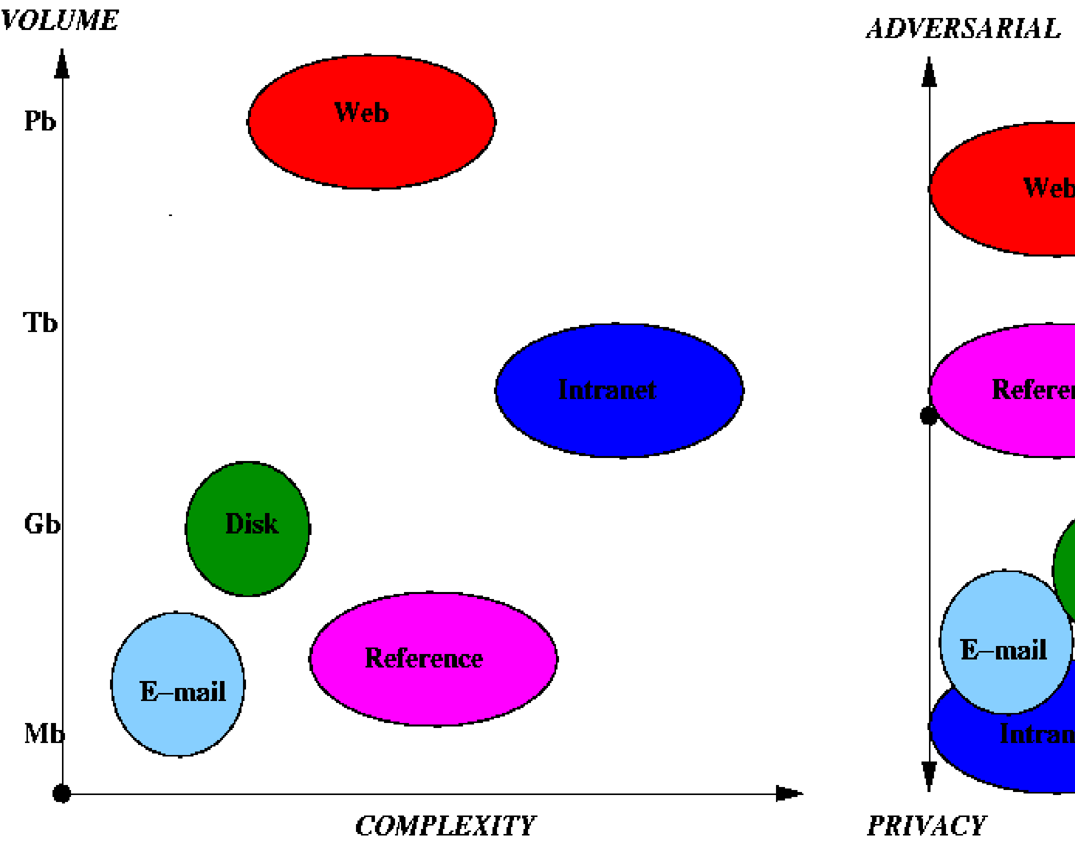
Web Mining

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain & Santiago, Chile

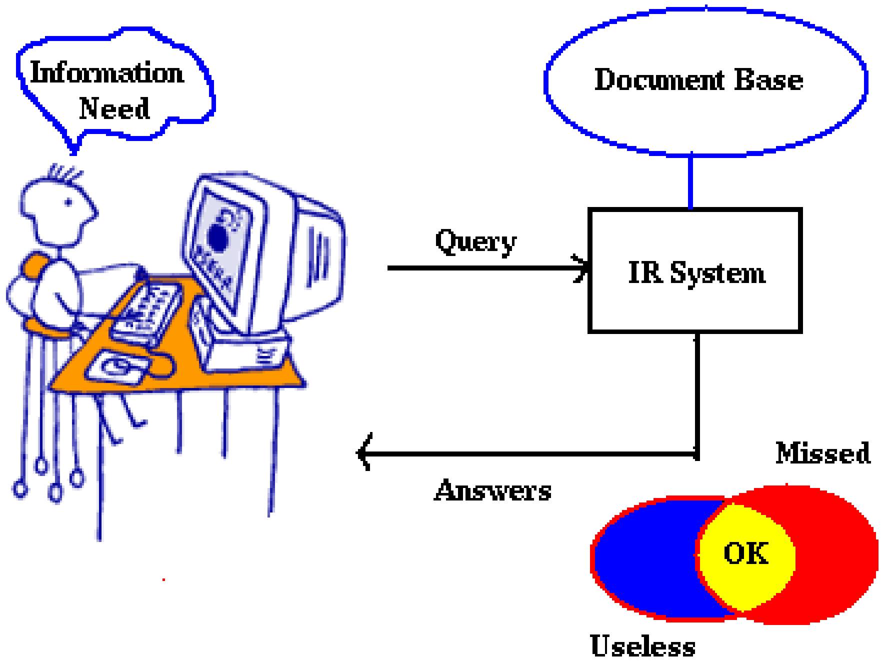
Agenda

- Introduction
- Web Information Retrieval (Web IR)
- Web Mining
- Case Study: Query Mining
- Concluding Remarks

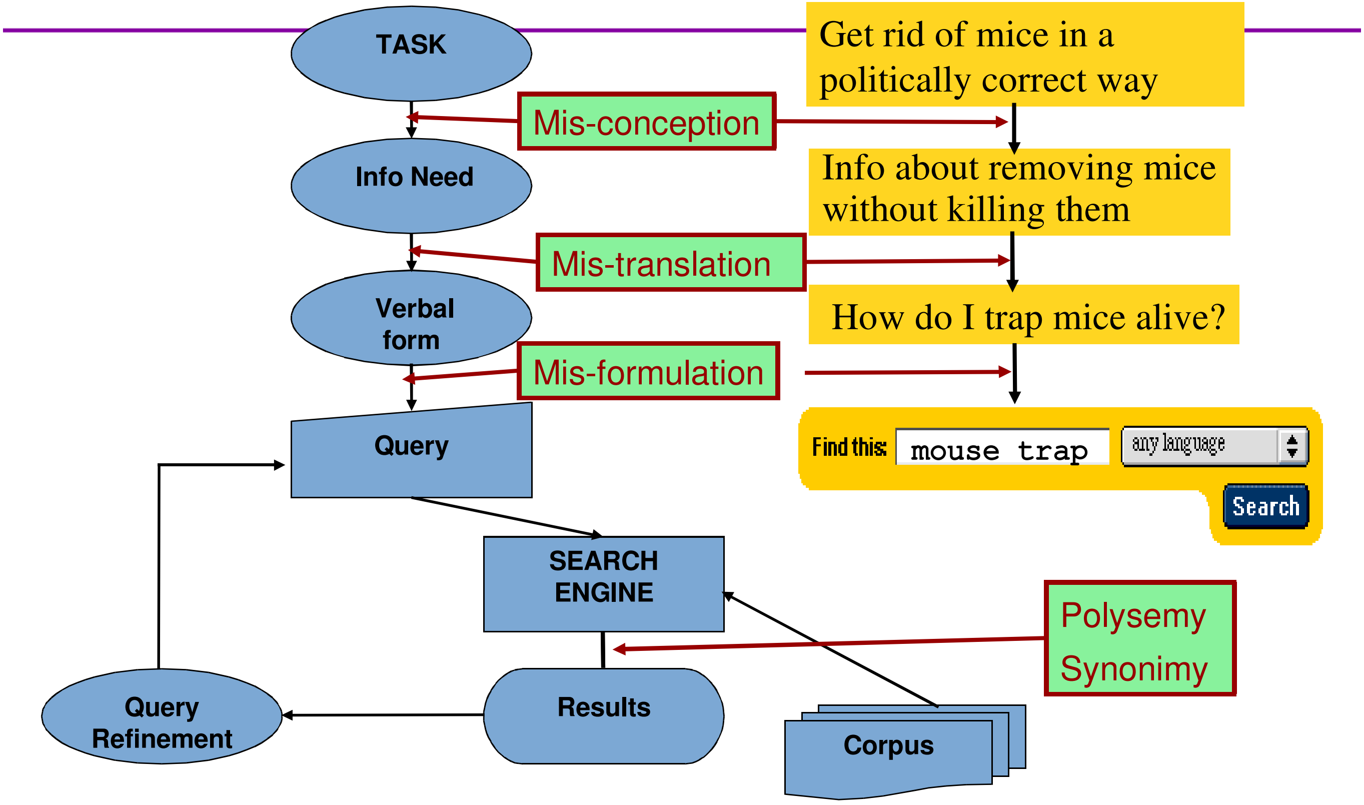
Different Views on Data



The IR Problem



The classic search model



Get rid of mice in a politically correct way

Info about removing mice without killing them

How do I trap mice alive?

Classic IR Goal

– Classic relevance

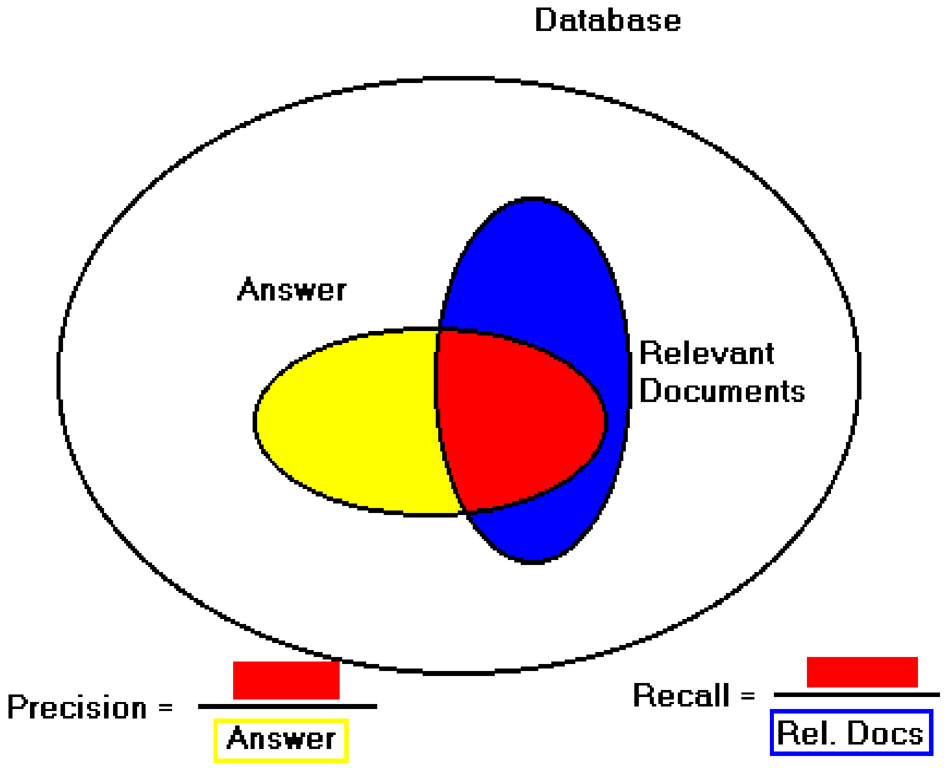
- For each query Q and stored document D in a given corpus assume there exists relevance Score(Q, D)
  - –Score is average over users U and contexts C
- Optimize Score(Q, D) as opposed to Score(Q, D, U, C)
- That is, usually:
  - –Context ignored
  - –Individuals ignored
  - –Corpus predetermined

Bad assumptions in the web context

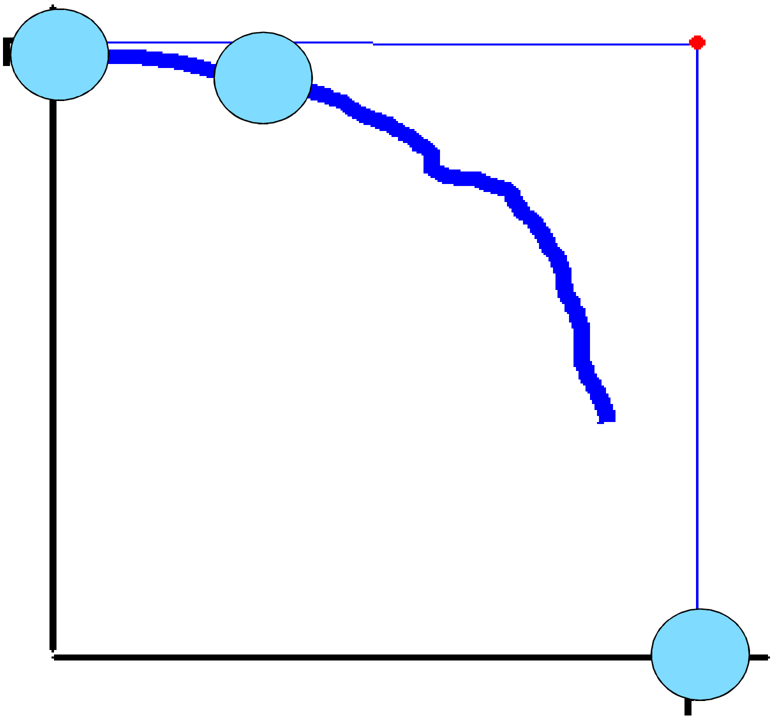
The Notion of Relevance

- Data retrieval: semantics tied to syntax
- Information retrieval: ambiguous semantics
- Relevance:
  - Depends on the user
  - Depends on the context (task, time, etc)
  - Corollary: The Perfect IR System does not exist

Evaluation: First Quality, next Efficiency



Evaluation: Comparing Systems

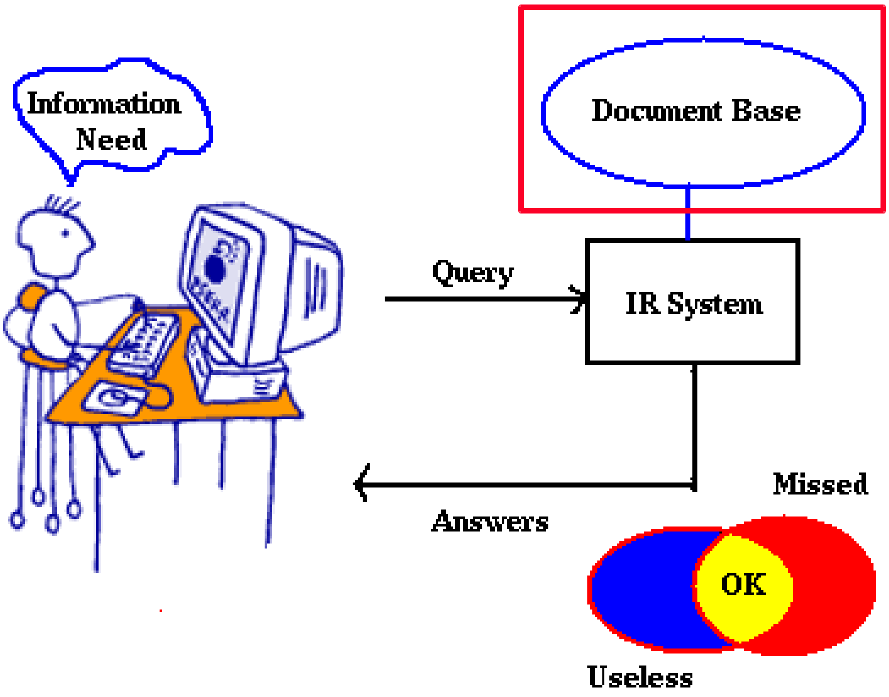


p-r normalized graph
(11 recall levels)

TREC:

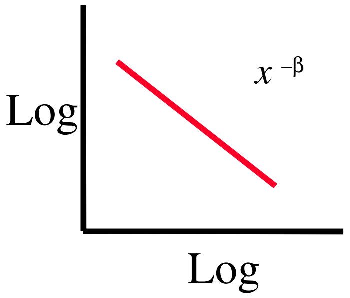
Collection
+
Queries
+
Answers

Challenges in Current IR Systems



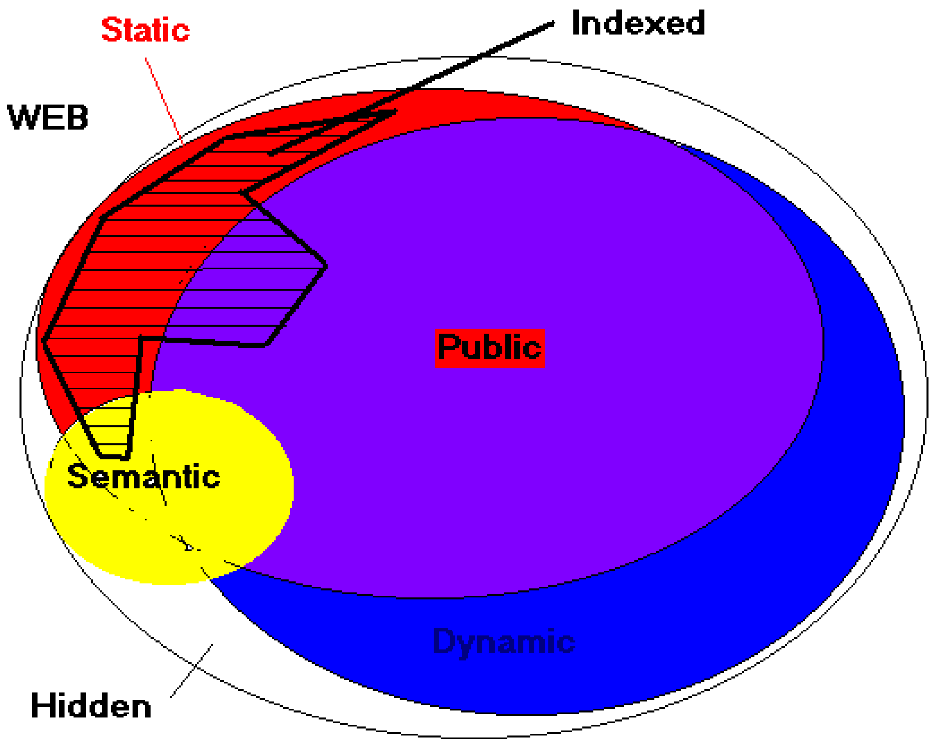
Document Base: Web

- Largest public repository of *data* (more than 20 billion static pages?)
- Today, there are more than 120 million Web servers
- Well connected graph with out-link and in-link power law distributions

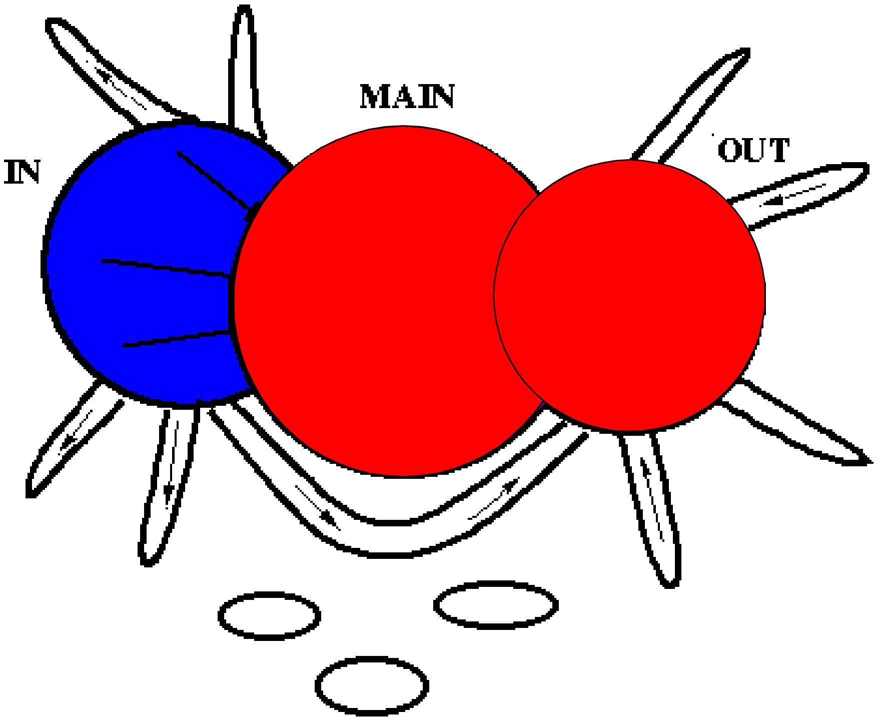


Self-similar &
Self-organizing

The Different Facets of the Web



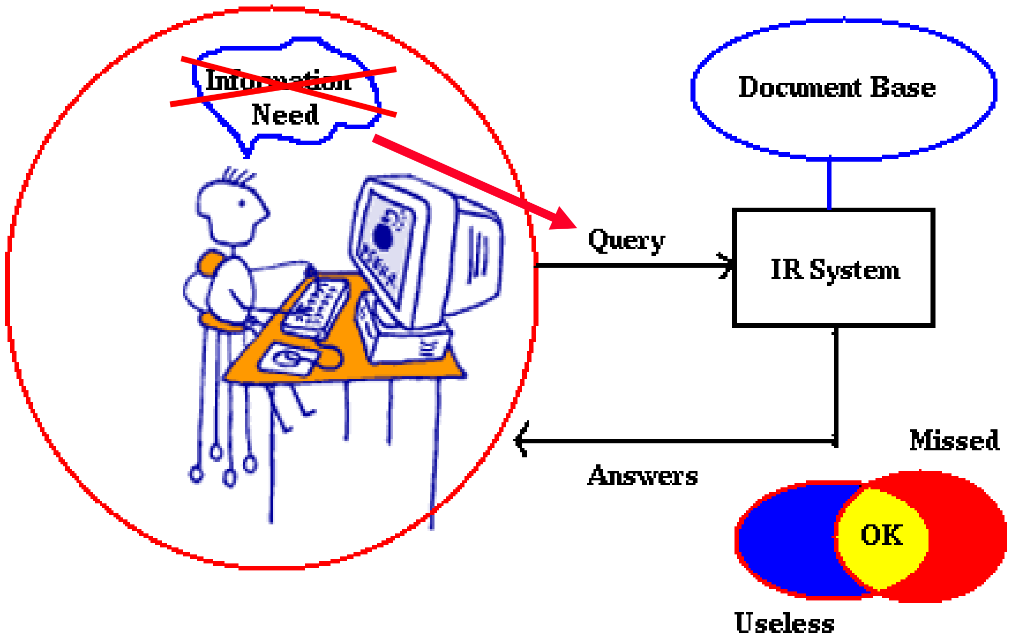
The Structure of the Web



Challenges posed by the data

- Integration of autonomous data sources
  - Data/information integration
- Supporting heterogeneous data
  - How to do effective querying in the presence of structured and text data
  - How to support IR-style querying on DBs
    - Because now users seem to know IR/keyword style querying more, even though structure is good because it supports structured querying!
  - How to support imprecise queries

The User Behind the Query



Web Search Queries

- Cultural and educational diversity
- Short queries & impatient interaction
  - few queries posed & few answers seen
- Smaller & different vocabulary
- Different user goals (Broder, 2000):
  - Information need
  - Navigational need
  - Transactional need
- Refined by Rose & Levinson, WWW 2004

User Needs

- Need (Broder 2002)
  - **Informational** – want to learn about something (~40% / 65%)

    `Low hemoglobin`

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

    `United Airlines`

  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service
    - Downloads
    - Shop

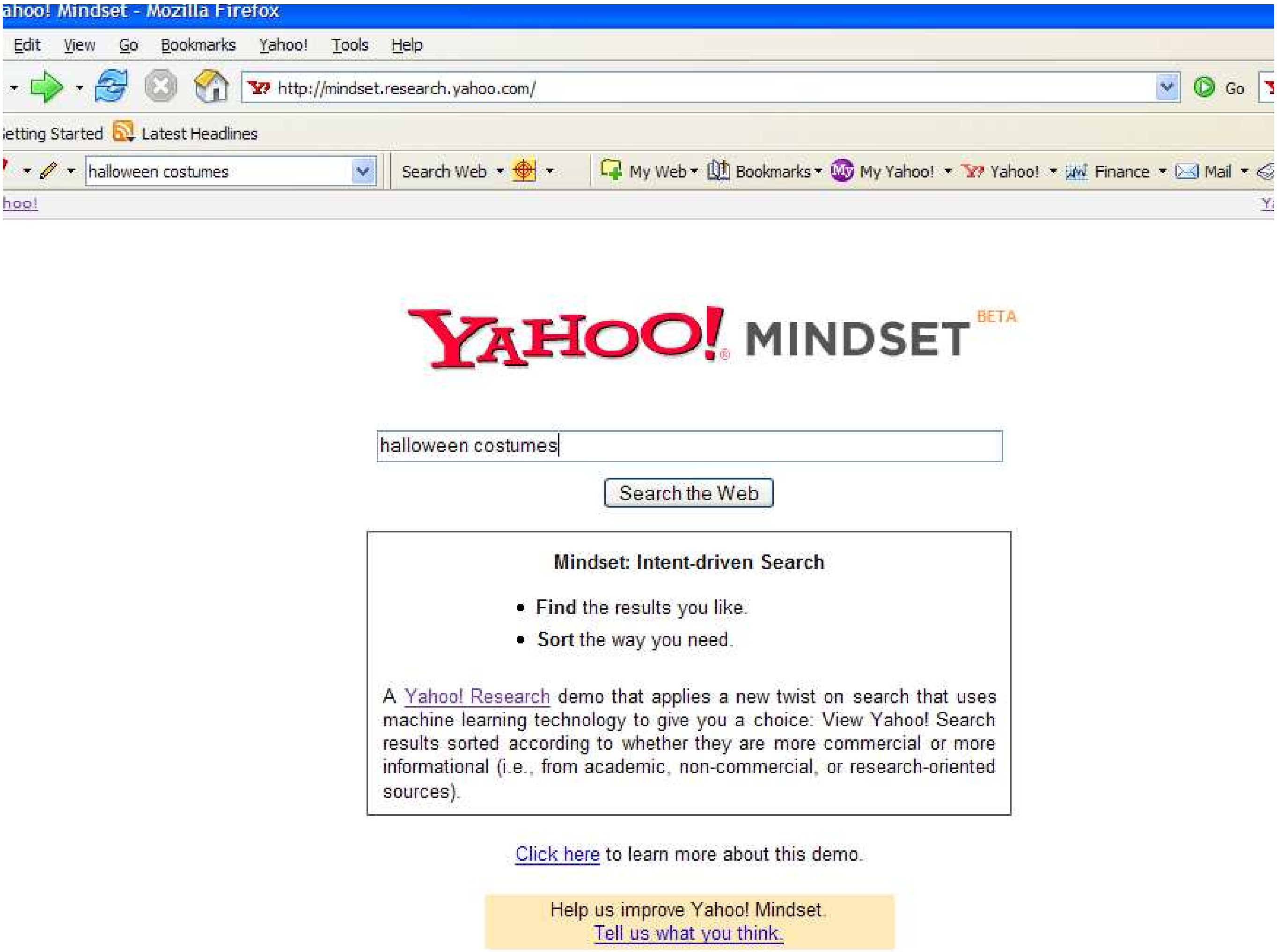
    `Edinburgh weather`

    `Mars surface images`

    `Canon S410`

  - Gray areas
    - Find a good hub
    - Exploratory search "see what's there"

    `Car rental Brasil`


ahoo! Mindset - Mozilla Firefox
Edit View Go Bookmarks Yahoo! Tools Help
http://mindset.research.yahoo.com/
Go
ietting Started Latest Headlines
halloween costumes
Search Web
My Web Bookmarks My Yahoo! Yahoo! Finance Mail
hoo!
YAHOO! MINDSET BETA
halloween costumes
Search the Web

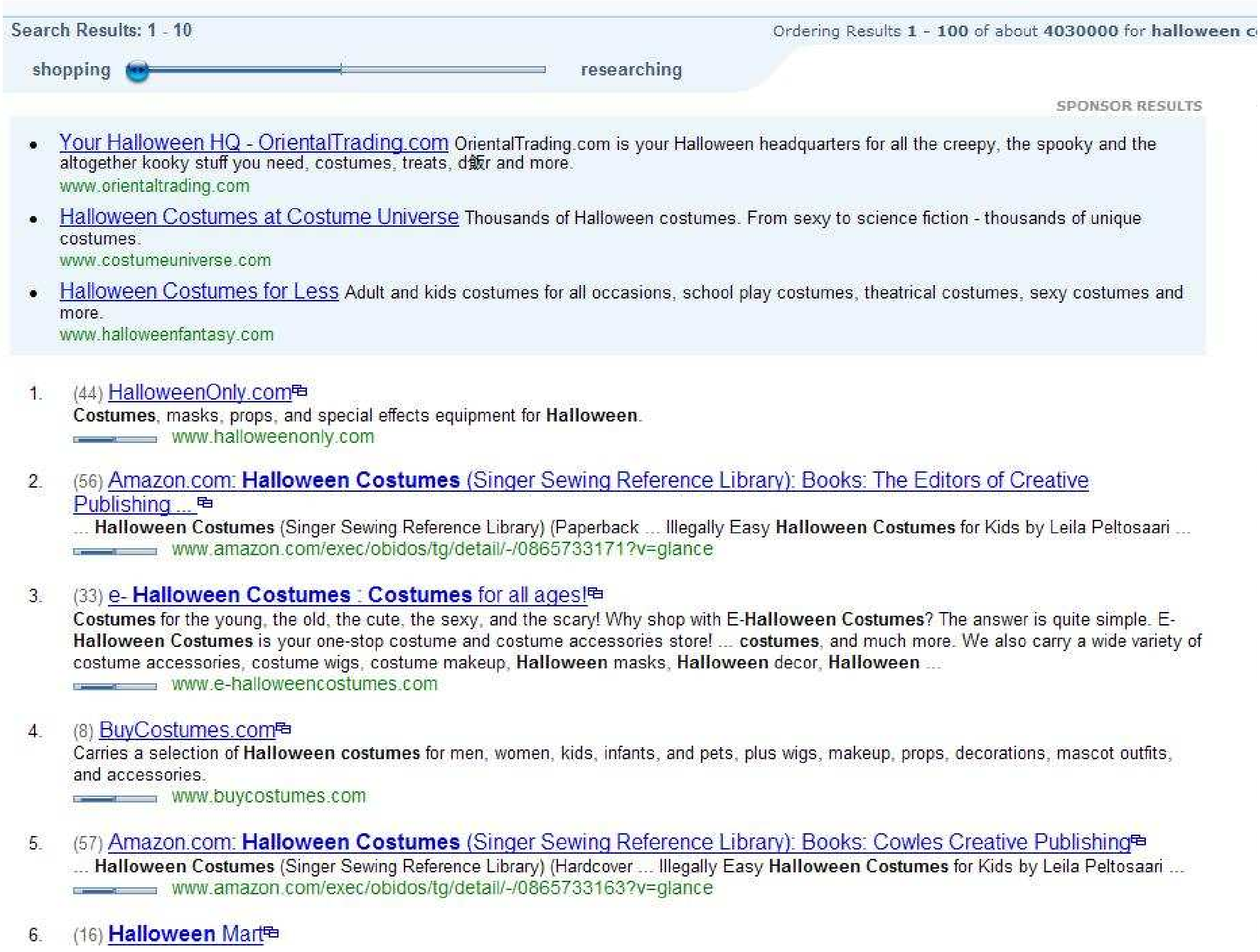
Mindset: Intent-driven Search

- **Find** the results you like.
- **Sort** the way you need.

