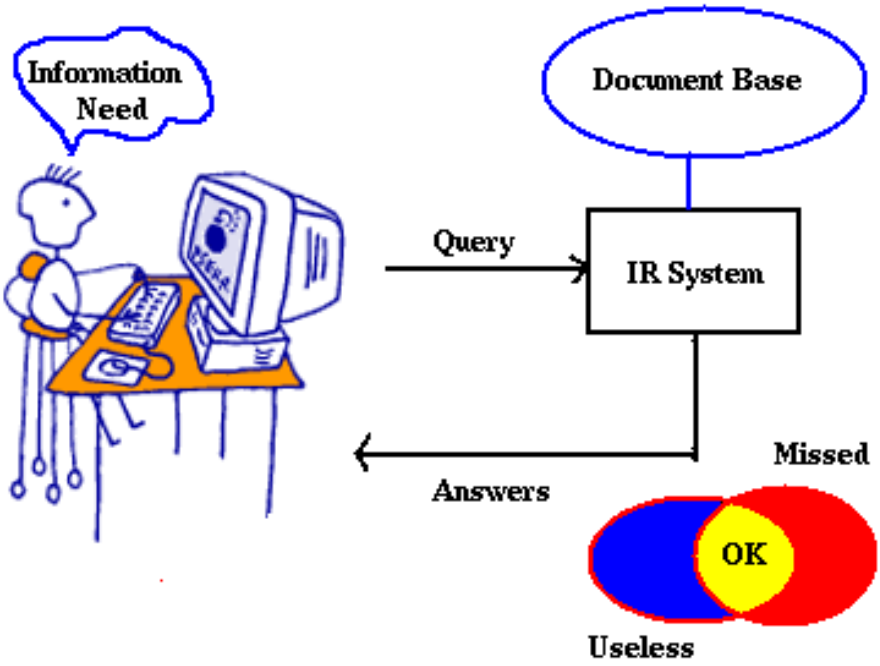
Web Mining
or
The Wisdom of Crowds

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

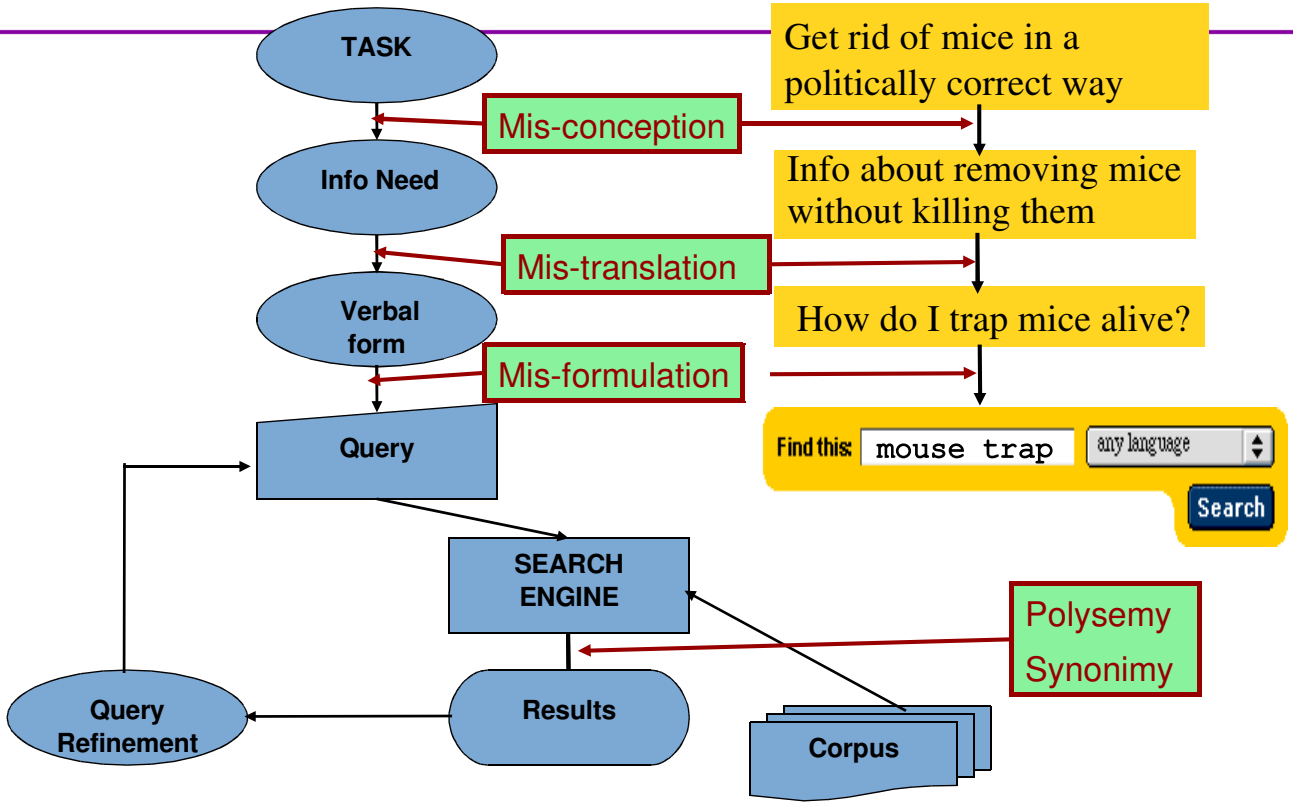
Agenda

- Motivation: Web Retrieval
- Web Mining as a Process
- Examples
- Case Study: Query Mining
- Concluding Remarks

