

# Web Information Retrieval

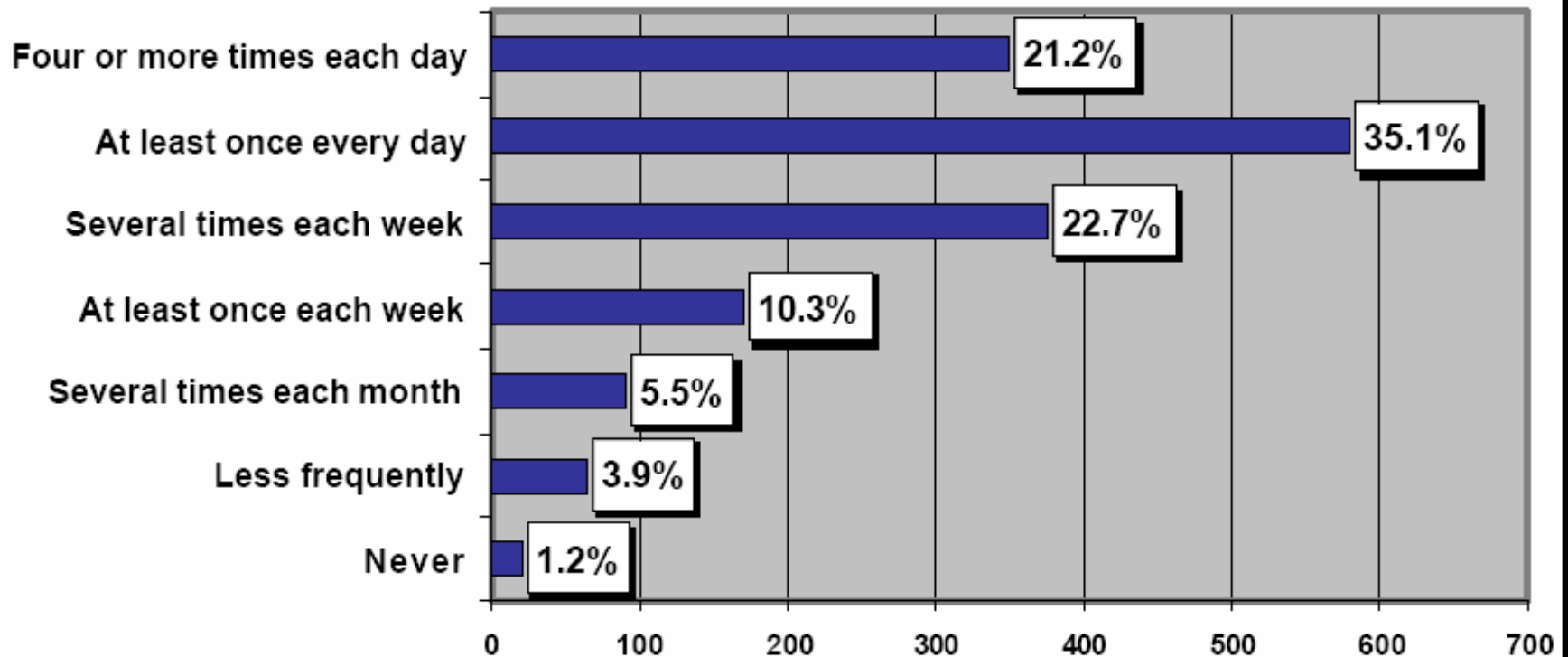
## Lecture 9

### Information Retrieval in the Web

# Search use ...

(iProspect Survey, 4/04)

## How often do you use search engines on the Internet?



# Without search engines the web wouldn't scale

---

1. No incentive in creating content unless it can be easily found – other finding methods haven't kept pace (taxonomies, bookmarks, etc)
2. The web is both a technology artifact and a social environment
  - “The Web has become the “new normal” in the American way of life; those who don't go online constitute an ever-shrinking minority.” – [Pew Foundation report, January 2005]
3. Search engines make aggregation of interest possible:
  - Create incentives for very specialized niche players
    - Economical – specialized stores, providers, etc
    - Social – narrow interests, specialized communities, etc
4. The acceptance of search interaction makes “unlimited selection” stores possible:
  - Amazon, Netflix, etc
5. Search turned out to be the best mechanism for advertising on the web, a \$15+ B industry.
  - Growing very fast but entire US advertising industry \$250B – huge room to grow
  - Sponsored search marketing is about \$10B

# Classical IR vs. Web IR

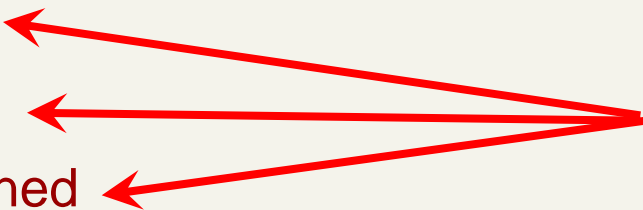
# Basic assumptions of Classical Information Retrieval

---

- **Corpus**: Fixed document collection
- **Goal**: Retrieve documents with information content that is relevant to user's **information need**

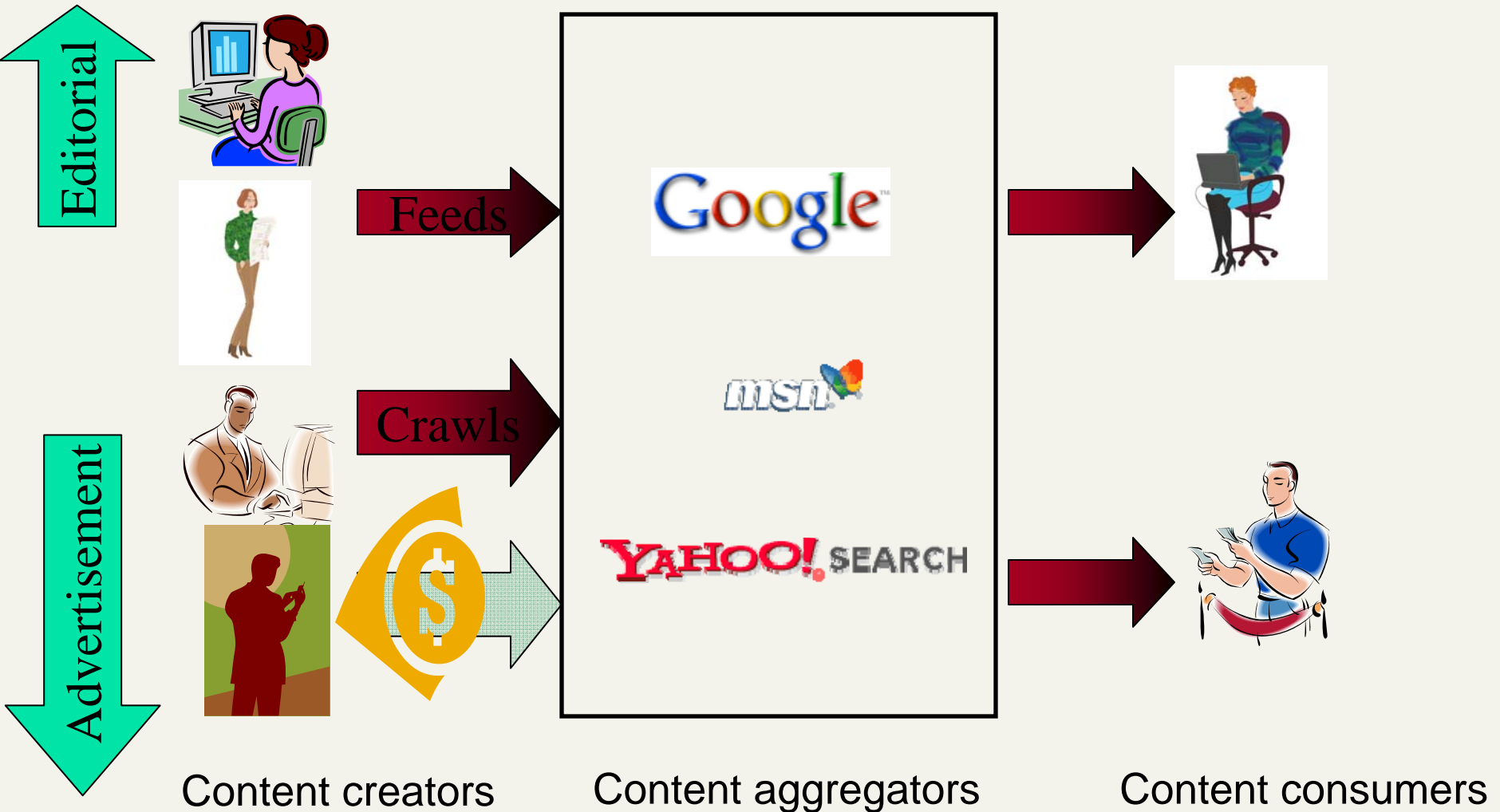
# Classic IR Goal

---

- Classic relevance
    - For each query  $Q$  and stored document  $D$  in a given corpus assume there exists relevance  $\text{Score}(Q, D)$ 
      - Score is average over users  $U$  and contexts  $C$
    - Optimize  $\text{Score}(Q, D)$  as opposed to  $\text{Score}(Q, D, U, C)$
    - That is, usually:
      - Context ignored
      - Individuals ignored
      - Corpus predetermined
- 
- Bad assumptions  
in the web context

