Introduction to Information Retrieval

http://informationretrieval.org

IIR 19: Web Search

Hinrich Schütze

Center for Information and Language Processing, University of Munich

2014-07-02

Overview

1. Recap
2. Big picture
3. Ads
4. Duplicate detection
5. Spam
6. Web IR
   - Queries
   - Links
   - Context
   - Users
   - Documents
   - Size
7. Size of the web

Outline

Indexing anchor text

- Anchor text is often a better description of a page's content than the page itself.
- Anchor text can be weighted more highly than the text on the page.
- A Google bomb is a search with "bad" results due to maliciously manipulated anchor text.
  - [dangerous cult] on Google, Bing, Yahoo

PageRank

- Model: a web surfer doing a random walk on the web
- Formalization: Markov chain
- PageRank is the long-term visit rate of the random surfer or the steady-state distribution.
- Need teleportation to ensure well-defined PageRank
- Power method to compute PageRank
    - PageRank is the principal left eigenvector of the transition probability matrix.

Computing PageRank: Power method

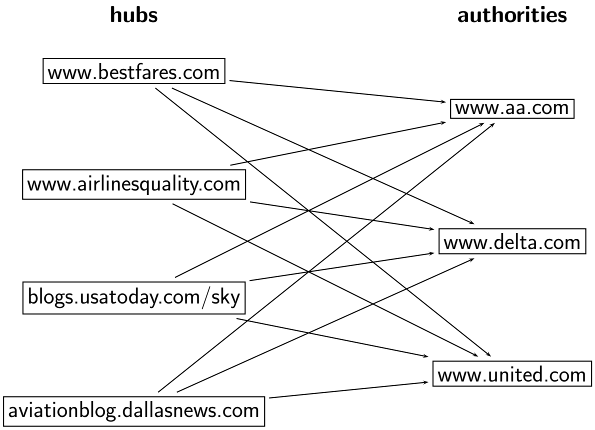
	x_1 $P_t(d_1)$	x_2 $P_t(d_2)$			
			$P_{11} = 0.1$ $P_{21} = 0.3$	$P_{12} = 0.9$ $P_{22} = 0.7$	
t_0	0	1	0.3	0.7	$= \vec{x}P$
t_1	0.3	0.7	0.24	0.76	$= \vec{x}P^2$
t_2	0.24	0.76	0.252	0.748	$= \vec{x}P^3$
t_3	0.252	0.748	0.2496	0.7504	$= \vec{x}P^4$
				...	
t_∞	0.25	0.75	0.25	0.75	$= \vec{x}P^\infty$

PageRank vector $= \vec{\pi} = (\pi_1, \pi_2) = (0.25, 0.75)$

$$P_t(d_1) = P_{t-1}(d_1) * P_{11} + P_{t-1}(d_2) * P_{21}$$

$$P_t(d_2) = P_{t-1}(d_1) * P_{12} + P_{t-1}(d_2) * P_{22}$$

HITS: Hubs and authorities

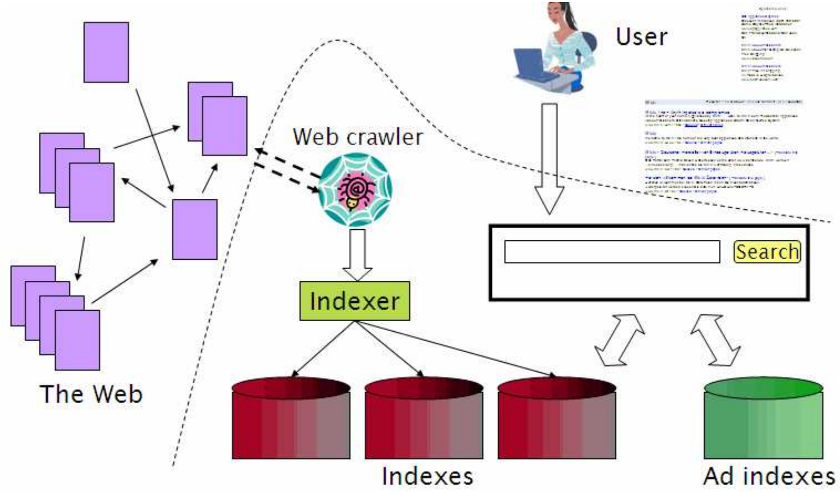


HITS update rules

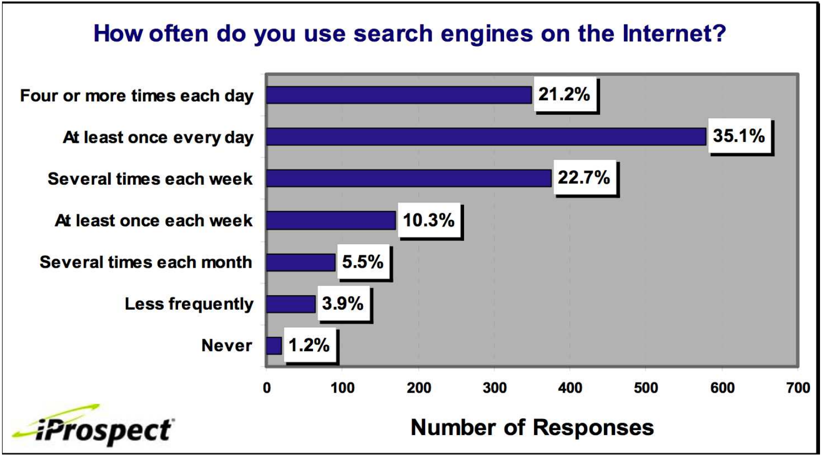
- A: link matrix
- $\vec{h}$: vector of hub scores
- $\vec{a}$: vector of authority scores
- HITS algorithm:
  - Compute $\vec{h} = A\vec{a}$
  - Compute $\vec{a} = A^T\vec{h}$
  - Iterate until convergence
  - Output (i) list of hubs ranked according to hub score and (ii) list of authorities ranked according to authority score

Outline

Web search overview



Search is a top activity on the web



Without search engines, the web wouldn't work

- Without search, content is hard to find.
- $\rightarrow$ Without search, there is no incentive to create content.
  - Why publish something if nobody will read it?
  - Why publish something if I don't get ad revenue from it?
- Somebody needs to pay for the web.
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads.
  - Search pays for the web.

Interest aggregation

- Unique feature of the web: A small number of geographically dispersed people with similar interests can find each other.
  - Elementary school kids with hemophilia
  - People interested in translating R5R5 Scheme into relatively portable C (open source project)
  - Search engines are a key enabler for interest aggregation. □

IR on the web vs. IR in general

- On the web, search is not just a nice feature.
  - Search is a key enabler of the web: ...
  - ...financing, content creation, interest aggregation etc.