A Yahoo! Research demo that applies a new twist on search that uses machine learning technology to give you a choice: View Yahoo! Search results sorted according to whether they are more commercial or more informational (i.e., from academic, non-commercial, or research-oriented sources).

Click here to learn more about this demo.

Help us improve Yahoo! Mindset.
Tell us what you think.



Search Results: 1 - 10

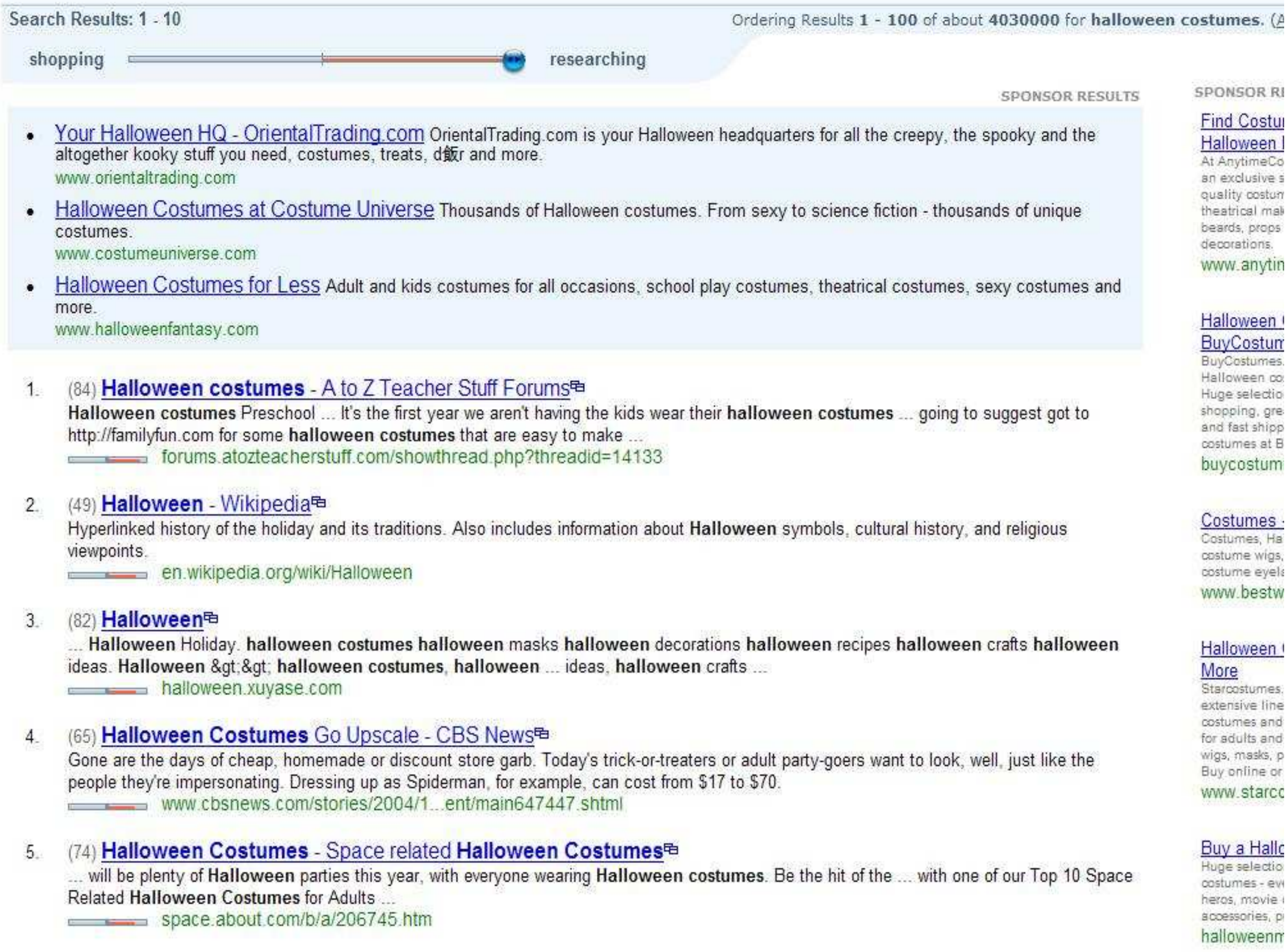
Ordering Results **1 - 100** of about **4030000** for **halloween c**

shopping — researching

SPONSOR RESULTS

- Your Halloween HQ - OrientalTrading.com OrientalTrading.com is your Halloween headquarters for all the creepy, the spooky and the altogether kooky stuff you need, costumes, treats, d飯r and more.
  www.orientaltrading.com
- Halloween Costumes at Costume Universe Thousands of Halloween costumes. From sexy to science fiction - thousands of unique costumes.
  www.costumeuniverse.com
- Halloween Costumes for Less Adult and kids costumes for all occasions, school play costumes, theatrical costumes, sexy costumes and more.
  www.halloweenfantasy.com

1. (44) HalloweenOnly.com
   Costumes, masks, props, and special effects equipment for **Halloween**.
   www.halloweenonly.com
2. (56) Amazon.com: **Halloween Costumes** (Singer Sewing Reference Library): Books: The Editors of Creative Publishing ...
   ... **Halloween Costumes** (Singer Sewing Reference Library) (Paperback ... Illegally Easy **Halloween Costumes** for Kids by Leila Peltosaari ...
   www.amazon.com/exec/obidos/tg/detail/-/0865733171?v=glance
3. (33) e- **Halloween Costumes** : **Costumes** for all ages!
   Costumes for the young, the old, the cute, the sexy, and the scary! Why shop with E-**Halloween Costumes**? The answer is quite simple. E-**Halloween Costumes** is your one-stop costume and costume accessories store! ... **costumes**, and much more. We also carry a wide variety of costume accessories, costume wigs, costume makeup, **Halloween** masks, **Halloween** decor, **Halloween** ...
   www.e-halloweencostumes.com
4. (8) BuyCostumes.com
   Carries a selection of **Halloween costumes** for men, women, kids, infants, and pets, plus wigs, makeup, props, decorations, mascot outfits, and accessories.
   www.buycostumes.com
5. (57) Amazon.com: **Halloween Costumes** (Singer Sewing Reference Library): Books: Cowles Creative Publishing
   ... **Halloween Costumes** (Singer Sewing Reference Library) (Hardcover ... Illegally Easy **Halloween Costumes** for Kids by Leila Peltosaari ...
   www.amazon.com/exec/obidos/tg/detail/-/0865733163?v=glance
6. (16) **Halloween** Mart



Search Results: 1 - 10

Ordering Results **1 - 100** of about **4030000** for **halloween costumes.** (A

shopping — researching

SPONSOR RESULTS

- Your Halloween HQ - OrientalTrading.com OrientalTrading.com is your Halloween headquarters for all the creepy, the spooky and the altogether kooky stuff you need, costumes, treats, d飯r and more.
  www.orientaltrading.com
- Halloween Costumes at Costume Universe Thousands of Halloween costumes. From sexy to science fiction - thousands of unique costumes.
  www.costumeuniverse.com
- Halloween Costumes for Less Adult and kids costumes for all occasions, school play costumes, theatrical costumes, sexy costumes and more.
  www.halloweenfantasy.com

1. (84) **Halloween costumes** - A to Z Teacher Stuff Forums
   Halloween costumes Preschool ... It's the first year we aren't having the kids wear their **halloween costumes** ... going to suggest got to http://familyfun.com for some **halloween costumes** that are easy to make ...
   forums.atozteacherstuff.com/showthread.php?threadid=14133
2. (49) **Halloween** - Wikipedia
   Hyperlinked history of the holiday and its traditions. Also includes information about **Halloween** symbols, cultural history, and religious viewpoints.
   en.wikipedia.org/wiki/Halloween
3. (82) **Halloween**
   ... **Halloween** Holiday. **halloween costumes halloween** masks **halloween** decorations **halloween** recipes **halloween** crafts **halloween** ideas. **Halloween** &gt;&gt; **halloween costumes**, **halloween** ... ideas, **halloween** crafts ...
   halloween.xuyase.com
4. (65) **Halloween Costumes** Go Upscale - CBS News
   Gone are the days of cheap, homemade or discount store garb. Today's trick-or-treaters or adult party-goers want to look, well, just like the people they're impersonating. Dressing up as Spiderman, for example, can cost from $17 to $70.
   www.cbsnews.com/stories/2004/1...ent/main647447.shtml
5. (74) **Halloween Costumes** - Space related **Halloween Costumes**
   ... will be plenty of **Halloween** parties this year, with everyone wearing **Halloween costumes**. Be the hit of the ... with one of our Top 10 Space Related **Halloween Costumes** for Adults ...
   space.about.com/b/a/206745.htm

SPONSOR RI

Find Costur
Halloween I
At AnytimeCo
an exclusive s
quality costum
theatrical mak
beards, props
decorations.
www.anytin

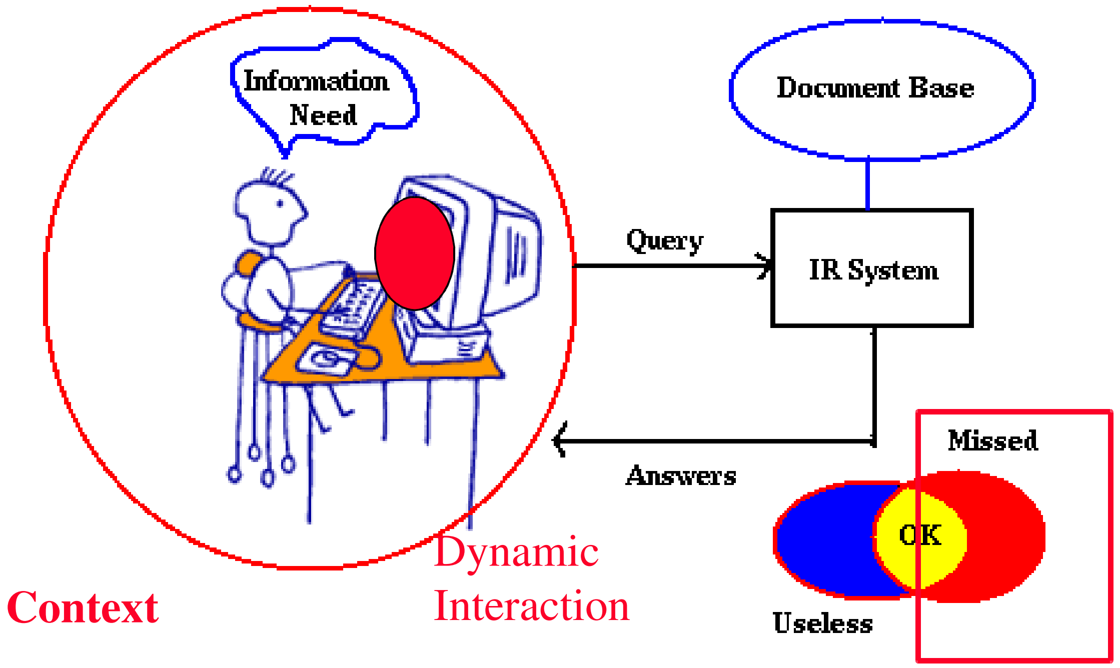
Halloween C
BuyCostum
BuyCostumes.
Halloween co
Huge selectio
shopping, gre
and fast shipp
costumes at B
buycostum

Costumes -
Costumes, Ha
costume wigs,
costume eyela
www.bestw

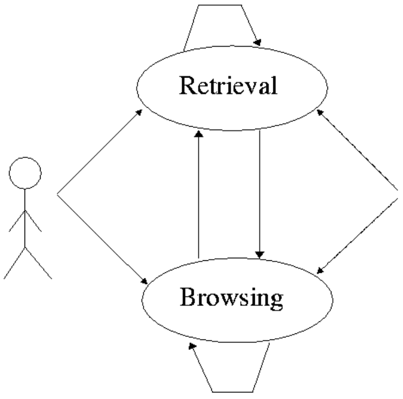
Halloween C
More
Starcostumes.
extensive line
costumes and
for adults and
wigs, masks, p
Buy online or
www.starcc

Buy a Hallo
Huge selectio
costumes - ev
heros, movie
accessories, p
halloweenm

Challenges in Current IR Systems



Interaction

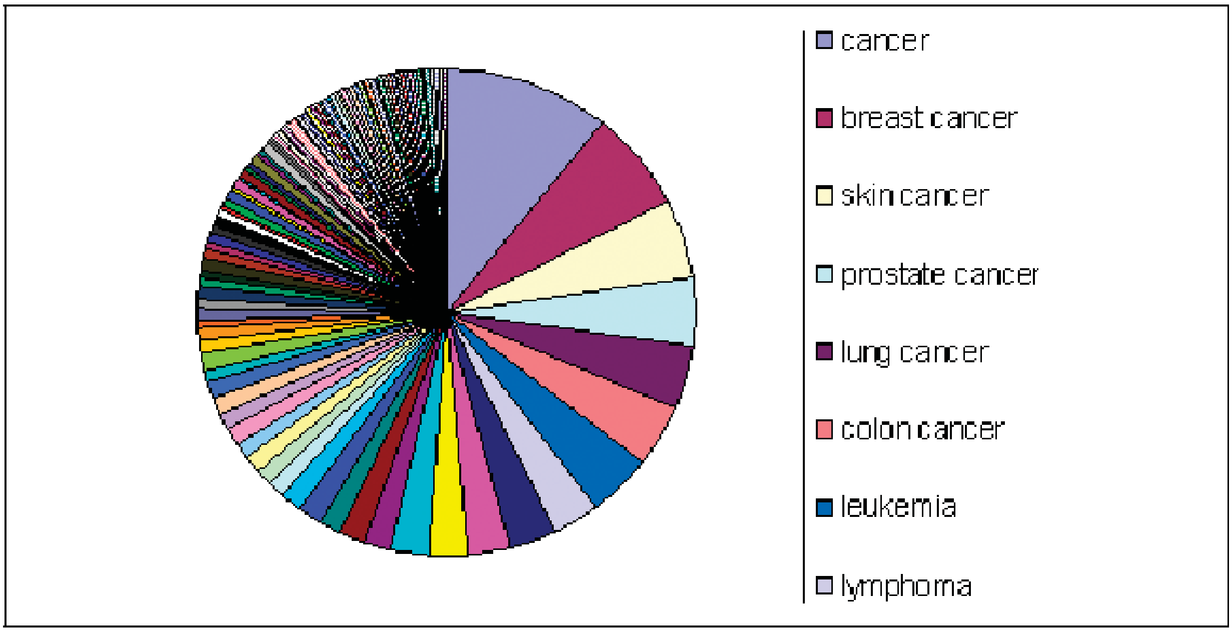


- Inexperienced users
- Dynamic information needs
- Varying task: querying, browsing

- No content overview
- Poor query language, no help

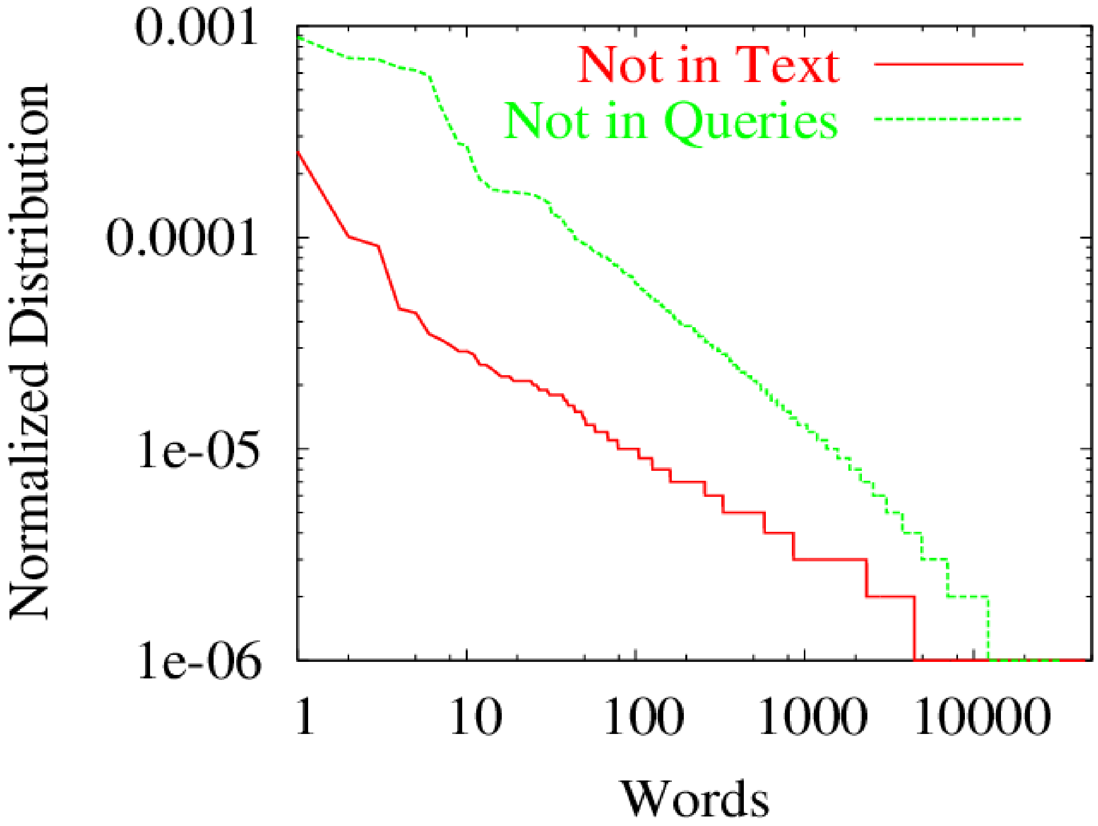
- Poor preview, no visualization
- Missing answers: partial Web coverage, invisible Web, different words or media, ...
- Useless answers

Query Distribution

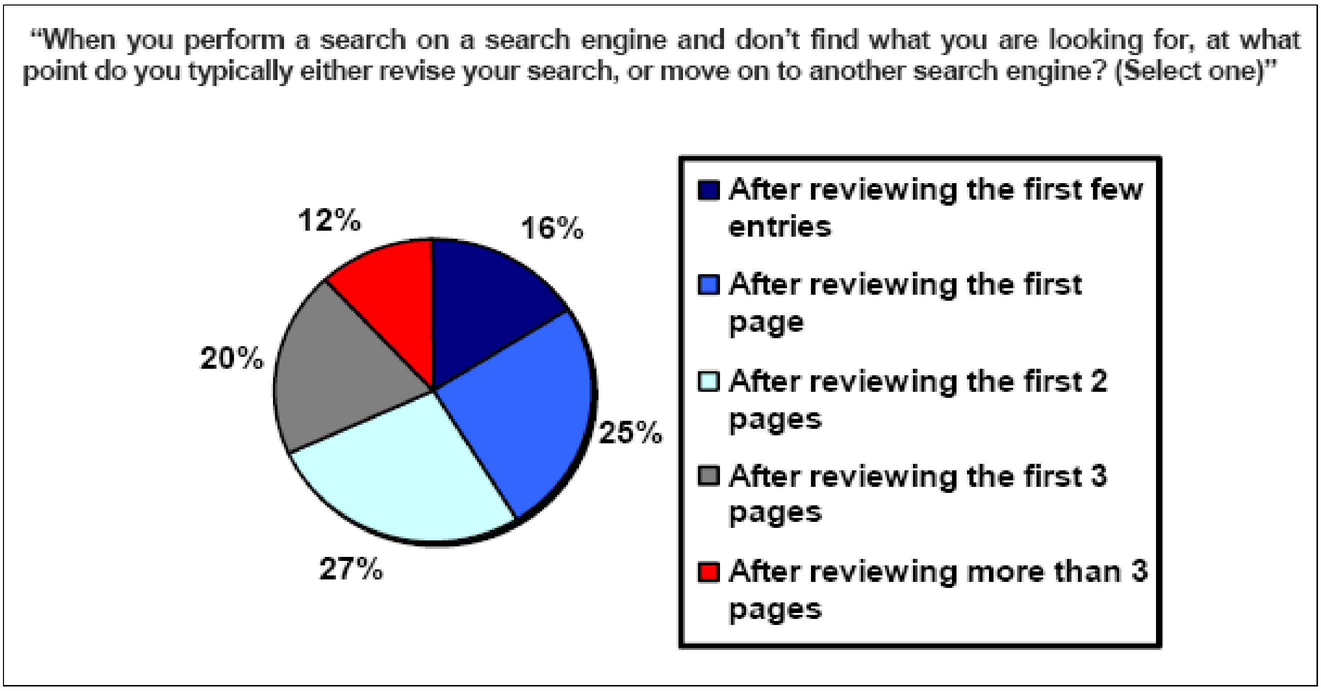


**Power law: few popular broad queries,
many rare specific queries**

Queries and Text



How far do people look for results?



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

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Typical Session

- Two queries of
- .. two words, looking at…
- .. two answer pages, doing
- .. two clicks per page

- What is the goal?

MP3

games

cars

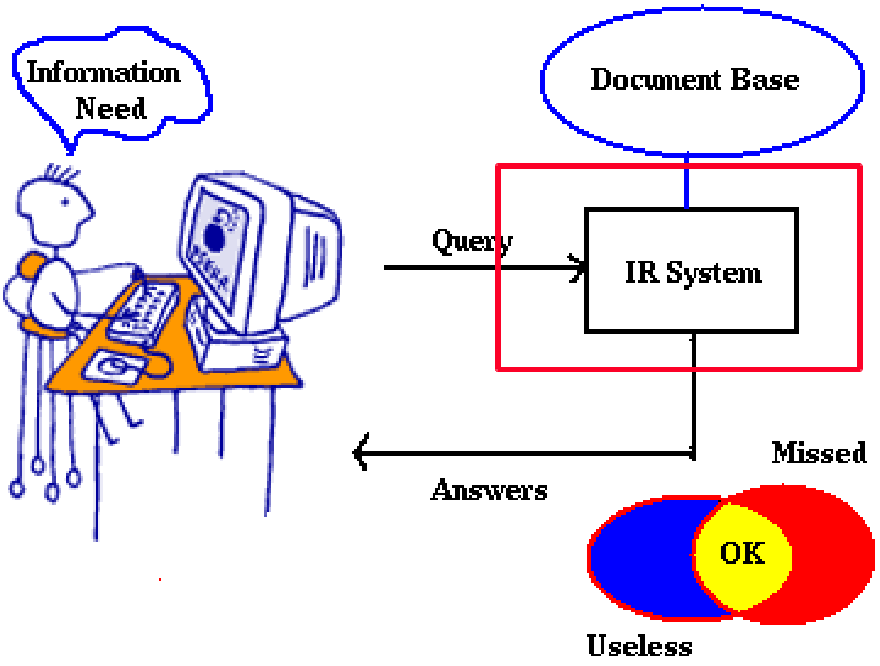
britney spears

pictures

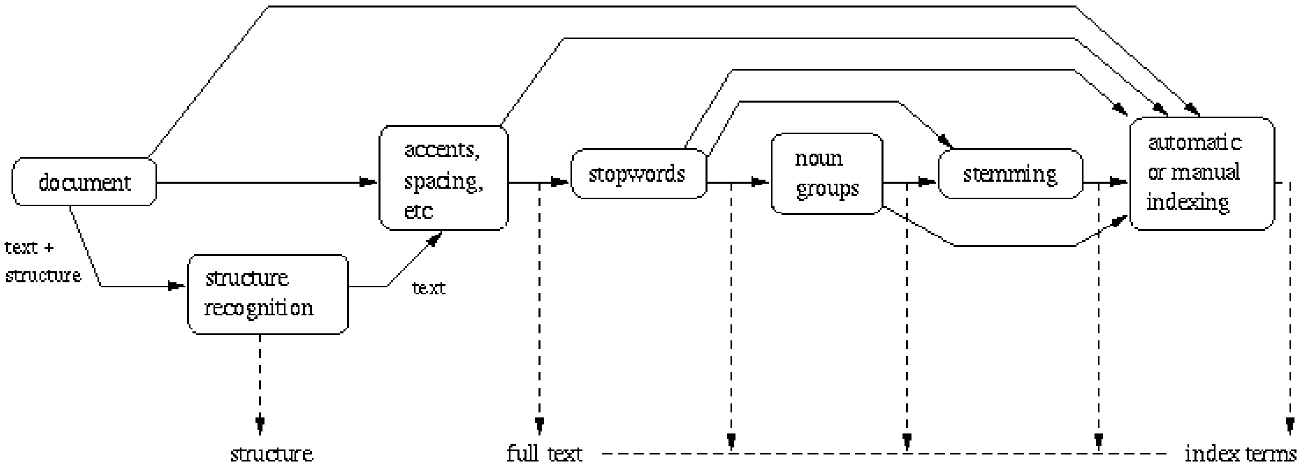
ski

U de Chile

Challenges in Current IR Systems



Bag-of-Words Representation

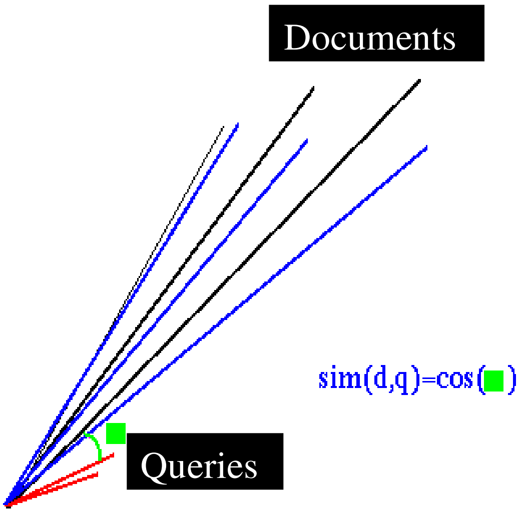


Full-text continuum:

ambiguity vs. completeness trade-off

Text Similarity Models

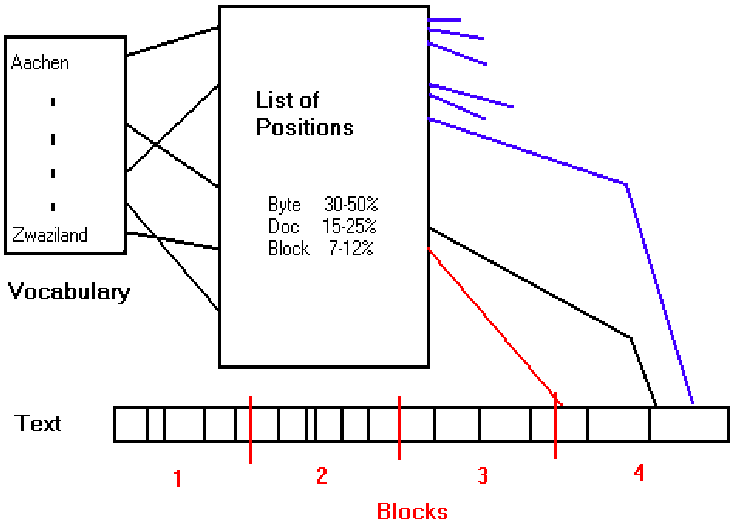
Vector model:
- **words are dimensions**
- ***tf-idf* is used for weights**
- **stopwords vs. rare words**