# Web IR

# The coarse-level dynamics





# Brief (non-technical) history

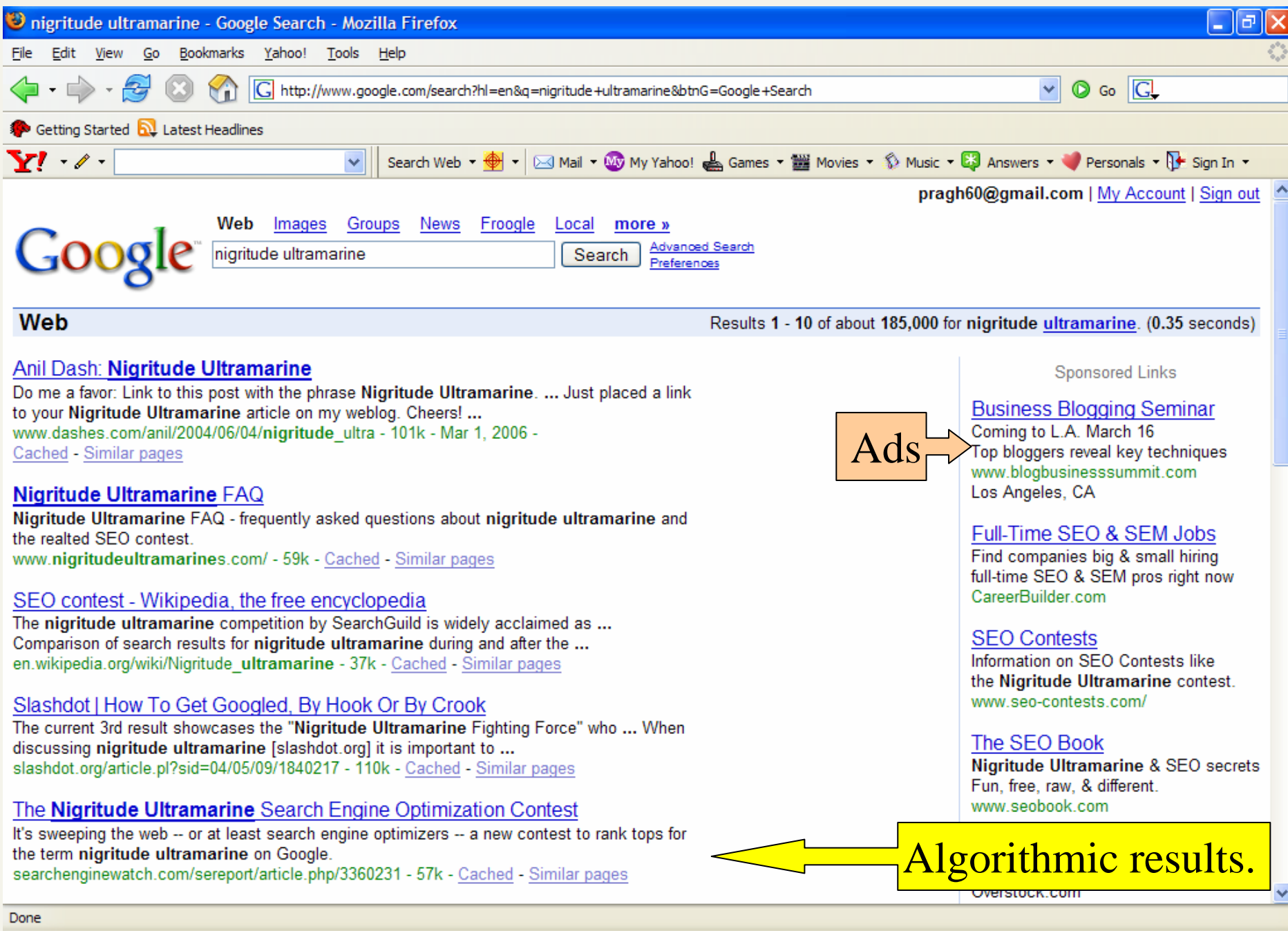
---

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

# Brief (non-technical) history

---

- 1998+: Link-based ranking pioneered by Google
  - Blew away all early engines save Inktomi
  - Great user experience in search of a business model
  - Meanwhile Goto/Overture's annual revenues were nearing \$1 billion
- Result: Google added paid-placement “ads” to the side, independent of search results
  - Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)



# Ads vs. search results

- Google has maintained that **ads** (based on vendors bidding for keywords) do not affect vendors' **rankings** in search results

Sponsored Links

[CG Appliance Express](#)  
Discount Appliances (650) 756-3931  
Same Day Certified Installation  
[www.cgappliance.com](http://www.cgappliance.com)  
San Francisco-Oakland-San Jose, CA

[Miele Vacuum Cleaners](#)  
**Miele** Vacuums- Complete Selection  
Free Shipping!  
[www.vacuums.com](http://www.vacuums.com)

[Miele Vacuum Cleaners](#)  
**Miele**-Free Air shipping!  
All models. Helpful advice.  
[www.best-vacuum.com](http://www.best-vacuum.com)

Search =  
*miele*

Web

Results 1 - 10 of about 7,310,000 for **miele**. (0.12 seconds)

[Miele, Inc -- Anything else is a compromise](#)

At the heart of your home, Appliances by **Miele**. ... USA. to **miele.com**. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System ...  
[www.miele.com/](http://www.miele.com/) - 20k - [Cached](#) - [Similar pages](#)

[Miele](#)

Welcome to **Miele**, the home of the very best appliances and kitchens in the world.  
[www.miele.co.uk/](http://www.miele.co.uk/) - 3k - [Cached](#) - [Similar pages](#)

[Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten ...](#) - [ [Translate this page](#) ]

Das Portal zum Thema Essen & Geniessen online unter [www.zu-tisch.de](http://www.zu-tisch.de). **Miele** weltweit ...ein Leben lang. ... Wählen Sie die **Miele** Vertretung Ihres Landes.  
[www.miele.de/](http://www.miele.de/) - 10k - [Cached](#) - [Similar pages](#)

[Herzlich willkommen bei Miele Österreich](#) - [ [Translate this page](#) ]

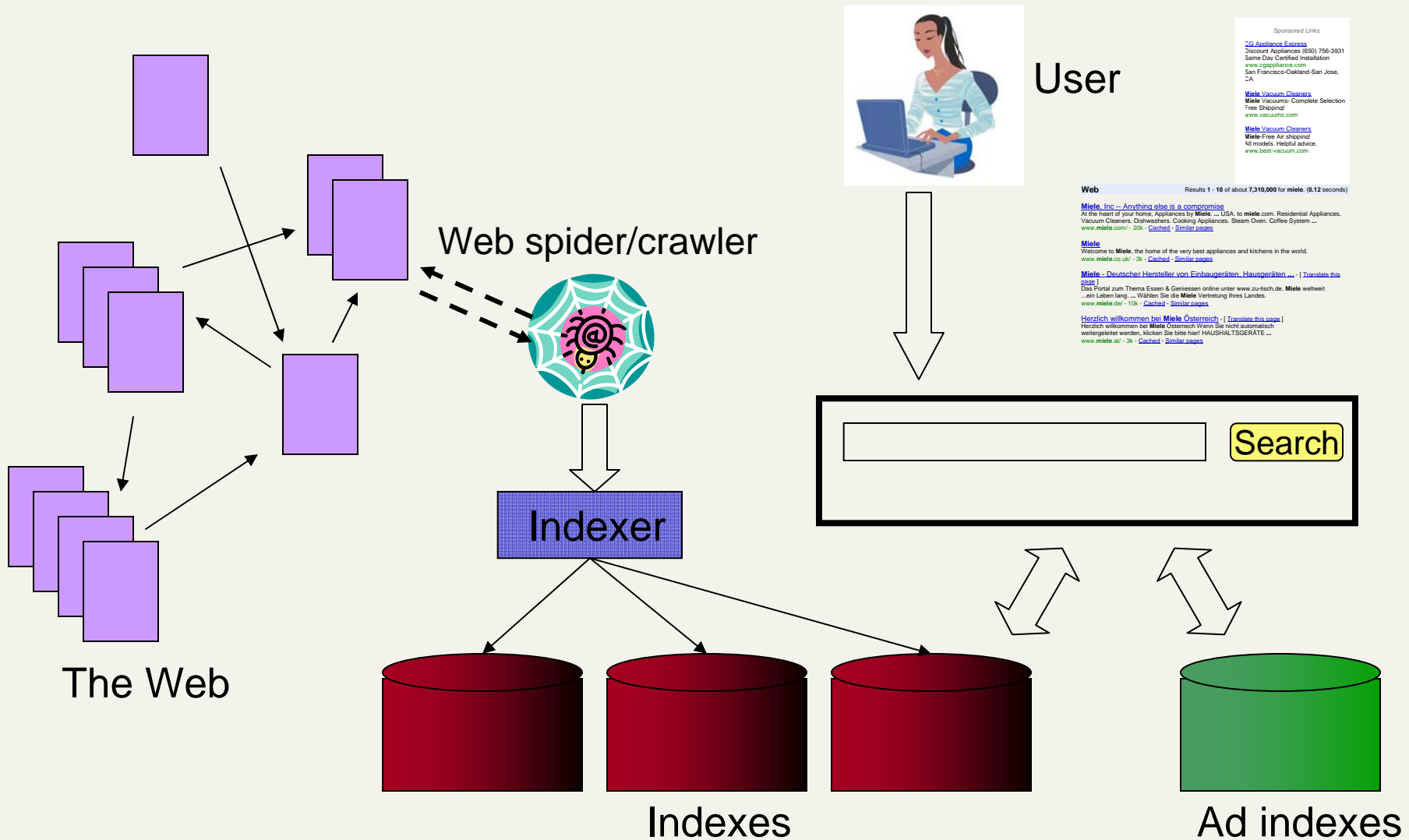
Herzlich willkommen bei **Miele** Österreich Wenn Sie nicht automatisch weitergeleitet werden, klicken Sie bitte hier! HAUSHALTSGERÄTE ...  
[www.miele.at/](http://www.miele.at/) - 3k - [Cached](#) - [Similar pages](#)