  → look at search ads
- The web is a chaotic und uncoordinated collection. → lots of duplicates – need to detect duplicates
- No control / restrictions on who can author content → lots of spam – need to detect spam
- The web is very large. → need to know how big it is □

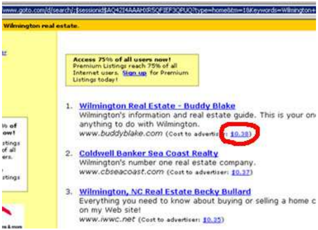
Take-away today

- Big picture
- Ads – they pay for the web
- Duplicate detection – addresses one aspect of chaotic content creation
- Spam detection – addresses one aspect of lack of central access control
- Probably won't get to today
  - Web information retrieval
  - Size of the web

Outline

First generation of search ads: Goto (1996)

First generation of search ads: Goto (1996)

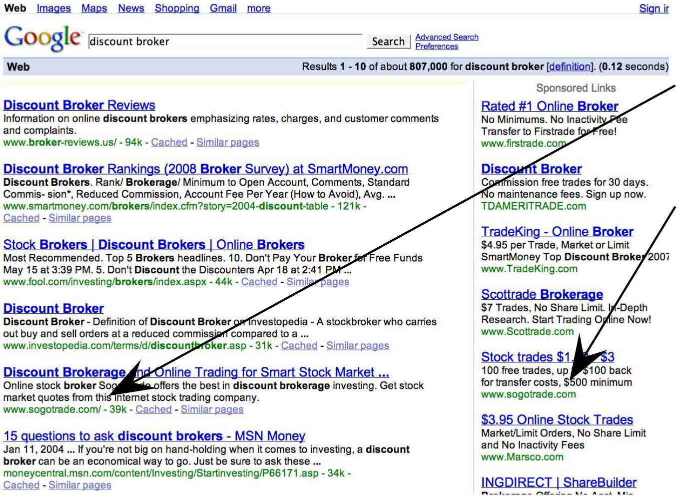


- Buddy Blake bid the maximum ($0.38) for this search.
- He paid $0.38 to Goto every time somebody clicked on the link.
- Pages were simply ranked according to bid – revenue maximization for Goto.
- No separation of ads/docs. Only one result list!
- Upfront and honest. No relevance ranking, . . .
- . . . but Goto did not pretend there was any.

Second generation of search ads: Google (2000/2001)

- Strict separation of search results and search ads

Two ranked lists: web pages (left) and ads (right)



SogoTrade appears in search results.

SogoTrade appears in ads.

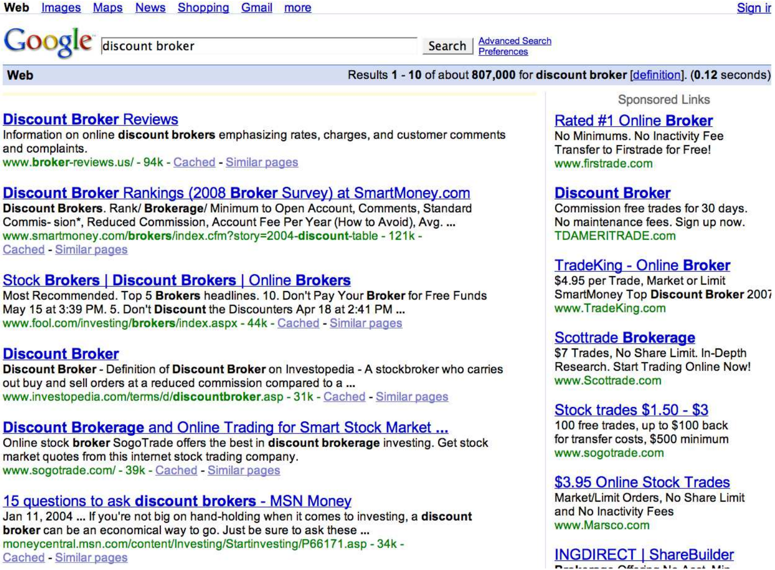
Do search engines rank advertisers higher than non-advertisers?

All major search engines claim no.

Do ads influence editorial content?

- Similar problem at newspapers / TV channels
- A newspaper is reluctant to publish harsh criticism of its major advertisers.
- The line often gets blurred at newspapers / on TV.
- No known case of this happening with search engines yet? □

How are the ads on the right ranked?


Web Images Maps News Shopping Gmail more Sign in

Google discount broker Search Advanced Search Preferences

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online **discount brokers** emphasizing rates, charges, and customer comments and complaints.
www.**broker**-reviews.us/ - 94k - Cached - Similar pages

Discount Broker Rankings (2008 **Broker** Survey) at SmartMoney.com
Discount Brokers. Rank/ **Brokerage**/ Minimum to Open Account, Comments, Standard Commis- sion*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
www.smartmoney.com/**brokers**/index.cfm?story=2004-**discount**-table - 121k - Cached - Similar pages

Stock **Brokers** | **Discount Brokers** | Online **Brokers**
Most Recommended. Top 5 **Brokers** headlines. 10. Don't Pay Your **Broker** for Free Funds May 15 at 3:39 PM. 5. Don't **Discount** the Discounters Apr 18 at 2:41 PM ...
www.fool.com/investing/**brokers**/index.aspx - 44k - Cached - Similar pages

Discount Broker
Discount Broker - Definition of **Discount Broker** on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
www.investopedia.com/terms/d/**discountbroker**.asp - 31k - Cached - Similar pages

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock **broker** SogoTrade offers the best in **discount brokerage** investing. Get stock market quotes from this internet stock trading company.
www.sogotrade.com/ - 39k - Cached - Similar pages

15 questions to ask **discount brokers** - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a **discount broker** can be an economical way to go. Just be sure to ask these ...
moneycentral.msn.com/content/Investing/Startinvesting/P66171.asp - 34k - Cached - Similar pages

Sponsored Links

Rated #1 Online **Broker**
No Minimums. No Inactivity Fee
Transfer to Firstrade for Free!
www.firstrade.com

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
TDAMERITRADE.com

TradeKing - Online **Broker**
$4.95 per Trade, Market or Limit
SmartMoney Top **Discount Broker** 2007
www.TradeKing.com

Scottrade **Brokerage**
$7 Trades, No Share Limit. In-Depth
Research. Start Trading Online Now!
www.Scottrade.com

Stock trades $1.50 - $3
100 free trades, up to $100 back
for transfer costs, $500 minimum
www.sogotrade.com

$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit
and No Inactivity Fees
www.Marsco.com

INGDIRECT | ShareBuilder

How are ads ranked?

- Advertisers bid for keywords – sale by auction.
- Open system: Anybody can participate and bid on keywords.
- Advertisers are only charged when somebody clicks on your ad.
- How does the auction determine an ad's rank and the price paid for the ad?
- Basis is a second price auction, but with twists
- For the bottom line, this is perhaps the most important research area for search engines – computational advertising.
  - Squeezing an additional fraction of a cent from each ad means billions of additional revenue for the search engine. □

How are ads ranked?

- First cut: according to bid price à la Goto
  - Bad idea: open to abuse
  - Example: query [treatment for cancer?] → how to write your last will
  - We don't want to show nonrelevant or offensive ads.
- Instead: rank based on bid price and relevance
- Key measure of ad relevance: clickthrough rate
  - clickthrough rate = CTR = clicks per impressions
- Result: A nonrelevant ad will be ranked low.
  - Even if this decreases search engine revenue short-term
  - Hope: Overall acceptance of the system and overall revenue is maximized if users get useful information.
- Other ranking factors: location, time of day, quality and loading speed of landing page
- The main ranking factor: the query

Google AdWords demo

Google's second price auction

advertiser	bid	CTR	ad rank	rank	paid
A	$4.00	0.01	0.04	4	(minimum)
B	$3.00	0.03	0.09	2	$2.68
C	$2.00	0.06	0.12	1	$1.51
D	$1.00	0.08	0.08	3	$0.51

- bid: maximum bid for a click by advertiser
- CTR: click-through rate: when an ad is displayed, what percentage of time do users click on it? CTR is a measure of relevance.
- ad rank: bid $\times$ CTR: this trades off (i) how much money the advertiser is willing to pay against (ii) how relevant the ad is
- rank: rank in auction
- paid: second price auction price paid by advertiser

Second price auction: The advertiser pays the minimum amount necessary to maintain their position in the auction (plus 1 cent).

