Introduction to Information Retrieval

(Manning, Raghavan, Schutze)

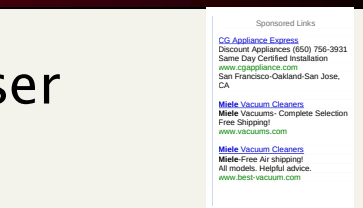
Chapter 19
Web search basics

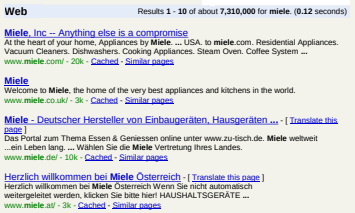
1. Brief history and overview

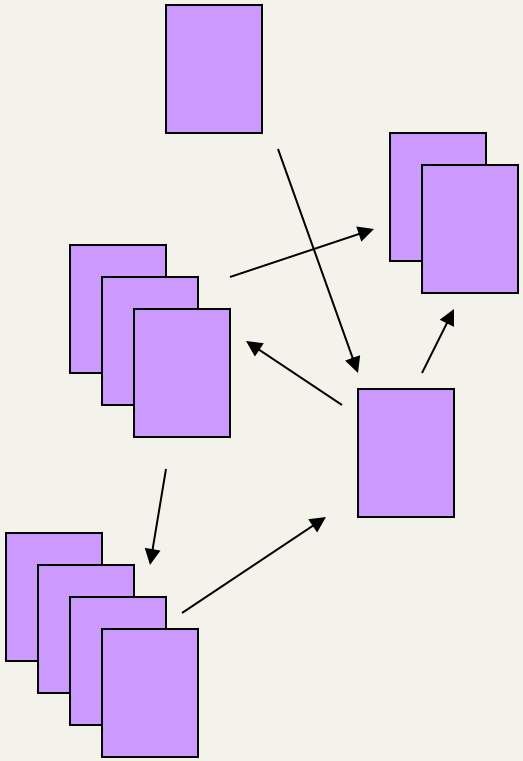
- Early keyword-based engines
  - Altavista, Excite, Infoseek, Inktomi, ca. 1995-1997
- A hierarchy of categories
  - Yahoo!
  - Many problems, popularity declined. Existing variants are About.com and **Open Directory Project**
- Classical IR techniques continue to be necessary for web search, by no means sufficient
  - E.g., classical IR measures relevancy, web search needs to measure relevancy + authoritativeness