The IR Problem



The classic search model

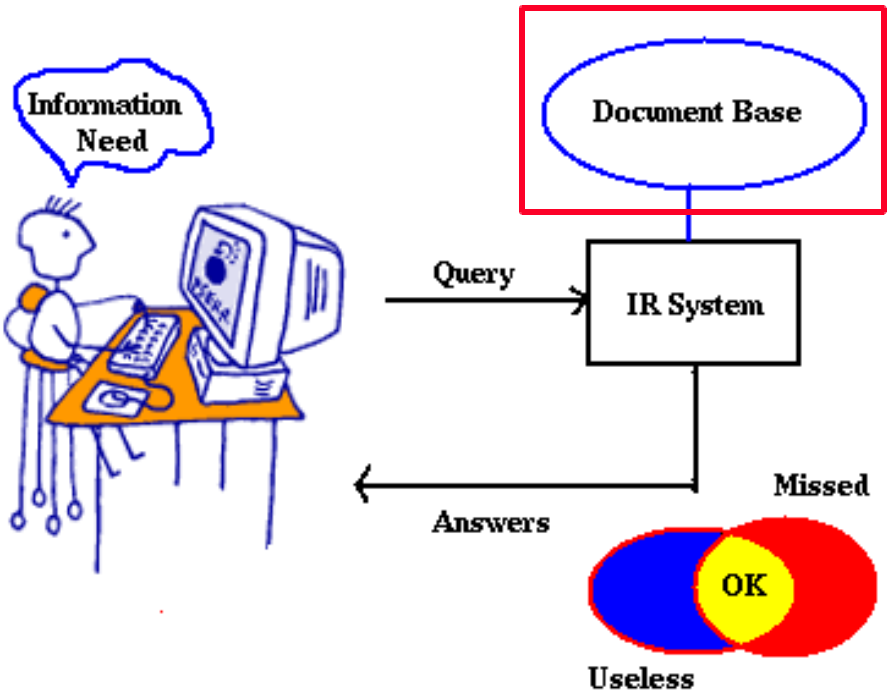


Classic IR Goal

- 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

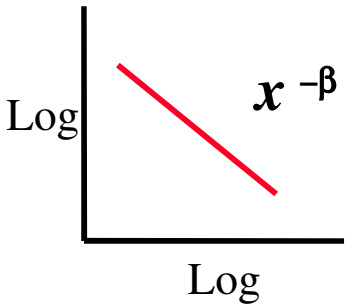
Bad assumptions in the web context

Challenges in Current IR Systems



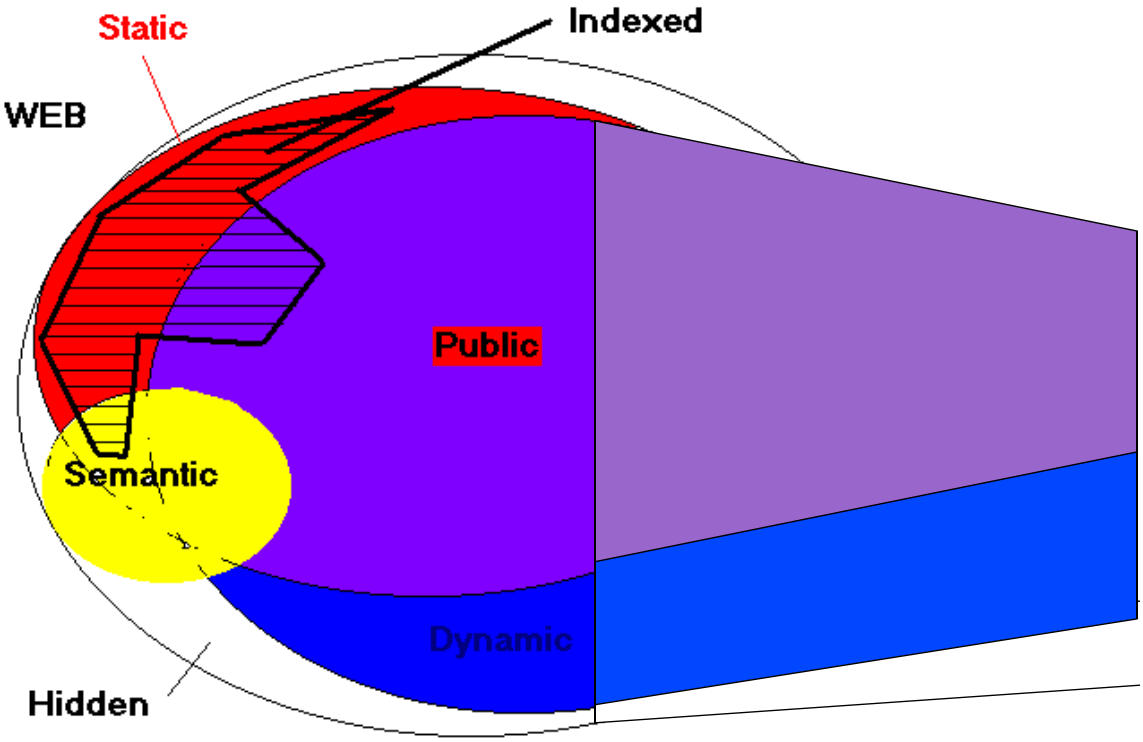
Document Base: Web