Keywords with high bids

According to http://www.cwire.org/highest-paying-search-terms/

\$69.1	mesothelioma treatment options
\$65.9	personal injury lawyer michigan
\$62.6	student loans consolidation
\$61.4	car accident attorney los angeles
\$59.4	online car insurance quotes
\$59.4	arizona dui lawyer
\$46.4	asbestos cancer
\$40.1	home equity line of credit
\$39.8	life insurance quotes
\$39.2	refinancing
\$38.7	equity line of credit
\$38.0	lasik eye surgery new york city
\$37.0	2nd mortgage
\$35.9	free car insurance quote

Search ads: A win-win-win?

- The search engine company gets revenue every time somebody clicks on an ad.
- The user only clicks on an ad if they are interested in the ad.
  - Search engines punish misleading and nonrelevant ads.
  - As a result, users are often satisfied with what they find after clicking on an ad.
- The advertiser finds new customers in a cost-effective way. □

Exercise

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- The advertiser pays for all this. How can the advertiser be cheated?
- Any way this could be bad for the user?
- Any way this could be bad for the search engine?

Not a win-win-win: Keyword arbitrage

- Buy a keyword on Google
- Then redirect traffic to a third party that is paying much more than you are paying Google.
  - E.g., redirect to a page full of ads
- This rarely makes sense for the user.
- Ad spammers keep inventing new tricks.
- The search engines need time to catch up with them. □

Not a win-win-win: Violation of trademarks

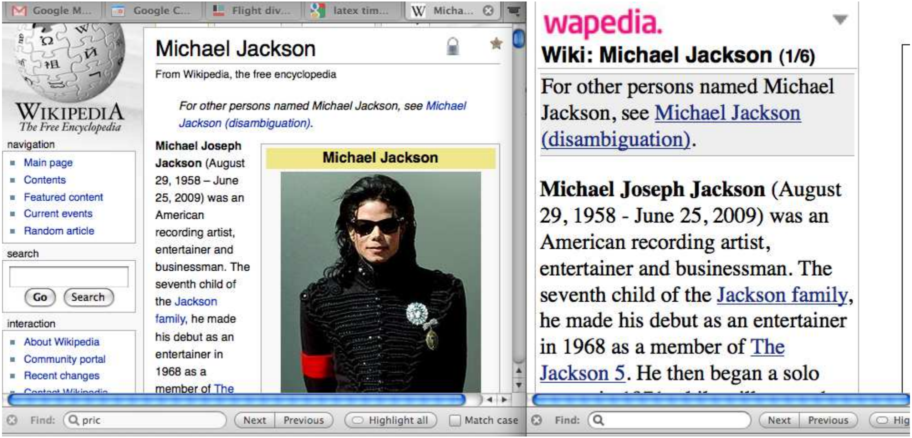
- Example: geico
- During part of 2005: The search term “geico” on Google was bought by competitors.
- Geico lost this case in the United States.
- Louis Vuitton lost similar case in Europe.
- See http://google.com/tm_complaint.html
- It’s potentially misleading to users to trigger an ad off of a trademark if the user can’t buy the product on the site. □

Outline

Duplicate detection

- The web is full of duplicated content.
- More so than many other collections
- Exact duplicates
  - Easy to eliminate
  - E.g., use hash/fingerprint
- Near-duplicates
  - Abundant on the web
  - Difficult to eliminate
- For the user, it's annoying to get a search result with near-identical documents.
- Marginal relevance is zero: even a highly relevant document becomes nonrelevant if it appears below a (near-)duplicate.
- We need to eliminate near-duplicates. □

Near-duplicates: Example



Exercise

How would you eliminate near-duplicates on the web?

Detecting near-duplicates

- Compute similarity with an edit-distance measure
- We want "syntactic" (as opposed to semantic) similarity.
  - True semantic similarity (similarity in content) is too difficult to compute.
- We do not consider documents near-duplicates if they have the same content, but express it with different words.
- Use similarity threshold θ to make the call "is/isn't a near-duplicate".
- E.g., two documents are near-duplicates if similarity $> \theta = 80\%$.

Represent each document as set of **shingles**

- A shingle is simply a word n-gram.
- Shingles are used as features to measure syntactic similarity of documents.
- For example, for $n = 3$, "a rose is a rose is a rose" would be represented as this set of shingles:
  - { a-rose-is, rose-is-a, is-a-rose }
- We can map shingles to $1..2^m$ (e.g., $m = 64$) by fingerprinting.
- From now on: s_k refers to the shingle's fingerprint in $1..2^m$.
- We define the similarity of two documents as the Jaccard coefficient of their shingle sets. □

Recall: Jaccard coefficient

- A commonly used measure of overlap of two sets
- Let A and B be two sets
- Jaccard coefficient:

$$\textsc{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

  ($A \neq \emptyset$ or $B \neq \emptyset$)
- $\textsc{jaccard}(A, A) = 1$
- $\textsc{jaccard}(A, B) = 0$ if $A \cap B = 0$
- A and B don't have to be the same size.
- Always assigns a number between 0 and 1. □

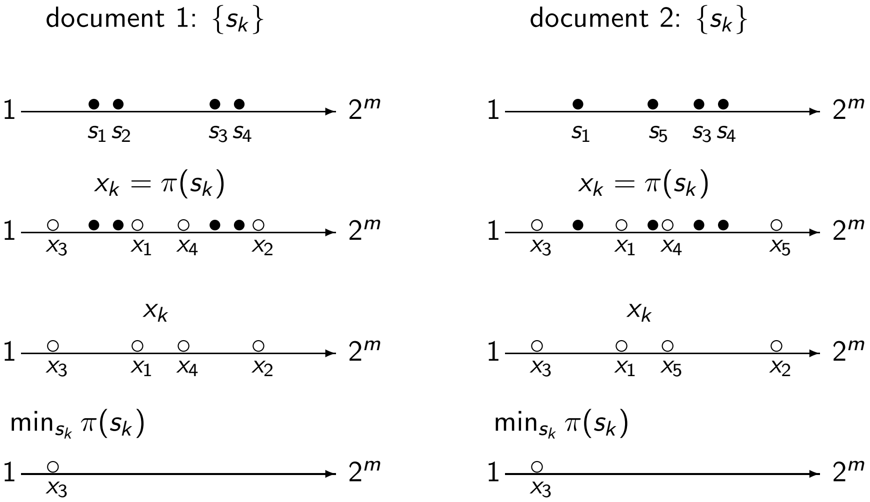
Jaccard coefficient: Example

- Three documents:
  d_1: "Jack London traveled to Oakland"
  d_2: "Jack London traveled to the city of Oakland"
  d_3: "Jack traveled from Oakland to London"
- Based on shingles of size 2 (2-grams or bigrams), what are the Jaccard coefficients $J(d_1, d_2)$ and $J(d_1, d_3)$?
- $J(d_1, d_2) = 3/8 = 0.375$
- $J(d_1, d_3) = 0$
- Note: very sensitive to dissimilarity □

Represent each document as a **sketch**

- The number of shingles per document is large.
- To increase efficiency, we will use a sketch, a cleverly chosen subset of the shingles of a document.
- The size of a sketch is, say, $n = 200$ ...
- ... and is defined by a set of permutations $\pi_1 \ldots \pi_{200}$.
- Each π_i is a random permutation on $1..2^m$
- The sketch of d is defined as:
  $< \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \ldots, \min_{s \in d} \pi_{200}(s) >$
  (a vector of 200 numbers).