Web search overview

User

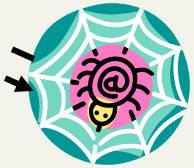




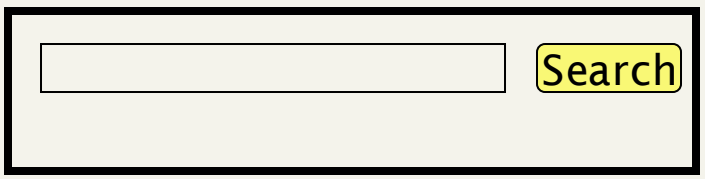


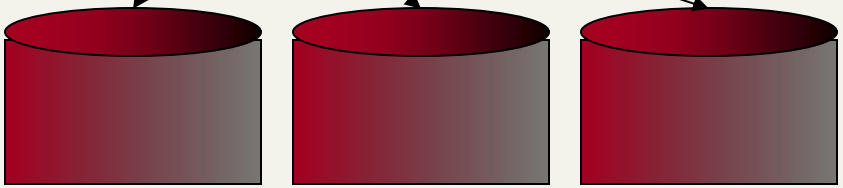
The Web

Web spider

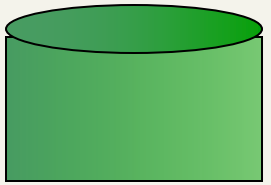


Indexer



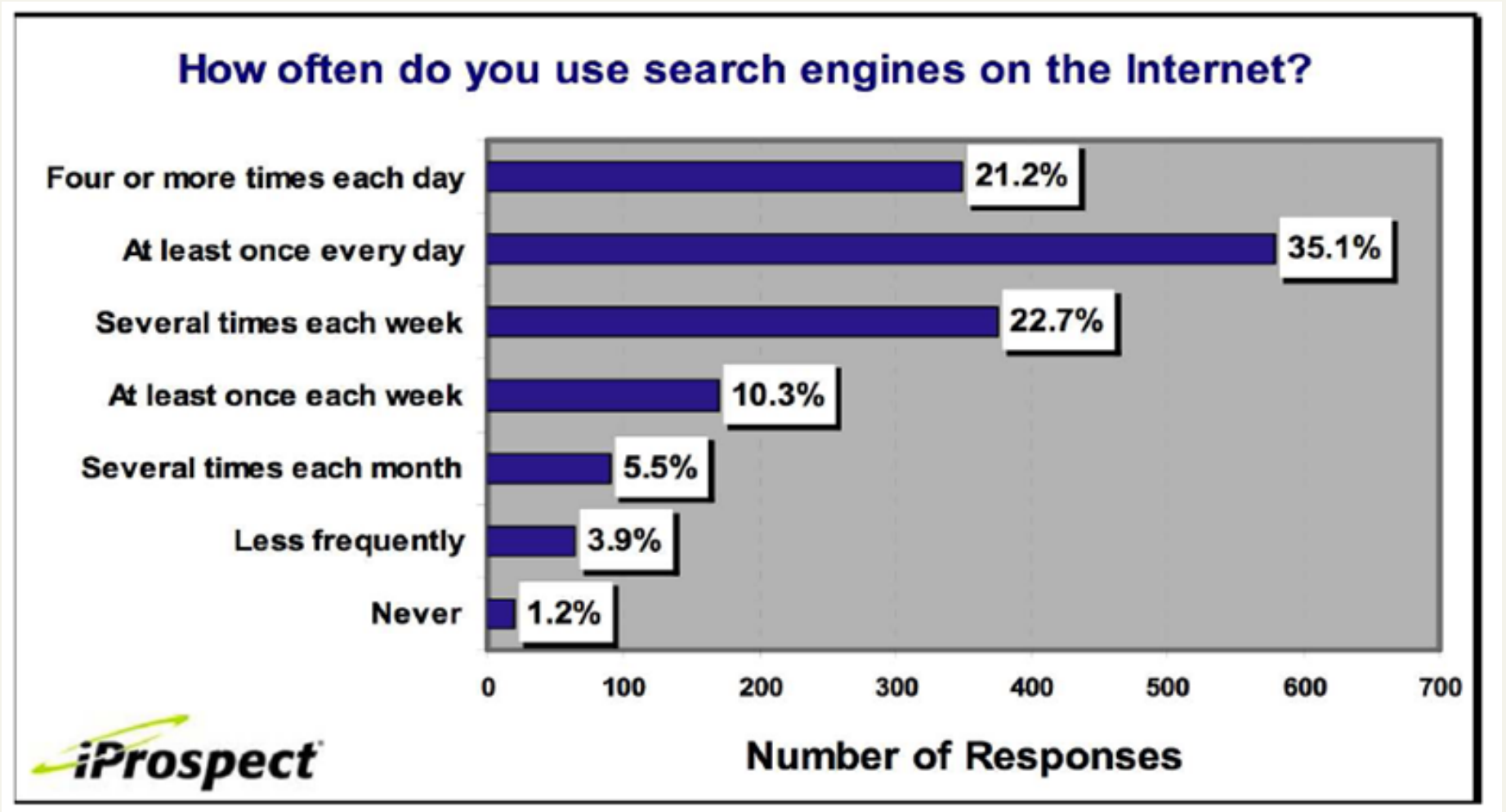


Indexes



Ad indexes

Search is a top activity on the Web



Without search engines, the web wouldn't work

- Without search, content is hard to find
- Without search, there's 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?
- 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)
  - Interest aggregation without search engines is not possible
- Somebody needs to pay for the web
  - Servers, web infrastructure, content creation
  - A large part today is paid by search ads