- Set Models:
  - Boolean, Fuzzy sets, ...
- Algebraic Models:
  - Vector, LSI, etc.
- Probabilistic Models:
  - Probabilistic, Inference & belief networks

Index

- Inverted index
- Lists sorted by weight
  - global (e.g. Pagerank)
  - local (e.g. word weights)
- Hashing + set operations
- Compressed
- Incremental updates



Web Retrieval

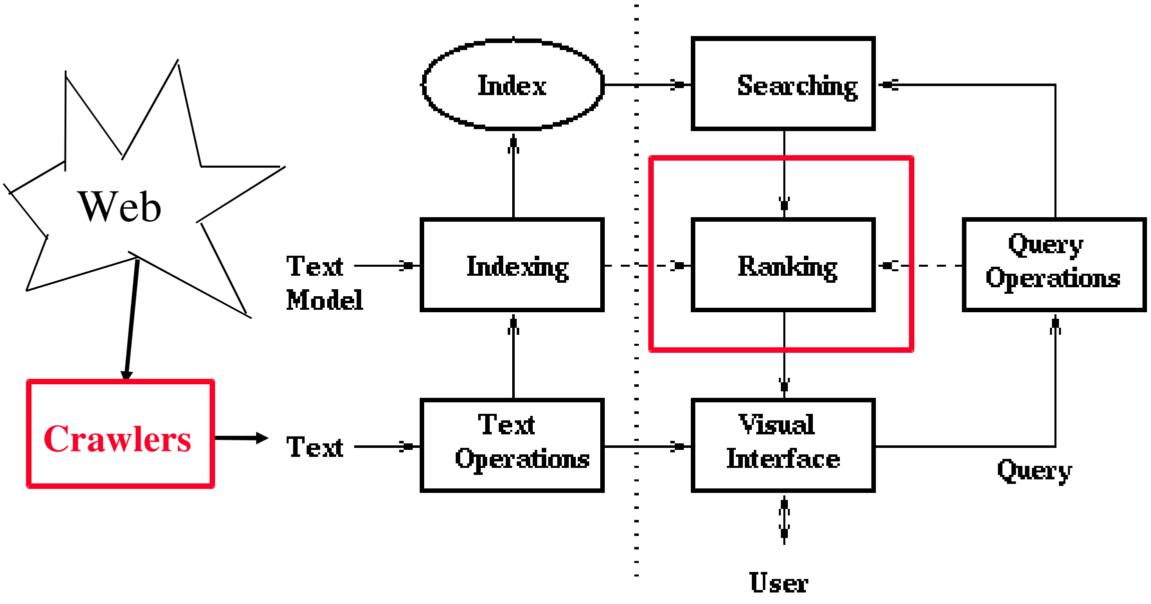
- Centralized Software Architecture
- Hypertext Structure
  - Allows to include link ranking
- On-line Quality Evaluation
- Distributed Data
  - Crawling
- Locally Distributed Index
  - Parallel Indexing
  - Parallel Query Processing
- Advertising Business Model
  - Word based and pay-per-click

Web Retrieval

- Problems:
  - volume
  - fast rate of change and growth
  - dynamic content
  - redundancy
  - organization and data quality
  - diversity
  - …..
- Deal with data overload

Web Retrieval Architecture

- Centralized parallel architecture



Algorithmic Challenges

- Crawling:
  - Quantity
  - Freshness
  - Quality
  - Politeness vs. Usage of Resources

Conflict

Adversarial IR

- Ranking
  - Words, links, usage logs, … , metadata
  - Spamming of all kinds of data
  - Good precision, unknown recall

Link Ranking

- Incoming links count & variations
  (Li /Marchiori / Carriere *et al.* 1997; Joo & Myaeng, 1998)
- HITS (Kleinberg, 1998)
  - Authorities: good pages - Hubs: good links
- PageRank (Page & Brin, 1998)
  - Random walk + random teleportation if "bored"
- Many variations of these ideas
- Good to find communities, spam, etc.
- Application to other problems
- Today: just a component of a search engine ranking

Fight Spam

- Adversarial Web Retrieval
- Text Spam (e.g. Cloaking)
- Link Spam (e.g. Link Farms)
- Metadata spam
- Ad spam (e.g. Clicks, Bids)

The Big Challenge

Meet the diverse user needs
given
their poorly made queries
and
the size and heterogeneity of the Web corpus

Web Mining

- Content: text & multimedia mining
- Structure: link analysis, graph mining
- Usage: log analysis, query mining
- Relate all of the above
  - Web characterization
  - Particular applications

Dynamic

Motivations for Web Mining

- The Dream of the Semantic Web
  - Hypothesis: Explicit Semantic Information
  - Obstacle: Us
- User Actions: Implicit Semantic Information
  - It's free!
  - Large volume!
  - It's unbiased!
  - Can we capture it?
  - Hypothesis: Queries are the best source

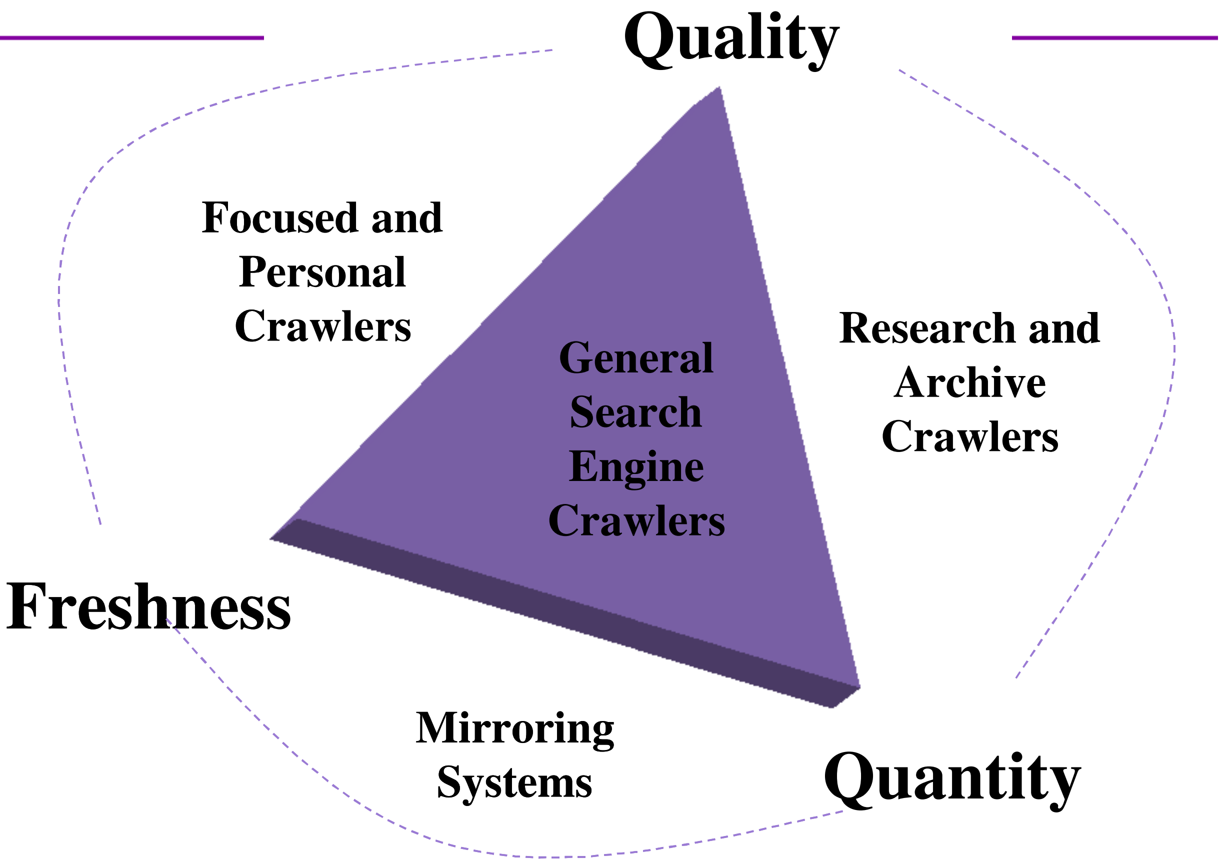
Data Recollection

- Content and structure: Crawling
- Usage: Logs
  - Web Server logs
  - Specific Application logs

Crawling

- NP-Hard Scheduling Problem
- Different goals
- Many Restrictions
- Difficult to define optimality
- No standard benchmark

Crawling Goals



Quality

Focused and Personal Crawlers

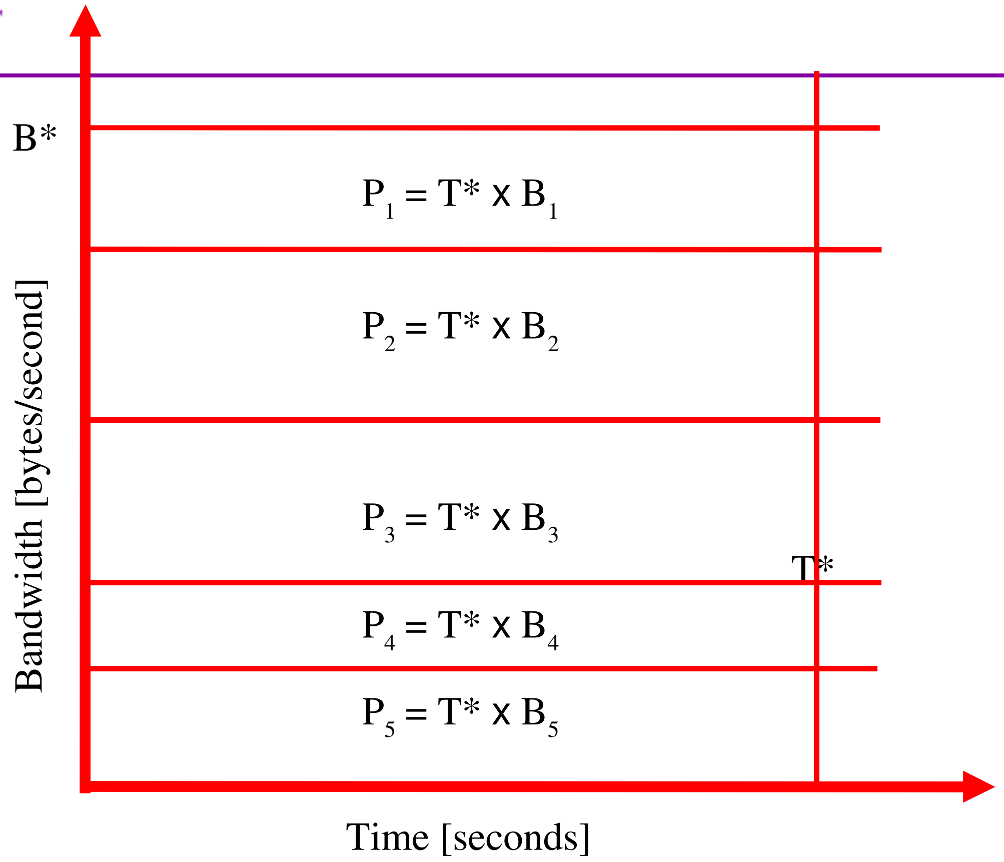
General Search Engine Crawlers

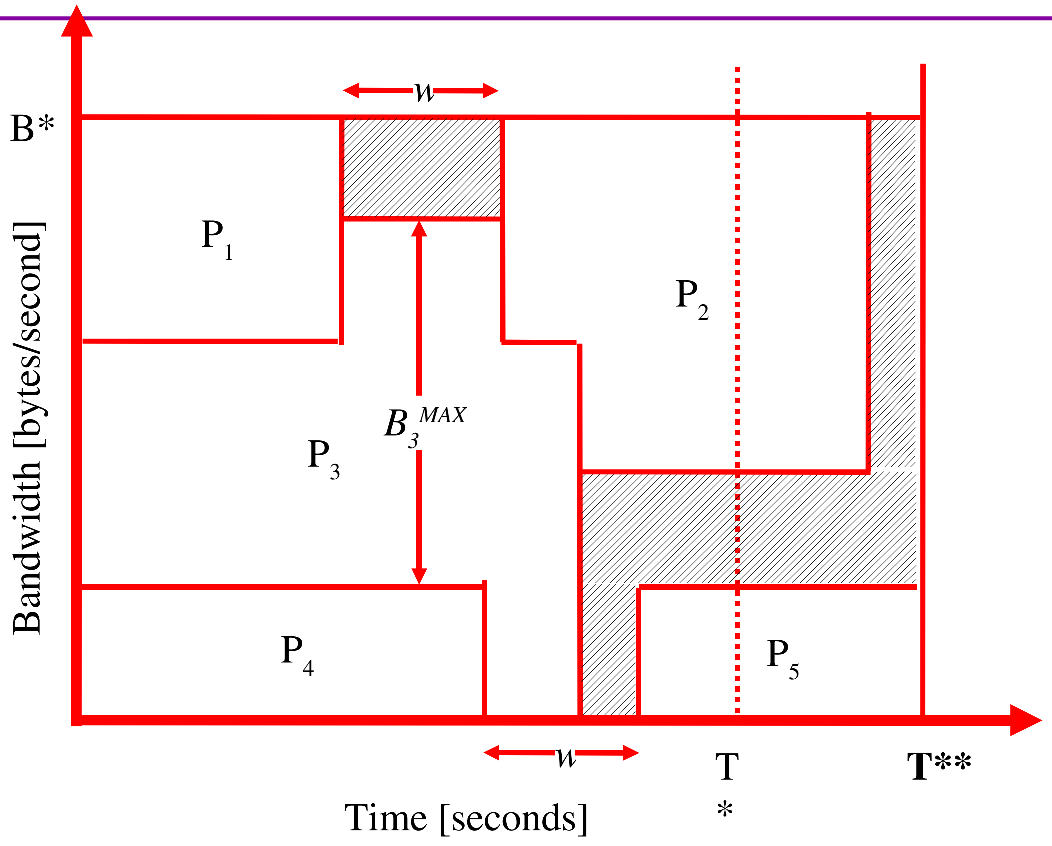
Research and Archive Crawlers

Freshness

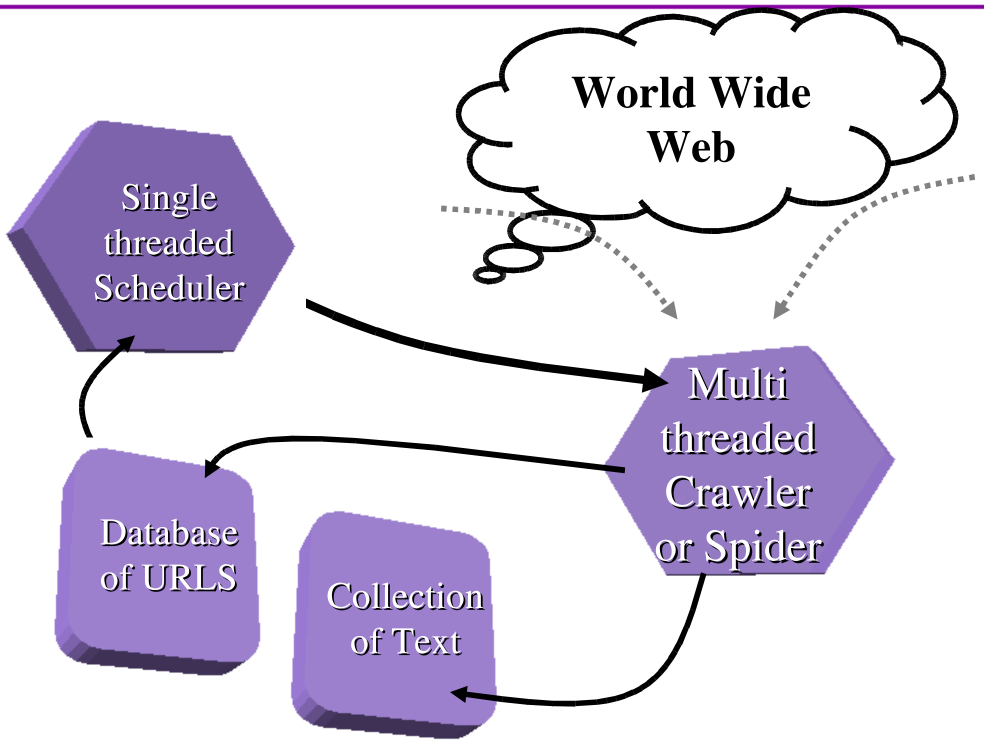
Mirroring Systems

Quantity


B*
$P_1 = T^* \times B_1$
$P_2 = T^* \times B_2$
$P_3 = T^* \times B_3$
T*
$P_4 = T^* \times B_4$
$P_5 = T^* \times B_5$
Bandwidth [bytes/second]
Time [seconds]


w
B*
P_1
P_2
B_3^{MAX}
P_3
P_4
P_5
w
T*
T**
Bandwidth [bytes/second]
Time [seconds]

Software Architecture



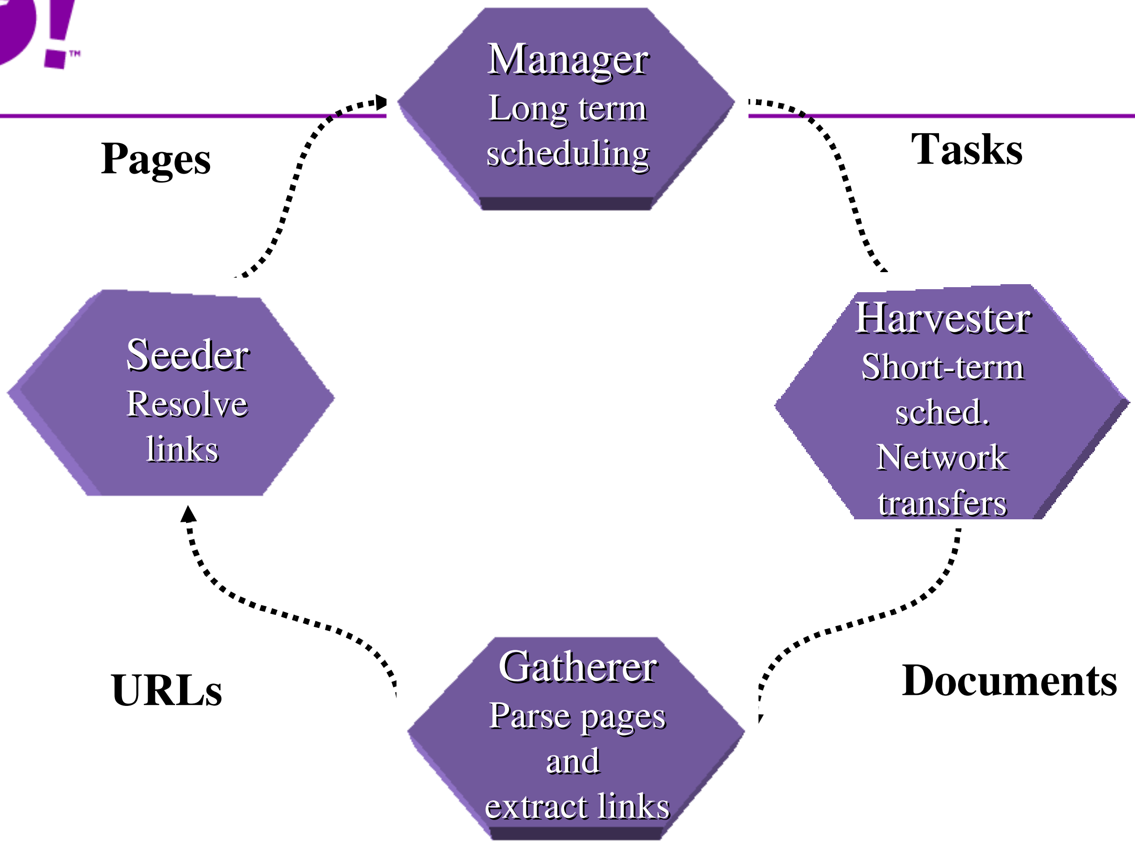
World Wide Web

Single threaded Scheduler

Multi threaded Crawler or Spider

Database of URLS

Collection of Text



Manager
Long term scheduling

Tasks

Harvester
Short-term sched.
Network transfers

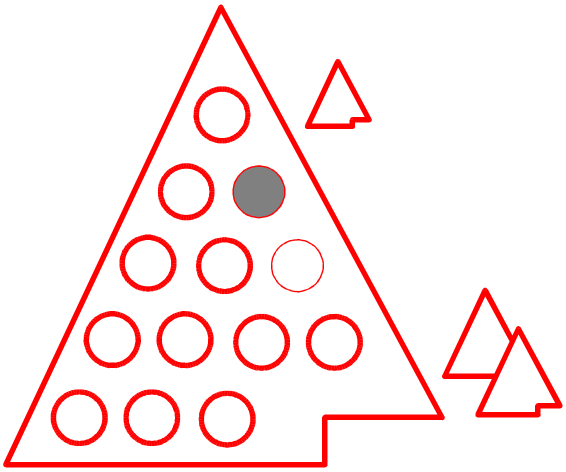
Documents

Gatherer
Parse pages and extract links

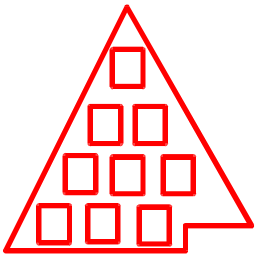
URLs

Seeder
Resolve links

Pages



Queue of Web sites
(long-term scheduling)



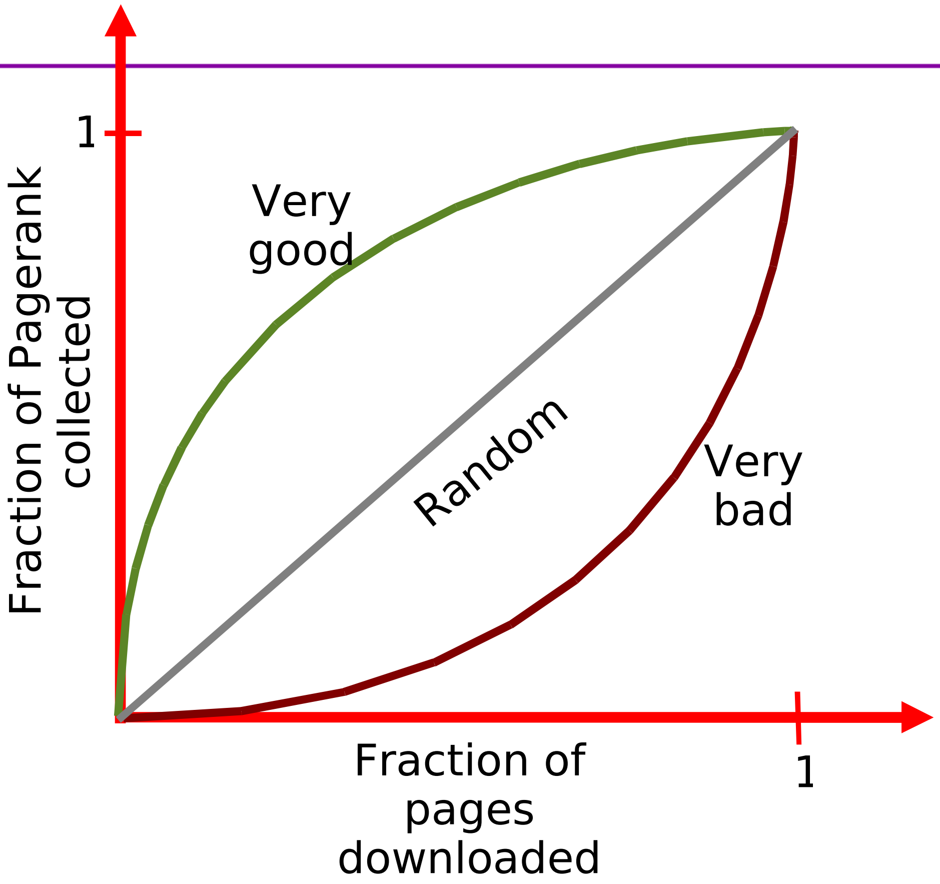
Queue of Web pages
for each site
(short-term scheduling)

Formal Problem

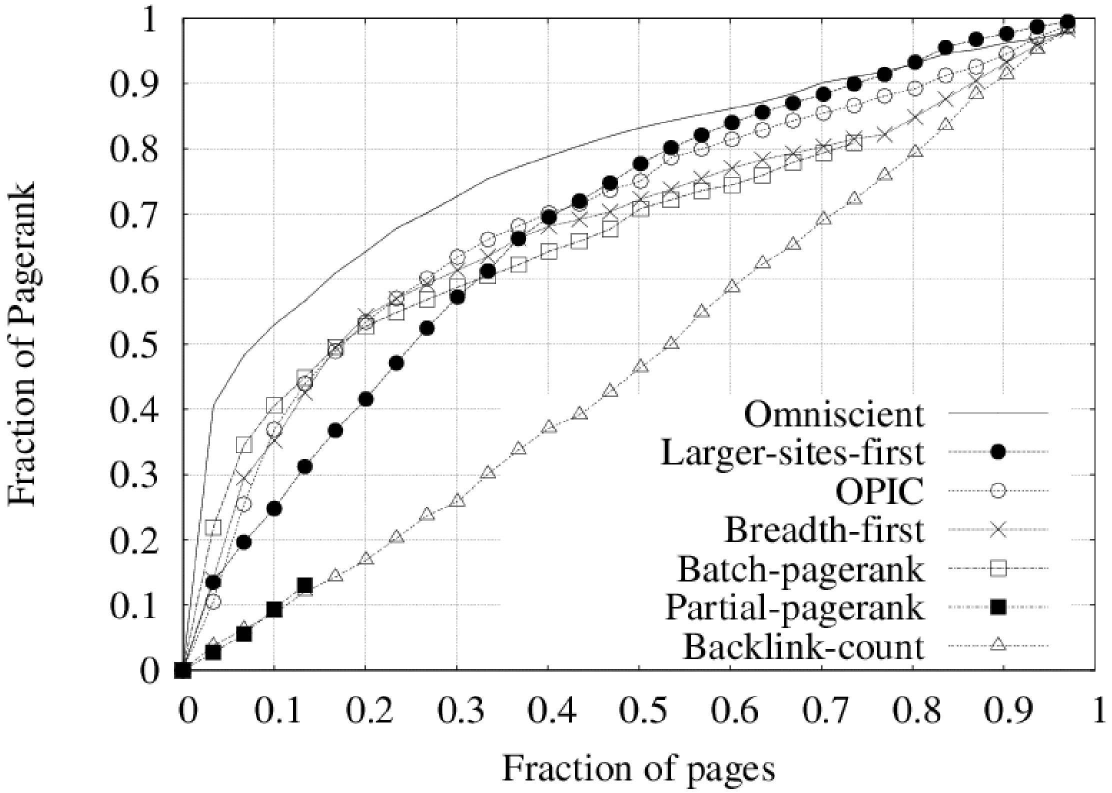
- Find a sequence of page requests *(p,t)* that:
  - Optimizes a function of the volume, quality and freshness of the pages
  - Has a bounded crawling time
  - Fulfils politeness
  - Maximizes the use of local bandwidth
- Must be on-line: how much knowledge?