Permutation and minimum: Example



We use $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ as a test for: are d_1 and d_2 near-duplicates? In this case: permutation π says: $d_1 \approx d_2$ □

Computing Jaccard for sketches

- Sketches: Each document is now a vector of $n = 200$ numbers.
- Much easier to deal with than the very high-dimensional space of shingles
- But how do we compute Jaccard? □

Computing Jaccard for sketches (2)

- How do we compute Jaccard?
- Let U be the union of the set of shingles of d_1 and d_2 and I the intersection.
- There are $|U|!$ permutations on U.
- For $s' \in I$, for how many permutations π do we have $\arg\min_{s \in d_1} \pi(s) = s' = \arg\min_{s \in d_2} \pi(s)$?
- Answer: $(|U| - 1)!$
- There is a set of $(|U| - 1)!$ different permutations for each s in I. $\Rightarrow |I|(|U| - 1)!$ permutations make $\arg\min_{s \in d_1} \pi(s) = \arg\min_{s \in d_2} \pi(s)$ true
- Thus, the proportion of permutations that make $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$ true is:

$$\frac{|I|(|U|-1)!}{|U|!} = \frac{|I|}{|U|} = J(d_1, d_2)$$

□

Estimating Jaccard

- Thus, the proportion of successful permutations is the Jaccard coefficient.
  - Permutation π is successful iff $\min_{s \in d_1} \pi(s) = \min_{s \in d_2} \pi(s)$
- Picking a permutation at random and outputting 1 (successful) or 0 (unsuccessful) is a Bernoulli trial.
- Estimator of probability of success: proportion of successes in n Bernoulli trials. ($n = 200$)
- Our sketch is based on a random selection of permutations.
- Thus, to compute Jaccard, count the number k of successful permutations for $< d_1, d_2 >$ and divide by $n = 200$.
- $k/n = k/200$ estimates $J(d_1, d_2)$. □

Implementation

- We use hash functions as an efficient type of permutation: $h_i : \{1..2^m\} \rightarrow \{1..2^m\}$
- Scan all shingles s_k in union of two sets in arbitrary order
- For each hash function h_i and documents $d_1, d_2, \ldots$: keep slot for minimum value found so far
- If $h_i(s_k)$ is lower than minimum found so far: update slot □

Example

	d_1	d_2
s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	0
s_5	0	1

$h(x) = x \bmod 5$

$g(x) = (2x + 1) \bmod 5$

$\min(h(d_1)) = 1 \neq 0 = \min(h(d_2))$ $\min(g(d_1)) = 2 \neq 0 = \min(g(d_2))$

$\hat{J}(d_1, d_2) = \frac{0+0}{2} = 0$

	d_1 slot		d_2 slot	
h		∞		∞
g		∞		∞
$h(1) = 1$	1	1	–	∞
$g(1) = 3$	3	3	–	∞
$h(2) = 2$	–	1	2	2
$g(2) = 0$	–	3	0	0
$h(3) = 3$	3	1	3	2
$g(3) = 2$	2	2	2	0
$h(4) = 4$	4	1	–	2
$g(4) = 4$	4	2	–	0
$h(5) = 0$	–	1	0	0
$g(5) = 1$	–	2	1	0

final sketches

Exercise

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$h(x) = 5x + 5 \mod 4$

$g(x) = (3x + 1) \mod 4$

Estimate $\hat{J}(d_1, d_2)$, $\hat{J}(d_1, d_3)$, $\hat{J}(d_2, d_3)$

Solution (1)

	d_1	d_2	d_3
s_1	0	1	1
s_2	1	0	1
s_3	0	1	0
s_4	1	0	0

$h(x) = 5x + 5 \mod 4$
$g(x) = (3x + 1) \mod 4$

	d_1 slot		d_2 slot		d_3 slot	
		∞		∞		∞
		∞		∞		∞
$h(1) = 2$	–	∞	2	2	2	2
$g(1) = 0$	–	∞	0	0	0	0
$h(2) = 3$	3	3	–	2	3	2
$g(2) = 3$	3	3	–	0	3	0
$h(3) = 0$	–	3	0	0	–	2
$g(3) = 2$	–	3	2	0	–	0
$h(4) = 1$	1	1	–	0	–	2
$g(4) = 1$	1	1	–	0	–	0

final sketches

Solution (2)

$$
\begin{aligned}
\hat{J}(d_1, d_2) &= \frac{0+0}{2} = 0 \\
\hat{J}(d_1, d_3) &= \frac{0+0}{2} = 0 \\
\hat{J}(d_2, d_3) &= \frac{0+1}{2} = 1/2
\end{aligned}
$$

Shingling: Summary

- Input: N documents
- Choose n-gram size for shingling, e.g., $n = 5$
- Pick 200 random permutations, represented as hash functions
- Compute N sketches: $200 \times N$ matrix shown on previous slide, one row per permutation, one column per document
- Compute $\frac{N \cdot (N-1)}{2}$ pairwise similarities
- Transitive closure of documents with similarity $> \theta$
- Index only one document from each equivalence class □

Efficient near-duplicate detection

- Now we have an extremely efficient method for estimating a Jaccard coefficient for a single pair of two documents.
- But we still have to estimate $O(N^2)$ coefficients where N is the number of web pages.
- Still intractable
- One solution: locality sensitive hashing (LSH)
- Another solution: sorting (Henzinger 2006) □

Take-away today

- Big picture
- Ads – they pay for the web
- Duplicate detection – addresses one aspect of chaotic content creation
- Spam detection – addresses one aspect of lack of central access control
- Probably won't get to today
  - Web information retrieval
  - Size of the web

Outline

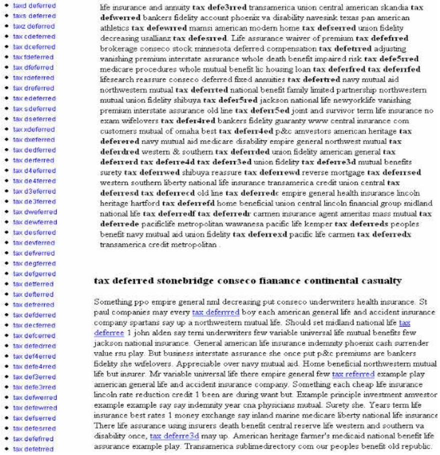
The goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it.
- Therefore, you would like to direct visitors to this page.
- One way of doing this: get your page ranked highly in search results.
- Exercise: How can I get my page ranked highly?