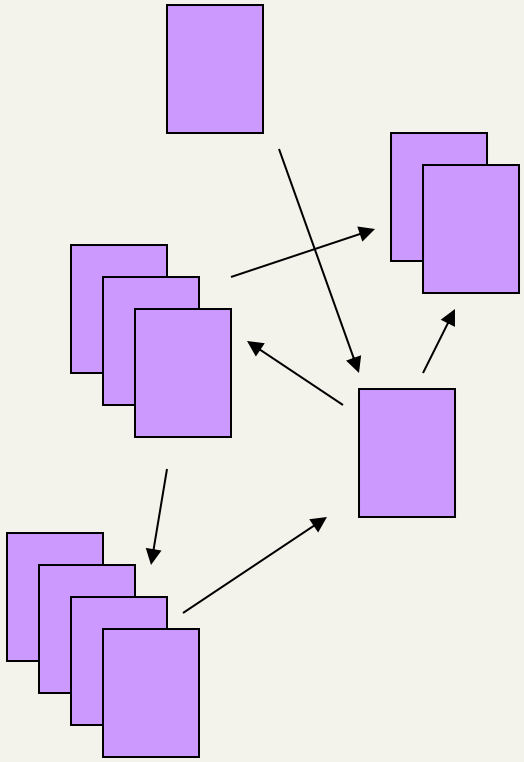
Web IR: Differences from traditional IR

- Links: The web is a hyperlinked document collection
- Queries: web queries are different, more varied and there are a lot of them
  - How many? 10^8 every day, approaching 10^9
- Users: users are different, more varied and there are a lot of them
  - How many? 10^9
- Documents: documents are different, more varied and there are a lot of them
  - How many? ~ 10^{11}. Indexed 10^{10}
- Context: context is more important on the web than in many other IR applications
- Ads and spam

2. Web characteristics

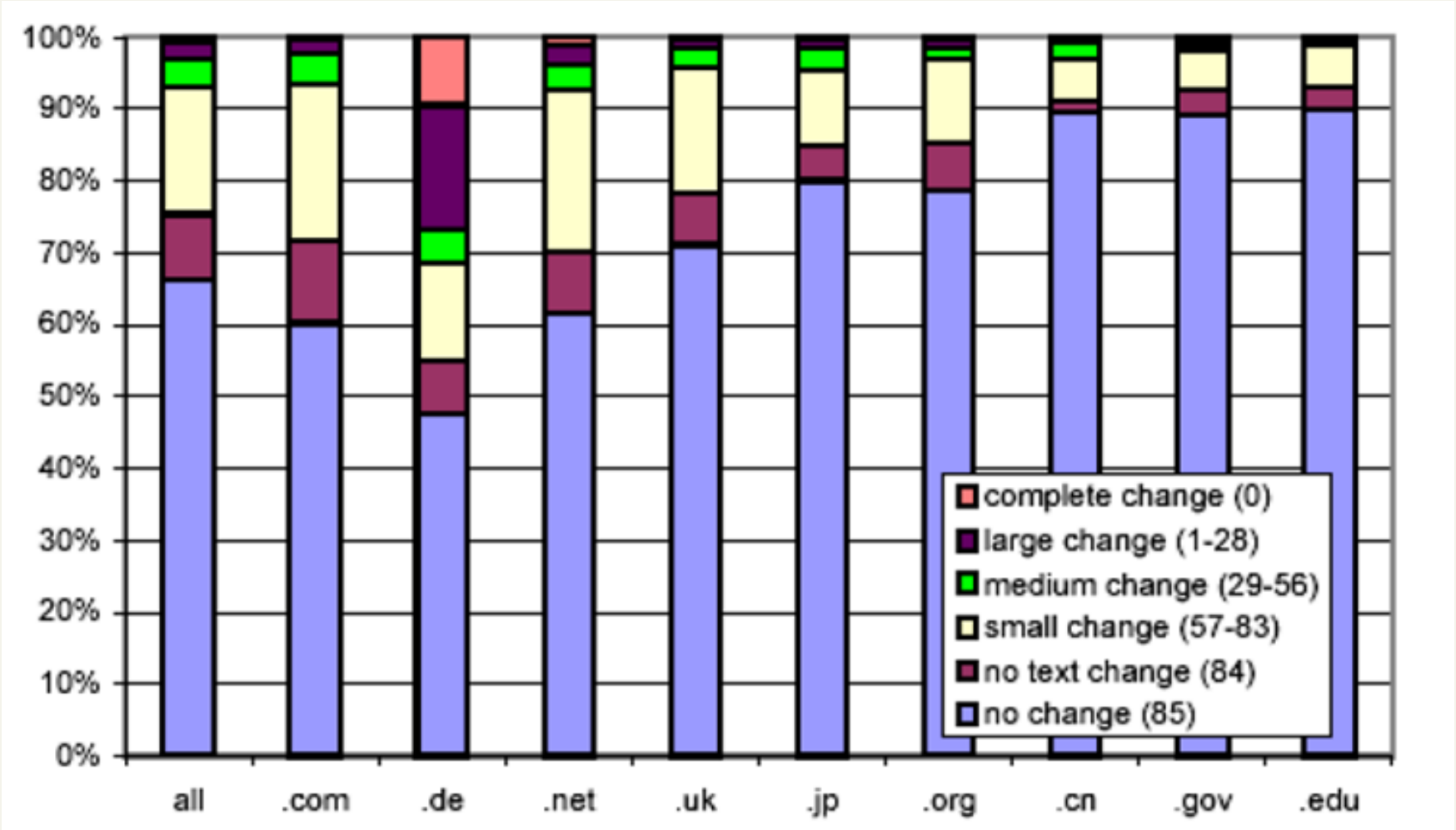
- Web document
- Size of the Web
- Web graph
- Spam

