

Search engines:

- Importance (Introna/Nissenbaum paper; IGNORE the Tech Overview section, see below)
- anatomy of a search engine
- business model of search engine companies
- new developments

Importance

- The search engine is the main tool through which the users see the web
- Web engines see only a fraction of the web
- How many webpages are there? One estimate: 10^{12}
- How many are indexed by search engines? 25%?

Anatomy of a search engine: google

- Google history
 - BackRub, stanford U. 1996 (Page/Brin)
 - googol = 10^{100}
 - Search, images, videos (youtube), chat, gmail, google earth, scholar,...

Page rank PR

- Idea: a page is important when it is referred to a lot, or referred to from an important page
- PR is used to prioritize; works well even with search is just on page titles
- This differs between engines: for 12,000 random queries, first ranked page was identical for 1.1%

Anatomy of a search engine

- Design criteria
- Architecture
- Data structures

Requirements

- Basic IR concepts:
 - Recall: what % of relevant docs are retrieved
 - Precision: what % of docs retrieved are relevant
- Quantity:
 - handle hundreds of thousands of queries/sec
- Quality
 - High precision (not with pres. engines)

Operation

- Crawling
- Searching
- Ranking

Page rank

- Idea: a page is important when it is referred to a lot, or referred to from an important page
- PR is used to prioritize; works well even with search is just on page titles
- This differs: only 1.1% of queries to AskJeeves, Yahoo, MSN, Google agree on the top page

PR details

- Pages T_1, \dots, T_n point to page A, $C(A)$ is a link fan-out of A

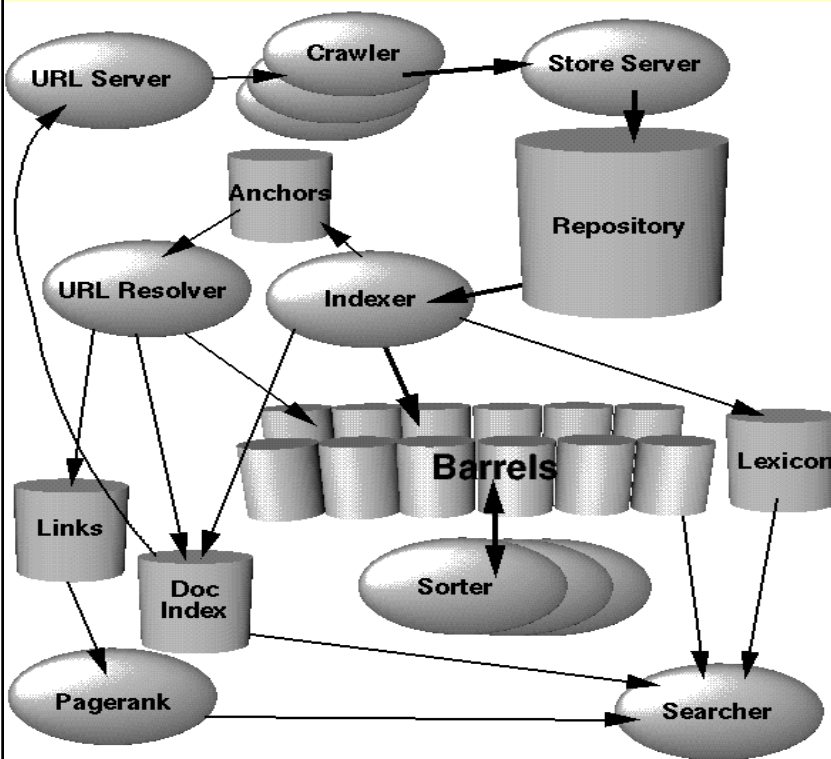
$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

d = dumping factor = .85

Model of random walk on the Web

$PR(p)$ = prob. that a “random” user will visit p

Google architecture



- URL server sends list of URLs to be fetched to crawlers
- StoreServer compresses and stores pages
- Indexer extracts words, their pos., size, capital.
- Anchors contain links and their text
- Sorter generates *inverted index*
- Searcher uses Lexicon, II, and PR

Some details

- Barrels store words (wordIDs); if a doc contains a word, doc`s ID and its wordID are stored with hitlist of this word in the doc
- Lexicon points to Inverted Barrels; ea word points to docid and hits

