The CMU machine learning protesters

http://www.flickr.com/photos/30686429@N07/3953914015/in/set-72157622330082619/

Web basics

David Kauchak
cs458
Fall 2012

adapted from:
http://www.stanford.edu/class/cs276/handouts/lecture13-webchar.ppt

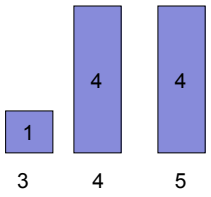
Administrative

- Schedule for the next two weeks
  - Sunday 10/21: assignment 3 (start working now!)
  - Friday 10/19 – Tuesday 10/23: midterm
    - 1.5 hours
    - take-home
    - can take it any time in that window
      - must NOT talk to anyone else about the midterm until after Tuesday
    - open book and open notes, though closed web

Course feedback

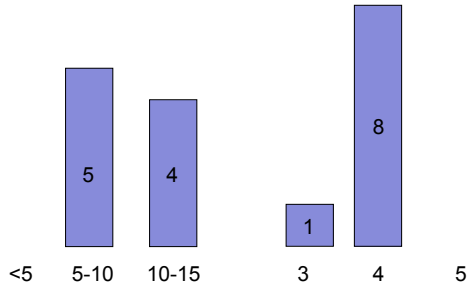
- Thanks!
- If you ever have other feedback…
- Assignments/homeworks
  - I do recognize that they are a lot of hard work
  - but they should be useful in learning (and fun in a love/hate sort of way)
  - will lighten up some in the final half/third of the course
- Course content
  - Lots of different IR systems (I understand sometimes we cover a lot of random topics)
  - Underneath the covers, a lot of it is engineering and trial and error

Course feedback



overall how is the class going

Course feedback



time spent /wk

difficulty

Informal quiz

Boolean queries

c OR a AND f
a AND f OR c

c b d

e d c

b d f

a f e

Outline

Brief overview of the web

Challenges with web IR:

- Web Spam
- Estimating the size of the web
- Detecting duplicate pages

Brief (non-technical) history

Early keyword-based engines

- Altavista, Excite, Infoseek, Inktomi, ca. 1995-1997

Sponsored search ranking: Goto.com (morphed into Overture.com)

- 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

Yahoo followed suit, acquiring Overture (for paid placement) and Inktomi (for search)

Why did Google win?

Relevance/link-based

Simple UI

Hardware – used commodity parts

- inexpensive
- easy to expand
- fault tolerance through redundancy

What's wrong (from the search engine's standpoint) of having a cost-per-click (CPC) model and ranking ads based only on CPC?

Post 2000

Lot's of start-ups have tried…

- Snap (2005): Overture's previous owner
- Cuil (2008): ex-google employees
- Powerset (2007): NLP folks (a lot from Xerox PARC) … bought by Microsoft
- Many more… http://en.wikipedia.org/wiki/Web_search_engine

2007	wikiseek	Inactive
	Sproose	Inactive
	Wikia Search	Inactive
	Blackle.com	Active
2008	Powerset	Inactive (redirects to Bing)
	Picollator	Inactive
	Viewzi	Inactive
	Boogami	Inactive
	LeapFish	Inactive
	Forestle	Inactive (redirects to Ecosia)
	VADLO	Active
	DuckDuckGo	Active, Aggregator
2009	Bing	Active, Launched as rebranded Live Search
	Yebol	Active
	Mugurdy	Inactive due to a lack of funding
	Goby	Active
2010	Blekko	Active
	Cuil	Inactive
	Yandex	Active, Launched global (English) search
	Yummly	Active
2011	Interred	Active
	Yandex	Active, Launched Turkey search
2012	Volunia	Active

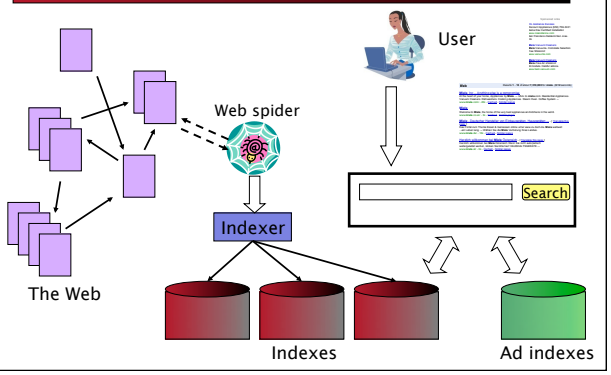
Current market share

Google:	67%
Bing:	16%
Yahoo:	13%
Ask:	3%
AOL:	1.5%
Rest:	9%

(comscore)

http://searchenginewatch.com/article/2205504/Bing-Gains-More-Ground-in-Search-Engine-Market-Share-Yahoo-Resumes-Downward-Slide

Web search basics



User needs/queries

Researchers/search engines often categorize user needs/queries into different types

For example…?