Crawling Heuristics

- Breadth-first
- Ranking-ordering
  - PageRank
- Largest Site-first
- Use of:
  - Partial information
  - Historical information
- No Benchmark for Evaluation

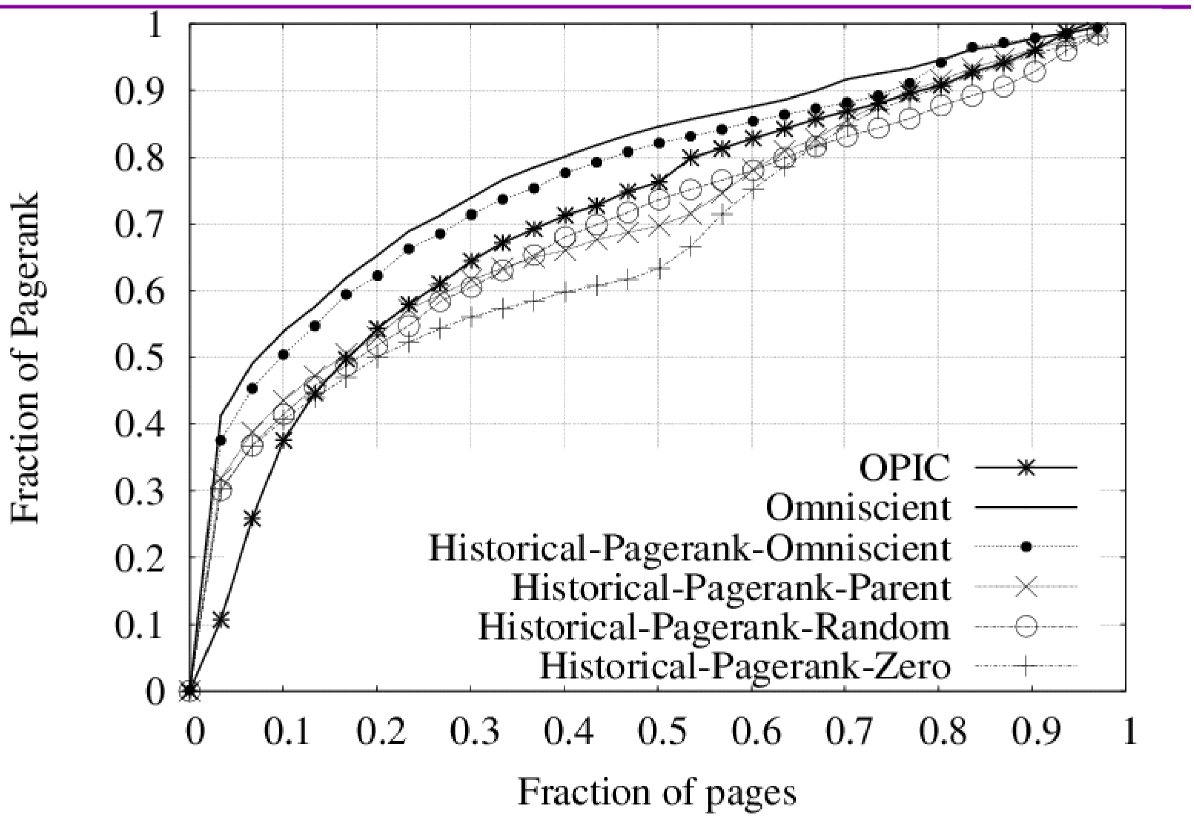


No Historical Information

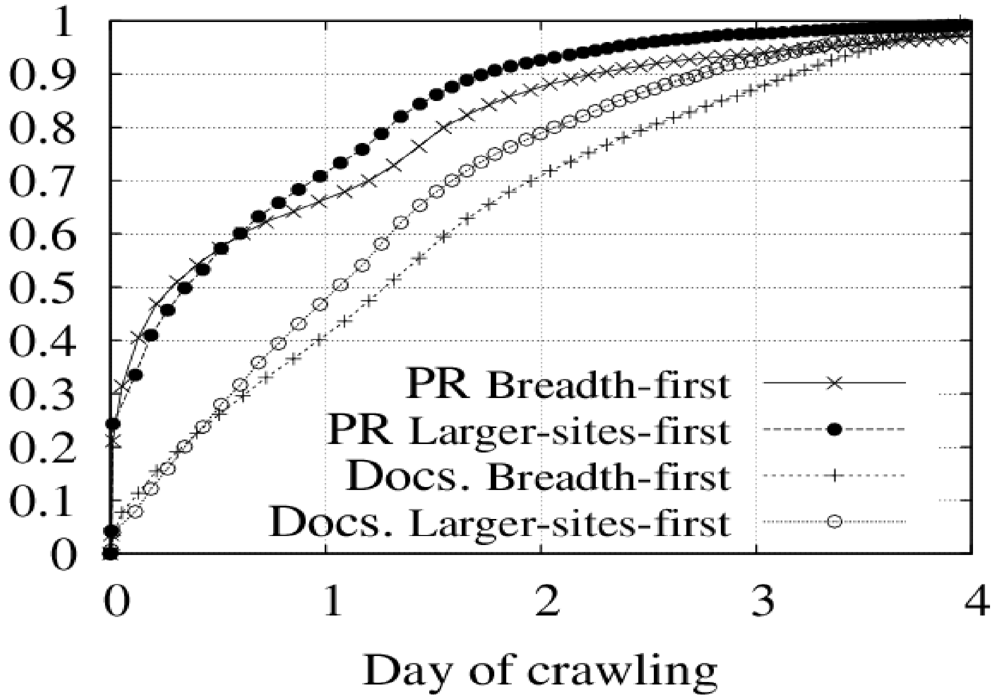


Baeza-Yates, Castillo, Marin & Rodriguez, WWW2005

Historical Information



Validation in the Greek domain



Data Cleaning

- Problem Dependent
- Content: Duplicate and spam detection
- Links: Spam detection
- Logs: Spam detection
  - Robots vs. persons

Data Processing

- Structure: content, links and logs
  - XML, relational database, etc.
- Usage mining:
  - Anonymize if needed
  - Define sessions

Data Characteristics

- Yahoo! as a Case Study
  - Data Volume
  - Data Types

Yahoo! World

- Search
  - Yahoo! Image,
  - Yahoo! Video,
  - Yahoo! Local,
  - Yahoo! News,
  - Yahoo! Shopping Search,
- Communication
  - Yahoo! Mail,
  - Yahoo! Messenger,
  - My Web,
  - Yahoo! Personals,
  - Yahoo! 360º,
  - Yahoo! Photos,
  - Flickr, Delicious,
  - Yahoo! Answers
- Content:
  - Yahoo! Sports,
  - Yahoo! Finance,
  - Yahoo! Music,
  - Yahoo! Movies,
  - Yahoo! News,
  - Yahoo! Games.
  - My Yahoo!
- Mobile:
  - Yahoo! Mobile
- Commerce:
  - Yahoo! Shopping,
  - Yahoo! Autos,
  - Yahoo! Auctions,
  - Yahoo! Travel,
- Small Business:
  - Yahoo! Small Business
  - Yahoo! Domains,
  - Yahoo! Web Hosting,
  - Yahoo! Merchant Solutions,
  - Yahoo! Business Email,
  - HotJobs
- Advertising:
  - Yahoo! Search Marketing
  - Yahoo! Publisher Network.

Yahoo! Numbers

(April '06, Oct'06)

24 languages, 20 countries

- > 4 billion page views per day (largest in the world)
- > 500 million unique users each month (half the Internet users!)
- > 250 million mail users (1 million new accounts a day)
- 95 million groups members
- 7 million moderators
- 4 billion music videos streamed in 2005

- 20 Pb of storage (20M Gb)
  - US Library of congress every day (28M books, 20TB)
- 12 Tb of data processed per day
- 7 billion song ratings
- 2 billion photos stored
- 2 billion Mail+Messenger sent per day

Crawled Data

- WWW
  - Web Pages & Links
  - Blogs
  - Dynamic Sites

heterogeneous,
large,
dangerous

- Sales Providers (Push)
  - Advertising
  - Items for sale: Shopping, Travel, etc.

very high quality
& structure,
expensive,
sparse,
safe

- News Index
  - RSS Feeds
  - Contracted information

high quality,
sparse,
redundant

Produced data

- Yahoo's Web
  - Ygroups
  - YCars, YHealth, Ytravel

homogeneous,
high quality,
safer,
highly structured

- Produced Content
  - Edited (news)
  - Purchased (news)

Trusted,
high quality,
sparse

- Direct Interaction:
  - Tagged Content
    - Object tagging (photos, pages, ?)
    - Social links
  - Question Answering

Ambiguous
semantics?
trust?
quality?

"Information Games"
(e..g. www.espgame.org)

Observed Data

- Query Logs
  - spelling, synonyms, phrases (named entities), substitutions

good quality, sparse, power law

- Click-Thru
  - relevance, intent, wording

good quality, sparse, mostly safe

- Advertising
  - relevance, value, terminology

Trusted, high quality, homogeneous, structured

- Social
  - links, communities, dialogues...

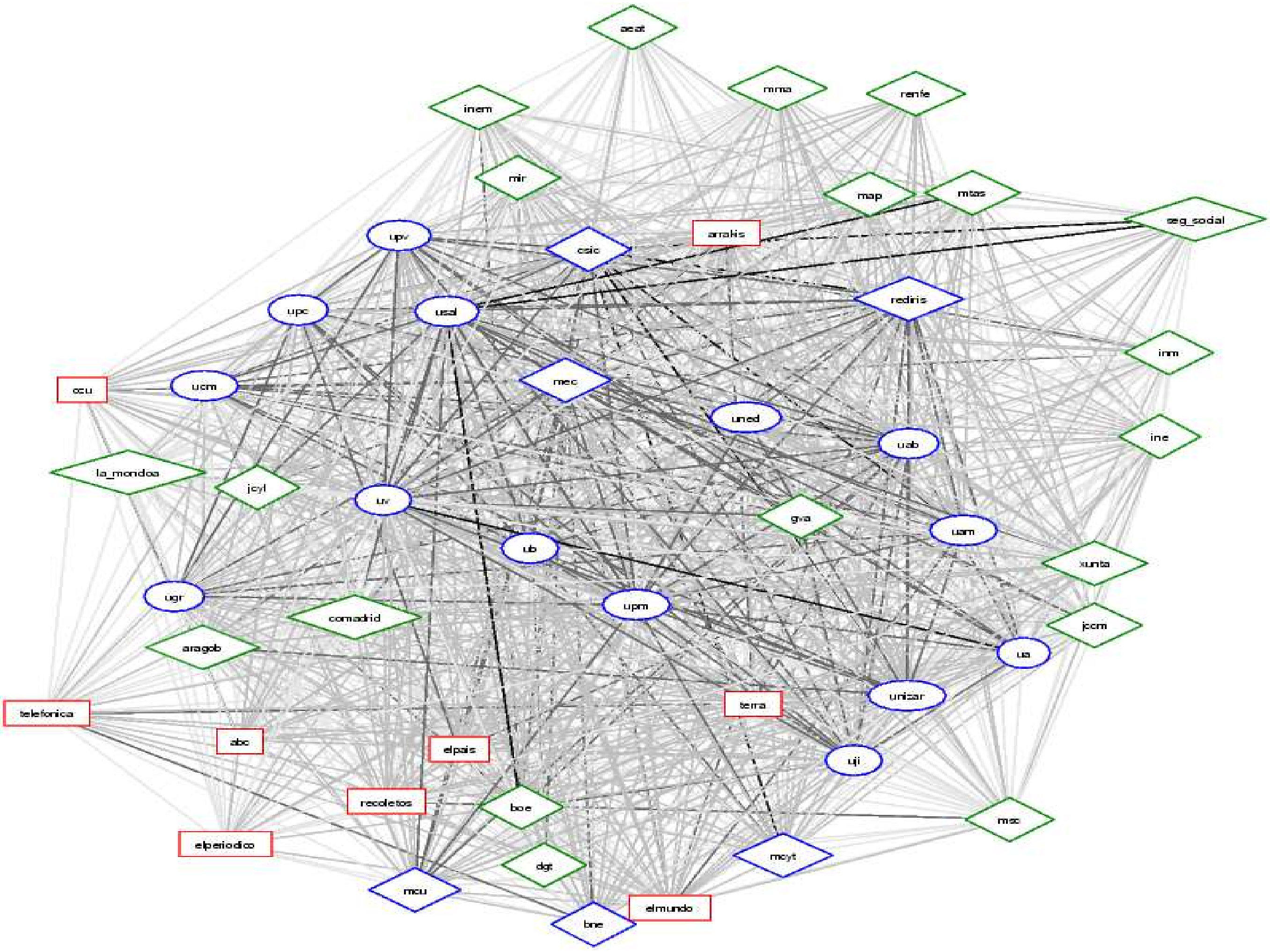
trust? quality?

Web Characterization

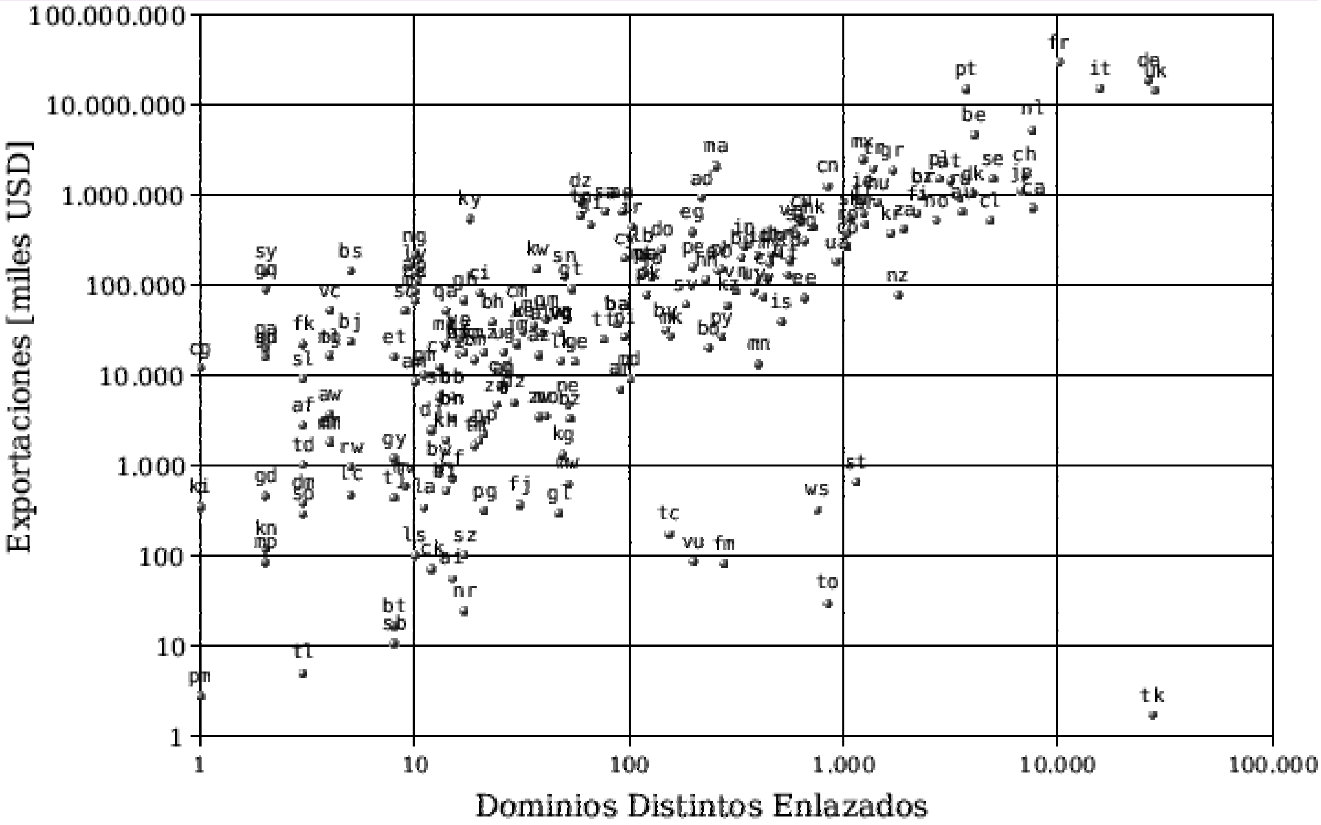
- Different scopes: global, country, etc.
- Different levels: pages, sites, domains
- Different content: text, images, etc.
- Different technologies: software, OS, etc.

A Few Examples

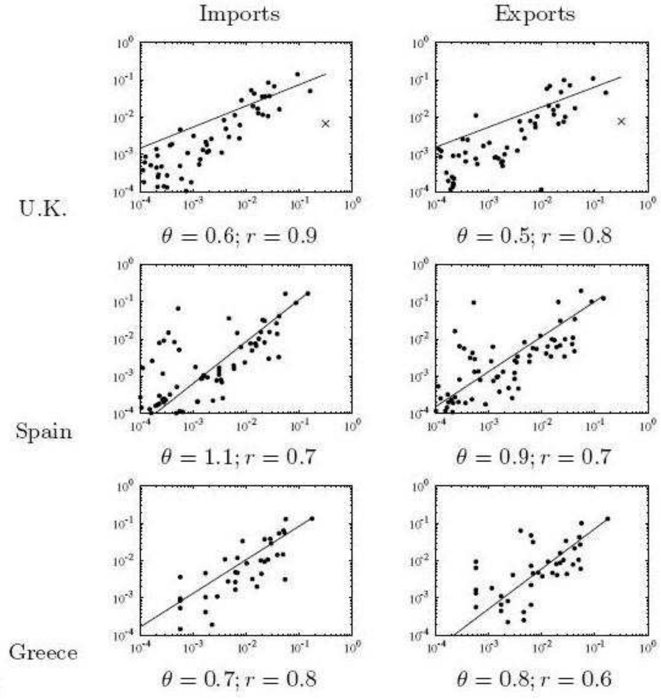
- Web Characterization of Spain
- Link Analysis
- Log Analysis
- Web Dynamics



Mirror of the Society



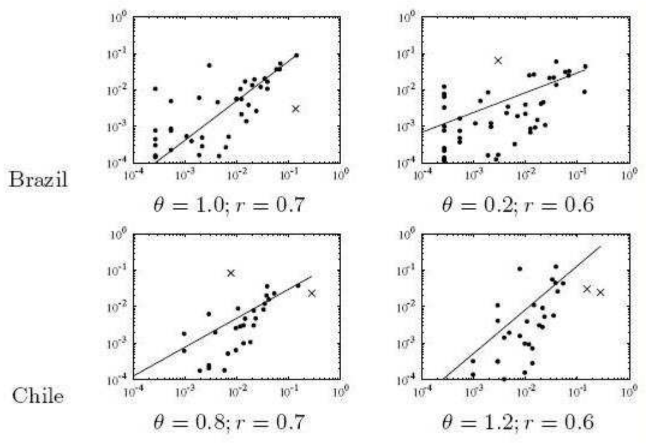
Exports/Imports vs. Domain Links



$\theta = 0.6; r = 0.9$ $\theta = 0.5; r = 0.8$

$\theta = 1.1; r = 0.7$ $\theta = 0.9; r = 0.7$

$\theta = 0.7; r = 0.8$ $\theta = 0.8; r = 0.6$

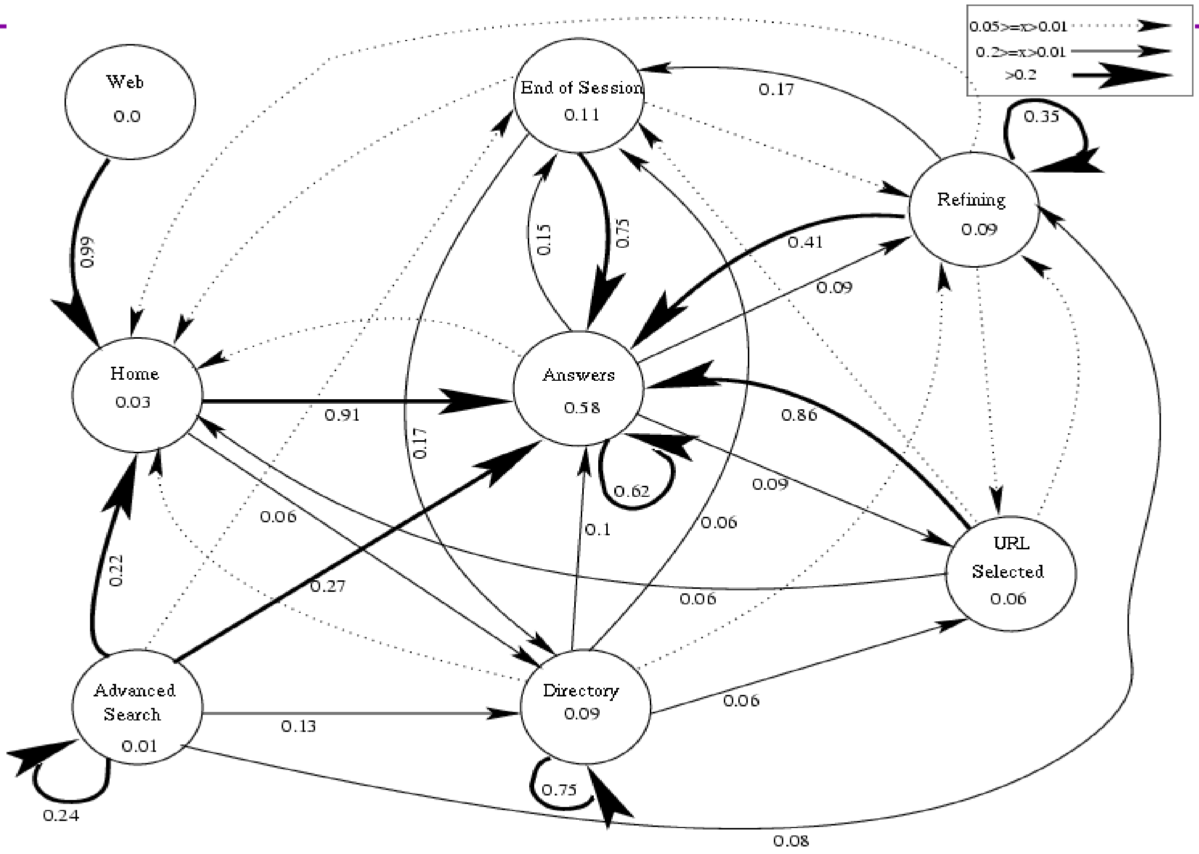


$\theta = 1.0; r = 0.7$ $\theta = 0.2; r = 0.6$

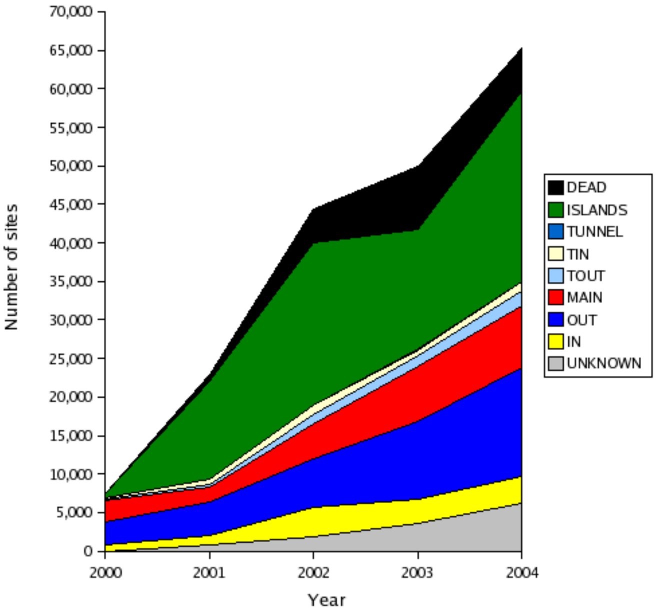
$\theta = 0.8; r = 0.7$ $\theta = 1.2; r = 0.6$