Crawling and indexing

- Parsing into anchors and words – error robustness (flex+stack)
- Indexing in parallel – hashing into barrels using the lexicon – the problem of new words shared

Searching

- 1 parse query
- 2 convert words into wordIDs
- 3 Identify a barrel for ea. word
- 4 scan doclists until a doc that matches all the search words is found


Ranking

- For a single word, identify the hit list and its type, count the # of hits of ea type, vector-multiply
- Combine with PR
- For multiple words, *take proximity into account*

Business models of search engines

- How do search engine companies make money?
- Google capitalization = \$130B (GM = \$16B)!
- Targeted advertising – sponsored links

Web Images Videos Shopping News Maps More | MSN Hotmail Sign in Canada (français) Preferences

bing 

Web Images Videos More ▾

RELATED SEARCHES
Cheap Hotels
Best Western Hotels
Hotels.ca
Holiday Inn
Niagara Falls Hotels
Hilton Hotels
Delta Hotels
Hotel Reservations

SEARCH HISTORY
Search more to see your history

See all
Clear all - Turn off

ALL RESULTS 1-11 of 530,000,000 results - [Advanced](#)

Hotels.com Official Site · www.hotels.com Sponsored sites

Get low CDN rates on top **hotels** from the **hotel** experts. Book today!

Hotel Deals at Expedia.ca · www.expedia.ca

Get traveller reviews, photos & exclusive rates on thousands of **hotels**

Niagara ON Hotel · www.GreatWolf.com/NiagaraFalls

Indoor waterpark & other attractions at the Great Wolf, Niagara Falls.

Hotels · www.booking.com/Hotels

Book now and save up to 75%. No reservation fee and pay at your **hotel**.

hotels.com | Hotel Bookings & Reservations Hotel Offers ...
www.hotels.com
Choose from luxury to cheap, B&B to 5 star **hotels**. Read reviews, detailed descriptions, maps and quality photos. Book today and save with our Price Match Guarantee.
Promotions/Last Minute Deals Bookings
Customer Service Jobs

Hotel - [Wikipedia, the free encyclopedia](#)
Etymology · Types · Management · Historic **hotels**
A **hotel** is an establishment that provides paid lodging on a short-term basis. The provision of basic accommodation, in times past, consisting only of a room with a bed, a ...
en.wikipedia.org/wiki/Hotel

[toronto.com, Toronto Hotels, Toronto Hotel, Accommodations, Hotel ...](#)
toronto.com is Toronto's leading online lodging guide for **hotels** in Toronto. Get all the information you need to know about any **hotel, hotels, downtown hotels, boutique hotels** ...
www.toronto.com/hotels

[Hotels, Rooms, Reservations - Choice Hotels Canada](#)
Official Site. Choice **Hotels** Canada provides **hotel** rooms at great rates and great value. Get free breakfast and Internet at most locations - Comfort, Comfort Suites, Quality ...
www.choice**hotels**.ca

[Winnipeg Hotels Guide: Hotels in Winnipeg, Manitoba \(MB\)](#)

Bing search for "hotel"

Sponsored sites

Niagara Falls Hotel
Reserve a Room Now & Enjoy All The Amenities Of Our Luxurious **Hotel**.
www.CairnCroft.com

Tremblant Winter Romance
Surprise your special someone. Romantic 2 night getaway to Tremblant.
www.rvmt.com

Bermuda Fall Special
Every 2 Nights Gets 3rd Night Free. Rates Starting at \$125 Per Night.
BermudaTourism.com

Cheap Hotels
Hotel Bookings - Compare & Save up to 70%
www.CompareOnlineHotels.com

Cheap Hotel Rates
Cheapest **Hotel** Rates at TripMama®. Compare & Save Up to 65%.
www.TripMama.com/Hotels

[See your message here](#)

Find: Next Previous Highlight all Match case

http://0.r.msn.com/?id=4vmfFzK7zIIEK0zQz86dzUngPOGNWGPQ81Xf9TBNz_4COdvnWZinOUPKKoae8a3tOB-82FUUD_3dtIGHWyxABr-17eMtoUmHjKGOw3ooBpCH2up5kMIlyon36AY836nkOAE2f3sEY2zrM-Sus4GJRQZ06F6btFvXTHFvCWnQJ8tgJ8H1fstA4tk...