The Web document collection



- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text collections
- Growth – slowed down from initial "volume doubling every few months" but still expanding
- Content can be *dynamically generated*
  - Mostly ignored by crawlers

Web pages change frequently (Fetterly 1997)



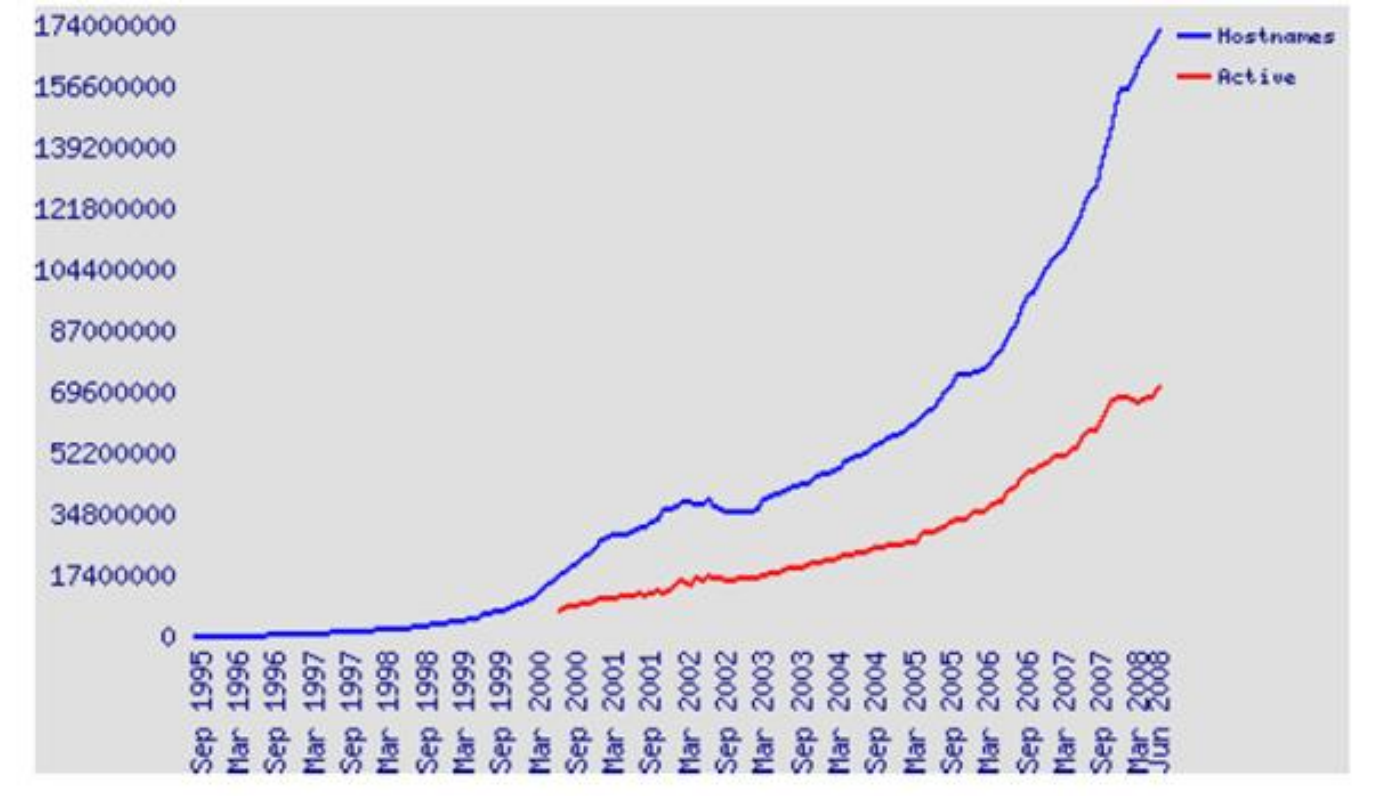
Duplicate documents

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

Duplicate detection

- The web is full of duplicated content
- Strict duplicate detection = exact match
  - Not as common
- But many, many cases of near duplicates
  - E.g., Last modified date the only difference between two copies of a page
- Various techniques
  - Fingerprint, shingles, sketch