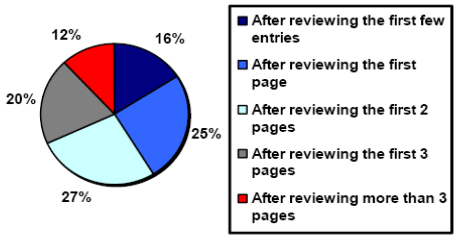
User Needs

Need [Brod02, RL04]

- **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 `Seattle weather`
  - Downloads `Mars surface images`
  - Shop `Canon S410`
- **Gray areas**
  - Find a good hub `Car rental Brasil`
  - Exploratory search "see what's there"

How far do people look for results?

"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Users' empirical evaluation of results

Quality of pages varies widely

- Relevance is not enough
- Other desirable qualities (non IR!!)
  - Content: Trustworthy, diverse, non-duplicated, well maintained
  - Web readability: display correctly & fast
  - No annoyances: pop-ups, etc

Users' empirical evaluation of results

Precision vs. recall

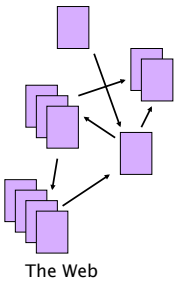
- On the web, recall seldom matters
- Recall matters when the number of matches is very small

What matters

- Precision at 1? Precision above the fold?
- Comprehensiveness – must be able to deal with obscure queries

User perceptions may be unscientific, but are significant over a large aggregate

How is the web unique?



No design/co-ordination

Content includes truth, lies, obsolete information, contradictions …

Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…

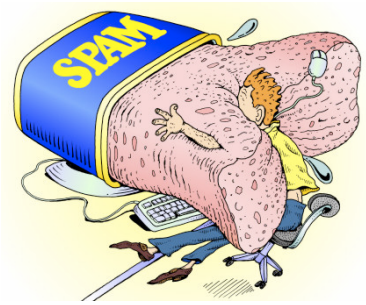
Financial motivation for ranked results

Scale much larger than previous text collections … but corporate records are catching up

Growth – slowed down from initial "volume doubling every few months" but still expanding

Content can be *dynamically generated*

Web Spam



http://blog.lib.umn.edu/wilsper/informationcentral/spam.jpg

The trouble with sponsored search …

It costs money. What's the alternative?

Search Engine Optimization:

- "Tuning" your web page to rank highly in the algorithmic search results for select keywords
- Alternative to paying for placement
- Intrinsically a marketing function

Performed by companies, webmasters and consultants ("Search engine optimizers") for their clients

Some perfectly legitimate, more very shady

Simplest forms

First generation engines relied heavily on *tf/idf*

What would you do as an SEO?

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

Pure word density cannot be trusted as an IR signal

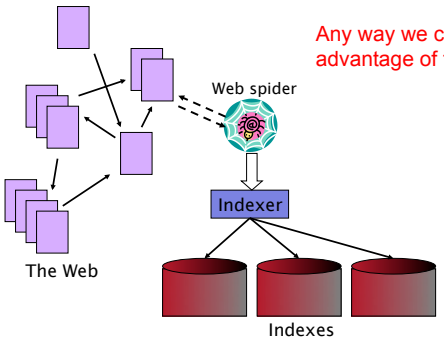
Variants of keyword stuffing

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, …"

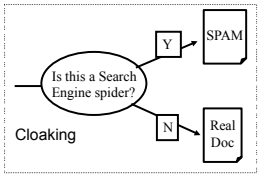
Spidering/indexing



Any way we can take advantage of this system?

Cloaking

Serve fake content to search engine spider



More spam techniques

Doorway pages

- Pages optimized for a single keyword that re-direct to the real target page

Link spamming/link farms

- 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

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
- Suspect pattern detection

More on spam

Web search engines have policies on SEO practices they tolerate/block
- http://help.yahoo.com/help/us/ysearch/index.html
- http://www.google.com/intl/en/webmasters/

Adversarial IR: the unending (technical) battle between SEO's and web search engines

Research http://airweb.cse.lehigh.edu/

Size of the web

http://www.stormforce31.com/wximages/www.jpg

What is the size of the web?

BIG!

http://www.worldwidewebsize.com/

What is the size of the web?

The web is really infinite

- Dynamic content, e.g., calendar
- Soft 404: www.yahoo.com/<anything> is a valid page

What about just the static web… issues?

