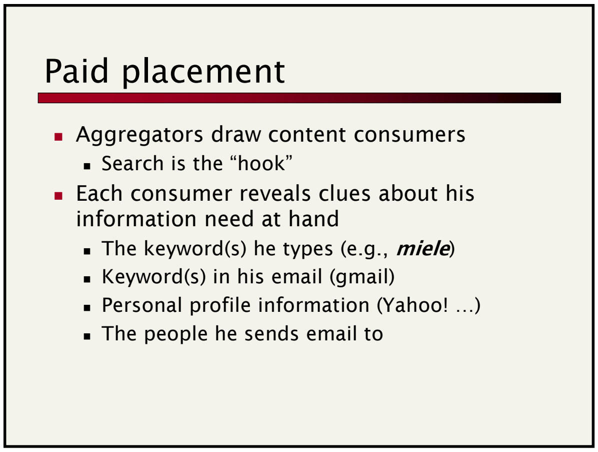
Users' evaluation

- Quality of pages varies widely
  - Relevance is not enough
  - Duplicate elimination
- Precision vs. recall
  - On the web, recall seldom matters
- What matters
  - Precision at 1? Precision above the fold?
  - Comprehensiveness – must be able to deal with obscure queries
    - Recall matters when the number of matches is very small
- User perceptions may be unscientific, but are significant over a large aggregate



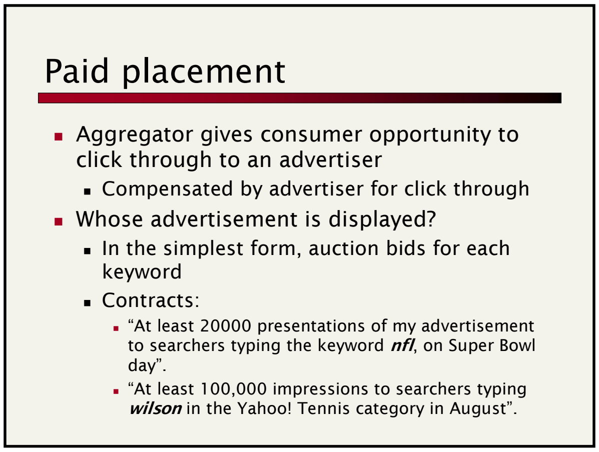
Paid placement

Brief summary



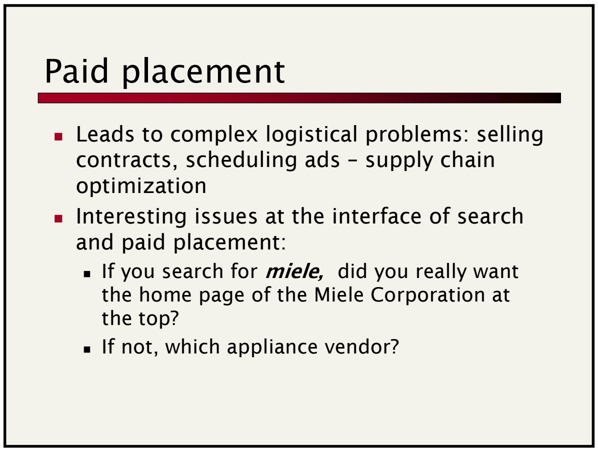
Paid placement

- 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".



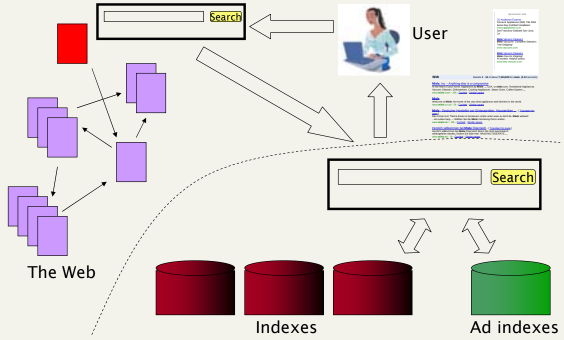
Paid placement

- 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

Affiliate search



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

(SEO, SEM …)

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
  - Also known as Search Engine Marketing
- Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients

Simplest forms

- Early engines relied on the density of terms
  - The top-ranked pages for the query ***maui resort*** were the ones containing the most ***maui***'s and ***resort***'s
- SEOs responded with dense repetitions of chosen terms
  - e.g., ***maui resort maui resort maui resort***
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers

Can't trust the words on a web page, for ranking.