# Ads vs. search results

---

- Other vendors (Yahoo, MSN) have made similar statements from time to time
  - Any of them can change anytime
- We will focus primarily on search results independent of paid placement ads
  - Although the latter is a fascinating technical subject in itself

# Web search basics



# User Needs

---

## ■ Needs

- Informational – want to learn about something (~40% / 65%)

Leukemia

- Navigational – want to go to that page (~25% / 15%)

Lufthansa

- Transactional – want to do something (web-mediated) (~35% / 20%)

- Access a service

Weather rome

- Downloads

Mars surface images

- Shop

Canon S410

- Gray areas

- Find a good hub

Car rental Brasil

- Exploratory search “see what’s there”

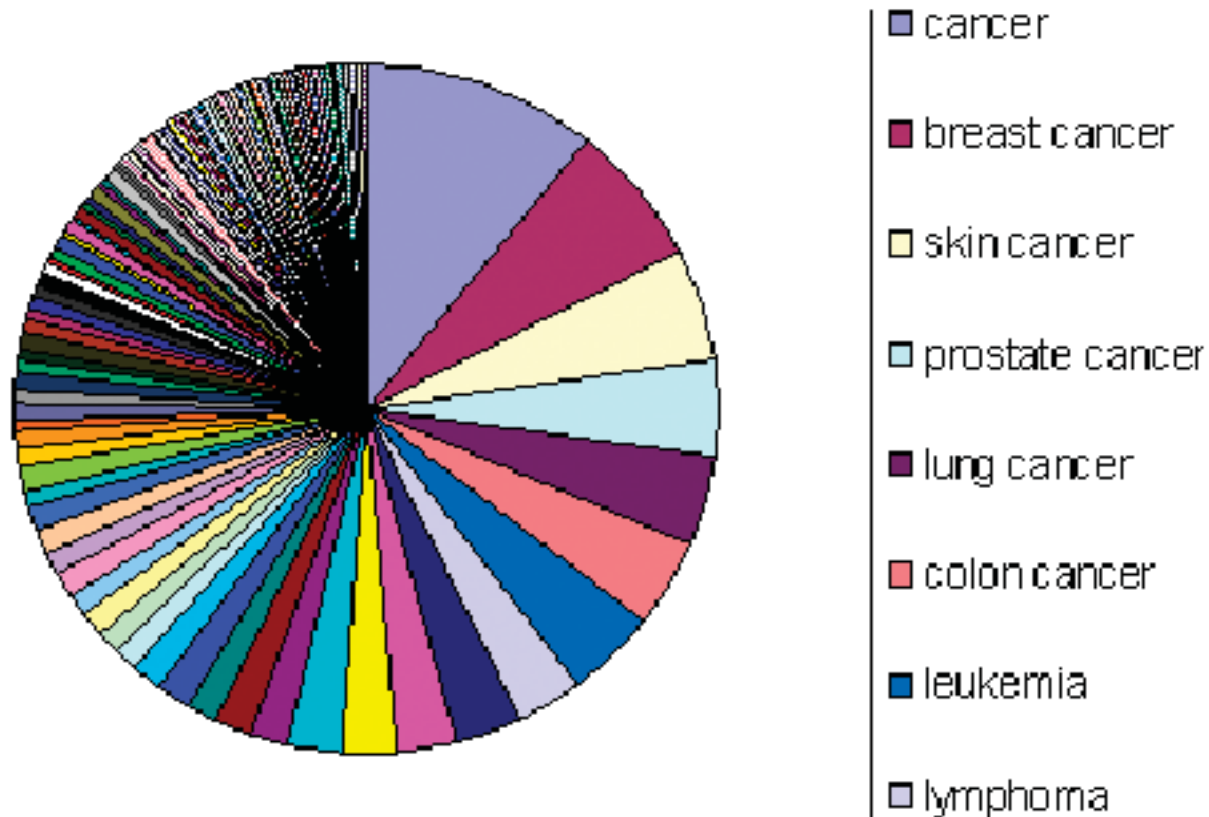
# Web search users

---

- Make ill defined queries
  - Short
    - AV 2001: 2.54 terms avg, 80% < 3 words)
    - AV 1998: 2.35 terms avg, 88% < 3 words
  - Imprecise terms
  - Sub-optimal syntax (most queries without operator)
  - Low effort
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth
- Specific behavior
  - 85% look over one result screen only (mostly above the fold)
  - 78% of queries are not modified (one query/session)
  - Follow links – “the scent of information” ...



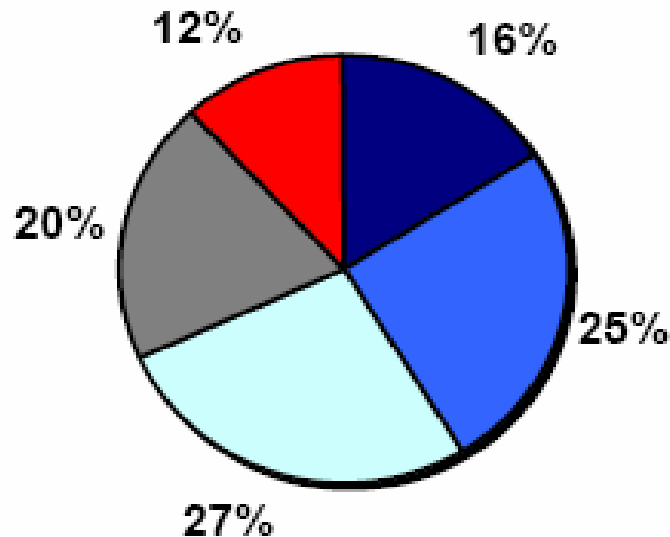
# Query Distribution



**Power law: few popular broad queries,  
many rare specific queries**

# How far do people look for results?

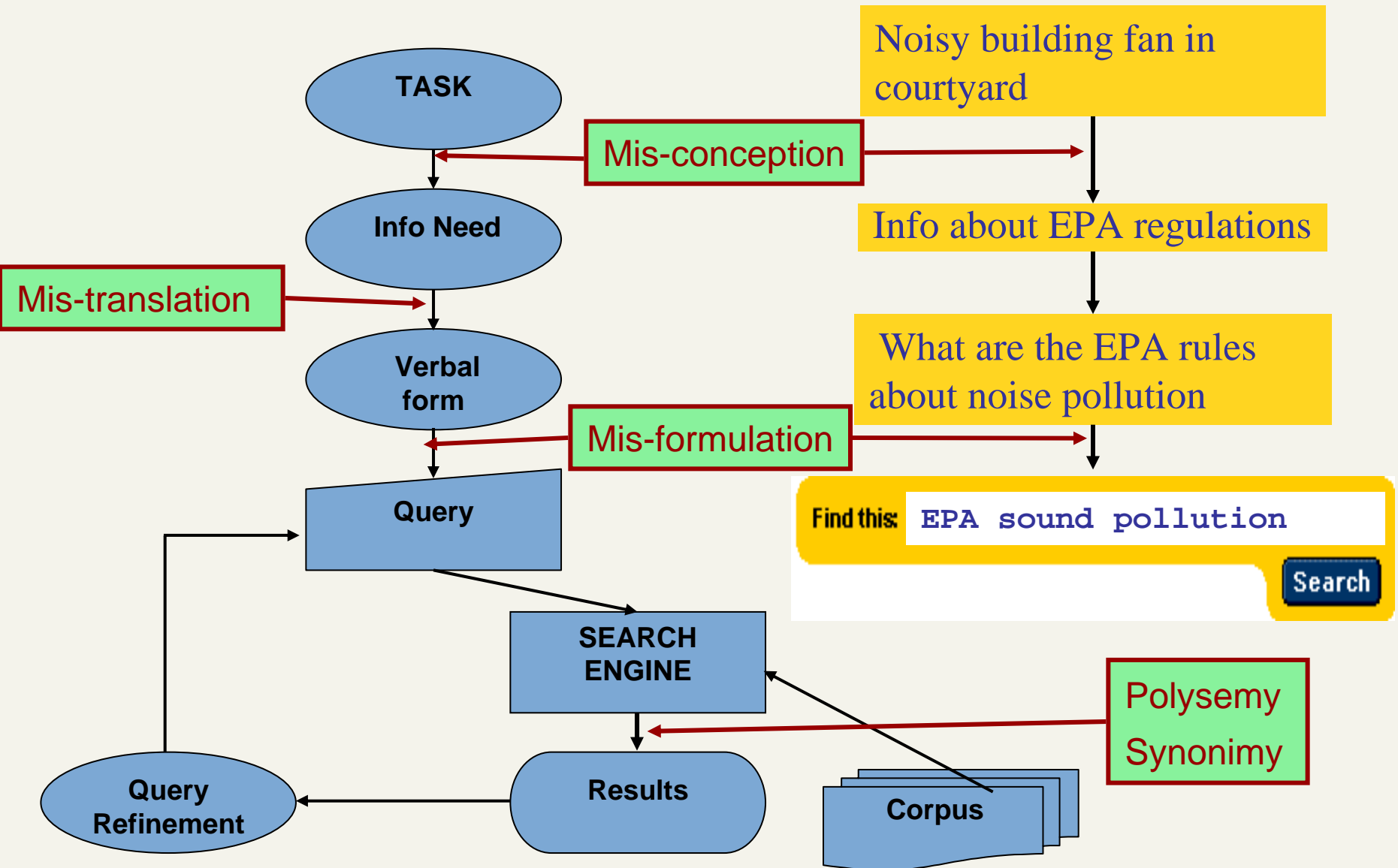
“When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)”