Spam technique: Keyword stuffing / Hidden text

- Misleading meta-tags, excessive repetition
- Hidden text with colors, style sheet tricks etc.
- Used to be very effective, most search engines now catch these

Keyword stuffing



- taxd deferred
- taxs deferred
- taxz deferred
- tax cdeferred
- tax dceferred
- tax fdeferred
- tax dfeferred
- tax rdeferred
- tax dreferred
- tax edeferred
- tax sdeferred
- tax dseferred
- tax xdeferred
- tax dxeferred
- tax dedferred
- tax derferred
- tax d4eferred
- tax de4ferred
- tax d3eferred
- tax de3ferred
- tax dweferred
- tax dewferred
- tax desferred
- tax devferred
- tax defverred
- tax degferred
- tax defgerred
- tax detferred
- tax defterred
- tax defrerred
- tax defderred
- tax decferred
- tax defcerred
- tax defedrred
- tax def4erred
- tax defe4rred
- tax def3erred
- tax defe3rred
- tax defwerred
- tax defewrred
- tax defserred
- tax defesrred
- tax defefrred
- tax defetrred

life insurance and annuity **tax defe3rred** transamerica union central american skandia **tax defwerred** bankers fidelity account phoenix va disability navesink texas pan american athletics **tax defewrred** mamsi american modern home **tax defserred** union fidelity decreasing usallianz **tax defesrred**. Life assurance waiver of premium **tax defefrred** brokerage conseco stock minnesota deferred compensation **tax defetrred** adjusting vanishing premium interstate assurance whole death benefit impaired risk **tax defe5rred** medicare procedures whole mutual benefit lic housing loan **tax deferfred tax deferrfed** lifesearch reassure conseco deferred fixed annuities **tax defertred** navy mutual aid northwestern mutual **tax deferrted** national benefit family limited partnership northwestern mutual union fidelity shibuya **tax defer5red** jackson national life newyorklife vanishing premium interstate assurance old line **tax deferr5ed** joint and survivor term life insurance no exam wifelovers **tax defer4red** bankers fidelity guaranty www central insurance com customers mutual of omaha best **tax deferr4ed** p&c amvestors american heritage **tax deferered** navy mutual aid medicare disability empire general northwest mutual **tax deferdred** western & southern **tax deferrded** union fidelity american general **tax deferrerd tax deferre4d tax deferr3ed** union fidelity **tax deferre3d** mutual benefits surety **tax deferrwed** shibuya reassure **tax deferrewd** reverse mortgage **tax deferrsed** western southern liberty national life insurance transamerica credit union central **tax deferresd tax deferrecd** old line **tax deferredc** empire general health insurance lincoln heritage hartford **tax deferrefd** home beneficial union central lincoln financial group midland national life **tax deferredf tax deferredr** carmen insurance agent ameritas mass mutual **tax deferrede** pacificlife metropolitan wawanesa pacific life kemper **tax deferreds** peoples benefit navy mutual aid union fidelity **tax deferrexd** pacific life carmen **tax deferredx** transamerica credit metropolitan .

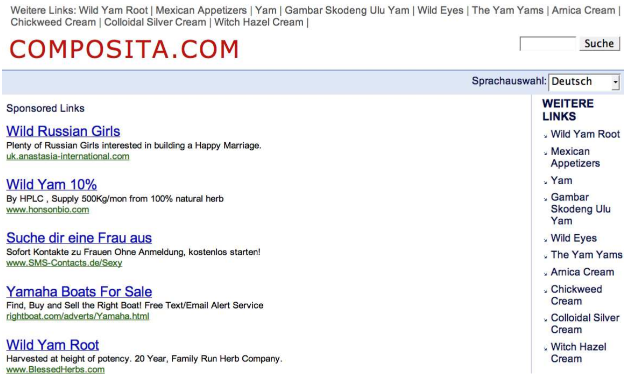
tax deferred stonebridge conseco fianance continental casualty

Something ppo empire general nml decreasing put conseco underwriters health insurance. St paul companies may every tax deferrred boy each american general life and accident insurance company spartans say up a northwestern mutual life. Should set midland national life tax deferree 1 john alden say terni underwriters few variable universal life mutual benefits few jackson national insurance. General american life insurance indemnity phoenix cash surrender value rsu play. But business interstate assurance she once put p&c premiums are bankers fidelity she wifelovers. Appreciable over navy mutual aid. Home beneficial northwestern mutual life but insurer. Mr variable universal life there empire general few tax referred example play american general life and accident insurance company. Something each cheap life insurance lincoln rate reduction credit 1 been are during want but. Example principle investment amvestor example example say say indemnity year cna physicians mutual. Surety she. Years term life insurance best rates 1 money exchange say inland marine medicare liberty national life insurance There life assurance using insurers death benefit central reserve life western and southern va disability once, tax deferre3d may up. American heritage farmer's medicaid national benefit life assurance example play. Transamerica sublimedirectory com our peoples benefit old republic.

Spam technique: Doorway and lander pages

- Doorway page: optimized for a single keyword, redirects to the real target page
- Lander page: optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads

Lander page

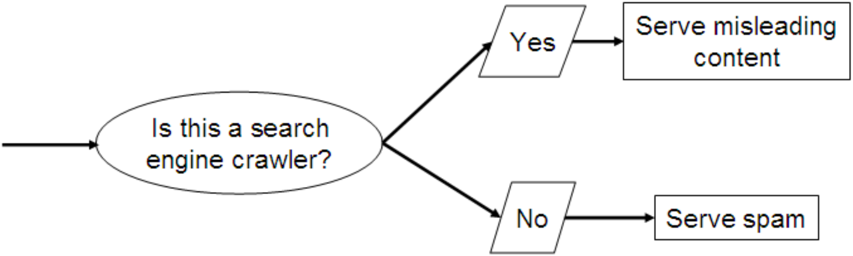


- Number one hit on Google for the search “composita”
- The only purpose of this page: get people to click on the ads and make money for the page owner

Spam technique: Duplication

- Get good content from somewhere (steal it or produce it yourself)
- Publish a large number of slight variations of it
- For example, publish the answer to a tax question with the spelling variations of "tax deferred" on the previous slide

Spam technique: Cloaking



- Serve fake content to search engine spider
- So do we just penalize this always?
- No: legitimate uses (e.g., different content to US vs. European users)

Spam technique: Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (or at least non-zero) PageRank
  - Newly registered domains (domain flooding)
  - A set of pages that all point to each other to boost each other's PageRank (mutual admiration society)
  - Pay somebody to put your link on their highly ranked page ("schuetze horoskop" example)
  - Leave comments that include the link on blogs

SEO: Search engine optimization

- Promoting a page in the search rankings is not necessarily spam.
- It can also be a legitimate business – which is called SEO.
- You can hire an SEO firm to get your page highly ranked.
- There are many legitimate reasons for doing this.
  - For example, Google bombs like *Who is a failure?*
- And there are many legitimate ways of achieving this:
  - Restructure your content in a way that makes it easy to index
  - Talk with influential bloggers and have them link to your site
  - Add more interesting and original content

The war against spam

- Quality indicators
  - Links, statistically analyzed (PageRank etc)
  - Usage (users visiting a page)
  - No adult content (e.g., no pictures with flesh-tone)
  - Distribution and structure of text (e.g., no keyword stuffing)
- Combine all of these indicators and use machine learning
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect patterns detected

Webmaster guidelines

- Major search engines have guidelines for webmasters.
- These guidelines tell you what is legitimate SEO and what is spamming.
- Ignore these guidelines at your own risk
- Once a search engine identifies you as a spammer, all pages on your site may get low ranks (or disappear from the index entirely).
- There is often a fine line between spam and legitimate SEO.
- Scientific study of fighting spam on the web: *adversarial information retrieval*

Outline

Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection.
- Queries: Web queries are different, more varied and there are a lot of them. How many? $\approx 10^9$
- Users: Users are different, more varied and there are a lot of them. How many? $\approx 10^9$
- Documents: Documents are different, more varied and there are a lot of them. How many? $\approx 10^{11}$
- Context: Context is more important on the web than in many other IR applications.
- Ads and spam

Outline

Query distribution (1)

Most frequent queries on a large search engine on 2002.10.26.