Growth of the web



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

Size of the web: issues

- How to define 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 (e.g., “2+4”)

What can we attempt to measure?

- The relative sizes of search engines
- Issues
  - Can I claim a page 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 would you estimate the number of pages indexed by a web search engine?

Simple methods 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

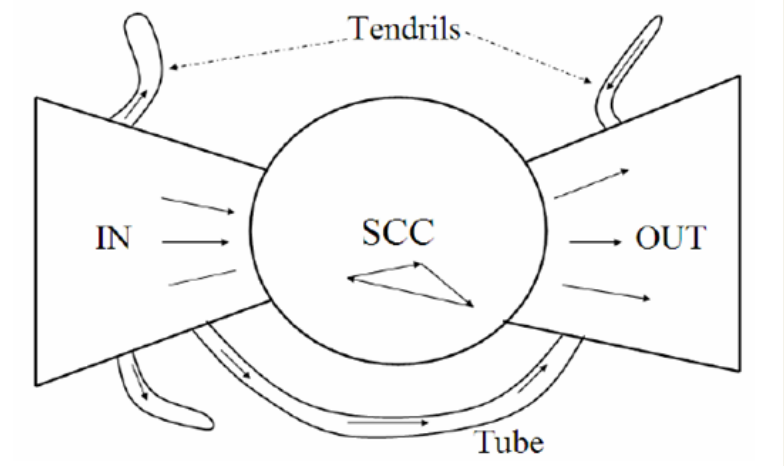
>= 21,450,000,000 on 2007.07.07 and

>= 25,350,000,000 on 2008.07.03

- But page counts of google search results are only rough estimates

web graph

- The Web is a directed graph
  - Not strongly connected, i.e., there are pairs of pages such that one cannot reach the other by following links
- Links are not randomly distributed, rather, power law
  - Total # of pages with in-degree *i* is proportional to $1/i^a$
- The web has a bowtie shape
  - Strongly connected component (SCC) in the center
  - Many pages that get linked to, but don't link (OUT)
  - Many pages that link to other pages, but don't get linked to (IN)
  - IN and OUT similar size, SCC somehow larger



Goal of spamming on the web

- You have a page that will generate lots of revenue for you if people visit it
- Therefore, you'd like to redirect visitors to this page
- One way of doing this: get your page ranked highly in search results

Simplest forms

- First generation engines relied heavily on *tf/idf*
- **Hidden text**: dense repetitions of chosen keywords
  - Often, the repetitions would be in the same color as the background of the web page. So that repeated terms got indexed by crawlers, but not visible to humans on browsers
- **Keyword stuffing**: misleading meta-tags with excessive repetition of chosen keywords

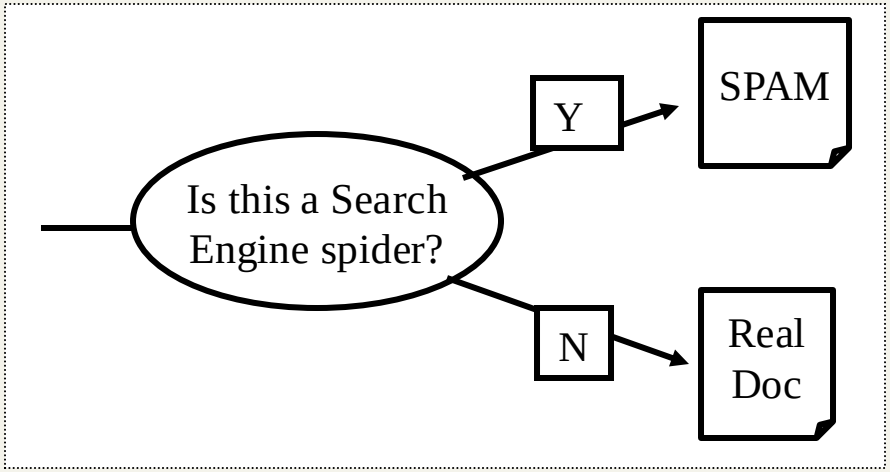
- Used to be effective, most search engines now catch these