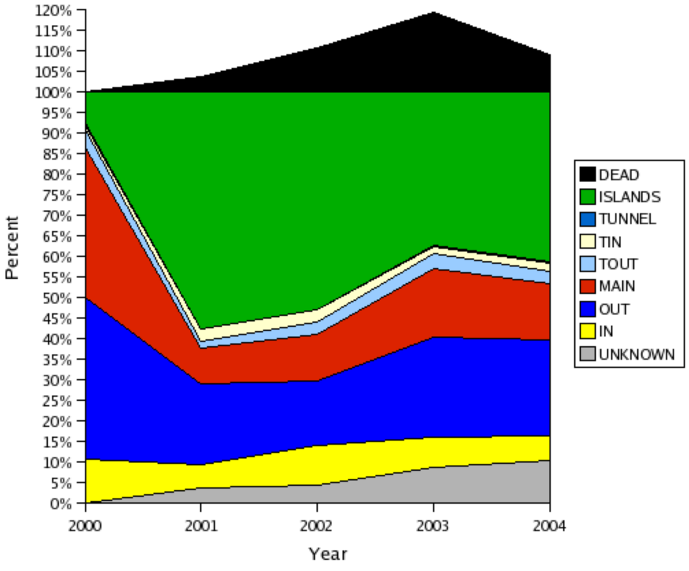
Baeza-Yates & Castillo, WWW2006

User Modeling

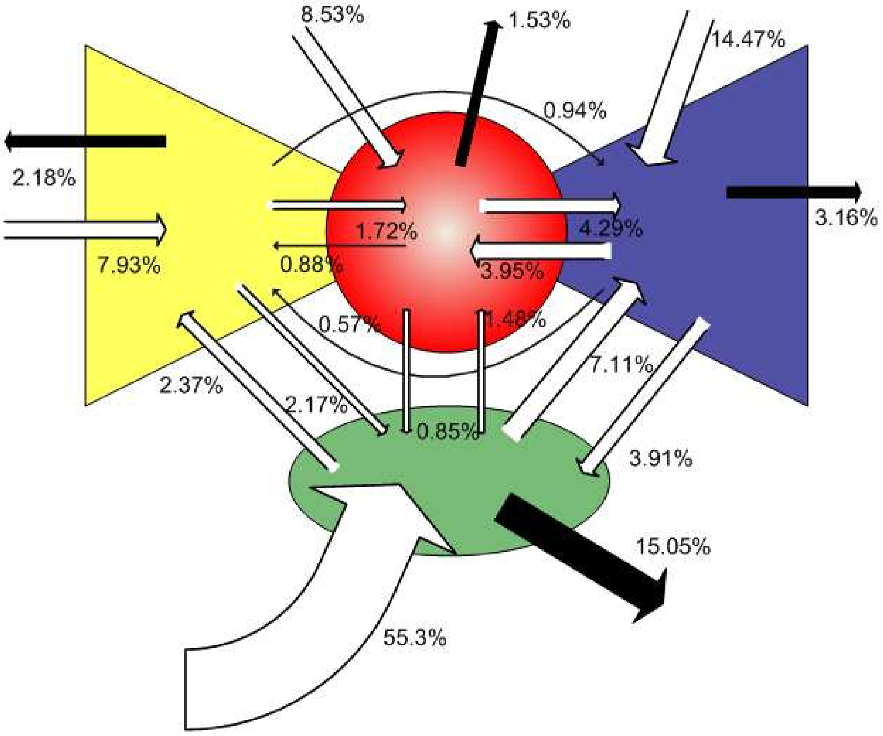


Size Evolution

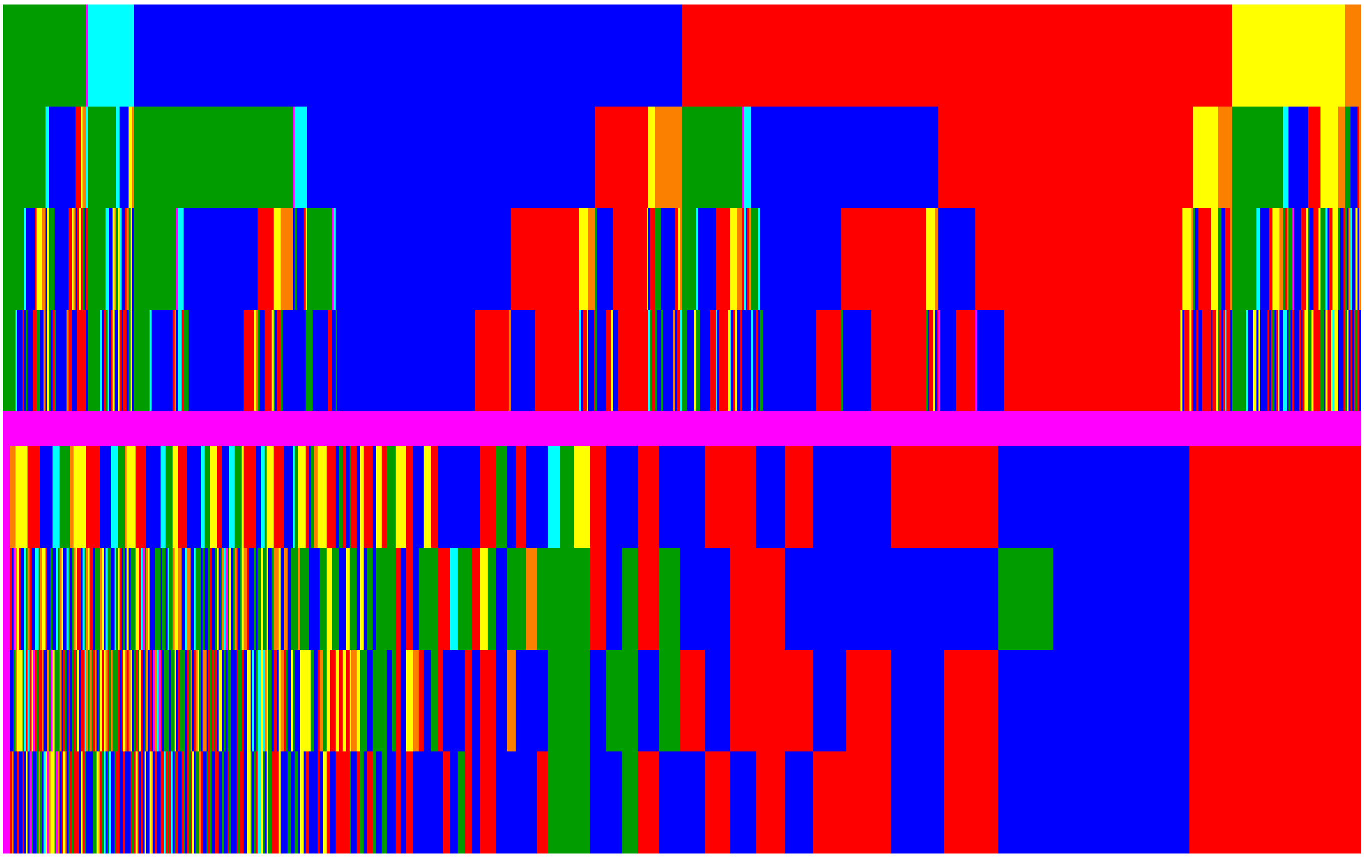


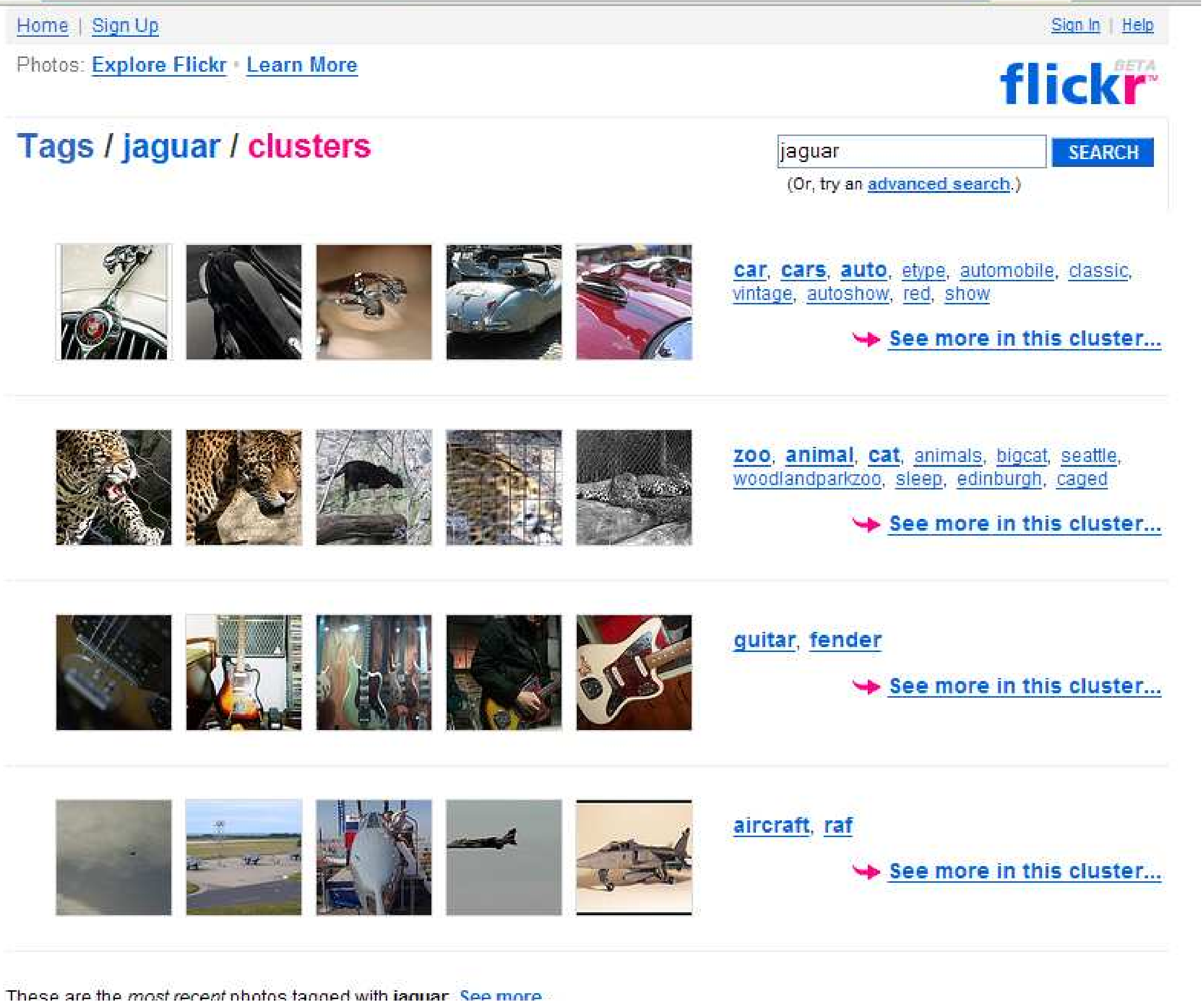


Structure Macro Dynamics



Structure Micro Dynamics





Home | Sign Up

Sign In | Help

Photos: Explore Flickr · Learn More

Tags / jaguar / clusters

jaguar SEARCH

(Or, try an advanced search.)

car, cars, auto, etype, automobile, classic, vintage, autoshow, red, show

See more in this cluster...

zoo, animal, cat, animals, bigcat, seattle, woodlandparkzoo, sleep, edinburgh, caged

See more in this cluster...

guitar, fender

See more in this cluster...

aircraft, raf

See more in this cluster...

These are the *most recent* photos tagged with **jaguar**. See more

The Power of Social Media

- Flickr – community phenomenon
- Millions of users share and tag each others' photographs (why???)
- The *wisdom of the crowds* can be used to search
- The principle is not new – anchor text used in "standard" search
- What about to generate pseudo-semantic resources?

The Wisdom of Crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
- Bottom line:

> *“large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future”.*

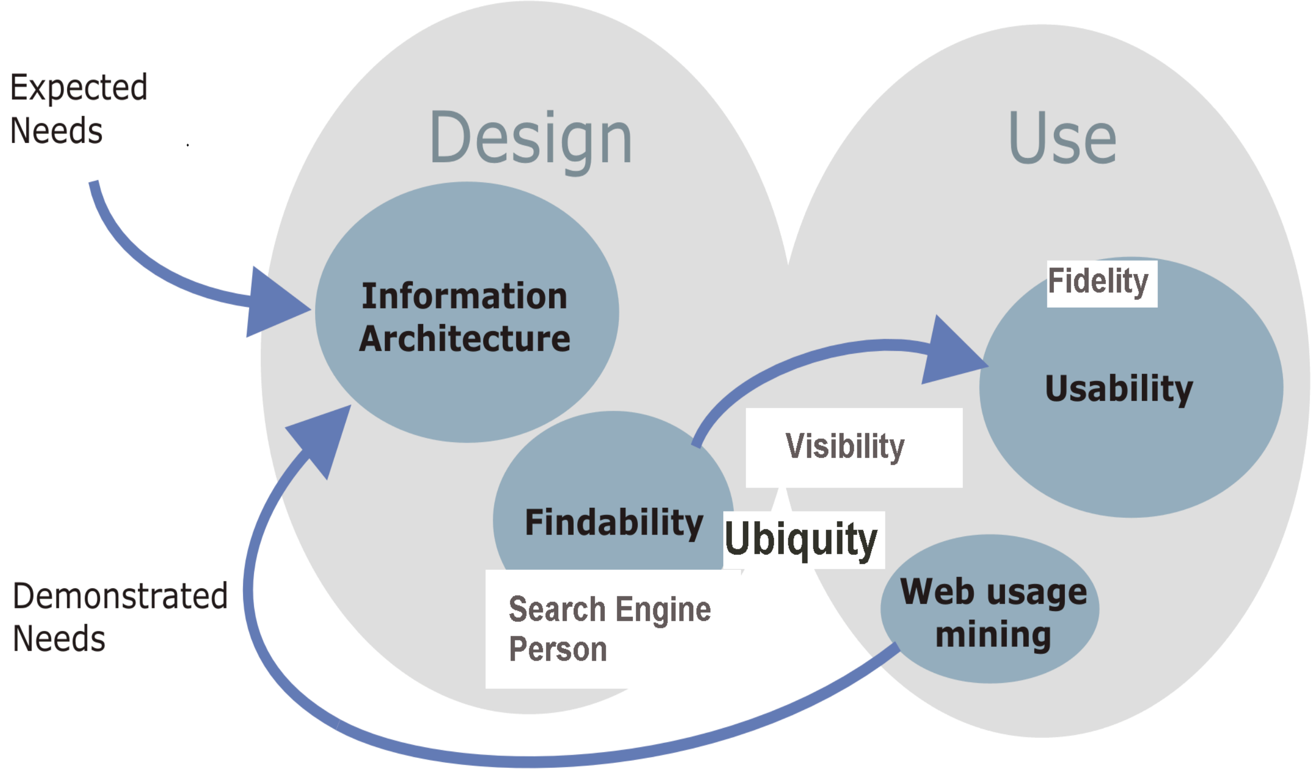
The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - –not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - –Queries and actions (or no action!)

Mining Queries for ...

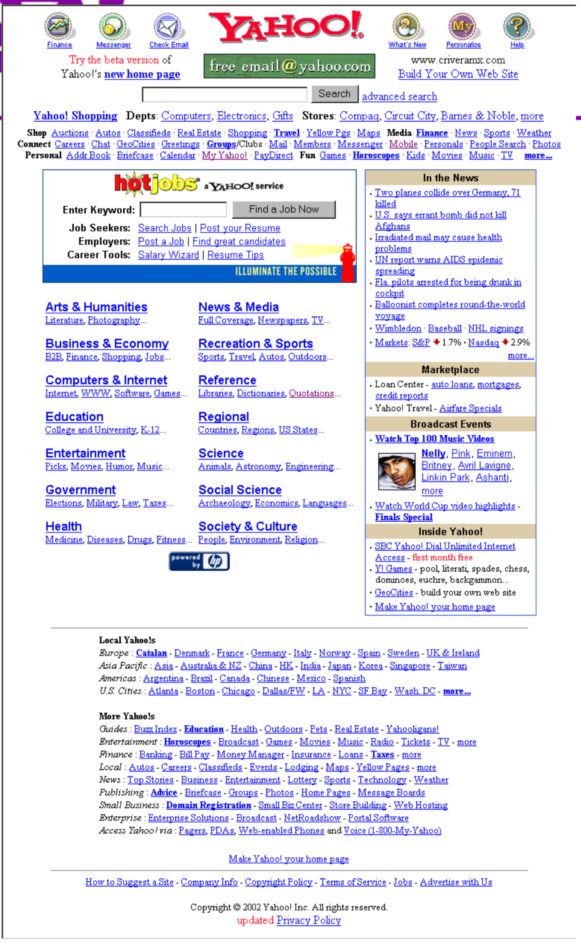
- Improved Web Search: index layout, ranking
- User Driven Design
  - Information Scent
  - The Web Site that the Users Want
  - The Web Site that You should Have
  - Improve content & structure
- Bootstrap of pseudo-semantic resources

Web Design



User Driven Design

- *User-driven design*
  - Best example: Yahoo!
- Navigational log analysis
  - Site reorganization
- Query log analysis
  - Information Scent
  - Content that is missing: market niches



Try the beta version of Yahoo!'s **new home page**

free_email@yahoo.com

www.criveramx.com
Build Your Own Web Site

Search advanced search

Yahoo! Shopping **Depts**: Computers, Electronics, Gifts **Stores**: Compaq, Circuit City, Barnes & Noble, more

Shop Auctions · Autos · Classifieds · Real Estate · Shopping · **Travel** · Yellow Pgs · Maps **Media** **Finance** · News · Sports · Weather
Connect Careers · Chat · GeoCities · Greetings · **Groups**/Clubs · Mail · Members · Messenger · Mobile · Personals · People Search · Photos
Personal Addr Book · Briefcase · Calendar · My Yahoo! · PayDirect **Fun** Games · **Horoscopes** · Kids · Movies · Music · TV **more...**

hotjobs a YAHOO! service
Enter Keyword: Find a Job Now
Job Seekers: Search Jobs | Post your Resume
Employers: Post a Job | Find great candidates
Career Tools: Salary Wizard | Resume Tips
ILLUMINATE THE POSSIBLE

Arts & Humanities
Literature, Photography...

News & Media
Full Coverage, Newspapers, TV...

Business & Economy
B2B, Finance, Shopping, Jobs...

Recreation & Sports
Sports, Travel, Autos, Outdoors...

Computers & Internet
Internet, WWW, Software, Games...

Reference
Libraries, Dictionaries, Quotations...

Education
College and University, K-12...

Regional
Countries, Regions, US States...

Entertainment
Picks, Movies, Humor, Music...

Science
Animals, Astronomy, Engineering...

Government
Elections, Military, Law, Taxes...

Social Science
Archaeology, Economics, Languages...

Health
Medicine, Diseases, Drugs, Fitness...

Society & Culture
People, Environment, Religion...

powered by hp

In the News
- Two planes collide over Germany, 71 killed
- U.S. says errant bomb did not kill Afghans
- Irradiated mail may cause health problems
- UN report warns AIDS epidemic spreading
- Fla. pilots arrested for being drunk in cockpit
- Balloonist completes round-the-world voyage
- Wimbledon · Baseball · NHL signings
- Markets: S&P ↓ 1.7% · Nasdaq ↓ 2.9%

more...

Marketplace
- Loan Center - auto loans, mortgages, credit reports
- Yahoo! Travel - Airfare Specials

Broadcast Events
- **Watch Top 100 Music Videos**
  Nelly, Pink, Eminem, Britney, Avril Lavigne, Linkin Park, Ashanti, more
- Watch World Cup video highlights - **Finals Special**

Inside Yahoo!
- SBC Yahoo! Dial Unlimited Internet Access - first month free
- Y! Games - pool, literati, spades, chess, dominoes, euchre, backgammon...
- GeoCities - build your own web site
- Make Yahoo! your home page

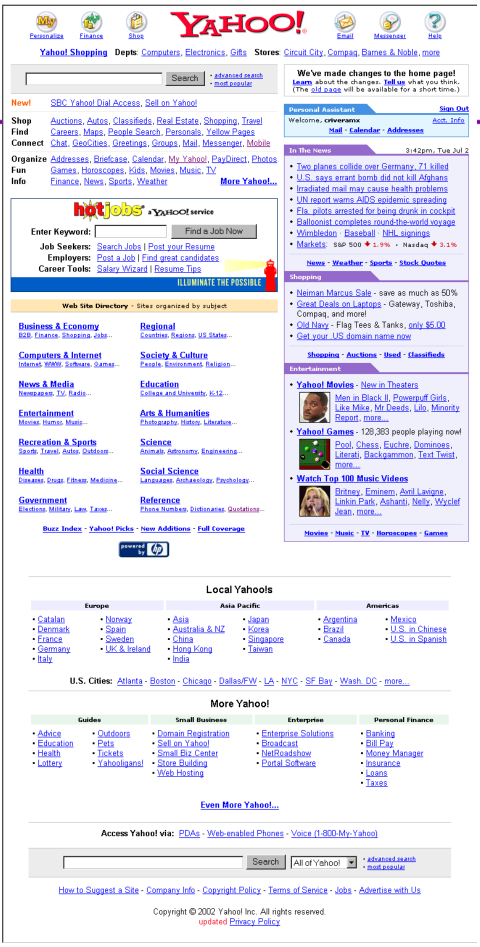
Local Yahoo!s
Europe : **Catalan** - Denmark - France - Germany - Italy - Norway - Spain - Sweden - UK & Ireland
Asia Pacific : Asia - Australia & NZ - China - HK - India - Japan - Korea - Singapore - Taiwan
Americas : Argentina - Brazil - Canada - Chinese - Mexico - Spanish
U.S. Cities : Atlanta - Boston - Chicago - Dallas/FW - LA - NYC - SF Bay - Wash. DC - **more...**

More Yahoo!s
Guides : Buzz Index - **Education** - Health - Outdoors - Pets - Real Estate - Yahooligans!
Entertainment : **Horoscopes** - Broadcast - Games - Movies - Music - Radio - Tickets - TV - more
Finance : Banking - Bill Pay - Money Manager - Insurance - Loans - **Taxes** - more
Local : Autos - Careers - Classifieds - Events - Lodging - Maps - Yellow Pages - more
News : Top Stories - Business - Entertainment - Lottery - Sports - Technology - Weather
Publishing : **Advice** - Briefcase - Groups - Photos - Home Pages - Message Boards
Small Business : **Domain Registration** - Small Biz Center - Store Building - Web Hosting
Enterprise : Enterprise Solutions - Broadcast - NetRoadshow - Portal Software
Access Yahoo! via : Pagers, PDAs, Web-enabled Phones and Voice (1-800-My-Yahoo)

Make Yahoo! your home page

How to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Jobs - Advertise with Us

Copyright © 2002 Yahoo! Inc. All rights reserved.
updated Privacy Policy



Yahoo! Shopping **Depts**: Computers, Electronics, Gifts **Stores**: Circuit City, Compaq, Barnes & Noble, more

Search · advanced search · most popular

We've made changes to the home page!
Learn about the changes. **Tell us** what you think.
(The old page will be available for a short time.)

New! SBC Yahoo! Dial Access, Sell on Yahoo!

Shop Auctions, Autos, Classifieds, Real Estate, Shopping, Travel
Find Careers, Maps, People Search, Personals, Yellow Pages
Connect Chat, GeoCities, Greetings, Groups, Mail, Messenger, Mobile
Organize Addresses, Briefcase, Calendar, My Yahoo!, PayDirect, Photos
Fun Games, Horoscopes, Kids, Movies, Music, TV
Info Finance, News, Sports, Weather **More Yahoo!...**

hotjobs a YAHOO! service
Enter Keyword: Find a Job Now
Job Seekers: Search Jobs | Post your Resume
Employers: Post a Job | Find great candidates
Career Tools: Salary Wizard | Resume Tips
ILLUMINATE THE POSSIBLE

Web Site Directory - Sites organized by subject

Business & Economy
B2B, Finance, Shopping, Jobs...

Regional
Countries, Regions, US States...

Computers & Internet
Internet, WWW, Software, Games...

Society & Culture
People, Environment, Religion...

News & Media
Newspapers, TV, Radio...

Education
College and University, K-12...

Entertainment
Movies, Humor, Music...

Arts & Humanities
Photography, History, Literature...

Recreation & Sports
Sports, Travel, Autos, Outdoors...

Science
Animals, Astronomy, Engineering...

Health
Diseases, Drugs, Fitness, Medicine...

Social Science
Languages, Archaeology, Psychology...

Government
Elections, Military, Law, Taxes...

Reference
Phone Numbers, Dictionaries, Quotations...

Buzz Index - **Yahoo! Picks** - **New Additions** - **Full Coverage**

powered by hp

Personal Assistant **Sign Out**
Welcome, **criveramx** Acct. Info
Mail - **Calendar** - **Addresses**

In The News 3:42pm, Tue Jul 2
- Two planes collide over Germany, 71 killed
- U.S. says errant bomb did not kill Afghans
- Irradiated mail may cause health problems
- UN report warns AIDS epidemic spreading
- Fla. pilots arrested for being drunk in cockpit
- Balloonist completes round-the-world voyage
- Wimbledon · Baseball · NHL signings
- Markets: S&P 500 ↓ 1.9% · Nasdaq ↓ 3.1%

News - **Weather** - **Sports** - **Stock Quotes**

Shopping
- Neiman Marcus Sale - save as much as 50%
- Great Deals on Laptops - Gateway, Toshiba, Compaq, and more!
- Old Navy - Flag Tees & Tanks, only $5.00
- Get your .US domain name now

Shopping - **Auctions** - **Used** - **Classifieds**

Entertainment
- **Yahoo! Movies** - New in Theaters
  Men in Black II, Powerpuff Girls, Like Mike, Mr Deeds, Lilo, Minority Report, more...
- **Yahoo! Games** - 128,383 people playing now!
  Pool, Chess, Euchre, Dominoes, Literati, Backgammon, Text Twist, more...
- **Watch Top 100 Music Videos**
  Britney, Eminem, Avril Lavigne, Linkin Park, Ashanti, Nelly, Wyclef Jean, more...

Movies - **Music** - **TV** - **Horoscopes** - **Games**

Local Yahoo!s

Europe		Asia Pacific		Americas	
Catalan	Norway	Asia	Japan	Argentina	Mexico
Denmark	Spain	Australia & NZ	Korea	Brazil	U.S. in Chinese
France	Sweden	China	Singapore	Canada	U.S. in Spanish
Germany	UK & Ireland	Hong Kong	Taiwan		
Italy		India			

U.S. Cities: Atlanta - Boston - Chicago - Dallas/FW - LA - NYC - SF Bay - Wash. DC - more...

More Yahoo!

Guides		Small Business	Enterprise	Personal Finance
Advice	Outdoors	Domain Registration	Enterprise Solutions	Banking
Education	Pets	Sell on Yahoo!	Broadcast	Bill Pay
Health	Tickets	Small Biz Center	NetRoadshow	Money Manager
Lottery	Yahooligans!	Store Building	Portal Software	Insurance
		Web Hosting		Loans
				Taxes

Even More Yahoo!...

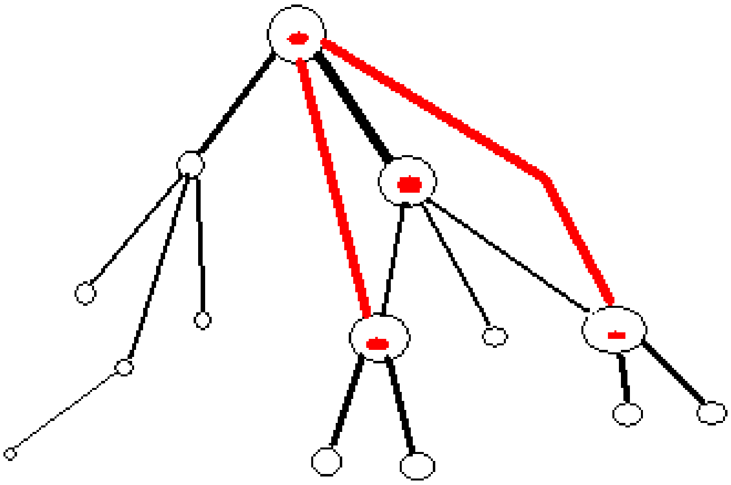
Access Yahoo! via: PDAs - Web-enabled Phones - Voice (1-800-My-Yahoo)

Search All of Yahoo! · advanced search · most popular

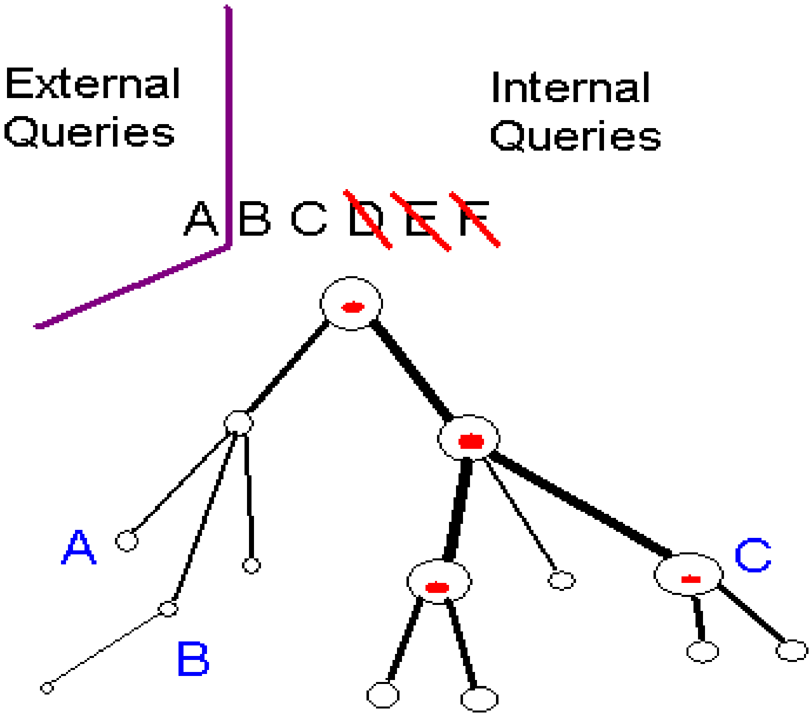
How to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Jobs - Advertise with Us