- After reviewing the first few entries
- After reviewing the first page
- After reviewing the first 2 pages
- After reviewing the first 3 pages
- After reviewing more than 3 pages

(Source: [iprospect.com](http://iprospect.com) WhitePaper\_2006\_SearchEngineUserBehavior.pdf)

# True example\*



\* To Google or to GOTO, Business Week Online,  
September 28, 2001

EPA = US Environmental Protection Agency

# Users' empirical evaluation of results

---

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, new info, non-duplicates, 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
- User perceptions may be unscientific, but are significant over a large aggregate

# Users' empirical evaluation of engines

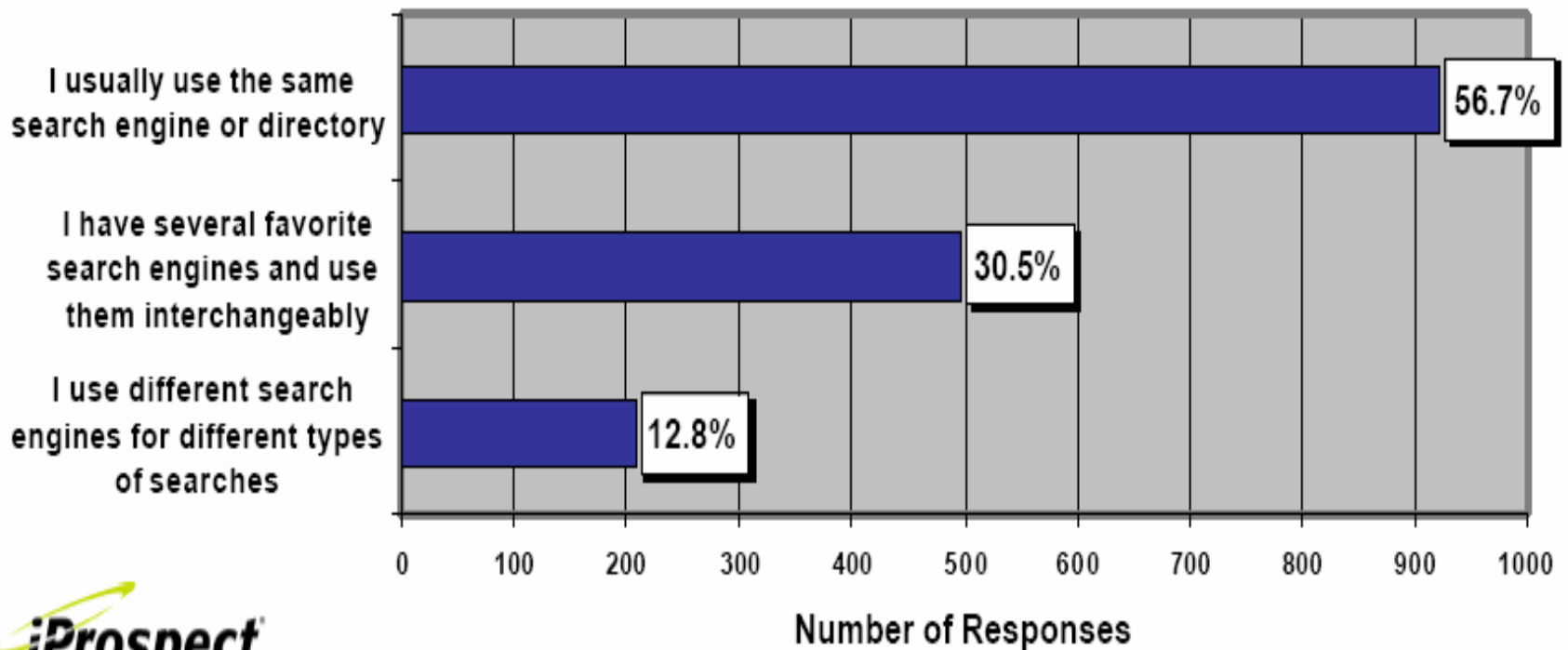
---

- Relevance and validity of results
- Speed
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective
- Coverage of topics for poly-semantic queries
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, syntax errors,...)
  - 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

# Loyalty to a given search engine

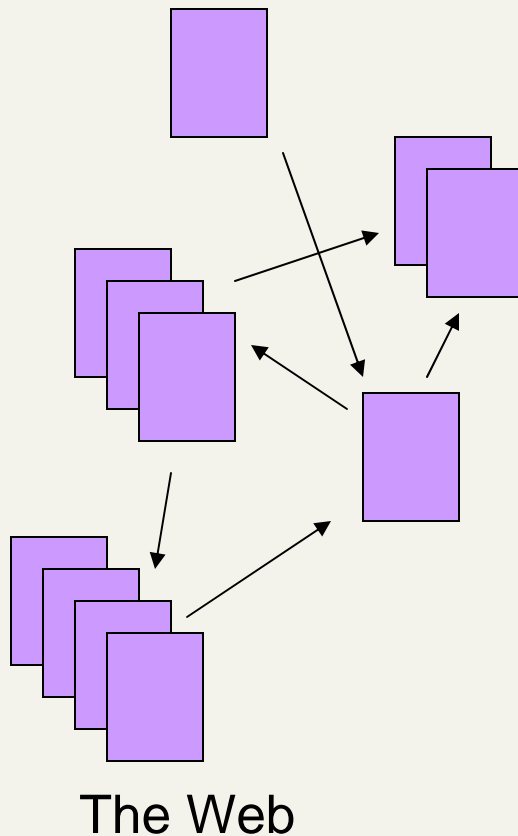
(iProspect Survey, 4/04)

Which would you say best describes how you use search engines?



# The Web corpus

---



- 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 corpora ... but corporate records are catching up.
- Growth – slowed down from initial “volume doubling every few months” but still expanding
- Content can be *dynamically generated*

# The Web: Dynamic content

---

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





# Dynamic content

---

- Most dynamic content is ignored by web spiders
  - Many reasons including malicious spider traps
- Some dynamic content (news stories from subscriptions) are sometimes delivered as static content
  - Application-specific spidering
- Spiders commonly view web pages just as Lynx (a text browser) would
- Note: even “static” pages are typically assembled on the fly (e.g., headers are common)

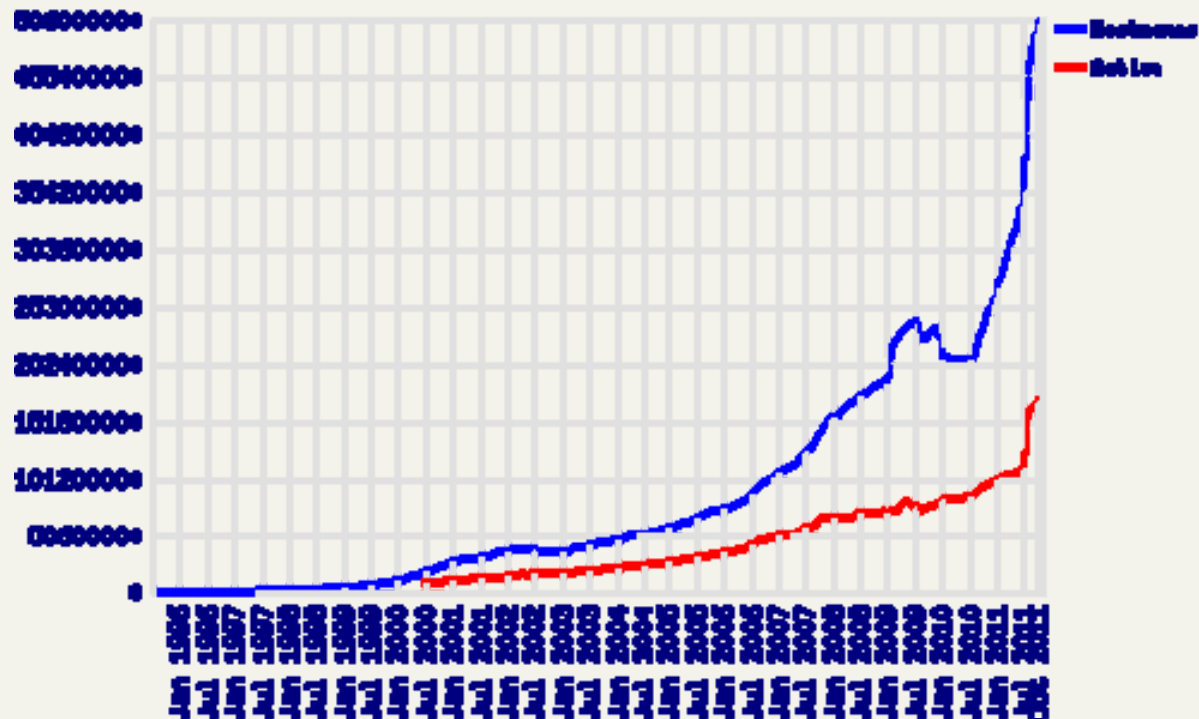
# The web: size

---

- What is being measured?
  - Number of hosts
  - Number of (static) html pages
    - Volume of data
- Number of hosts – netcraft survey
  - [http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html)
  - Monthly report on how many web hosts & servers are out there
- Number of pages – numerous estimates (will discuss later)