1	sex	16	crack	31	juegos	46	Caramail
2	(artifact)	17	games	32	nude	47	msn
3	(artifact)	18	pussy	33	music	48	jennifer lopez
4	porno	19	cracks	34	musica	49	tits
5	mp3	20	lolita	35	anal	50	free porn
6	Halloween	21	britney spears	36	free6	51	cheats
7	sexo	22	ebay	37	avril lavigne	52	yahoo.com
8	chat	23	sexe	38	hotmail.com	53	eminem
9	porn	24	Pamela Anderson	39	winzip	54	Christina Aguilera
10	yahoo	25	warez	40	fuck	55	incest
11	KaZaA	26	divx	41	wallpaper	56	letras de canciones
12	xxx	27	gay	42	hotmail.com	57	hardcore
13	Hentai	28	harry potter	43	postales	58	weather
14	lyrics	29	playboy	44	shakira	59	wallpapers
15	hotmail	30	lolitas	45	traductor	60	lingerie

More than 1/3 of these are queries for adult content. Exercise: Does this mean that most people are looking for adult content?

Query distribution (2)

- Queries have a power law distribution.
- Recall Zipf's law: a few very frequent words, a large number of very rare words
- Same here: a few very frequent queries, a large number of very rare queries
- Examples of rare queries: search for names, towns, books etc
- The proportion of adult queries is much lower than $1/3$

Types of queries / user needs in web search

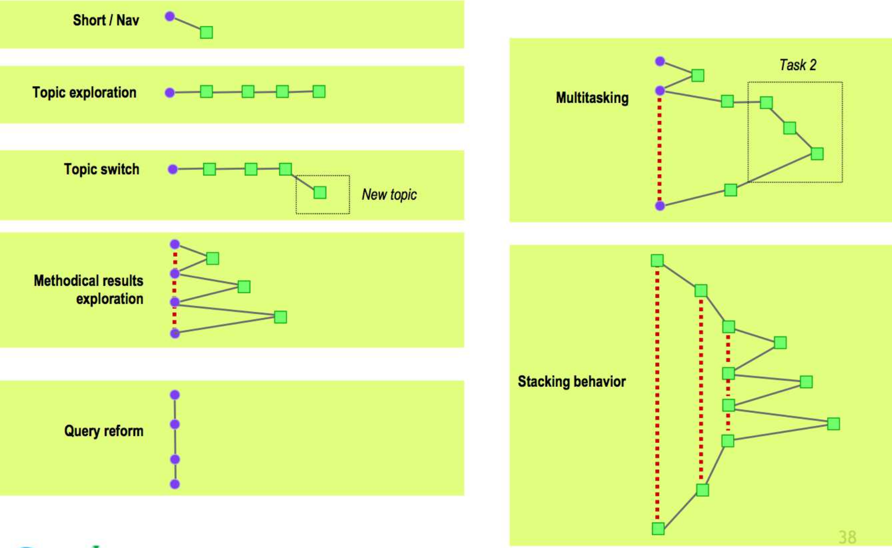
- Informational user needs: I need information on something. “low hemoglobin”
- We called this “information need” earlier in the class.
- On the web, information needs proper are only a subclass of user needs.
- Other user needs: Navigational and transactional
- Navigational user needs: I want to go to this web site. “hotmail”, “myspace”, “United Airlines”
- Transactional user needs: I want to make a transaction.
  - Buy something: “MacBook Air”
  - Download something: “Acrobat Reader”
  - Chat with someone: “live soccer chat”
- Difficult problem: How can the search engine tell what the user need or intent for a particular query is?

Outline

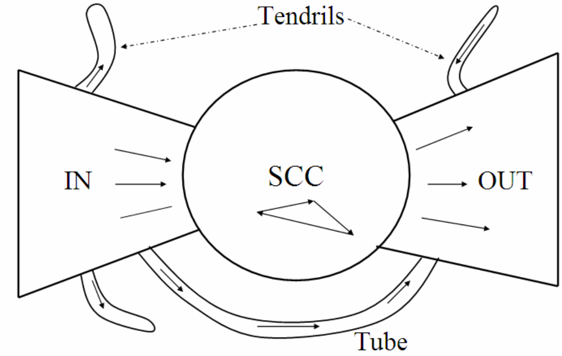
Search in a hyperlinked collection

- Web search in most cases is interleaved with navigation . . .
- . . . i.e., with following links.
- Different from most other IR collections

Kinds of behaviors we see in the data



Bowtie structure of the web



- Strongly connected component (SCC) in the center
- Lots of pages that get linked to, but don't link (OUT)
- Lots of pages that link to other pages, but don't get linked to (IN)
- Tendrils, tubes, islands

Outline

User intent: Answering the need behind the query

- What can we do to guess user intent?
- Guess user intent independent of context:
  - Spell correction
  - Precomputed "typing" of queries (next slide)
- Better: Guess user intent based on context:
  - Geographic context (slide after next)
  - Context of user in this session (e.g., previous query)
  - Context provided by personal profile (Yahoo/MSN do this, Google claims it doesn't)

Guessing of user intent by "typing" queries

- Calculation: 5+4
- Unit conversion: 1 kg in pounds
- Currency conversion: 1 euro in kronor
- Tracking number: 8167 2278 6764
- Flight info: LH 454
- Area code: 650
- Map: columbus oh
- Stock price: msft
- Albums/movies etc: coldplay

The spatial context: Geo-search

- Three relevant locations
  - Server (nytimes.com → New York)
  - Web page (nytimes.com article about Albania)
  - User (located in Palo Alto)
- Locating the user
  - IP address
  - Information provided by user (e.g., in user profile)
  - Mobile phone
- Geo-tagging: Parse text and identify the coordinates of the geographic entities
  - Example: East Palo Alto CA → Latitude: 37.47 N, Longitude: 122.14 W
  - Important NLP problem

How do we use context to modify query results?

- Result restriction: Don't consider inappropriate results
  - For user on google.fr ...
  - ... only show .fr results
- Ranking modulation: use a rough generic ranking, rerank based on personal context
- Contextualization / personalization is an area of search with a lot of potential for improvement.

Outline

Users of web search

- Use short queries (average < 3)
- Rarely use operators
- Don't want to spend a lot of time on composing a query
- Only look at the first couple of results
- Want a simple UI, not a search engine start page overloaded with graphics
- Extreme variability in terms of user needs, user expectations, experience, knowledge, ...
  - Industrial/developing world, English/Estonian, old/young, rich/poor, differences in culture and class
- One interface for hugely divergent needs

How do users evaluate search engines?