Copyright © 2002 Yahoo! Inc. All rights reserved.
updated Privacy Policy

Navigation Mining



Web Site Query Mining

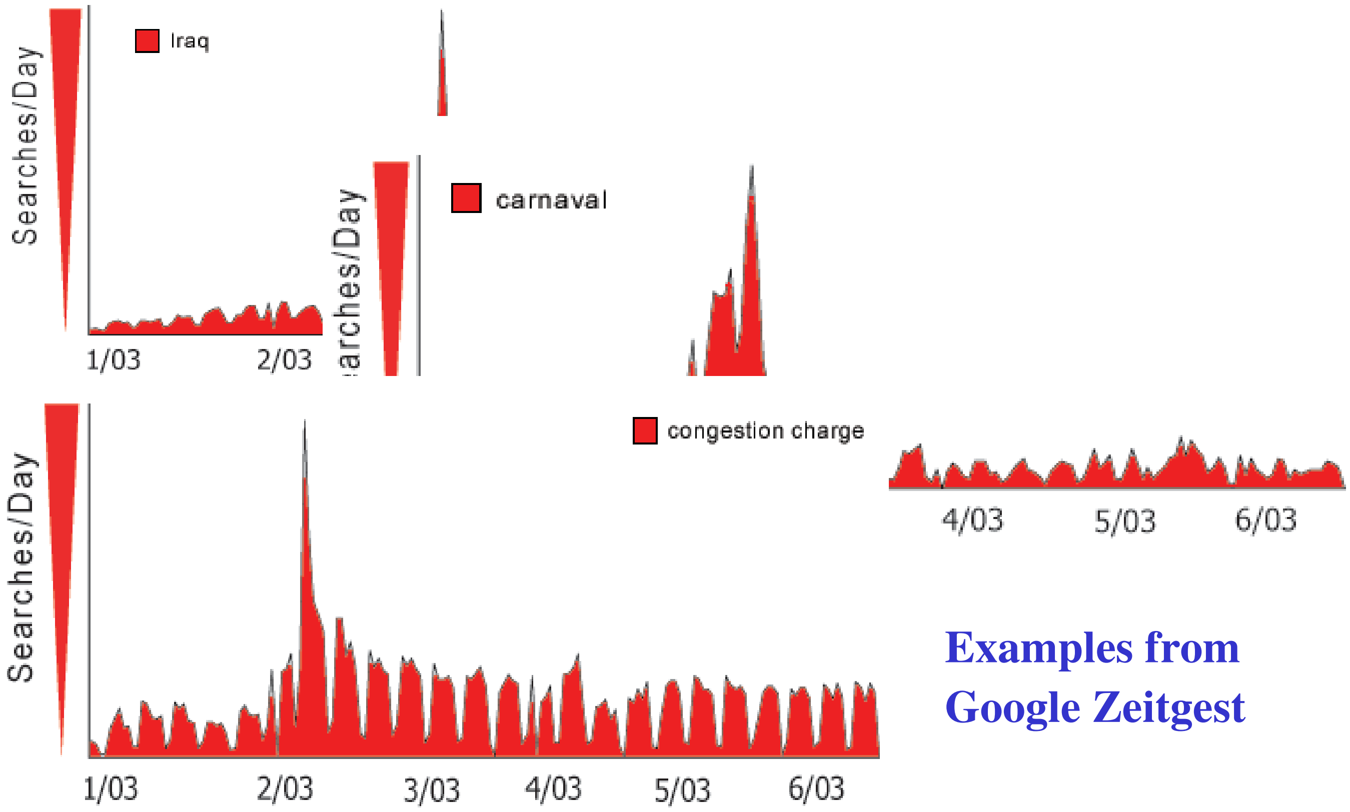


External Queries

Internal Queries

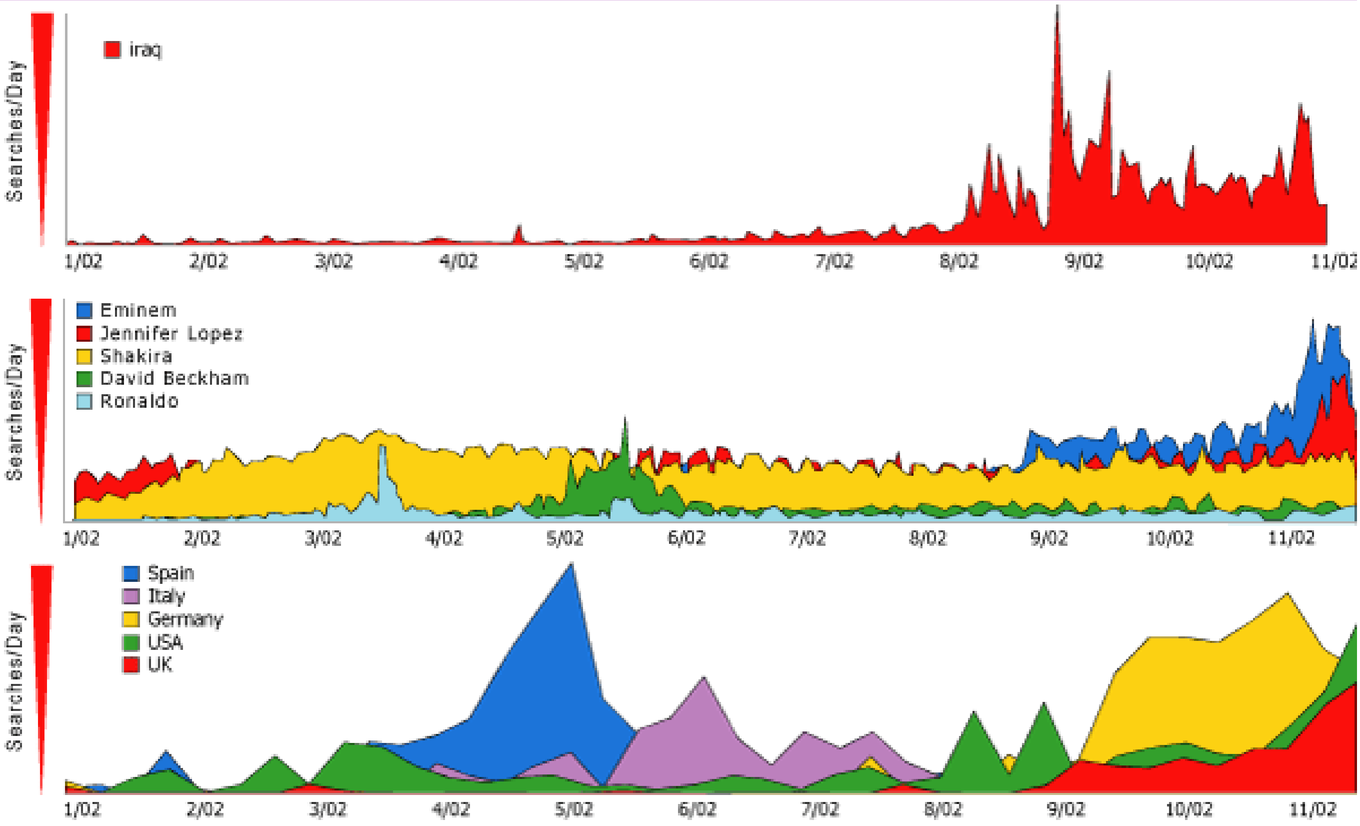
A B C ~~D~~ ~~E~~ ~~F~~

Social Mining (2003)



Examples from Google Zeitgest

Social Mining (2002)



Relevance of the Context

- There is no information without context
- Context and hence, content, will be implicit
- Balancing act: information vs. form
- Brown & Diguid: *The social life of information* (2000)
  - Current trend: less information, more context
- News highlights are similar to Web queries
  - E.g.: *Spell Unchecked (Indian Express, July 24, 2005)*

Context

- *Who you are*: age, gender, profession, etc.
- *Where you are and when*: time, location, speed and direction, etc.
- *What you are doing*: interaction history, task in hand, searching device, etc.

- *Issues*: privacy, intrusion, will to do it, etc.
- *Other sources*: Web, CV, usage logs, computing environment, ...
- *Goals*: personalization, localization, better ranking in general, etc.

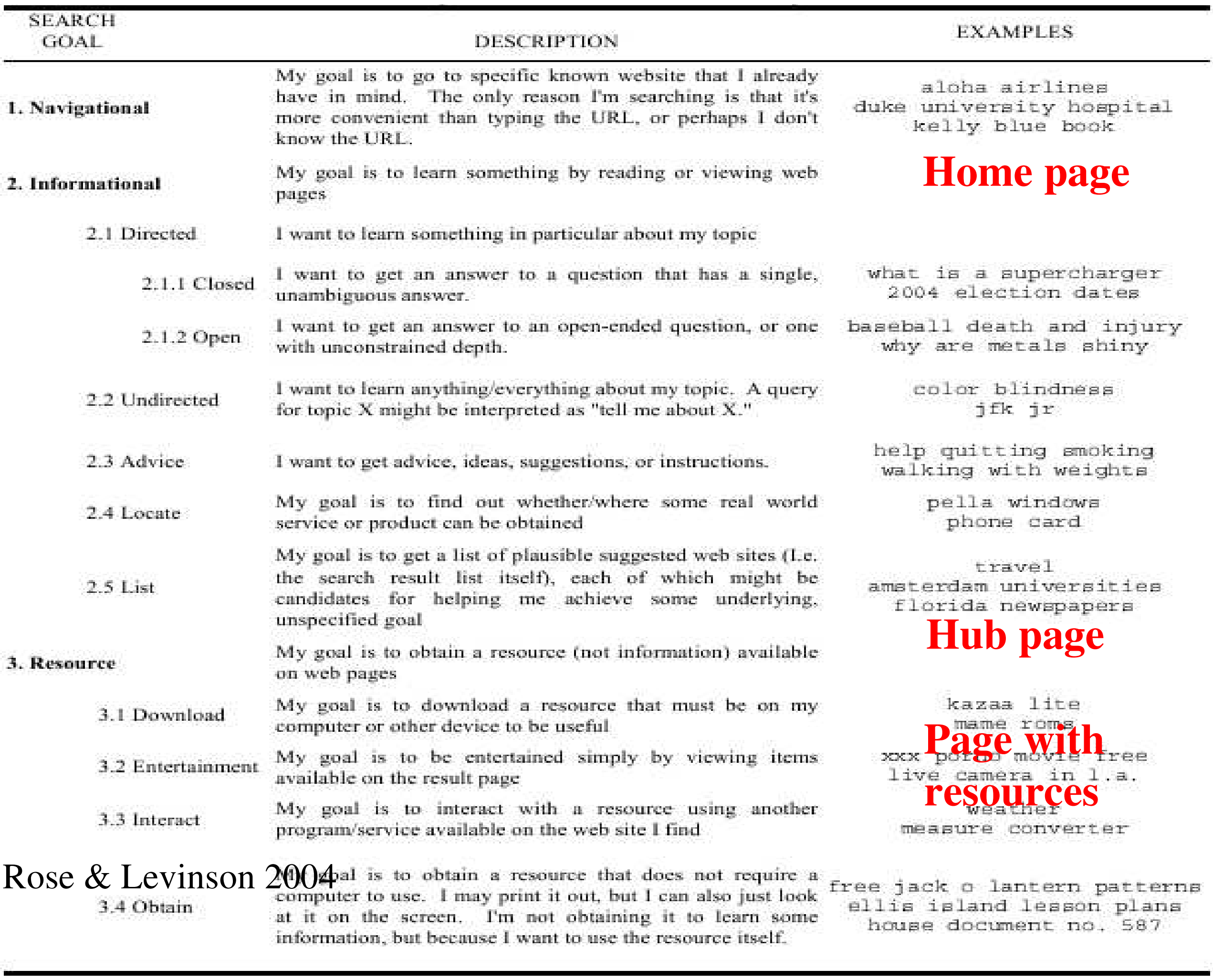
Using the Context

Example: *I want information about Santiago*

- **Context**
  - Family in Chile
  - Catholic
  - Travelling to Cuba
  - Lives in Argentina
  - Located in Santo Domingo
  - Architect
  - Spanish movies fan
  - Baseball fan
- **Probable Answer**
  - *Santiago de Chile*
  - *Santiago de Compostela*
  - *Santiago de Cuba*
  - *Santiago del Estero*
  - *Santiago de los Caballeros*
  - *Santiago Calatrava*
  - *Santiago Segura*
  - *Santiago Benito*

Context in Web Queries

- *Session:* $(q, (URL, t)^{*})^{+}$
- *Who you are*: age, gender, profession (IP), etc.
- *Where you are and when*: time, location (IP), speed and direction, etc.
- *What you are doing*: interaction history, task in hand, etc.
- *What you are using*: searching device (operating system, browser, ...)



SEARCH GOAL	DESCRIPTION	EXAMPLES
1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	My goal is to learn something by reading or viewing web pages	
2.1 Directed	I want to learn something in particular about my topic	
2.1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	what is a supercharger 2004 election dates
2.1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	baseball death and injury why are metals shiny
2.2 Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	color blindness jfk jr
2.3 Advice	I want to get advice, ideas, suggestions, or instructions.	help quitting smoking walking with weights
2.4 Locate	My goal is to find out whether/where some real world service or product can be obtained	pella windows phone card
2.5 List	My goal is to get a list of plausible suggested web sites (I.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal	travel amsterdam universities florida newspapers
3. Resource	My goal is to obtain a resource (not information) available on web pages	
3.1 Download	My goal is to download a resource that must be on my computer or other device to be useful	kazaa lite mame roms
3.2 Entertainment	My goal is to be entertained simply by viewing items available on the result page	xxx porno movie free live camera in l.a.
3.3 Interact	My goal is to interact with a resource using another program/service available on the web site I find	weather measure converter
3.4 Obtain	My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.	free jack o lantern patterns ellis island lesson plans house document no. 587

Home page

Hub page

Page with resources

Rose & Levinson 2004

Kang & Kim, SIGIR 2003

- **Features:**
  - Anchor usage rate
  - Query term distribution in home pages
  - Term dependence
- **Not effective: 60%**

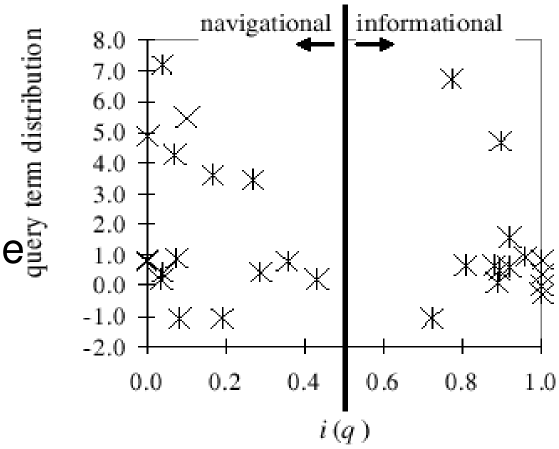


Figure 16: Query term distribution

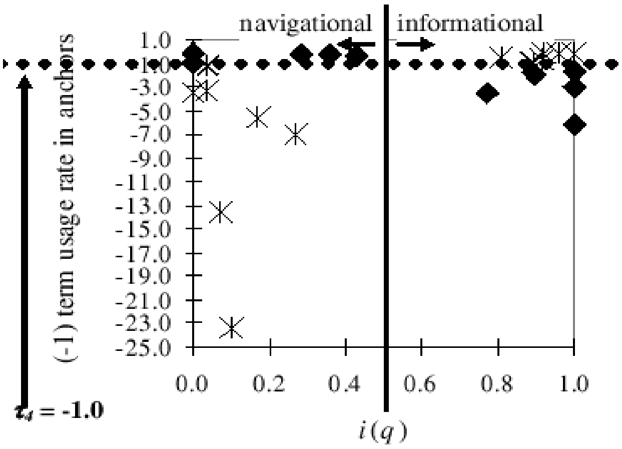


Figure 15: Anchor usage rate

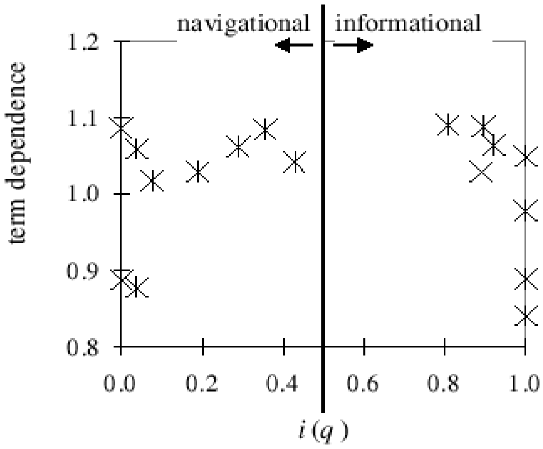


Figure 17: Term dependence

User Goals

- Liu, Lee & Cho, WWW 2005
- Top 50 CS queries
- Manual Query Classification: 28 people
- Informational goal *i(q)*
- *Remove software & person-names*
- *30 queries left*

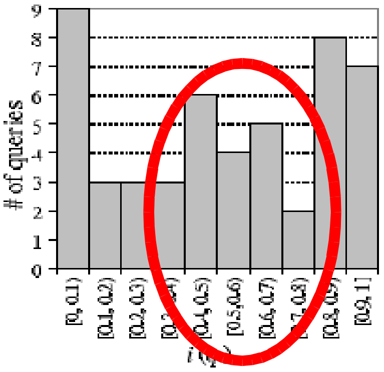


Figure 1: Query distribution along the $i(q)$ axis

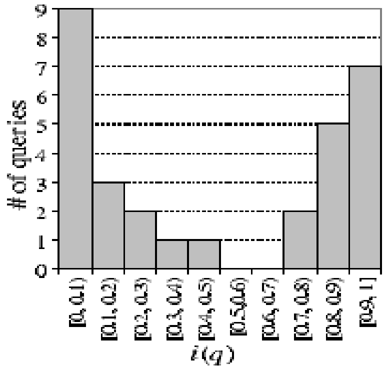


Figure 2: After removing software and person-name queries

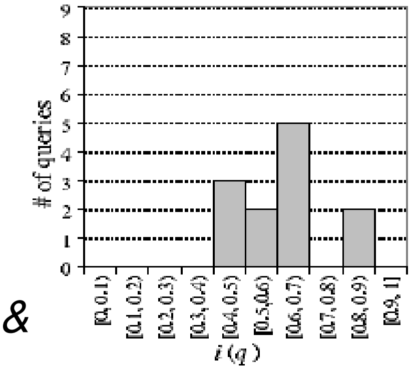


Figure 3: Distribution of the 12 software queries

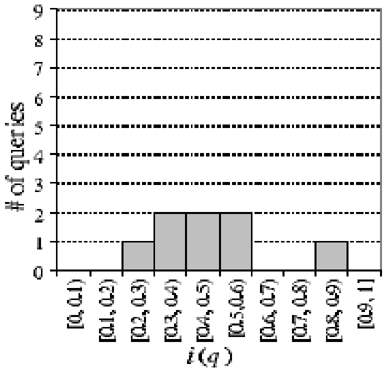
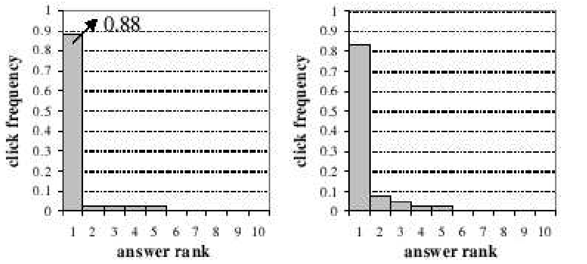


Figure 4: Distribution of the 8 person-name queries

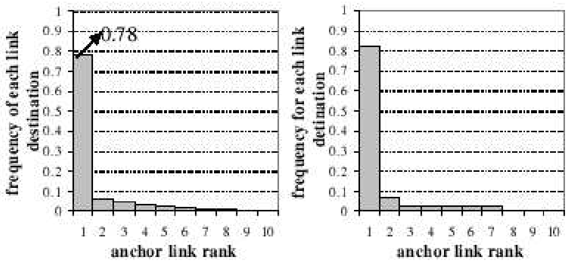
Features

- **Click & anchor text distribution**



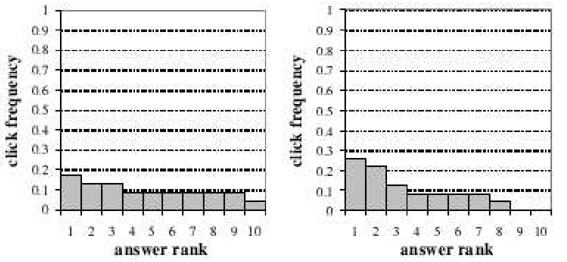
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 5: Click distributions for sample navigational queries



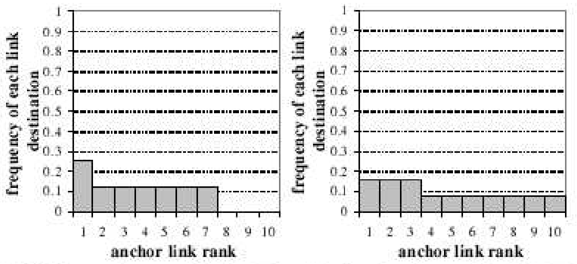
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 7: Anchor-link distributions for sample navigational queries



(a) hidden markov model ($i(q)$=1) (b) simulated annealing ($i(q)$=1)

Figure 6: Click distributions for sample informational queries



(a) hidden markov model ($i(q)$=1) (b) simulated annealing ($i(q)$=1)

Figure 8: Anchor-link distributions for sample informational queries

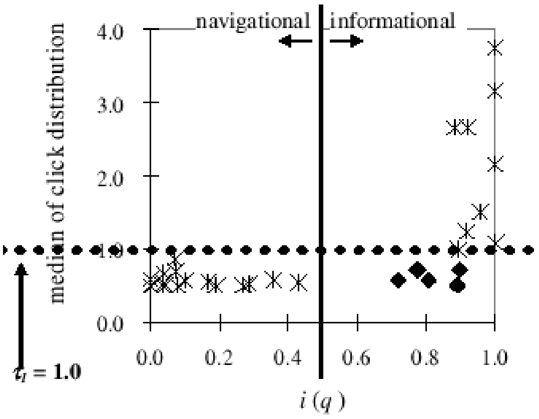


Figure 11: Median of click distribution

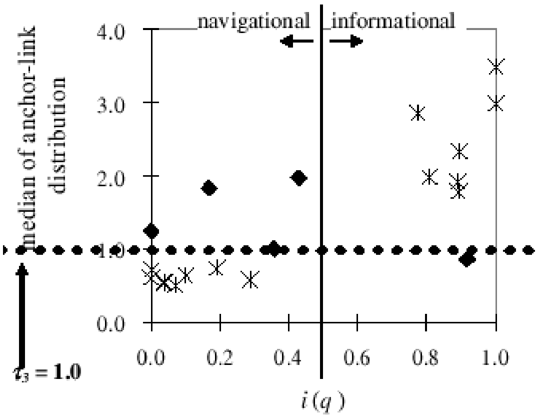


Figure 13: Median of anchor-link distribution

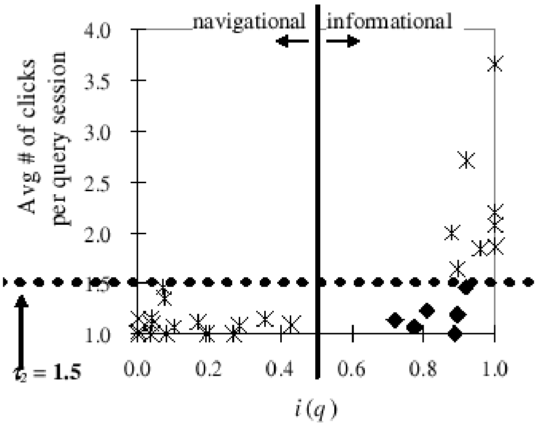


Figure 12: Avg # of clicks per query

Prediction power:

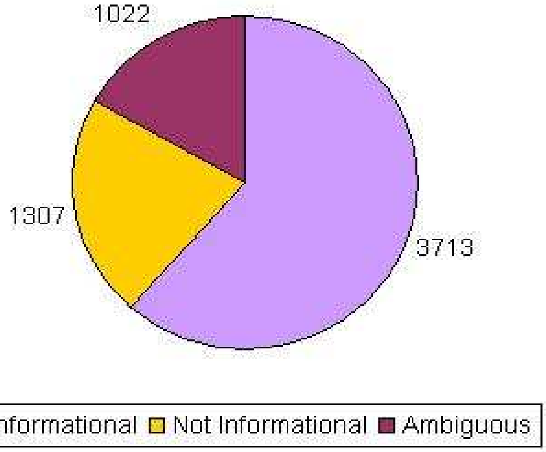
- **Single features: 80%**
- **Mixed features: 90%**

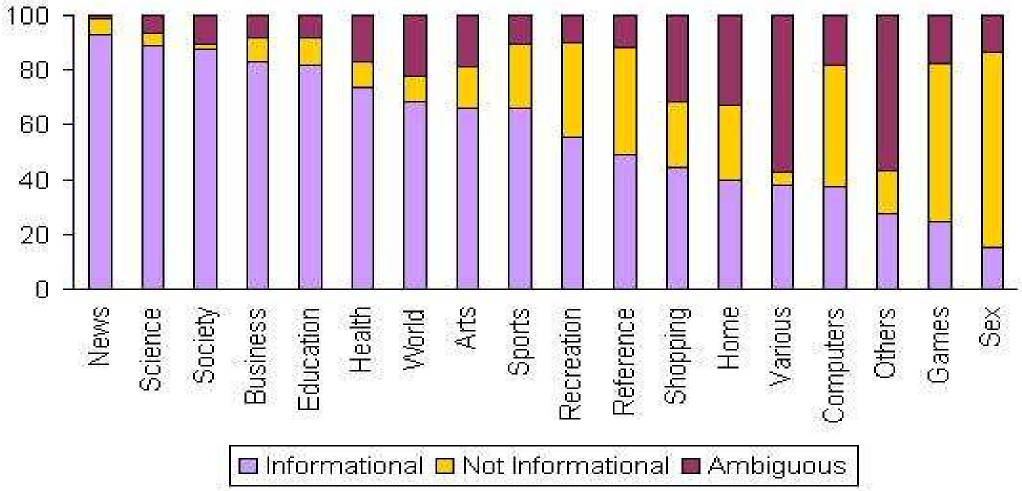
- **Drawbacks: Small evaluation, a posteriori feature**

User Intention

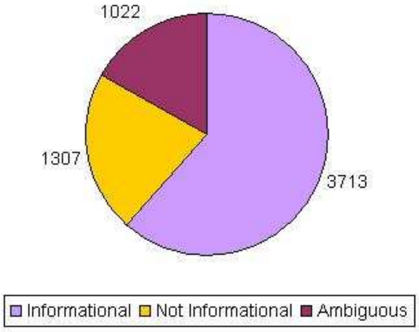
- **Manual classification of more than 6,000 popular queries**
- **Query Intention & topic**
- **Classification & Clustering**
- **Machine Learning on all the available attributes**
- **Baeza-Yates, Calderon & Gonzalez (SPIRE 2006)**