- Spammers responded with a richer set of spam techniques

Cloaking

- Serve fake content to search engine spider
  - Causing web page to be indexed under misleading keywords
  - When user searches for these keywords and elects to view the page, he receives a page with totally different content

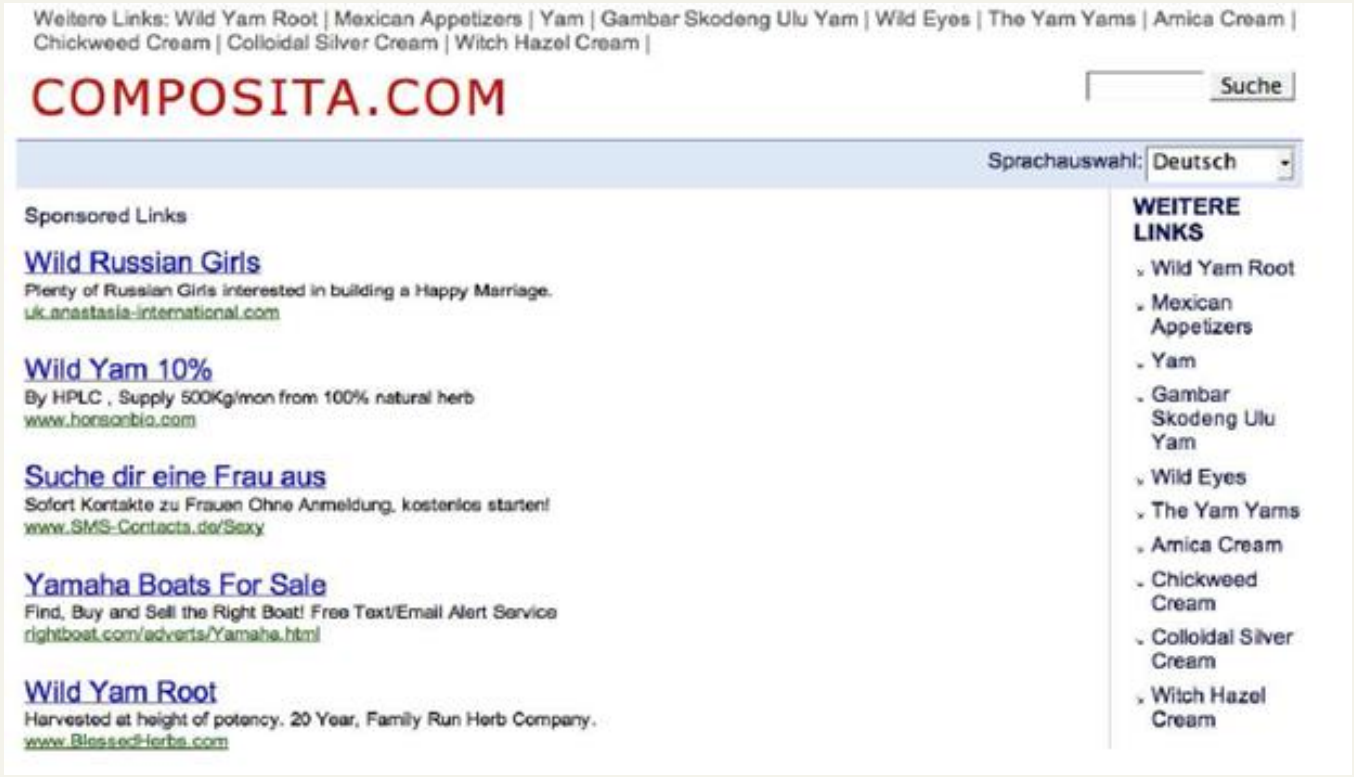
- So do we just penalize this anyways?
- No: legitimate uses, e.g., different contents to US and European users



More spam techniques

- **Doorway page**
  - Contains text/metadata carefully chosen to rank highly on selected keywords
  - When a browser requests the doorway page, it is redirected to a page containing content of a more commercial nature
- **Lander page**
  - Optimized for a single keyword or a misspelled domain name, designed to attract surfers who will then click on ads
- **Duplication**
  - Get good content from somewhere (steal it or produce it by 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” …