start | 6 Firefox | Inbox - Mozill... | F:\courses\29... | Microsoft Pow... | BenoitTrudelC... | EN Desktop | 59% | 8:47 PM

Google coverage

- Nissenbaum paper: “if you’re not in google, you don’t exist”
- What percentage of the Internet does google cover – estimates vary between 16% and 50%
- 35% of US searches and 65% of international searches

- Internet democratizes access to information and decentralizes it
- Internet contents are increasingly seen only through Google
- Irrefutable signs of gradual centralization and commercialization
- Is there systematic under-representation of some sites (short of systematic censorship)

- “...[crawlers] are guided by a set of criteria that steer them in a systematic way [...] not to select [certain types of sites and pages]
- Self-exclusion (by forbidding crawlers)
- Importance of ranking
- “[people] are most likely to find popular, large sites whose designers have enough technical savvy to succeed in the ranking game”:
keywords in comments, appropriate page titles, repeating hidden keywords, etc
- Do ads intervene in the ranking algorithm?

- Advocates public knowledge of the underlying algorithms
- And public support for developing more egalitarian and inclusive search engines (but Quaero just failed)

- She contends that “high percentage of search requests [...] are directed to a small percentage of the big markets” and the other way round; uses 80/20% as an example
- This would make the 20% of searches underserved
- Market argument around this and its alleged fallacies: Web as a special kind of public place: Hyde Park of the electronic age

What info does Google collect about its users?

IP address

browser

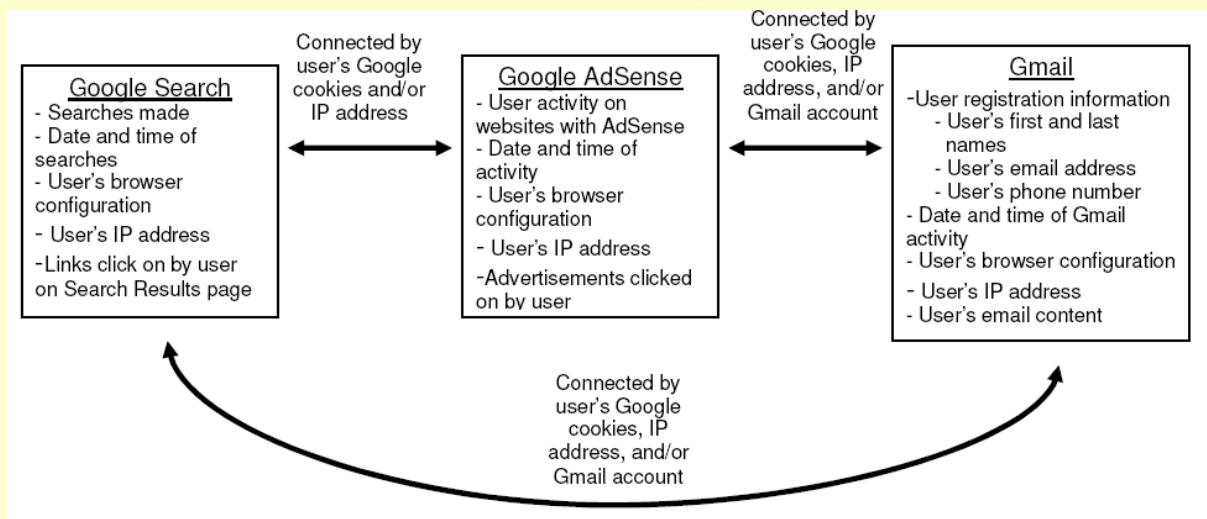
- Delichatsios/Sonuyi paper
- Data collected by Google Search: server log:
18.127.42.66 - 5/Dec/2005 9:20:46 -
http://www.google.com/search?q= dictionary - Firefox 1.0.7; Windows NT
5.1 - 740674ce123969
- Note IP address
 - www.Whatismyip.com
 - www.Whois.sc
- A cookie is placed on the user machine (740674ce123969)
- Helps link information about searches and other activities

- Also *links clicked from the search results page* are recorded
- AdSense delivers targeted ads
 - By type of webpage in the search result
 - By info about the user (i.e. their browser)
 - Each click for the CPC/CPM payment methods: CPC = charge per click (e.g. \$0.10), CPM=charge per “impression”; e.g. pay \$0.01 per 1K impressions (ad shown)
- Read Advertising and Search Engines article re CPC
 - Bidding for words (e.g. mesothelioma) to rank ads
 - Click fraud