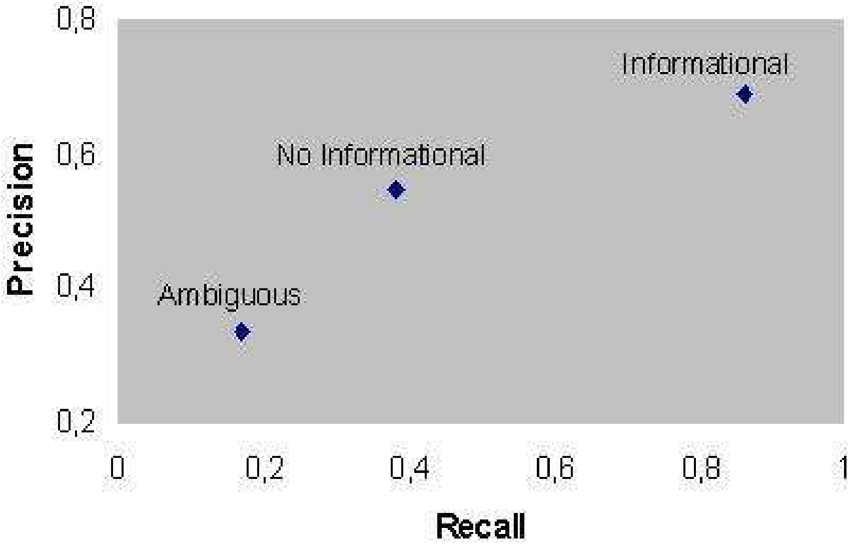
Classified Queries





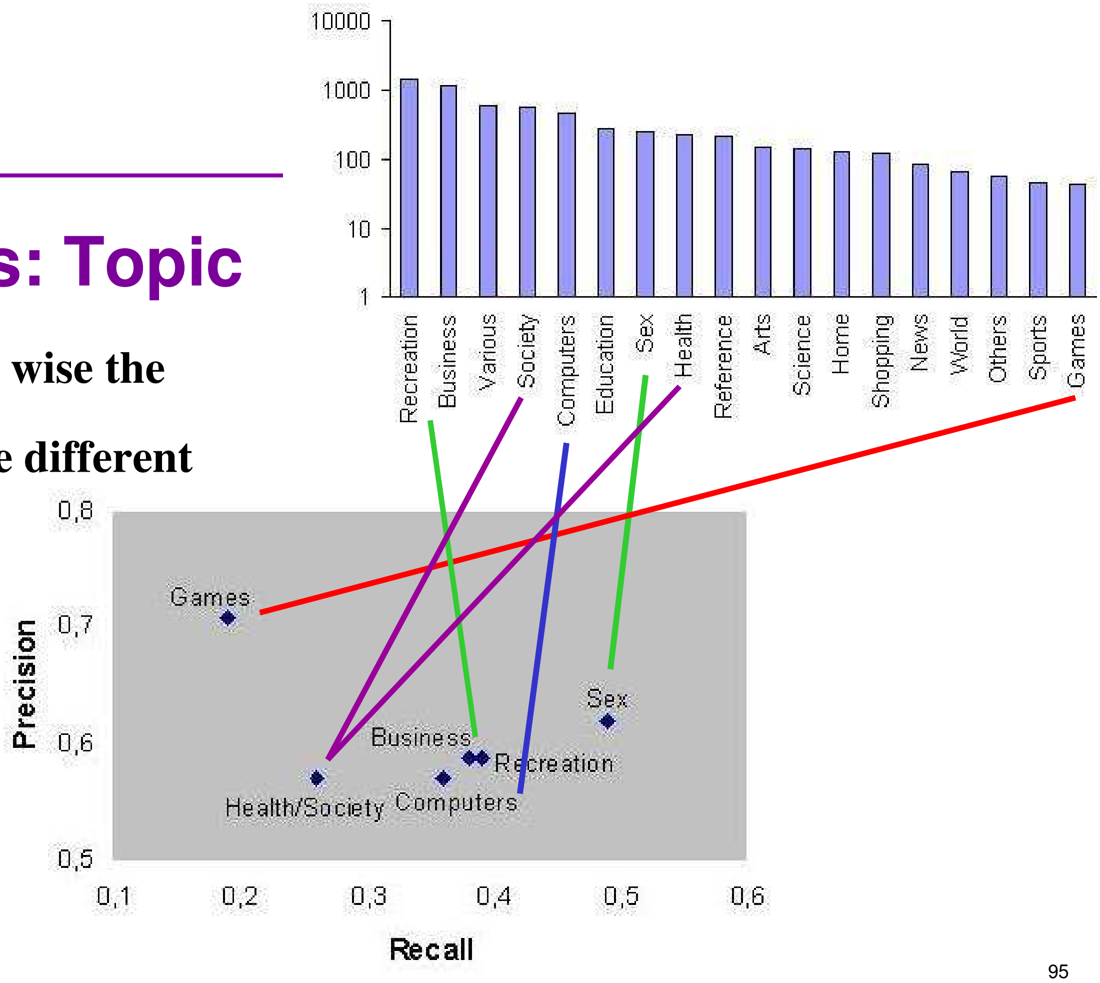
Results: User Intention

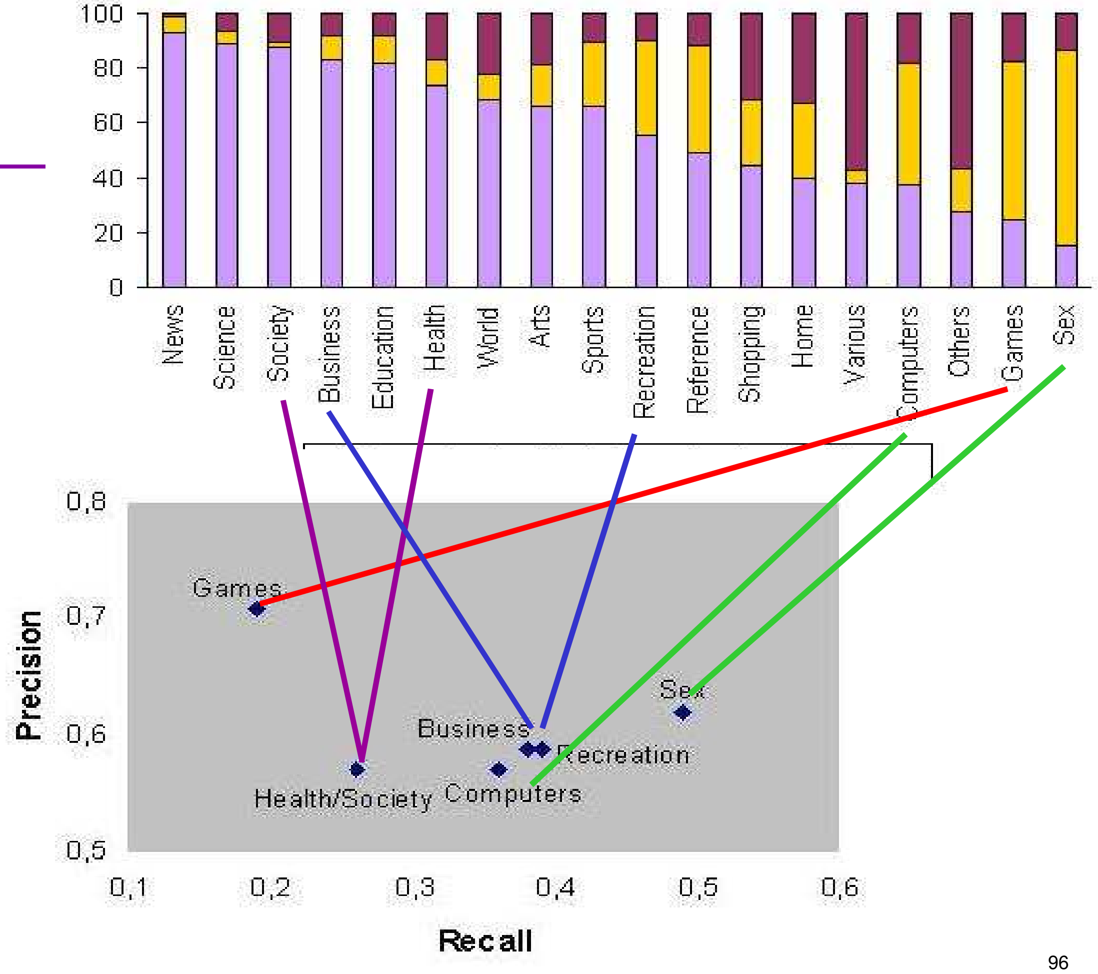




Results: Topic

- Volume wise the results are different



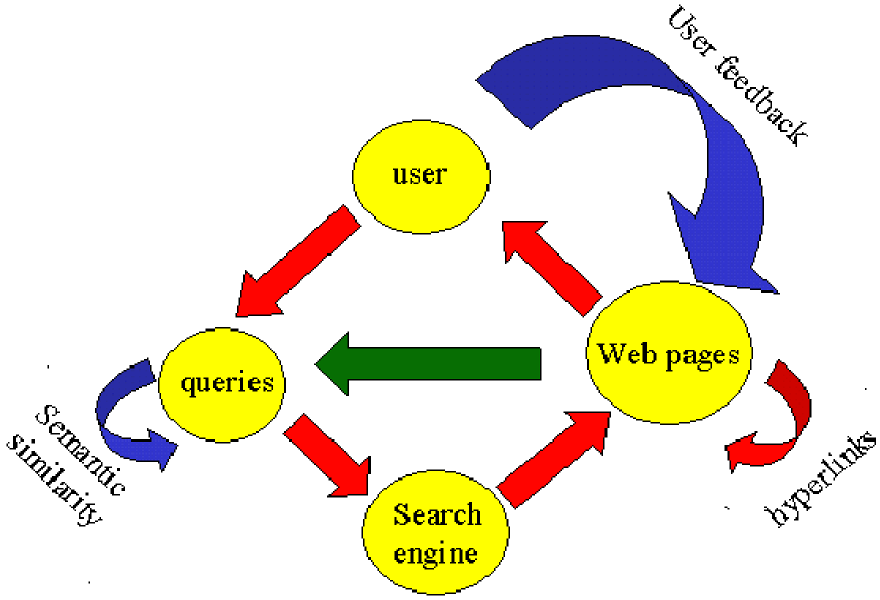


Clustering Queries

- Define relations among queries
  - Common words: sparse set
  - Common clicked URLs: better
  - Natural clusters
- Define distance function among queries
  - Content of clicked URLs
    (Baeza-Yates, Hurtado & Mendoza, 2004)
  - Summary of query answers (Sahami, 2006)

Goals

- Can we cluster queries well?
- Can we assign user goals to clusters?



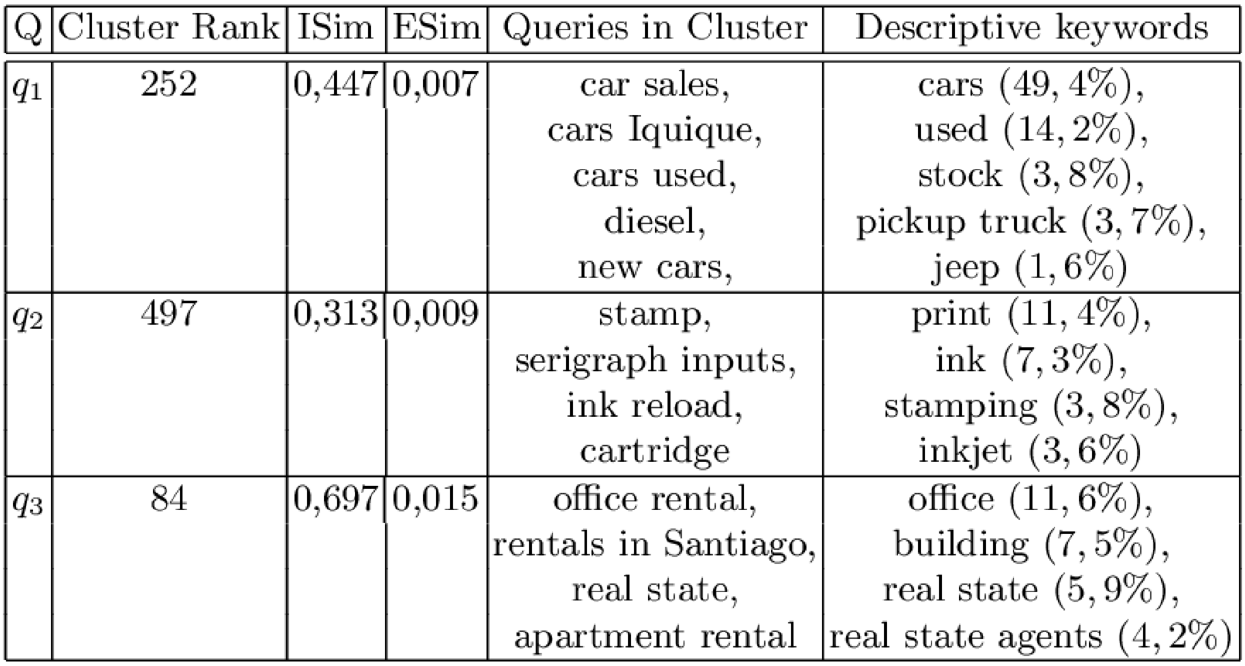
Our Approach

- Cluster text of clicked pages
  - Infer query clusters using a vector model

$$q[i] = \sum_{URLu} \frac{\mathtt{Pop}(q,u) \times \mathtt{Tf}(t_i,u)}{\max_t \mathtt{Tf}(t,u)}$$

- Pseudo-taxonomies for queries
  - Real language (slang?) of the Web
  - Can be used for classification purposes

Clusters Examples



Q	Cluster Rank	ISim	ESim	Queries in Cluster	Descriptive keywords
q_1	252	0,447	0,007	car sales, cars Iquique, cars used, diesel, new cars,	cars (49, 4%), used (14, 2%), stock (3, 8%), pickup truck (3, 7%), jeep (1, 6%)
q_2	497	0,313	0,009	stamp, serigraph inputs, ink reload, cartridge	print (11, 4%), ink (7, 3%), stamping (3, 8%), inkjet (3, 6%)
q_3	84	0,697	0,015	office rental, rentals in Santiago, real state, apartment rental	office (11, 6%), building (7, 5%), real state (5, 9%), real state agents (4, 2%)

Using the Clusters

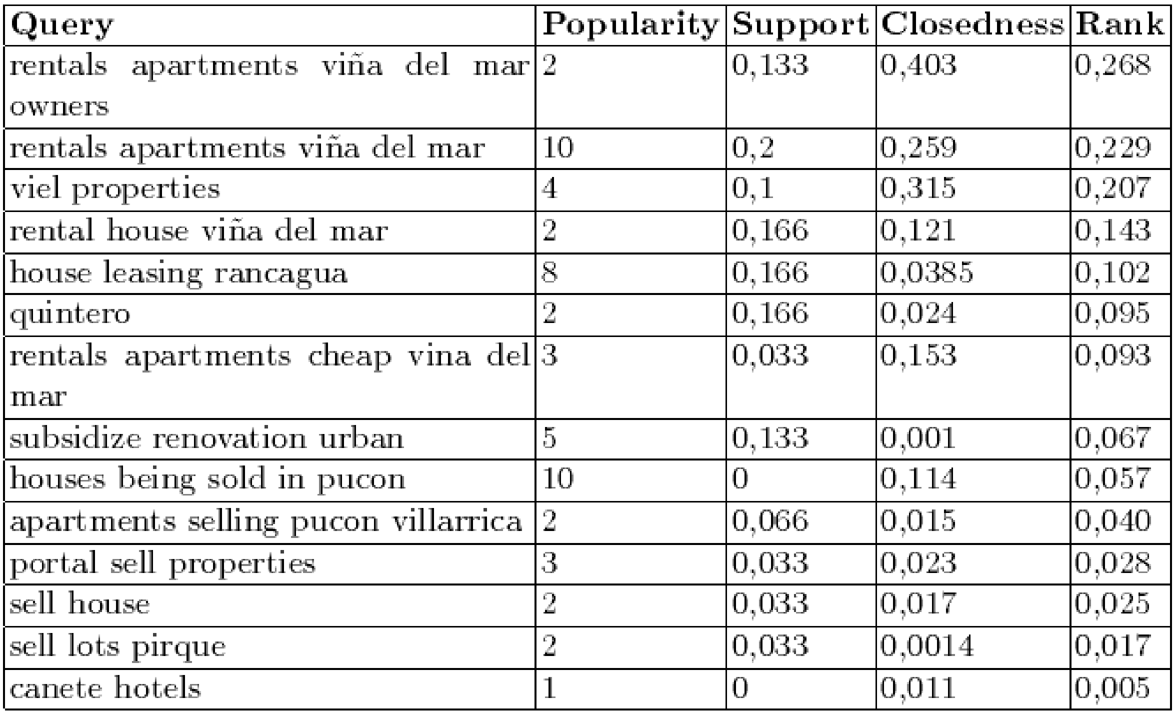
- Improved ranking
- Word classification
  - Synonyms & related terms are in the same cluster
  - Homonyms (polysemy) are in different clusters
- Query recommendation (ranking queries!)
  - Real queries, not query expansion

Baeza-Yates, Hurtado & Mendoza

Journal of ASIST 2007

$$\mathtt{Rank}(q) = \gamma \times \mathtt{Sup}(q, q_{ini}) + (1 - \gamma) \times \mathtt{Clos}(q)$$

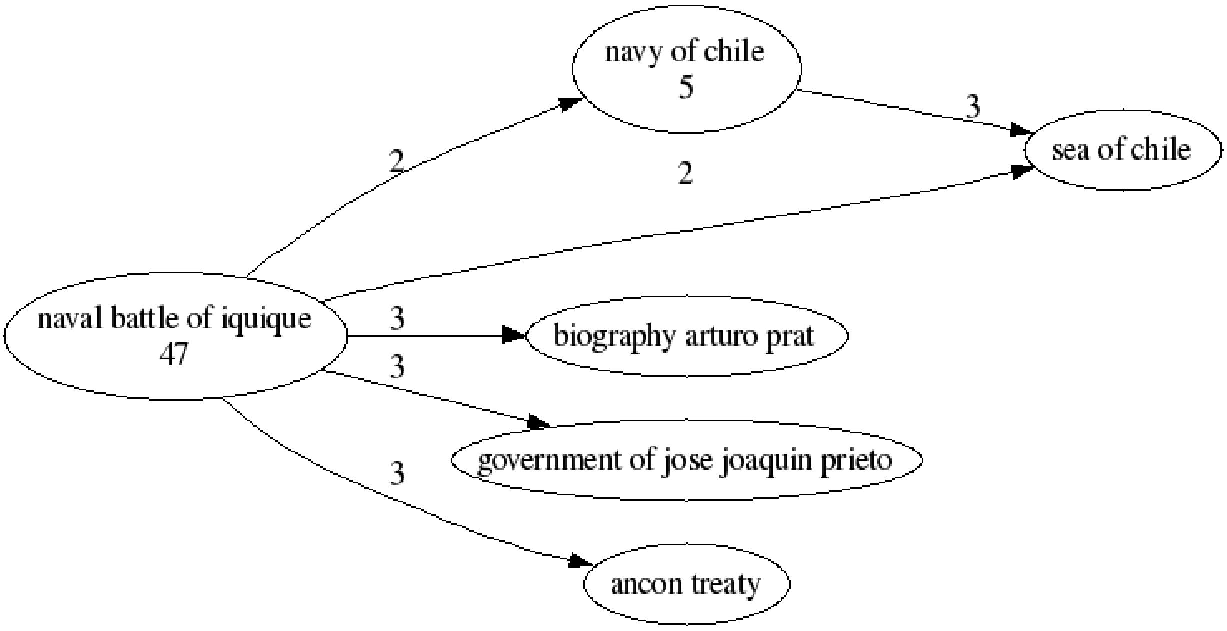
Query Recommendation



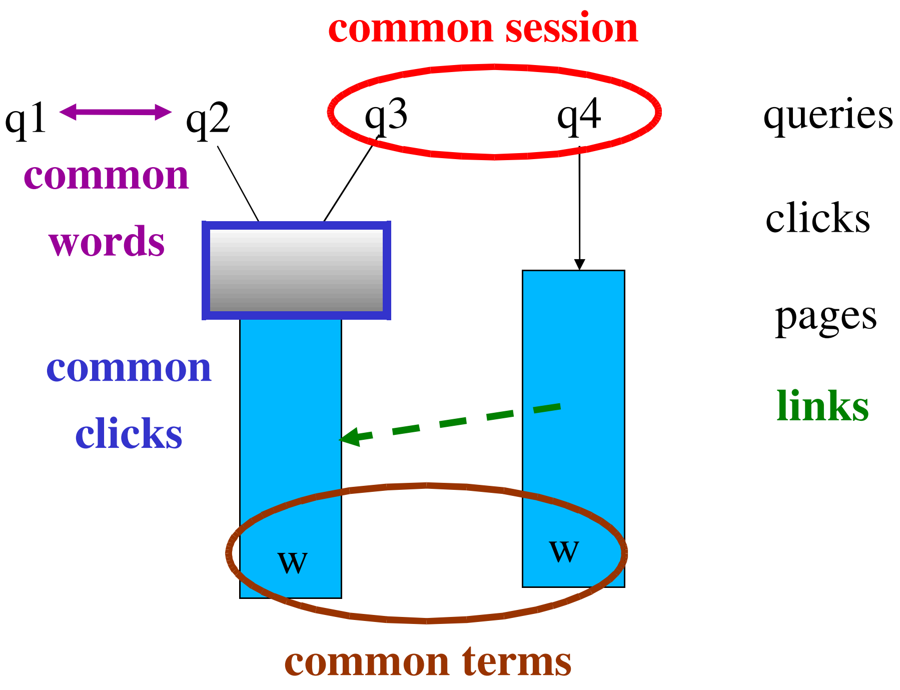
Query	Popularity	Support	Closedness	Rank
rentals apartments viña del mar owners	2	0,133	0,403	0,268
rentals apartments viña del mar	10	0,2	0,259	0,229
viel properties	4	0,1	0,315	0,207
rental house viña del mar	2	0,166	0,121	0,143
house leasing rancagua	8	0,166	0,0385	0,102
quintero	2	0,166	0,024	0,095
rentals apartments cheap vina del mar	3	0,033	0,153	0,093
subsidize renovation urban	5	0,133	0,001	0,067
houses being sold in pucon	10	0	0,114	0,057
apartments selling pucon villarrica	2	0,066	0,015	0,040
portal sell properties	3	0,033	0,023	0,028
sell house	2	0,033	0,017	0,025
sell lots pirque	2	0,033	0,0014	0,017
canete hotels	1	0	0,011	0,005

Simple Related Terms

- Query dominance based on clicked pages



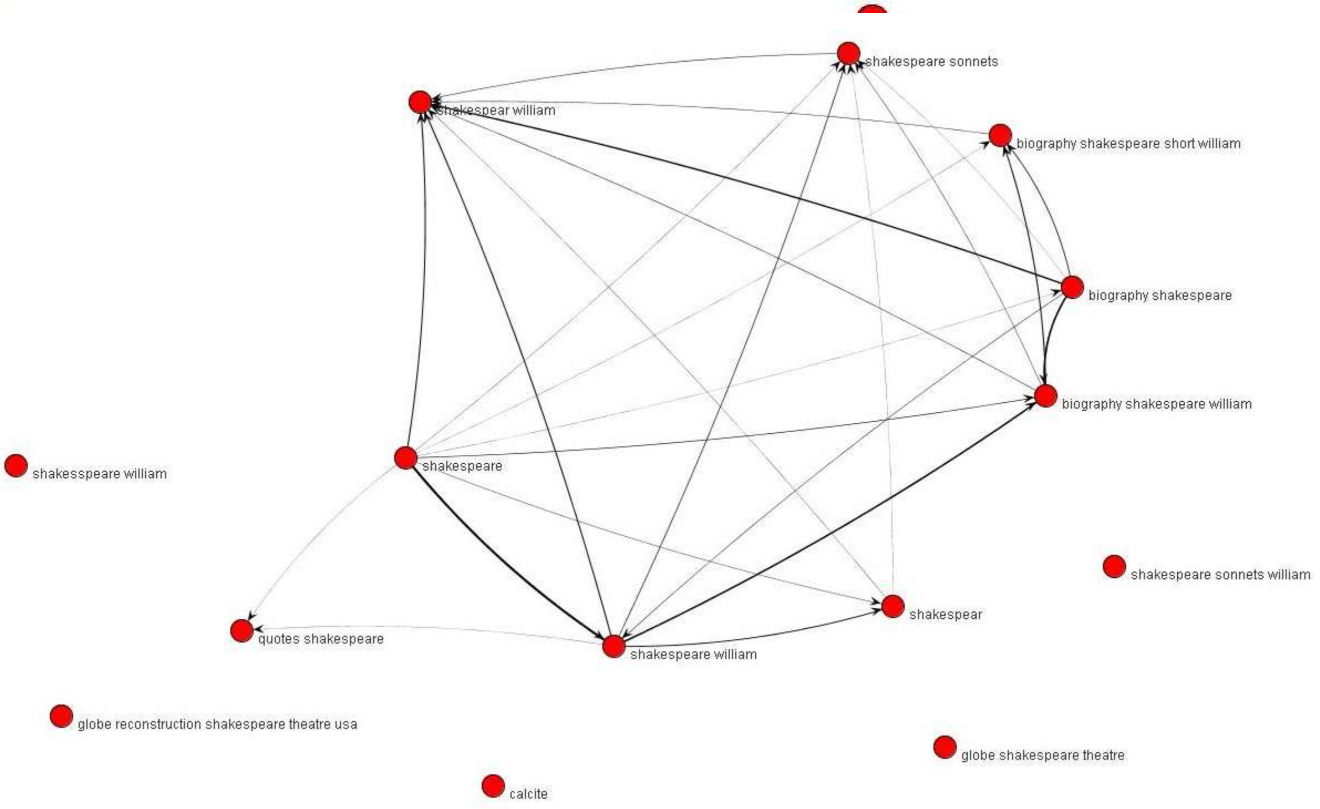
Relating Queries (Baeza-Yates, 2007)



Qualitative Analysis

Graph	Strength	Sparsity	Noise
Word	Medium	High	Polysemy
Session	Medium	High	Physical sessions
Click	High	Medium	**Multitopic pages** **Click spam**
Link	Weak	Medium	Link spam
Term	Medium	Low	Term spam

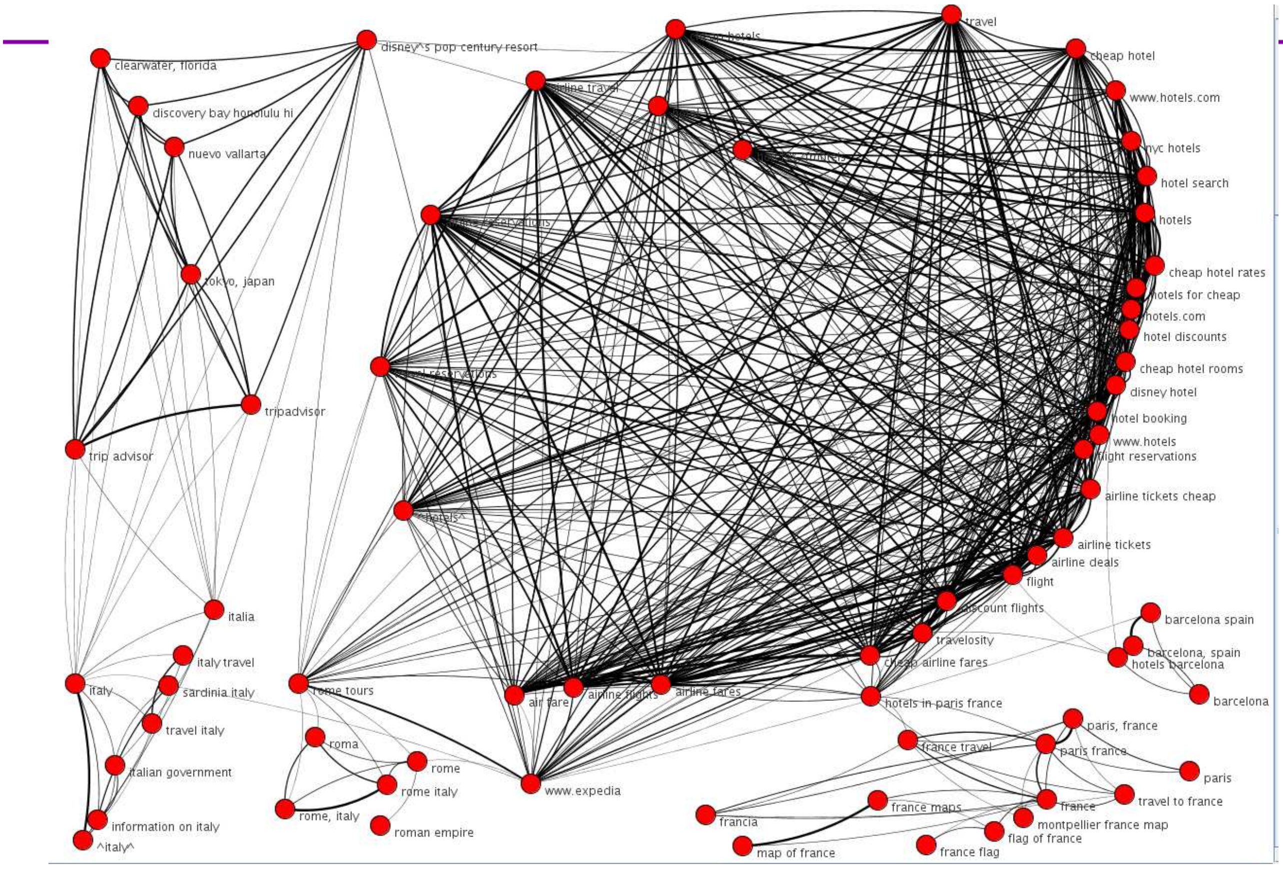
Words, Sessions and Clicks


shakespeare sonnets
shakespear william
biography shakespeare short william
biography shakespeare
biography shakespeare william
shakesspeare william
shakespeare
shakespeare sonnets william
shakespear
quotes shakespeare
shakespeare william
globe reconstruction shakespeare theatre usa
globe shakespeare theatre
calcite

Contributions

- Characterization of a large click graph
- Proposed specific distance and relations
- Hint the amount of implicit knowledge
- Evaluate the quality of the results