# Netcraft Web Server Survey

[http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html)



# The web: evolution

---

- All of these numbers keep changing
- Relatively few scientific studies of the evolution of the web [Fetterly & al, 2003]
  - <http://research.microsoft.com/research/sv/sv-pubs/p97-fetterly/p97-fetterly.pdf>
- Sometimes possible to extrapolate from small samples (fractal models) [Dill & al, 2001]
  - <http://www.vldb.org/conf/2001/P069.pdf>

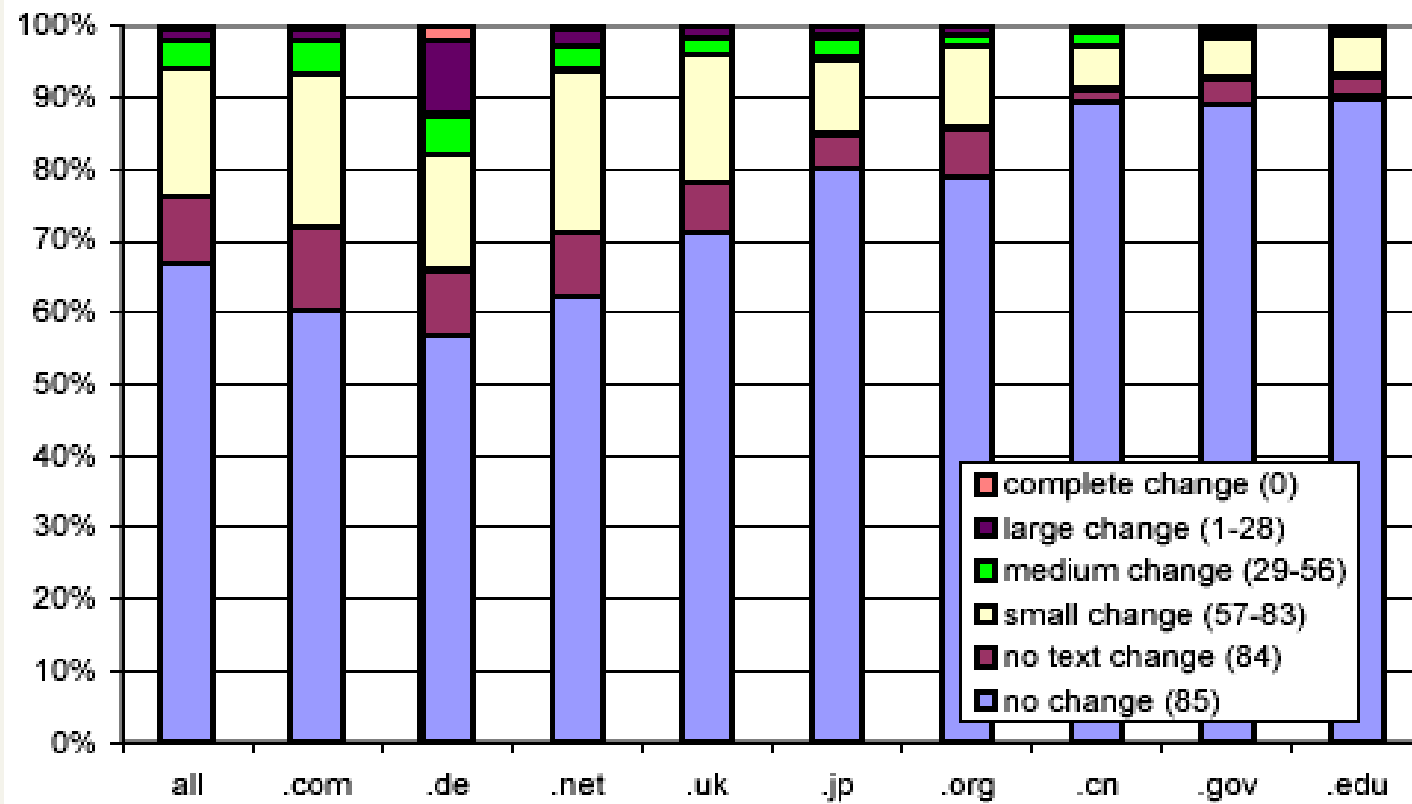
# Rate of change

---

- [Cho00] 720K pages from 270 popular sites sampled daily from Feb 17 – Jun 14, 1999
  - Any changes: 40% weekly, 23% daily
- [Fett02] Massive study 151M pages checked over few months
  - Significant changed -- 7% weekly
  - Small changes – 25% weekly
- [Ntul04] 154 large sites re-crawled from scratch weekly
  - 8% new pages/week
  - 8% die
  - 5% new content
  - 25% new links/week

# Static pages: rate of change

- Fetterly et al. study (2002): several views of data, 150 million pages over 11 weekly crawls
  - Bucketed into 85 groups by extent of change



# Other characteristics

---

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

# Spam

## Search Engine Optimization



# The trouble with paid placement...

---

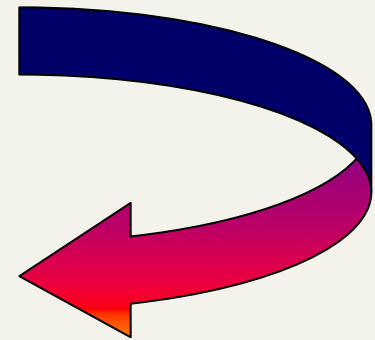
- It costs money. What's the alternative?
- *Search Engine Optimization:*
  - “Tuning” your web page to rank highly in the search results for select keywords
  - Alternative to paying for placement
  - Thus, intrinsically a marketing function
- Performed by companies, webmasters and consultants (“Search engine optimizers”) for their clients
- Some perfectly legitimate, some very shady

# Simplest forms

---

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

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, ..."

# Search engine optimization (Spam)

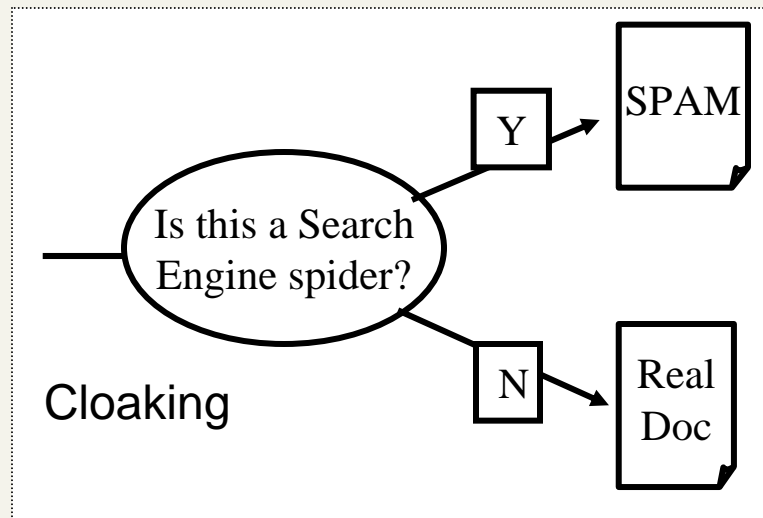
---

- Motives
  - Commercial, political, religious, lobbies
  - Promotion funded by advertising budget
- Operators
  - Contractors (Search Engine Optimizers) for lobbies, companies
  - Web masters
  - Hosting services
- Forums
  - E.g., Web master world ( [www.webmasterworld.com](http://www.webmasterworld.com) )
    - Search engine specific tricks
    - Discussions about academic papers ☺

# Cloaking

---

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




# The spam industry

## Search Engine Cloaker

Need more search engine listings?

### OUTSMART SEARCH ENGINES TO GET MORE HITS

Search Engine Cloaker is used by hundreds of top-ranked Webmasters to increase their search engine rankings.



#### Web Guide

Our hand-picked directory of the best business links on the web.

##### Cloaking

Category Path

[Home](#) > [Guide Topics](#) > [Technology](#) > [Internet](#) > [Search Technology](#) > [Search Engines](#) > [Search Engine Placement](#) > [Cloaking](#)


Links 1-8 of 8

[David: To Cloak or Not to Cloak?](#)

... at the "Cloaking & Doorways" ... one of Internet.com's ... cloaking

[News](#) [Best Keywords!](#) [SE](#)

#### phantomLine™ — the ultimate stealth



### Understanding Cloaking

#### Tutorial: Cloaking and Stealth Technology

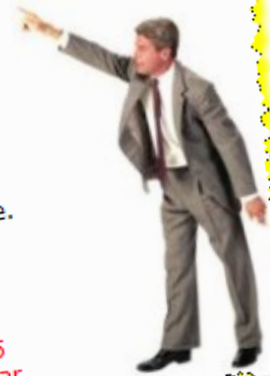
[Page 1](#) | [Page 2](#) | [Page 3](#) | [Page 4](#) | [Page 5](#)

Cloaking, stealth or phantom page technology constitutes the most sophisticated and efficient approach towards search engine optimization. A mystique surrounding cloaking or stealth tech

#### Free Domain Forwarding - Domain Cloaking - DNS Forwarding

Web site is cloaked when the web address of a web site is hidden from viewers in their browser window.

For example your user would type in  
www.yourname.com into their browser window.  
They are then automatically redirected to your web site:  
(http://www.someisp.com/~users/yourname/yoursite.html) or any where you like.  
However your users would continue to  
www.yourname.com as they browsed.