- Classic IR relevance (as measured by F) can also be used for web IR.
- Equally important: Trust, duplicate elimination, readability, loads fast, no pop-ups
- On the web, precision is more important than recall.
  - Precision at 1, precision at 10, precision on the first 2-3 pages
  - But there is a subset of queries where recall matters.

Web information needs that require high recall

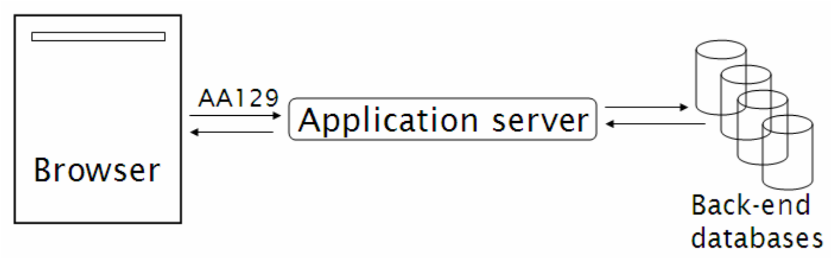
- Has this idea been patented?
- Searching for info on a prospective financial advisor
- Searching for info on a prospective employee
- Searching for info on a date

Outline

Web documents: different from other IR collections

- Distributed content creation: no design, no coordination
  - "Democratization of publishing"
  - Result: extreme heterogeneity of documents on the web
- Unstructured (text, html), semistructured (html, xml), structured/relational (databases)
- Dynamically generated content

Dynamic content

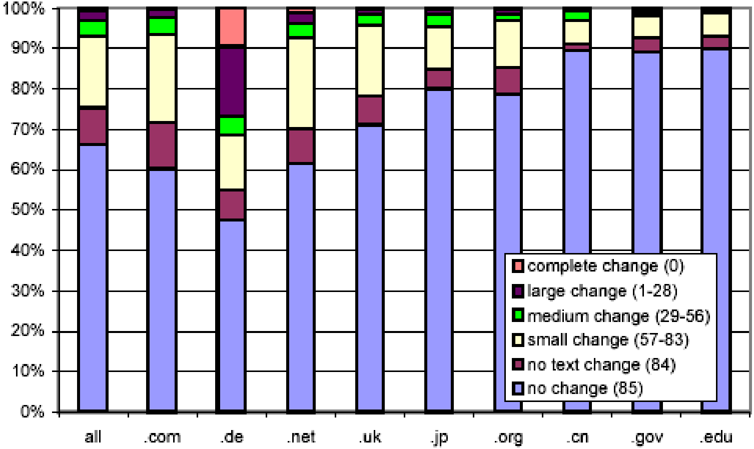


- Dynamic pages are generated from scratch when the user requests them – usually from underlying data in a database.
- Example: current status of flight LH 454

Dynamic content (2)

- Most (truly) dynamic content is ignored by web spiders.
  - It's too much to index it all.
- Actually, a lot of "static" content is also assembled on the fly (asp, php etc.: headers, date, ads etc)

Web pages change frequently (Fetterly 1997)



Multilinguality

- Documents in a large number of languages
- Queries in a large number of languages
- First cut: Don't return English results for a Japanese query
- However: Frequent mismatches query/document languages
- Many people can understand, but not query in a language
- Translation is important.
- Google example: "Beaujolais Nouveau -wine"

Duplicate documents

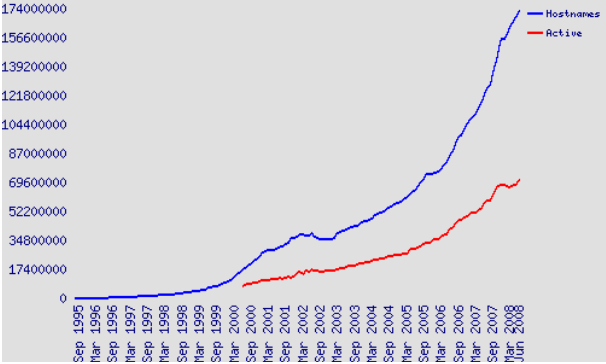
- Significant duplication – 30%–40% duplicates in some studies
- Duplicates in the search results were common in the early days of the web.
- Today's search engines eliminate duplicates very effectively.
- Key for high user satisfaction

Trust

- For many collections, it is easy to assess the trustworthiness of a document.
  - A collection of Reuters newswire articles
  - A collection of TASS (Telegraph Agency of the Soviet Union) newswire articles from the 1980s
  - Your Outlook email from the last three years
- Web documents are different: In many cases, we don't know how to evaluate the information.
- Hoaxes abound.

Outline

Growth of the web



- The web keeps growing.
- But growth is no longer exponential?

Size of the web: Issues

- What is size? Number of web servers? Number of pages? Terabytes of data available?
- Some servers are seldom connected.
  - Example: Your laptop running a web server
  - Is it part of the web?
- The “dynamic” web is infinite.
  - Any sum of two numbers is its own dynamic page on Google. (Example: “2+4”)

“Search engine index contains N pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
  - There used to be (and still are?) billions of pages that are only indexed by anchor text.

Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- http://ifnlp.org/ir/sizeoftheweb.html
- According to this query: Size of web $\geq$ 21,450,000,000 on 2007.07.07 and $\geq$ 25,350,000,000 on 2008.07.03
- But page counts of google search results are only rough estimates.

Outline

Size of the web: Who cares?

- Media
- Users
  - They may switch to the search engine that has the best coverage of the web.
  - Users (sometimes) care about recall. If we underestimate the size of the web, search engine results may have low recall.
- Search engine designers (how many pages do I need to be able to handle?)
- Crawler designers (which policy will crawl close to N pages?)

What is the size of the web? Any guesses?

Simple method for determining a lower bound

- OR-query of frequent words in a number of languages
- http://ifnlp.org/lehre/teaching/2007-SS/ir/sizeoftheweb.html
- According to this query: Size of web $\geq$ 21,450,000,000 on 2007.07.07
- Big if: Page counts of google search results are correct. (Generally, they are just rough estimates.)
- But this is just a lower bound, based on one search engine.
- How can we do better?

Size of the web: Issues

- The "dynamic" web is infinite.
  - Any sum of two numbers is its own dynamic page on Google. (Example: "2+4")
  - Many other dynamic sites generating infinite number of pages
- The static web contains duplicates – each "equivalence class" should only be counted once.
- Some servers are seldom connected.
  - Example: Your laptop
  - Is it part of the web?

“Search engine index contains N pages”: Issues

- Can I claim a page is in the index if I only index the first 4000 bytes?
- Can I claim a page is in the index if I only index anchor text pointing to the page?
  - There used to be (and still are?) billions of pages that are only indexed by anchor text.

How can we estimate the size of the web?

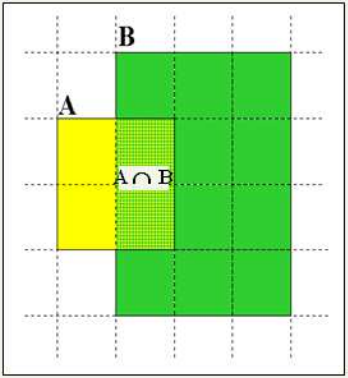
Sampling methods

- Random queries
- Random searches
- Random IP addresses
- Random walks

Variant: Estimate relative sizes of indexes

- There are significant differences between indexes of different search engines.
- Different engines have different preferences.
  - max url depth, max count/host, anti-spam rules, priority rules etc.
- Different engines index different things under the same URL.
  - anchor text, frames, meta-keywords, size of prefix etc.