Gmail

- since it requires registration, it records the name and secondary email address; in some cases, a cell number
- Scans email contents to place ads
- Server logs are recorded

Synthesized data: google, adsense, gmail

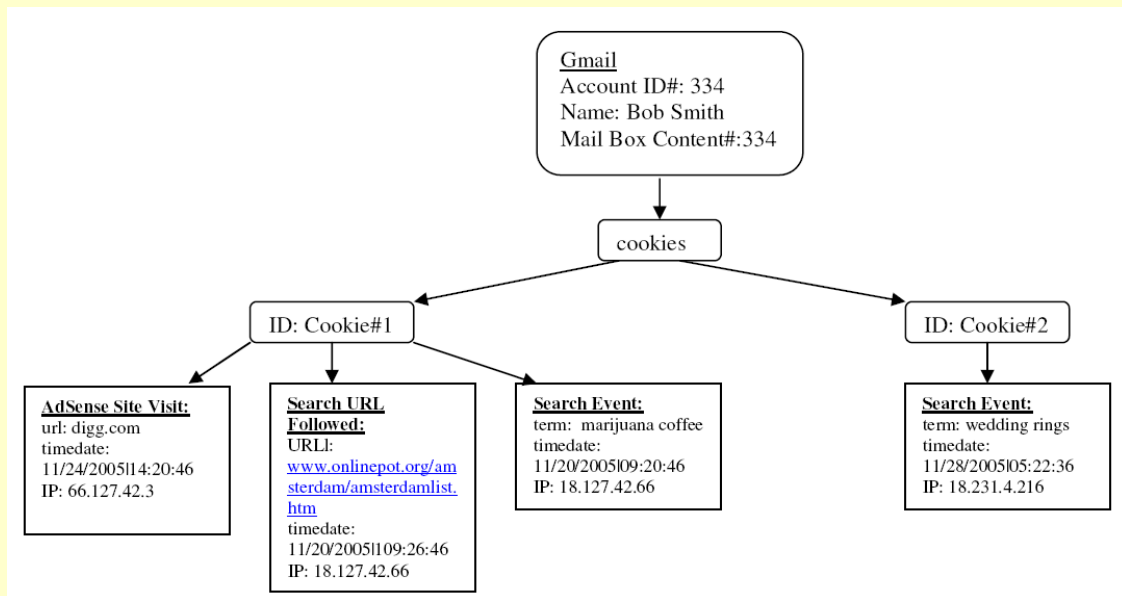


- Connected by cookies, IP address

Data synthesis scenario

- Bob does search at home in Boston
- Clicking a page in search results
- Uses gmail
- Travels to SF and uses laptop to visit digg.com
- Checks gmail
- Returns to MIT and erases cookies
- Searches for “wedding ring”
- Cookie is placed on his machine
- Link is logged, with cookie
- Activity is logged, incl. cookie
- SF IP is logged
- Activity is logged, incl. cookie
- Boston IP is logged; new cookie is issued
- Search is logged

Hypothetical profile



- Searches (what for), sites, IP s, gmail activity

- Also: +Google is subject to gov't subpoenas, e.g. under the Patriot Act
- All info google has can be transferred in case google is acquired

New developments in search engines

Work on understanding user context

- “...When we seek, we are not interested in “information” in general; rather we are interested in specific information related to our specific interests and needs.” (Introna/Nissenbaum paper)

Consider the *movie* relation with schema (*title, actor, genre, language*):

1. Nicole Kidman > Penelope Cruz | Drama
2. Penelope Cruz > Nicole Kidman | Drama and Spanish

These preferences illustrate that the ranking of the tuples of a relation is subjective. Nicole Kidman has played in many more movies than Penelope Cruz and has been nominated for many more academy awards. However, in the context of Spanish dramas the latter actress is considered more important.

Challenge: how to incorporate preferences into query answering? See 2006 paper by R. Agrawal

New developments in search engines

- Bing – try
 - Clustering
- “Deep web”

Discussion topic

- The proposed CRTC change in internet service charges: pros and cons
- Introduction
- Do own research
- Participate on line