Cloaking Services: Included Branded Email Services 5  
Mail boxes mailboxname@yourDomain.com \$49/Year

[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](http://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/](http://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](http://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](http://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](http://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](http://www.blogbusinesssummit.com)

Los Angeles, CA

[Full-Time SEO & SEM Jobs](#)

Find companies big &amp; small hiring full-time SEO &amp; SEM pros right now

[CareerBuilder.com](http://CareerBuilder.com)[SEO Contests](#)Information on SEO Contests like the **Nigritude Ultramarine** contest.[www.seo-contests.com/](http://www.seo-contests.com/)[The SEO Book](#)**Nigritude Ultramarine** & SEO secrets

Fun, free, raw, &amp; different.

[www.seobook.com](http://www.seobook.com)[Ultramarine - Companion](#)

Music - Dance - Electronic

[Overstock.com](http://Overstock.com)

# More spam techniques

---

- **Doorway pages**

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

- **Link spamming**

- Mutual admiration societies, hidden links, awards – more on these later
- *Domain flooding*: numerous domains that point or re-direct to a target page

- **Robots**

- Fake query stream – rank checking programs
  - “Curve-fit” ranking programs of search engines
- Millions of submissions via Add-Url



# The war against spam

---

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

# Answering “the need behind the query”

---

- Semantic analysis
  - Query language determination
    - Auto filtering
    - Different ranking (if query in Japanese do not return English)
  - Hard & soft (partial) matches
    - Personalities (triggered on names)
    - Cities (travel info, maps)
    - Medical info (triggered on names and/or results)
    - Stock quotes, news (triggered on stock symbol)
    - Company info
    - Etc.
- Natural Language reformulation
- Integration of Search and Text Analysis

# The spatial context -- geo-search

---

- Two aspects
  - Geo-coding -- encode geographic coordinates to make search effective
  - Geo-parsing -- the process of identifying geographic context.
- Geo-coding
  - Geometrical hierarchy (squares)
  - Natural hierarchy (country, state, county, city, zip-codes, etc)
  - Geo-parsing
  - Pages (infer from phone nos, zip, etc). About 10% can be parsed.
  - Queries (use dictionary of place names)
  - Users
    - Explicit (tell me your location -- used by NL, registration, from ISP)
    - From IP data
  - Mobile phones
    - In its infancy, many issues (display size, privacy, etc)

# Yahoo!: britney spears

Web | Images | Video | Directory | Local | News | Shopping

**YAHOO!** SEARCH

[My Web](#) BETA

Search Results Results 1 - 10 of about 66,800,000


Also try: [britney spears pictures](#), [britney spears lyrics](#) [More...](#)

**Y!** [Britney Spears Artist Page](#) - [Downloads](#) - [Videos](#) - [Photos](#) - [Buy CDs](#)  
Yahoo! Shortcut - [About](#)


1. [Britney Spears](#) - [Translate this page](#)  
official site with chat, email, tour information, merchandise, and more.  
Category: [Rock and Pop](#) > [Britney Spears](#)  
[www.britneyspears.com](#) - 2k - [Cached](#) - [More from this site](#) - [Save](#) - [Block](#)
2. [Britney.com](#)   
Jive Records' official site.  
Category: [Rock and Pop](#) > [Britney Spears](#)  
[www.britney.com](#) - 10k - [Cached](#) - [More from this site](#) - [Save](#) - [Block](#)
3. [World of Britnev](#)

# Ask Jeeves: las vegas

[Ask.com](#) | [MyJeeves](#)<sup>BETA</sup>


 [Web](#) | [Pictures](#) | [News](#) | [Local](#) | [Products](#) | [More »](#)

**Web Search:** las vegas

[Map of Las Vegas, NV 89101](#) [About](#)

**Local Information for Las Vegas, NV**  
**Find:** [Map](#) | [Jobs](#) | [Current Weather](#) | [Seasonal Climate](#) | [Local Time](#) | [Yellow Pages](#)  
**Go To:** [Official Site](#) | [Chamber of Commerce](#) | [Facts](#) | [Schools](#)

[Other matches: Las Vegas, NM](#)


 **Latest News: Las Vegas** [About](#)  
[Pacific Coast League Standings](#) AP via San Francisco Chronicle 27 minutes ago  
[New STACKED Trailer Reveals Exciting Poker Room Drama](#) gamesindustry.biz 1 hour ago

**Sponsored Web Results**

[Las Vegas Hotel Deals](#)  
Up to 70% Off Everything **Vegas**; Rooms, Shows, Golf & Tours.  
[www.TripReservations.com](#)

[Vegas Hotels & Shows](#)  
Great Rates on Hotels, Shows & More **VEGAS.com** Best **Vegas** Rate Guarantee  
[www.VEGAS.com](#)


# Yahoo! salvador hotels



 **SEARCH**

Web | [Images](#) | [Video](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#)

[My Web](#) **BETA**

Search Results Results 1 - 10 of about 17,900,00

 [Hotels in Salvador, Brazil - photos, reviews and deals](#)  
[Pestana Bahia Hotel - Salvador, Brazil](#) - ★★★★★☆ - from \$95.00 - [availability](#) - [rate it](#)  
[Sol Victoria Marina - Salvador, Brazil](#) - ★★★★★★ - from \$85.00 - [availability](#) - [rate it](#)  
[Praia do Forte Resort - Salvador, Brazil](#) - ★★★★★☆ - from \$170.00 - [rate it](#)  
[Yahoo! Shortcut - About](#)

- [San Salvador hotels: Read San Salvador hotel, motel, lodging reviews and compare prices](#)  
[hotels in San ...](#)   
San Salvador hotels, motels, resorts, inns and bed and breakfast: Find reviews, travel articles, guidebook lists, availability, price, deals, photos, class, amenities and more about hotels in San Salvador, El Salvador at [Y travel.yahoo.com/p-hotel-482915-san\\_salvador\\_hotels-i](#) - [More from this site](#) - [Save](#) - [Block](#)
- [Salvador Brazil Hotels ... Jake.com](#)   
Jake.com offers a cool new way to search for the perfect hotel for your trip to Salvador, Brazil. If you need a Business, Family or just want to stay in the hippest hotels Jake.com will help you find what your looking for.  
[Salvador Hotels](#)    [Airport \(0\)](#) [Salvador Hotels](#)    [Bahia Othon Palace](#)






# Yahoo shortcuts

- Various types of queries that are “understood”

## Yahoo! Shortcuts

### Shortcuts Categories


most popular: [Images](#) - [Maps](#) - [Weather](#) Suggest a new Shortcut. [Click Here](#)

 <h3><u>Local</u></h3> <p><a href="#">Maps</a> <a href="#">Weather</a> <a href="#">Local Listings</a> <a href="#">more...</a></p>	 <h3><u>News &amp; Information</u></h3> <p><a href="#">Sports Scores</a> <a href="#">Stock Quotes</a> <a href="#">Images</a> <small>NEW!</small> <a href="#">more...</a></p>	 <h3><u>Travel</u></h3> <p><a href="#">Airport Information</a> <a href="#">Flight Tracker</a> <a href="#">Hotel Finder</a> <a href="#">more...</a></p>	 <h3><u>Reference</u></h3> <p><a href="#">Dictionary</a> <a href="#">Definitions</a> <a href="#">Encyclopedia</a> <a href="#">Lookup</a> <a href="#">Synonym Finder</a> <a href="#">more...</a></p>	 <h3><u>Calculators</u></h3> <p><a href="#">Calculator</a> <a href="#">Time Zones</a> <a href="#">Weights and Measures</a> <a href="#">Converter</a> <a href="#">more...</a></p>
---	---	---	--	---



# Google

## andrei broder new york

 [Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#) [Desktop](#) [more »](#)

**Web** Results 1 - 10 of about 13,200 for **andrei broder** [new york](#). (0.22 seconds)

### [Phonebook results for \*\*andrei broder new york\*\*](#)



**Andrei Broder**, (718) 432-6973, 630 W 246th St, Bronx, NY 10471

[Google Maps](#) [Yahoo! Maps](#) [MapQuest](#)

### [A taxonomy of web search](#)

Publisher. ACM Press **New York**, NY, USA ... Aris Anagnostopoulos , **Andrei Z. Broder** , David Carmel, Sampling search-engine results, Proceedings of the 14th ...

[portal.acm.org/citation.cfm?id=792552](http://portal.acm.org/citation.cfm?id=792552) - [Similar pages](#)

### [Optimal plans for aggregation](#)

**Andrei Broder**, IBM Research Division. Michael Mitzenmacher, Harvard University  
... John Wiley and Sons. **New York** 1983. 16 M. Shaked and J.G. Shanthikumar ...