Variants of keyword stuffing

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks, etc.

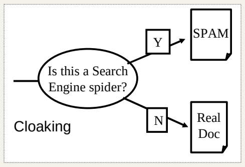
Meta-Tags =
"… London hotels, hotel, holiday inn, hilton, discount, booking, reservation, sex, mp3, britney spears, viagra, …"

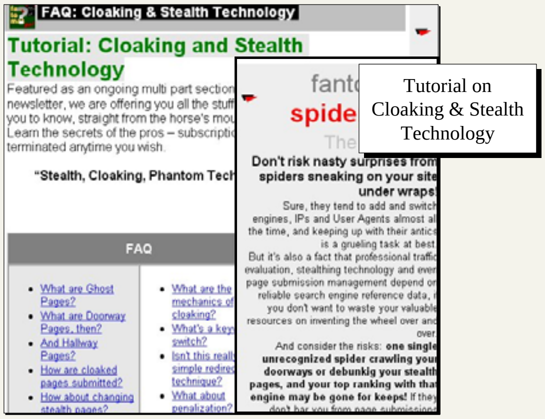
Search engine optimization (Spam)

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forum
  - Web master world (www.webmasterworld.com)
    - Search engine specific tricks
    - Discussions about academic papers ☺
    - More pointers in the Resources

More spam techniques

- **Cloaking**
- Serve fake content to search engine spider
- *DNS cloaking:* Switch IP address. Impersonate





FAQ: Cloaking & Stealth Technology

Tutorial: Cloaking and Stealth Technology

Featured as an ongoing multi part section newsletter, we are offering you all the stuff you to know, straight from the horse's mou Learn the secrets of the pros – subscriptio terminated anytime you wish.

"Stealth, Cloaking, Phantom Tech

FAQ

- What are Ghost Pages?
- What are Doorway Pages, then?
- And Hallway Pages?
- How are cloaked pages submitted?
- How about changing stealth pages?

- What are the mechanics of cloaking?
- What's a key switch?
- Isn't this reall simple redirec technique?
- What about penalization?

fanto

spide

The

Don't risk nasty surprises from spiders sneaking on your site under wraps!

Sure, they tend to add and switch engines, IPs and User Agents almost all the time, and keeping up with their antics is a grueling task at best. But it's also a fact that professional traffic evaluation, stealthing technology and even page submission management depend on reliable search engine reference data, you don't want to waste your valuable resources on inventing the wheel over and over.

And consider the risks: **one single unrecognized spider crawling your doorways or debunkig your stealth pages, and your top ranking with that engine may be gone for keeps!** If they don't bar you from page submission

Tutorial on Cloaking & Stealth Technology

More spam techniques

- **Doorway pages**
  - Pages optimized for a single keyword that re-direct to the real target page
- **Link spamming**
  - Mutual admiration societies, hidden links, awards – more on these later
  - *Domain flooding:* numerous domains that point or re-direct to a target page
- **Robots**
  - Fake query stream – rank checking programs
    - "Curve-fit" ranking programs of search engines
  - Millions of submissions via Add-Url

The war against spam

- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)
- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed

More on these in upcoming lectures.

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/

Preview of Web lectures

- Spidering issues
- Web size estimation
  - Search engine index estimation
- Duplicate and mirror detection
- Link analysis and ranking
  - Infrastructure for link indexes
- Behavioral ranking
- Other applications

Resources

- www.seochat.com/
- www.google.com/webmasters/seo.html
- www.google.com/webmasters/faq.html
- www.smartmoney.com/bn/ON/index.cfm?story=ON-20041215-000871-1140
- research.microsoft.com/research/sv/sv-pubs/p97-fetterly/p97-fetterly.pdf
- news.com.com/2100-1024_3-5491704.html
- www.jupitermedia.com/corporate/press.html