- Static web contains syntactic duplication, mostly due to mirroring (~30%)
- Some servers are seldom connected
- What do we count? A url? A frame? A section? A pdf document? An image?

Who cares about the size of the web?

It is an interesting question, but beyond that, who cares and why?

Media, and consequently the user

Search engine designer (crawling, indexing)

Researchers

What can we measure?

Besides absolute size, what else might we measure?

Users interface is through the search engine

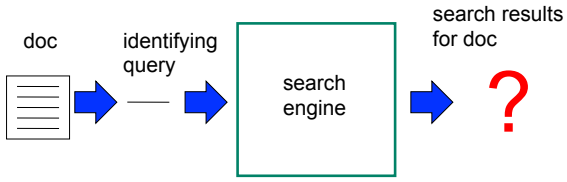
- Proportion of the web a particular search engine indexes
- The size of a particular search engine's index
- Relative index sizes of two search engines

Challenges with these approaches?

Biggest one: search engines don't like to let people know what goes on under the hood

Search engines as a black box

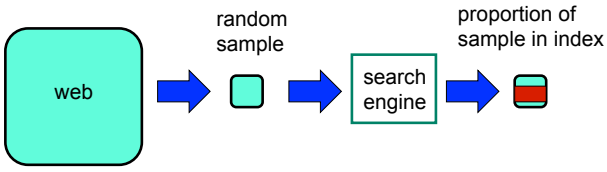
Although we can't ask how big a search engine's index is, we can often ask questions like "does a document exist in the index?"



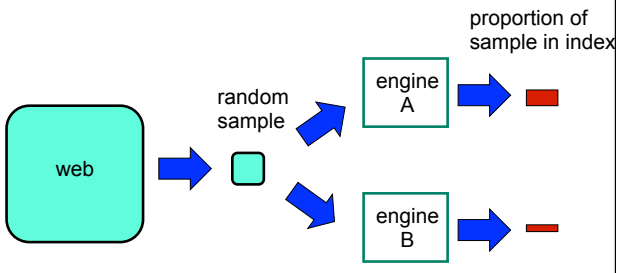
Proportion of the web indexed

We can ask if a document is in an index

How can we estimate the proportion indexed by a particular search engine?



Size of index A relative to index B



Sampling URLs

Both of these questions require us to have a random set of pages (or URLs)

Problem: Random URLs are hard to find!

Ideas?

Approach 1: Generate a random URL contained in a given engine

- Suffices for the estimation of relative size

Approach 2: Random pages/ IP addresses

- In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexes)

Random URLs from search engines

Issue a random query to the search engine

- Randomly generate a query from a lexicon and word probabilities (generally focus on less common words/queries)
- Choose random searches extracted from a query log (e.g. all queries from Middlebury College)

From the first 100 results, pick a random page/URL

Things to watch out for

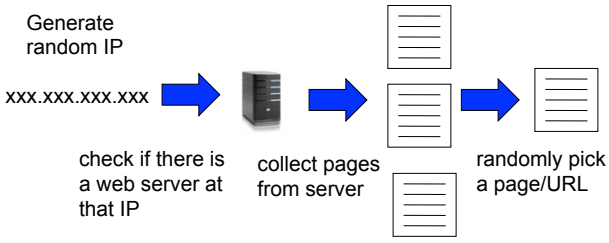
Biases induced by random queries

- Query Bias: Favors content-rich pages in the language(s) of the lexicon
- Ranking Bias: Use conjunctive queries & fetch all
- Checking Bias: Duplicates, impoverished pages omitted
- Malicious Bias: Sabotage by engine
- Operational Problems: Time-outs, failures, engine inconsistencies, index modification

Biases induced by query log

- Samples are correlated with source of log

Random IP addresses



Random IP addresses

[Lawr99] Estimated 2.8 million IP addresses running crawlable web servers (16 million total) from observing 2500 servers

OCLC using IP sampling found 8.7 M hosts in 2001

Netcraft [Netc02] accessed 37.2 million hosts in July 2002

Random walks

View the Web as a directed graph

Build a random walk on this graph

- Includes various "jump" rules back to visited sites
  - Does not get stuck in spider traps!
  - Can follow all links!
- Converges to a stationary distribution
  - Must assume graph is finite and independent of the walk.
  - Conditions are not satisfied (cookie crumbs, flooding)
  - Time to convergence not really known
- Sample from stationary distribution of walk
- Use the "strong query" method to check coverage by SE

Conclusions

No sampling solution is perfect

Lots of new ideas ...

....but the problem is getting harder

Quantitative studies are fascinating and a good research problem