# Answering “the need behind the query”: Context

---

- Context determination
  - spatial (user location/target location)
  - query stream (previous queries)
  - personal (user profile)
  - explicit (user choice of a vertical search, )
  - implicit (use Google from France, use google.fr)
- Context use
  - Result restriction
    - Kill inappropriate results
  - Ranking modulation
    - Use a “rough” generic ranking, but personalize later

# Google: dentists bronx



[Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#) [Desktop](#) [more »](#)

**Web**

Local results for **dentists** near **Bronx, NY**



[Montefiore Medical Ctr](#) - 2.0 miles NE - 3448 Boston Rd, Bronx, 10469 - (718) 547-6111  
[Montefiore Medical Group](#) - 3.2 miles SW - 305 E 161st St, Bronx, 10451 - (718) 579-2500  
[Bronx Park Dental Group](#) - 0.1 miles NE - 2016 Bronxdale Ave # 303, Bronx, 10462 - (718) 792-7972

**Bronx Dentists & Dental Insurance @ Dental Plans**  
**Bronx Dentists @ Dental Plans** - Individual and Family Discount Dental Plans and Insurance, Affordable Dental Coverage Starting at \$79.95 a Year!  
[static.dentalplans.com/newyork/bronx/](#) - 37k - [Cached](#) - [Similar pages](#)

**Bronx, NY - Dentist Reviews, Ratings & Recommendations.**  
DR.Oogle: **Dentist** Reviews, **Dentist** Ratings, **Dentist** Recommendations and **Dentist** Help. Post your case to receive second opinion from another **dentist** or ...  
[new-york.doctoroogle.com/ dentist\\_reviews.cfm/pageID/8/index/E](#) - 81k - Aug 6, 2005 - [Cached](#) - [Similar pages](#)

**New York, NY - Dentist Reviews, Ratings & Recommendations.**  
**Dentist**, New York, NY Abanto, Laarni **Dentist**, **Bronx**, NY Abati, Mario **Dentist**, **Bronx**, NY Abbadessa, Joseph **Dentist**, Staten Island, NY Abbaticchio, Stephen ...  
[new-york.doctoroogle.com/dentist\\_reviews.cfm/pageID/8](#) - 129k

# Yahoo!: dentists (bronx)

**YAHOO!** SEARCH

Web | [Images](#) | [Video](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#)

dentists 10471

Search

My Web **BETA**

Search Results


**Y!** [Local Results for \*\*dentists\*\* near \*\*Bronx\*\* - Map All Results](#)

[Quinnones, Madelyn - Citident Family Dentists](#) (718) 378-5030 - 1581 Westchester Ave, **Bronx**, NY - 1.31mi [map](#)

[Chumsky, Jay R DDS - Citident Family Dentists](#) (718) 378-5030 - 1581 Westchester Ave, **Bronx**, NY - 1.31mi [map](#)


[Gold, Andrew DDS - Andrew Gold Association Dentists](#) (718) 299-3600 - 505 Claremont Pkwy, **Bronx**, NY - 2.18mi [map](#)

[Yahoo! Shortcut](#) - [About](#)

1. [Bronx, NY Dentists at SuperPages.com](#) 

Find the Best **Dentists** in **Bronx**, New York at SuperPages.com. SuperPages from Verizon has listings for many more **Bronx** businesses.

[yellowpages.superpages.com/listings.jsp?C=dentists&CID=493578&...](#) - [More from this site](#) - [Save](#) - [Block](#)

2. [Dentists and Clinics in Bronx, New York - Directory USA](#) 

... **Dentists** and Clinics in **Bronx**, NY ... 729 Burke Ave. **Bronx**, NY 10467-6638 ...

Search for:

dentists


Address, City & State, or Zip [Search Tips](#)

Bronx, NY


☐ Make this my default Yahoo! location

Search

 **dentists in**  Bronx NY

 Zoom In

- [1](#)
- [2 street](#)
- [3](#)
- [4 city](#)
- [5](#)
- [6](#)
- [7](#)
- [8 state](#)
- [9](#)
- [10 country](#)

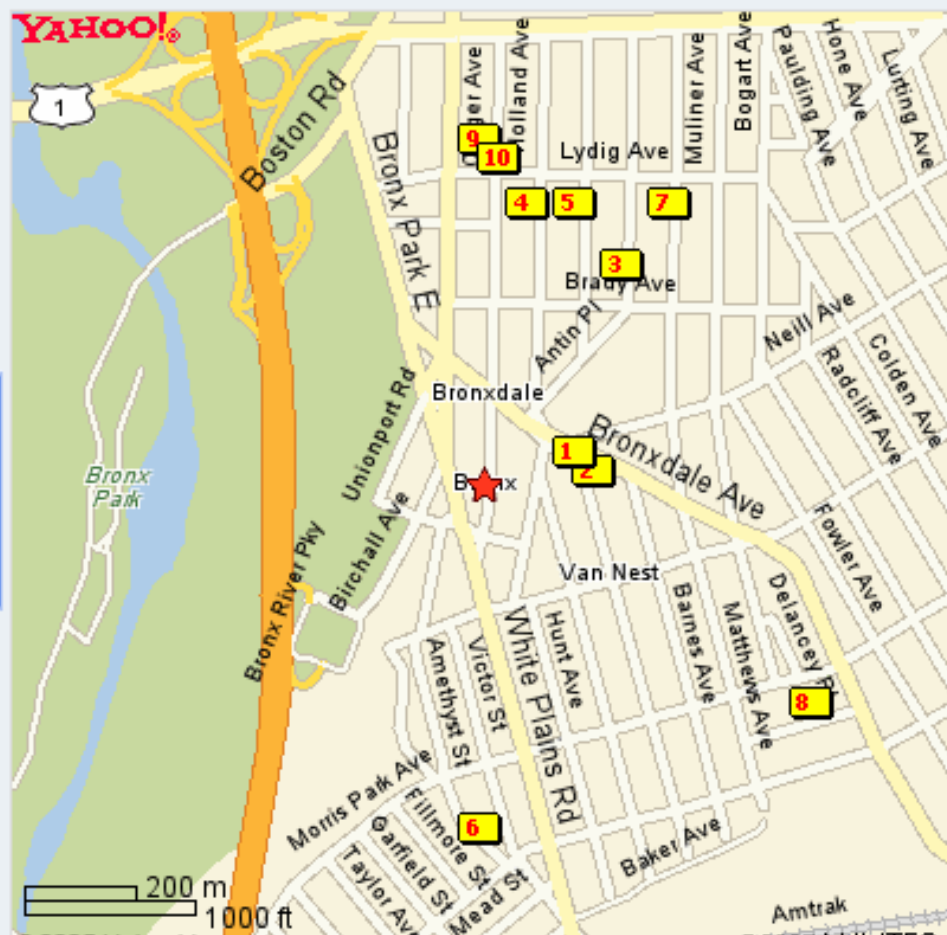
 Zoom Out

Clicking on Map:

- ☒ Zoom in & Re-Center
- ☐ Re-Center Only

[Map Legend](#)

SmartView™



 **dentists**

**Local Results:**

1. [Wilson, Fred - Br...](#)
2. [Belluzzi, Felice ...](#)
3. [Schaffer, Samuel ...](#)
4. [Ratner, Ej Dds](#)
5. [Magideko, Roman D...](#)
6. [Kasparian, Harry ...](#)
7. [Piltser, Yakov Dd...](#)
8. [Chu, Eric K Dds](#)
9. [Kirshbaum, David ...](#)
10. [Peruso, William D...](#)

Results 1-10 of 24

[Next](#) 

[View Detailed Results List](#)


**Map Nearby**

Also find:

e.g. restaurant, museum  
near current results.

[Map it](#)

# Query recommendation

 **TEOMA**  
AN ASK JEEVES SEARCH TECHNOLOGY

Search Tips


rock

Search


[Advanced Search](#)  
[Preferences](#)


☐ Find this phrase

**Sponsored Links**  
[Rock Music](#)  
Download **Rock** Music now. Get your own MP3's, movies & more.  
mp3-downloadhq.com/rock  
[Rock](#)  
Download Your Favorite Songs 100% Free And Legal!  
www.GetMusicFree.com  
[Rock N Roll T Shirts](#)  
Clothing Apparel at Shopping.com  
www.Shopping.com

 **Results**  
Relevant web pages


**Showing 1-10 of about 110,980,000:**  
[Rocks and Minerals](#)  
To navigate your way through this slide show you need to click on the mineral and **rock** groups located directly under the text.  
volcano.und.nodak.edu/vwdocs/vwlessons/les... | [Cached](#)  
[\[Related Pages\]](#)  
  
[Welcome to the \*\*Rock\*\* and Roll Hall of Fame and Museum](#)  
Why should I download Macromedia Flash? Copyright © 2005. The **Rock** and Roll Hall of Fame and Museum, Inc. All Rights Reserved.

 **Refine**  
Suggestions to narrow your search  
  
[Progressive Rock](#)  
[Official Website](#)  
[Rock Music](#)  
[Rock Shop](#)  
[Punk Rock](#)  
[Pop Punk](#)  
  
[\[Show All Refinements\]](#)

 **Resources**  
Link collections from experts and enthusiasts  
  
[Punk Music Explorer](#)  
www.punkrock.org/...  
  
[Rock y Metal Progresivo - 1 - Es.vg](#)  
[Dominios GRATI...](#)  
www.rockprogresivo.es.vg/



# Context transfer

[Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)


[Web](#) [Images](#) [Groups](#) [Directory](#)

Searched the web for **brass boot**. Results **1 - 10** of about **122,000**. Search took **0.44** seconds.