Lander page



- Number of 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

Link spam

- Create lots of links pointing to the page you want to promote
- Put these links on pages with high (at least non-zero) pagerank
  - Newer registered domains (domain flooding)
  - A set of pages pointing 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")
    - http://www-csli.stanford.edu/~hinrich/horoskop-schuetze.html
  - Leave comments that include the link on blogs
- Link farm

Search engine optimization

- Promoting a page 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
- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world (www.webmasterworld.com)

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/

The war against spam

- Quality indicators - prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
  - Distribution and structure of text (e.g., no keyword stuffing)
- 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

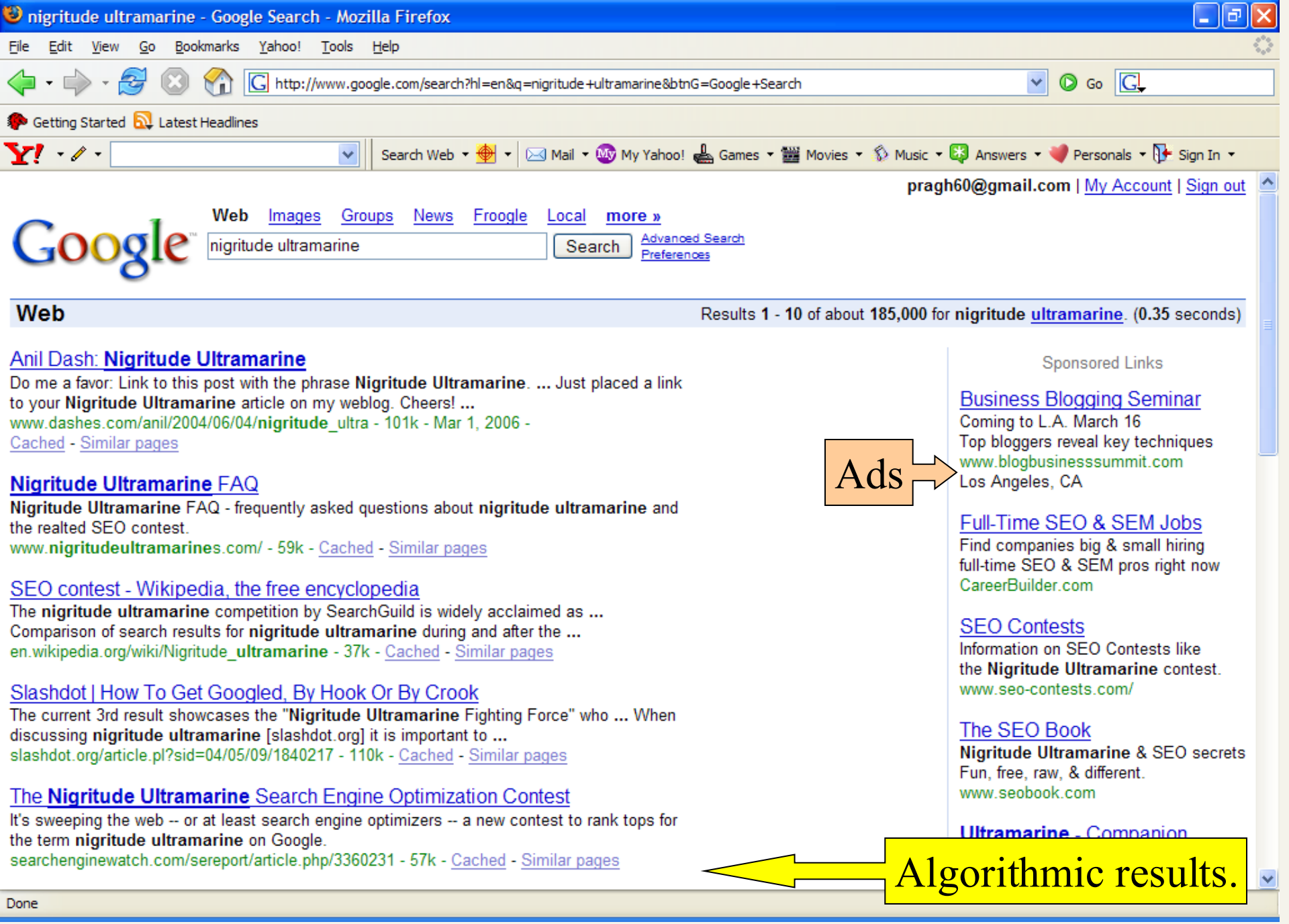
3. Advertising as economic model

- Sponsored search ranking: Goto.com (morphed into Overture.com → Yahoo!)
  - Your search ranking depended on how much you paid
  - Auction for keywords: ***casino*** was expensive!
  - No separation of ads/docs