Click Graph



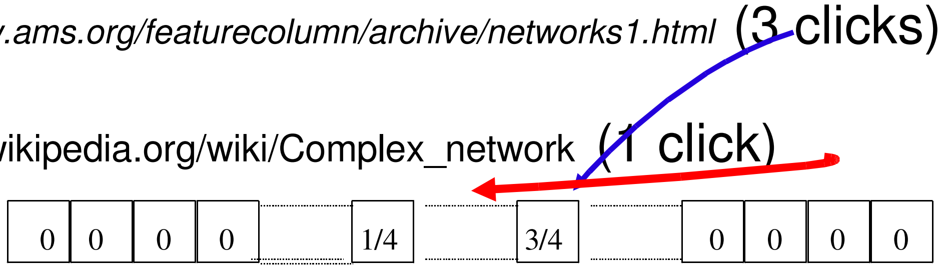
Formal Definition

- There is an edge between two queries *q* and *q'* if:
  - There is at least one URL clicked by both
- Edges can be weighted (for filtering)
  - We used the cosine similarity in a vector space defined by URL clicks

$$W(e) = \frac{\bar{q} \cdot \bar{q}'}{|\bar{q}|\ |\bar{q}'|} = \frac{\sum_{i \leq D} q(i) \cdot q'(i)}{\sqrt{\sum_{i \leq D} q(i)^2} \cdot \sqrt{\sum_{i \leq D} q'(i)^2}}$$

URL based Vector Space

- Consider the query *"complex networks"*
- Suppose for that query the clicks are:



- *www.ams.org/featurecolumn/archive/networks1.html* (3 clicks)
- en.wikipedia.org/wiki/Complex_network (1 click)

0	0	0	0		1/4		3/4		0	0	0	0

"Complex networks"

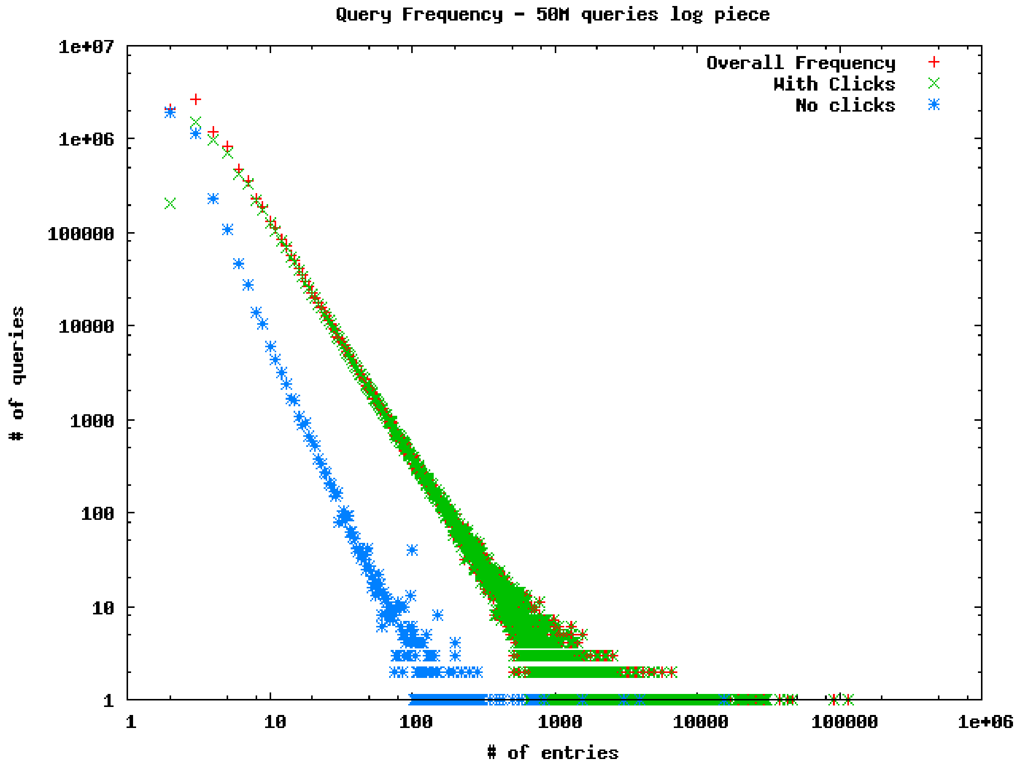
Building the Graph

- The graph can be built efficiently:
  - Consider the tuples (query, clicked url)
  - Sort by the second component
  - Each block with the same URL *u* gives the edges induced by *u*
  - Complexity: *O(max {M*|E|, n log n})* where *M* is the maximum number of URLs between two queries, and *n* is the number of nodes

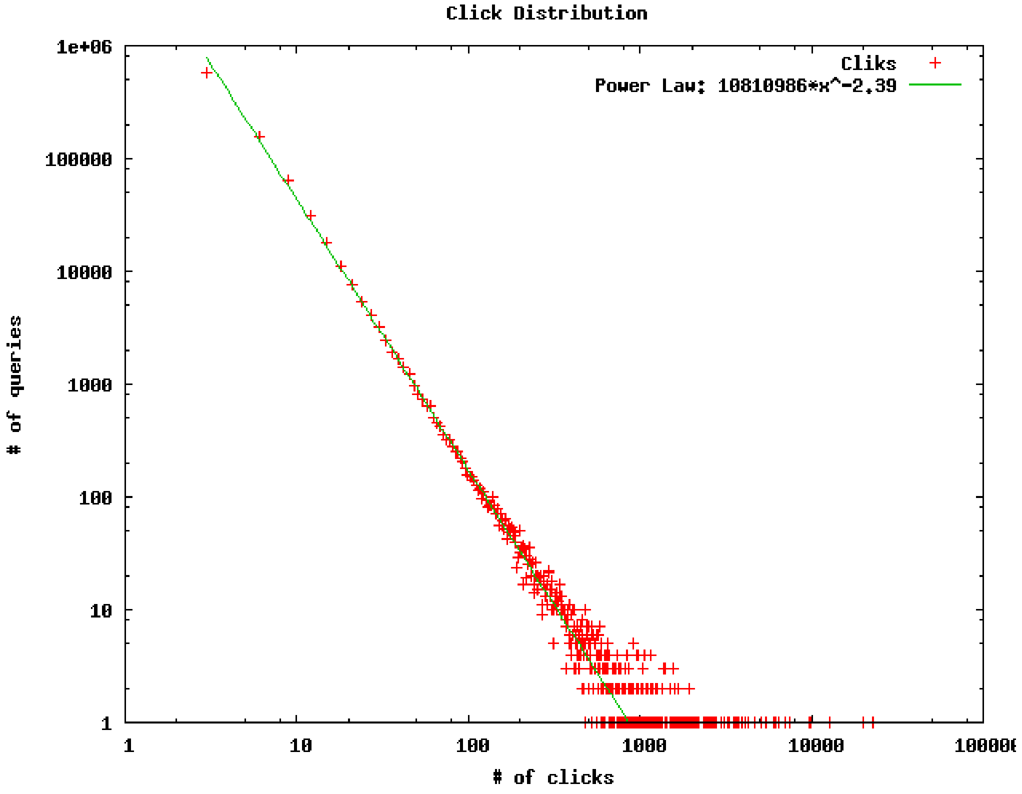
Anatomy of a Click Graph

- We built graphs using logs with up to 50 millions queries
  - For all the graphs we studied our findings are qualitatively the same (*scale-free network?*)
- Here we present the results for the following graph
  - 20M query occurrences
  - 2.8M distinct queries (nodes)
  - 5M distinct URLs
  - 361M edges

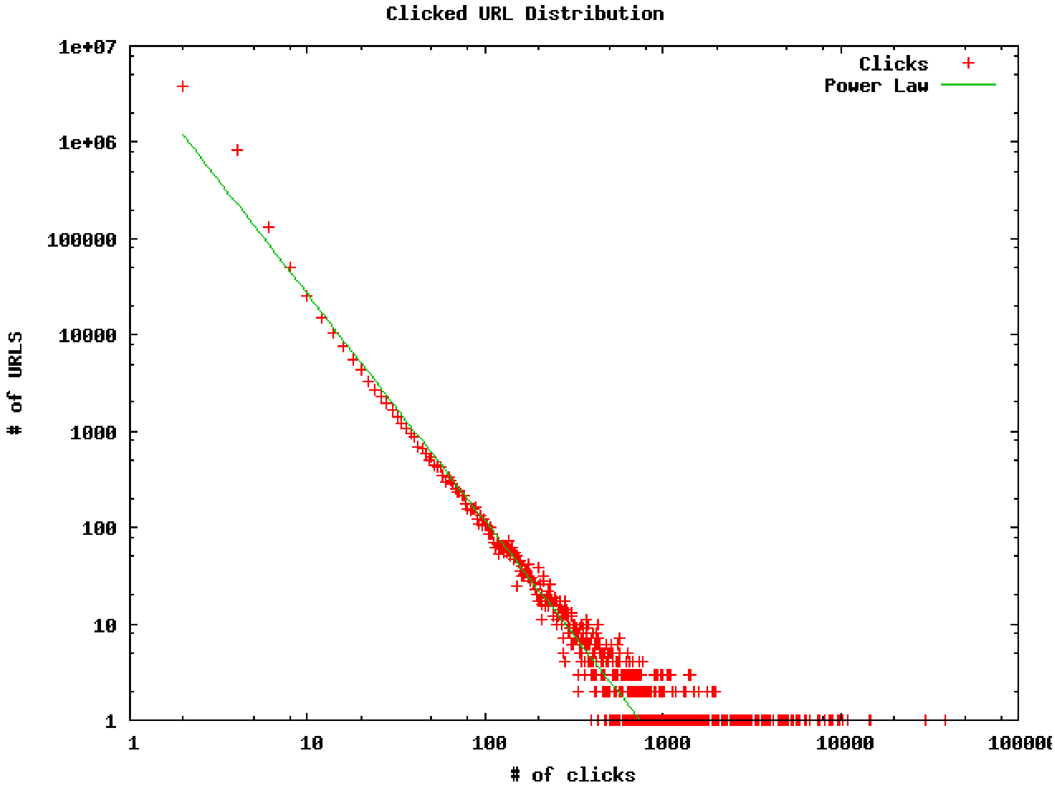
Query Frequency



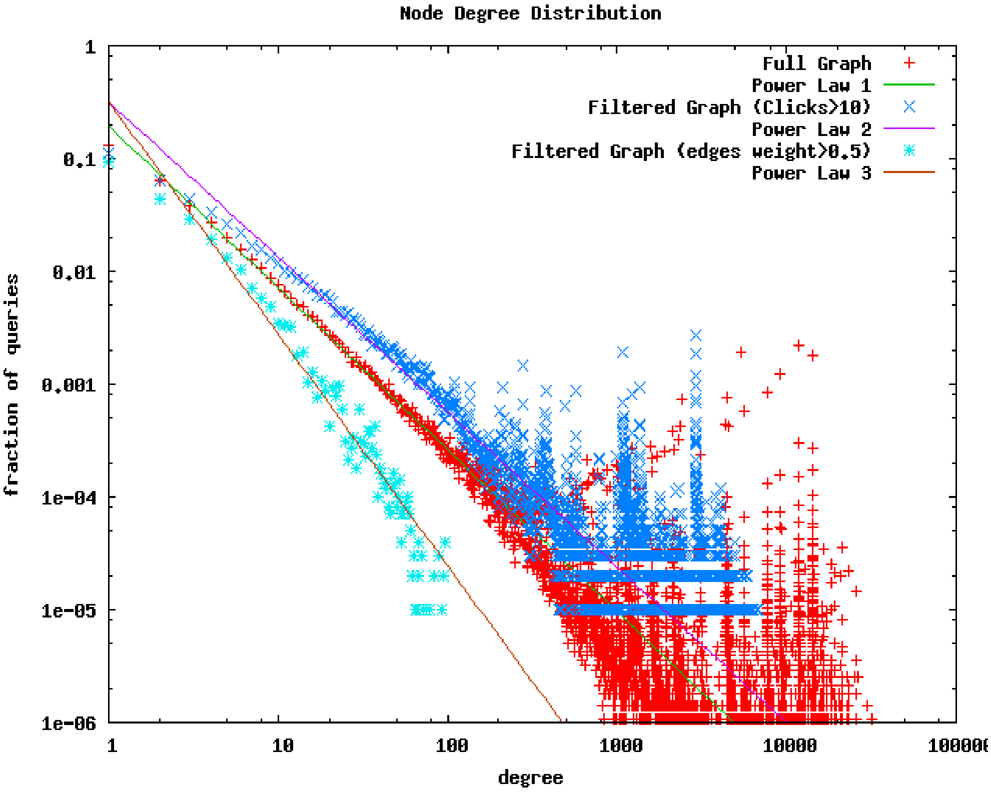
Click Distribution



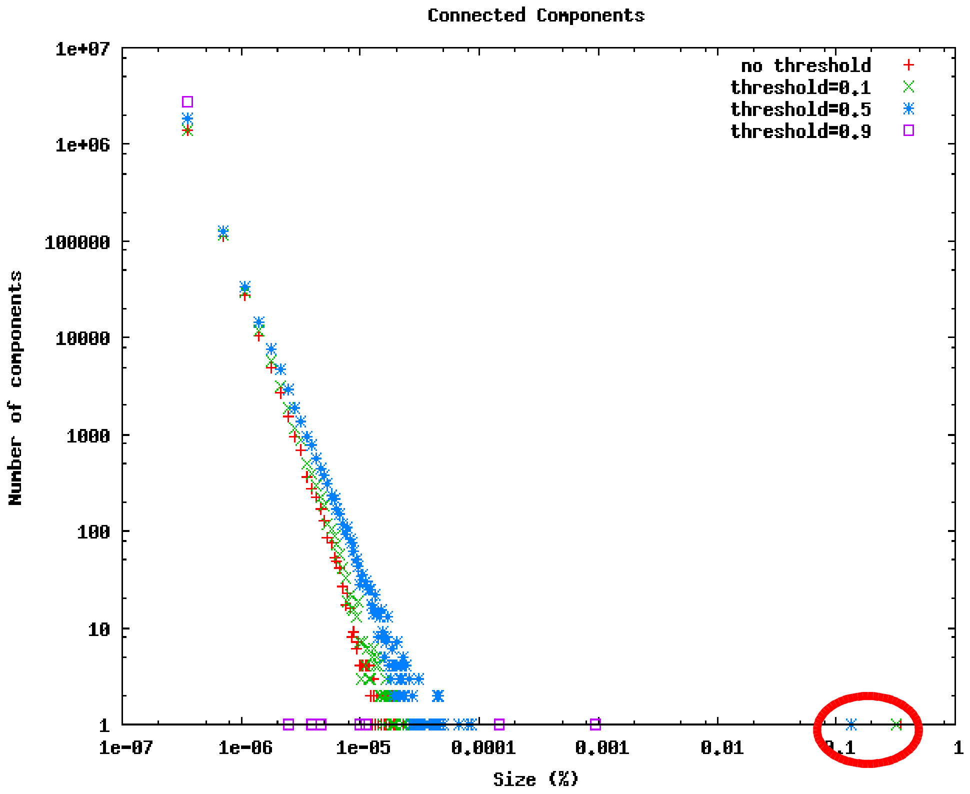
Clicked URL DIstribution



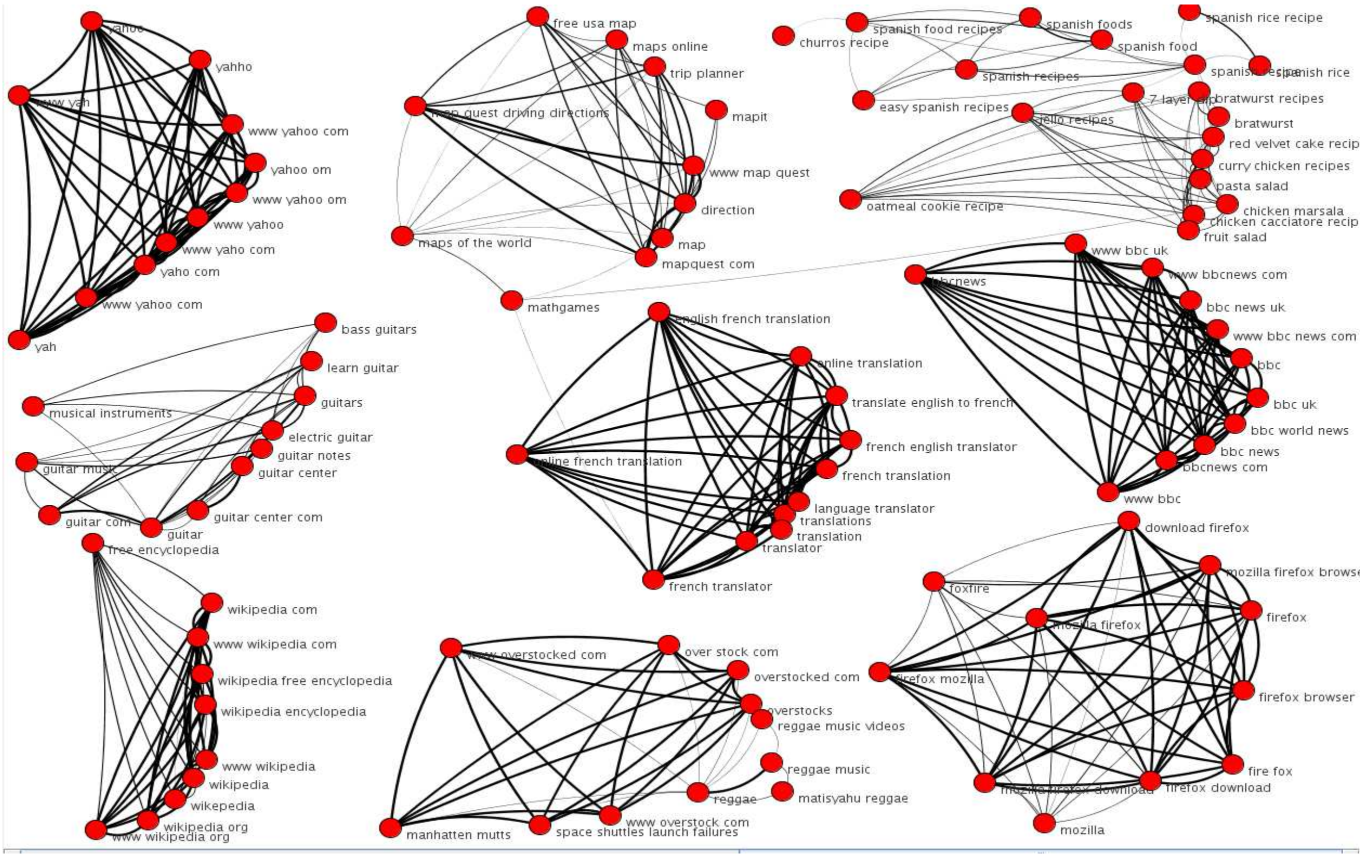
Node Degree Distribution



Connected Components



Implicit Folksonomy?



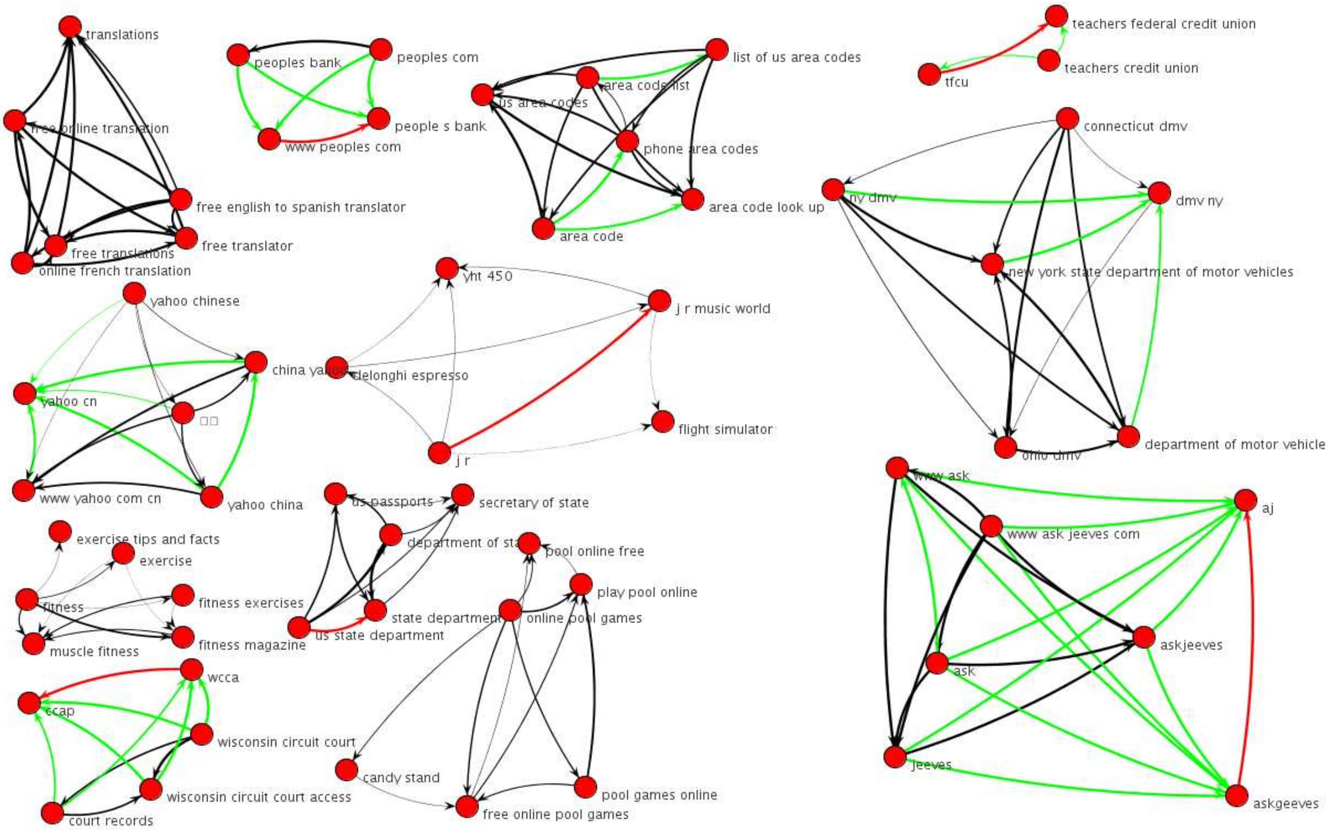
Set Relations and Graph Mining

- Identical sets: **equivalence**
- Subsets: **specificity**
  - directed edges
- Non empty intersections (with threshold)
  - degree of relation
- Dual graph: URLs related by queries
  - High degree: multi-topical URLs

Baeza-Yates & Tiberi

ACM KDD 2007

Implicit Knowledge? Webslang!



Evaluation: ODP Similarity

- A simple measure of similarity among queries using ODP categories
  - Define the similarity between two categories as the length of the longest shared path over the length of the longest path
  - Let *c_1,.., c_k* and *c'_1,.., c'_k* be the top *k* categories for two queries. Define the similarity (@*k*) between the two queries as *max{ sim(c_i,c'_j) | i,j=1,..,K }*

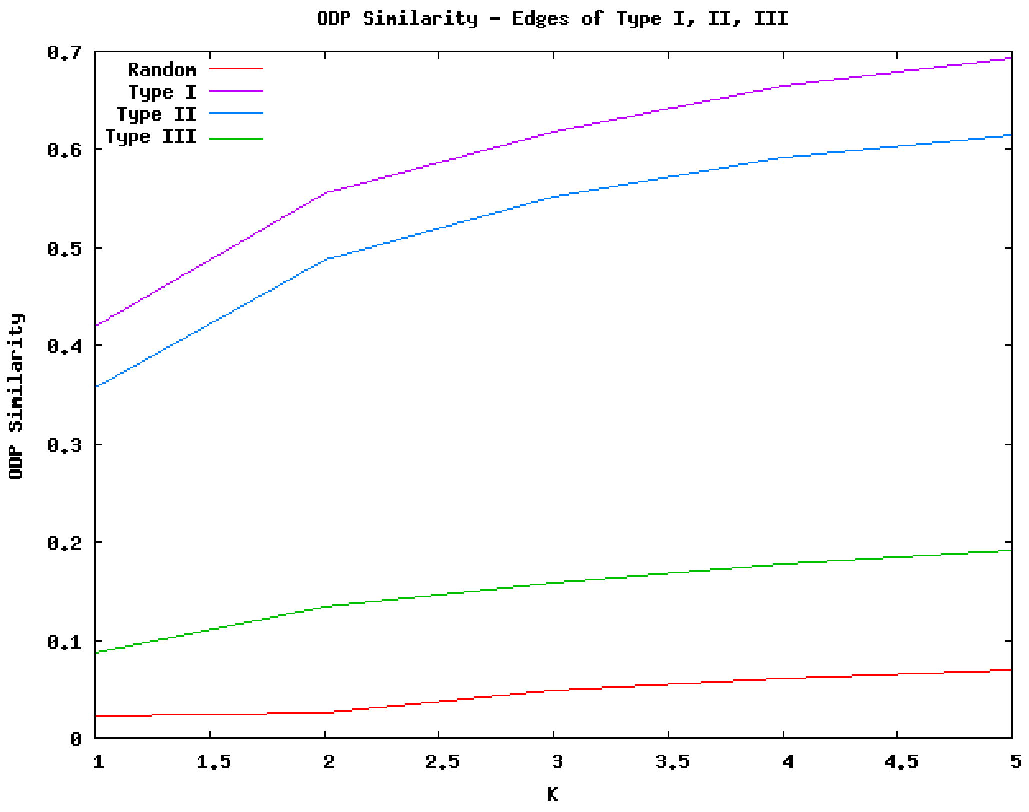
ODP Similarity

- Suppose you submit the queries "*Spain"* and "*Barcelona*" to ODP.
- The first category matches you get are:
  - Regional/ Europe/ Spain
  - Regional/ Europe/ Spain/ Autonomous Communities/ Catalonia/ Barcelona
- Similarity @1 is 1/2 because the longest shared path is "Regional/ Europe/ Spain" and the length of the longest is 6

Experimental Evaluation

- We evaluated a 1000 thousand edges sample for each kind of relation
- We also evaluated a sample of random pairs of not adjacent queries (baseline)
- We studied the similarity as a function of *k* (the number of categories used)

Experimental Evaluation



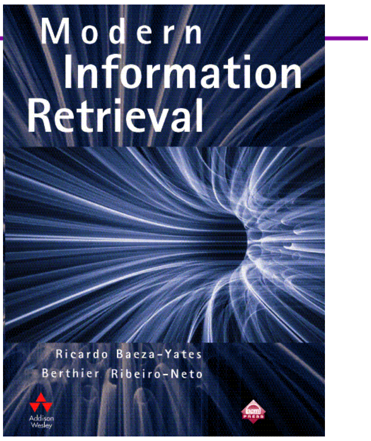
Open Issues

- Implicit social network
  - Any fundamental similarities?
- How to evaluate with partial knowledge?
  - Data volume amplifies the problem
- User aggregation vs. personalization
  - Optimize common tasks
  - Move away from privacy issues

Conclusions

- Web Mining: Potential for many different goals
- A fast prototyping platform is needed to explore
- Plenty of open problems:
  - Predict user goal + query recommendation
  - Take in account other query attributes
  - Generate topical metadata for documents based in queries that select that documents
  - Generate topical metadata for sites based on the above
  - Adaptive maintenance of the above

Bibliography – General



- **Modern Information Retrieval**
  by R. Baeza-Yates & B. Ribeiro-Neto, Addison-Wesley, 1999. Second edition in preparation.
- **Managing Gigabytes: Compressing and Indexing Documents and Images**
  by I.H. Witten, A. Moffat, and T.C. Bell. Morgan Kaufmann, San Francisco, second edition, 1999.
- **Mining the Web: Analysis of Hypertext and Semi Structured Data**
  by Soumen Chakrabarti. Morgan Kaufmann; August 15, 2002.
- **The Anatomy of a Large-scale Hypertextual Web Search Engine**
  by S. Brin and L. Page. 7th International WWW Conference, Brisbane, Australia; April 1998.
- **Websites:**
  - http://www.searchenginewatch.com/
  - http://www.searchengineshowdown.com/