**[BOOT - Dunham Boots - all sizes and widths HERE](#)** Sponsored  
[www.newbalancecatalogcenter.com](#) CLICK and SAVE! - Enter code" FIVE" at checkout

**[Brass Boot Shoes at Shoebuy.com! Free Shipping and No Sales Tax!](#)** Sponsored  
[www.shoebuy.com](#) Enjoy our huge selection, great prices & live customer service! Link

**[Brass Boot](#)**  
Amazing Shoedini! (Shoedini.com) The Easiest Way To Buy Shoes.  
[shop.store.yahoo.com/shoedini/v3.html](#) - 24k - [Cached](#) - [Similar pages](#)

Sponsored Links  
**[Brass Boots at Zappos](#)**  
Free shipping on Brass Boots  
Zappos guarantees your satisfaction  
[www.zappos.com](#)  
Interest: 

# No transfer

Google™ Advanced Search Preferences Language Tools

Web Images Groups Directory

Searched the web for **brass boot** Results 1 - 10 of about 122,000 Search

**BOOT - Dunham Boots - all sizes and widths HERE**  
[www.newbalancecatalogcenter.com](http://www.newbalancecatalogcenter.com) CLICK and SAVE! - Enter code" FIV

**Brass Boot Shoes at Shoebuy.com! Free Shipping and No Sales Tax**  
[www.shoebuy.com](http://www.shoebuy.com) Enjoy our huge selection, great prices & live customer service

**Brass Boot**  
Amazing Shoedini! (Shoedini.com) The Easiest Way To Buy Shoes.  
[shop.store.yahoo.com/shoedini/v3.html](http://shop.store.yahoo.com/shoedini/v3.html) - 24k - [Cached](#) - [Similar pages](#)

**SHOEbuy.com™** FREE SHIPPING. OVER 20 YEARS

HOME | MEN | WOMEN | TEENS | CHILDREN

View All Brands Advanced Search

**Quicksearch™**

Search by Size, Width, Color & More!

View All Brands

**featured COLLECTIONS**

- [Accessories Shop](#)
- [Back2School Shop](#)
- [Boot Shop](#)
- [Bridal Boutique](#)
- [Classics Shop](#)
- [Handbag Shop](#)
- [Sandal Shop](#)
- [Slipper Shop](#)

**featured BRANDS**

**MEN**  
[Dress, Casual, Athletic Boots, Slippers, Sandals](#)

**WOMEN**  
[Dress, Casual, Athletic Boots, Slippers, Sandals](#)

**TEENS**  
[Guys, Girls](#)

**CHILDREN**  
[Boys, Girls, Infants](#)

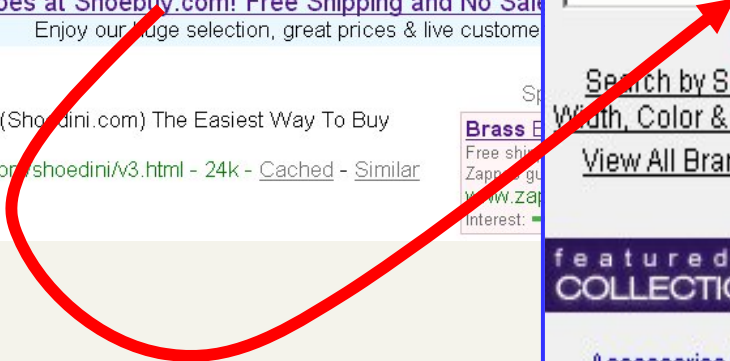
**Welcome to**  
the Net's largest footwear store!  
Free Shipping, No Sales Tax

visit our **back2school**

**kids teens col**

**Rockport**

**Aerosoles**





# Context transfer

Google™ Advanced Search Preferences Language Tools Search Tips


Web Images Groups Directory

Searched the web for **brass boot** Results 1 - 10 of about 122,000 Search took

**BOOT - Dunham Boots - all sizes and widths HERE**  
[www.newbalancecatalogcenter.com](http://www.newbalancecatalogcenter.com) CLICK and SAVE! - Enter code" FIVE" at c

**Brass Boot Shoes at Shoebuy.com! Free Shipping and No Sales Tax**  
[www.shoebuy.com](http://www.shoebuy.com) Enjoy our huge selection, great prices & live customer service

**Brass Boot**  
Amazing Shoedini! (Shoedini.com) The Easiest Way To Buy Shoes.  
[shop.store.yahoo.com/shoedini/v3.html](http://shop.store.yahoo.com/shoedini/v3.html) - 24k - Cached - Similar pages

Sponsored  
**Brass Boots at Zappos.com**  
Free shipping on Brass Boots at Zappos.com  
[www.zappos.com](http://www.zappos.com)  
Interest: 

**Zappos.com** The Web's Most Popular Shoe Store! Shopping Cart | My Account

Shoes Brands Search by Size On Sale

**BRASS BOOT**

Click on a link below to view the collection.


	styles
<a href="#">Brass Boot Entire Collection</a>	33
<a href="#">Brass Boot Impact</a>	2
<a href="#">Brass Boot Neo-Classic</a>	19

Would you like to be notified when we add new styles of **Brass Boot**? Just enter your e-mail address below and we'll let you know!

**About...**

**B**rass Boot offers a range of high-quality men's footwear handcrafted in Italy and Spain. The versatile Contemporary Collection is perfect for the casual office or paired with a suit. The NeoClassic Collection offers professional yet individual styles with distinctive details. The Metro Collection is both fashion-conscious and elegant.

**Did you know?**  
You can search for Brass Boot shoes by color, size, and width below!



# Transfer from search results

[Web](#) | [Images](#) | [Video](#) | [Directory](#) | [Local](#) | [News](#)



**YAHOO!** SEARCH

naive bayes performance

[My Web](#) BETA | [My Search History](#) OFF | [On](#) | [Subscriptions](#) (New) | [Shortcuts](#)

Search Results

Results 1 - 10 of about 75,200 for **naive bayes perform**

- [1. An analysis of data characteristics that affect \*\*naive Bayes performance\*\* \(PDF\)](#)   
... An analysis of data characteristics that affect **naive Bayes performance** ... ally a better indicator of **naive Bayes performance** than the ...  
[www.research.ibm.com/PM/icml01.pdf](#) - 357k - [View as html](#) - [More from this site](#) - [Save](#) - [Block](#)
- [2. Improving the \*\*Performance\*\* of \*\*Naive Bayes\*\* for Text Classification \(PDF\)](#)   
Improving the **Performance** of **Naive Bayes** for. Text Classification. Yirong Shen and Jing Jiang. CS224N Spring 2003. Abstract. We seek to improve the **performance** of the **naive Bayes** classifier  
[www-nlp.stanford.edu/courses/cs224n/2003/fp/yirong99/report.pdf](#) - 135k - [View as html](#) - [More from this site](#) - [Save](#) - [Block](#)



## An analysis of data characteristics that affect naive Bayes performance

Irina Rish  
Joseph Hellerstein  
Jayram Thathachar

RISH@US.IBM.COM  
HELLERS@US.IBM.COM  
JAYRAM@US.IBM.COM

IBM T.J. Watson Research Center 30 Saw Mill River Road, Hawthorne, NY 10532

### Abstract

Despite its unrealistic independence assumption, the naive Bayes classifier is remarkably successful in practice. This paper identifies some data characteristics for which naive Bayes works well, such as certain deterministic and almost-deterministic dependencies (i.e., low-entropy distributions). First, we address zero-Bayes-risk problems, proving naive Bayes optimality for any two-class concept that assigns class 0 to exactly one example (i.e.,  $H(P(x_i|0)) = 0$ ). We demonstrate empirically that the entropy of  $P(x_i|0)$  is a better predictor of the naive Bayes error than the class-conditional mutual information between features. Next, we consider a broader class of non-zero Bayes risk problems, further pursuing the study of low-entropy distributions. We derive error bounds for approximating the joint distribution by the product of marginals in case of nearly-deterministic class-conditional feature distributions  $P(x_i|C)$ , and we demonstrate how the performance of naive Bayes improves with decreasing entropy of such distributions. Finally, we consider functional dependencies between features and prove naive Bayes optimality in certain cases. Using Monte Carlo simulations, we show that naive

simplest Bayesian classifier is the widely used naive Bayes classifier. It greatly simplifies learning by assuming that features are independent given class, that is,  $P(x, c) = \prod_{i=1}^n P(x_i|c)$ , where  $x = (x_1, \dots, x_n)$  is a feature vector and  $c$  is a class. Although feature independence is generally a poor assumption, naive Bayes is surprisingly successful in practice (Langley et al., 1992; Domingos & Pazzani, 1997; Mitchell, 1997; Hellerstein et al., 2000). Naive Bayes has proven effective in text classification, medical diagnosis, and computer performance management, among many other applications.

Why does naive Bayes often work well even though its independence assumption is violated? A central observation is the following: optimality in terms of zero-one loss (classification error) is not necessarily related to the quality of the fit to a probability distribution (i.e., the appropriateness of the independence assumption). Rather, an optimal classifier is obtained as long as both the actual and estimated distributions agree on the most-probable class (Domingos & Pazzani, 1997). For example, (Domingos & Pazzani, 1997) prove naive Bayes optimality for some problems classes that have a high degree of feature dependencies, such as disjunctive and conjunctive concepts.

Herein, we probe further into the data characteristics that make naive Bayes work well. For zero-Bayes-risk problems, we prove naive Bayes optimality for any two-class concept with nominal features where only one example

Search PDF

Hide

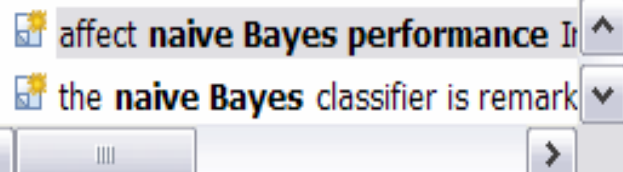
Finished searching for:  
**naive bayes performance**

Total instances found:  
**126**



New Search

Results:



[Done](#)



[Save and View this PDF in Reader](#)



[Find a word in the current PDF document](#)

# Resources

---

- IIR Chapter 19 – 19.4