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines
  - Google added paid-placement "ads" to the side, independent of search results
  - Strict separation of ads and results

First generation of search ads: Goto (1996)

- No separation of ads/docs. Just one results!
- Buddy Blake bid the maximum ($0.38) for this search
- He paid $0.38 to Goto every time somebody clicked on the link
- Upfront and honest. No relevance ranking, but Goto did not pretend there was any.


nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Getting Started Latest Headlines
Search Web Mail My Yahoo! Games Movies Music Answers Personals Sign In
pragh60@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 185,000 for nigritude ultramarine. (0.35 seconds)
Anil Dash: Nigritude Ultramarine
Do me a favor: Link to this post with the phrase Nigritude Ultramarine. ... Just placed a link to your Nigritude Ultramarine article on my weblog. Cheers! ...
www.dashes.com/anil/2004/06/04/nigritude_ultra - 101k - Mar 1, 2006 -
Cached - Similar pages
Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about nigritude ultramarine and the realted SEO contest.
www.nigritudeultramarines.com/ - 59k - Cached - Similar pages
SEO contest - Wikipedia, the free encyclopedia
The nigritude ultramarine competition by SearchGuild is widely acclaimed as ...
Comparison of search results for nigritude ultramarine during and after the ...
en.wikipedia.org/wiki/Nigritude_ultramarine - 37k - Cached - Similar pages
Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "Nigritude Ultramarine Fighting Force" who ... When discussing nigritude ultramarine [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages
The Nigritude Ultramarine Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for the term nigritude ultramarine on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages
Sponsored Links
Business Blogging Seminar
Coming to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA
Full-Time SEO & SEM Jobs
Find companies big & small hiring full-time SEO & SEM pros right now
CareerBuilder.com
SEO Contests
Information on SEO Contests like the Nigritude Ultramarine contest.
www.seo-contests.com/
The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
www.seobook.com
Ultramarine - Companion
Ads
Algorithmic results.
Done

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.
  - Being willing to pay for ads on a search engine is a quality signal (one of many) that users take into account.
- The advertiser finds new customers in a cost-effective way

The appeal of search ads to advertisers

- Why is web search potentially more attractive for advertisers than TV spots, newspaper ads or radio spots?
- Someone who just searched for “Saturn Aura Sport Sedan” is infinitely more likely to buy one than a random person watching TV.
- Most importantly, the advertiser only pays if the customer took an action indicating interest (i.e., clicking on the ad)

But frequently it's not a win-win-win

- Example: keyword arbitrage
  - Buy a keyword at Google
  - Then redirect traffic to a third party that is paying much more than you have to pay to Google
  - This rarely makes sense for the user
- Ad spammers keep inventing new tricks
  - The search engines need time to catch up with them

- Click spam: refers to clicks on sponsored search results not from bona fide search users
  - E.g., a devious advertiser may attempt to exhaust the advertising budget of a competitor by clicking repeatedly (through robotic click generator) on his sponsored search ads.

4. Search user experiences

- Users
- User queries
- Query distribution
- User’s empirical evaluations

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

User query needs

- Need [Brod02, RL04]
  - **Informational** – want to learn about something (~40% / 65%)
    - Not a single page containing the info
  - **Navigational** – want to go to that page (~25% / 15%)
  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service
    - Downloads
    - Shop
  - **Gray areas**
    - Find a good hub
    - Exploratory search "see what's there"

Low hemoglobin

United Airlines

Seattle weather

Mars surface images

Canon S410

Car rental Brasil

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	www.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.

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 very few 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

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
- 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

Users' empirical evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, search assist,…)
  - Explicit: Search within results, more like this, refine ...
  - Anticipative: related searches
- Deal with idiosyncrasies
  - Web specific vocabulary
    - Impact on stemming, spell-check, etc
  - Web addresses typed in the search box
  - …