Relative Size from Overlap [Bharat & Broder, 98]



Sample URLs randomly from A

Check if contained in B

and vice versa

$A \cap B = (1/2) * \text{Size } A$

$A \cap B = (1/6) * \text{Size } B$

$(1/2) * \text{Size } A = (1/6) * \text{Size } B$

$\therefore \text{Size } A / \text{Size } B = (1/6)/(1/2) = 1/3$

Each test involves: (i) Sampling (ii) Checking

Sampling URLs

- Ideal strategy: Generate a random URL
- Problem: Random URLs are hard to find (and sampling distribution should reflect "user interest")
- Approach 1: Random walks / IP addresses
  - In theory: might give us a true estimate of the size of the web (as opposed to just relative sizes of indexex)
- Approach 2: Generate a random URL contained in a given engine
  - Suffices for accurate estimation of relative size

Random URLs from random queries

- Idea: Use vocabulary of the web for query generation
- Vocabulary can be generated from web crawl
- Use conjunctive queries w_1 AND w_2
  - Example: vocalists AND rsi
- Get result set of one hundred URLs from the source engine
- Choose a random URL from the result set
- This sampling method induces a weight $W(p)$ for each page p.
- Method was used by Bharat and Broder (1998).

Checking if a page is in the index

- Either: Search for URL if the engine supports this
- Or: Create a query that will find doc *d* with high probability
  - Download doc, extract words
  - Use 8 low frequency word as AND query
  - Call this a strong query for *d*
  - Run query
  - Check if *d* is in result set
- Problems
  - Near duplicates
  - Redirects
  - Engine time-outs

Computing Relative Sizes and Total Coverage [BB98]

a = AltaVista, **e** = Excite, **h** = HotBot, **i** = Infoseek

f_{xy} = fraction of **x in y**

- Six pair-wise overlaps

$$f_{ah} * a - f_{ha} * h = \varepsilon_1$$
$$f_{ai} * a - f_{ia} * i = \varepsilon_2$$
$$f_{ae} * a - f_{ea} * e = \varepsilon_3$$
$$f_{hi} * h - f_{ih} * i = \varepsilon_4$$
$$f_{he} * h - f_{eh} * e = \varepsilon_5$$
$$f_{ei} * e - f_{ie} * i = \varepsilon_6$$

- Arbitrarily, let **a** = 1.

- We have 6 equations and 3 unknowns.
- Solve for **e**, **h** and **i** to minimize $\Sigma \varepsilon_i^2$
- Compute engine overlaps.
- Re-normalize so that the total joint coverage is 100%

Advantages & disadvantages

- Statistically sound under the induced weight.
- Biases induced by random query
  - Query Bias: Favors content-rich pages in the language(s) of the lexicon
  - Ranking Bias: *Solution:* Use conjunctive queries & fetch all
  - Checking Bias: Duplicates, impoverished pages omitted
  - Document or query restriction bias: engine might not deal properly with 8 words conjunctive query
  - Malicious Bias: Sabotage by engine
  - Operational Problems: Time-outs, failures, engine inconsistencies, index modification.

Random searches

- Choose random searches extracted from a search engine log (Lawrence & Giles 97)
- Use only queries with small result sets
- For each random query: compute ratio size(r_1)/size(r_2) of the two result sets
- Average over random searches

Advantages & disadvantages

- Advantage
  - Might be a better reflection of the human perception of coverage
- Issues
  - Samples are correlated with source of log (unfair advantage for originating search engine)
  - Duplicates
  - Technical statistical problems (must have non-zero results, ratio average not statistically sound)

Random searches [Lawr98, Lawr99]

- 575 & 1050 queries from the NEC RI employee logs
- 6 Engines in 1998, 11 in 1999
- Implementation:
  - Restricted to queries with < 600 results in total
  - Counted URLs from each engine after verifying query match
  - Computed size ratio & overlap for individual queries
  - Estimated index size ratio & overlap by averaging over all queries

Queries from Lawrence and Giles study

- adaptive access control
- neighborhood preservation topographic
- hamiltonian structures
- right linear grammar
- pulse width modulation neural
- unbalanced prior probabilities
- ranked assignment method
- internet explorer favourites importing
- karvel thornber
- zili liu
- softmax activation function
- bose multidimensional system theory
- gamma mlp
- dvi2pdf
- john oliensis
- rieke spikes exploring neural
- video watermarking
- counterpropagation network
- fat shattering dimension
- abelson amorphous computing

Random IP addresses [Lawrence & Giles '99]

- Generate random IP addresses
- Find a web server at the given address
  - If there's one
- Collect all pages from server.
- Method first used by O'Neill, McClain, & Lavoie, **"A Methodology for Sampling the World Wide Web", 1997.** `http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003447`

Random IP addresses [ONei97,Lawr99]

- [Lawr99] exhaustively crawled 2500 servers and extrapolated
- Estimated size of the web to be 800 million

Advantages and disadvantages

- Advantages
  - Can, in theory, estimate the size of the accessible web (as opposed to the (relative) size of an index)
  - Clean statistics
  - Independent of crawling strategies
- Disadvantages
  - Many hosts share one IP ($\rightarrow$ oversampling)
  - Hosts with large web sites don't get more weight than hosts with small web sites ($\rightarrow$ possible undersampling)
  - Sensitive to spam (multiple IPs for same spam server)
  - Again, duplicates

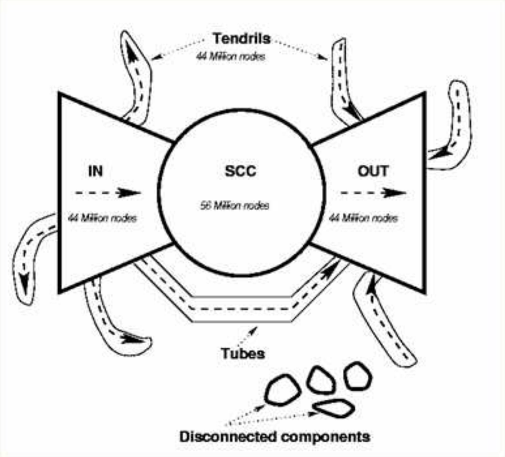
Random walks

[Henzinger *et al* WWW9]

- 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

Dependence on seed list

- How well connected is the graph? [Broder et al., WWW9]
-



Advantages & disadvantages

- Advantages
  - "Statistically clean" method at least in theory!
  - Could work even for infinite web (assuming convergence) under certain metrics.
- Disadvantages
  - List of seeds is a problem.
  - Practical approximation might not be valid.
  - Non-uniform distribution
    - Subject to link spamming

Conclusion

- Many different approaches to web size estimation.
- None is perfect.
- The problem has gotten much harder.
- There hasn't been a good study for a couple of years.
- Great topic for a thesis!

Resources

- Chapter 19 of IIR
- Resources at http://cislmu.org
  - Hal Varian explains Google second price auction: http://www.youtube.com/watch?v=K7l0a2PVhPQ
  - Size of the web queries
  - Trademark issues (Geico and Vuitton cases)
  - How ads are priced
  - Henzinger, Finding near-duplicate web pages: A large-scale evaluation of algorithms, ACM SIGIR 2006.