- Largest public repository of *data*
  (more than 20 billion static pages?)
- Today, there are 150 million Web servers (Nov 07)
  and more than 500 million hosts (Jul 07)
- Well connected graph with out-link and in-link power law distributions

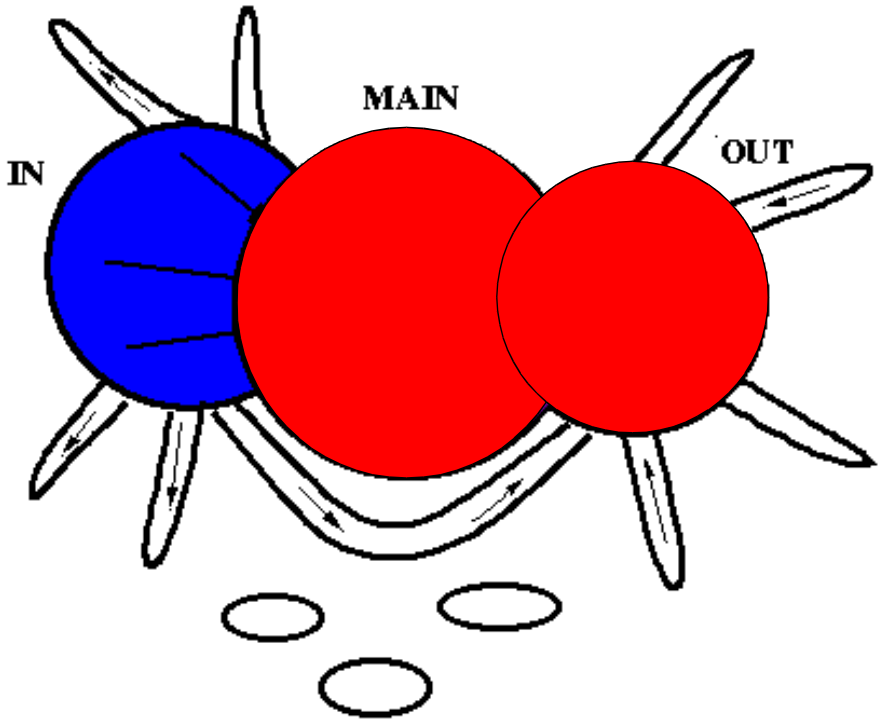


Self-similar &
Self-organizing

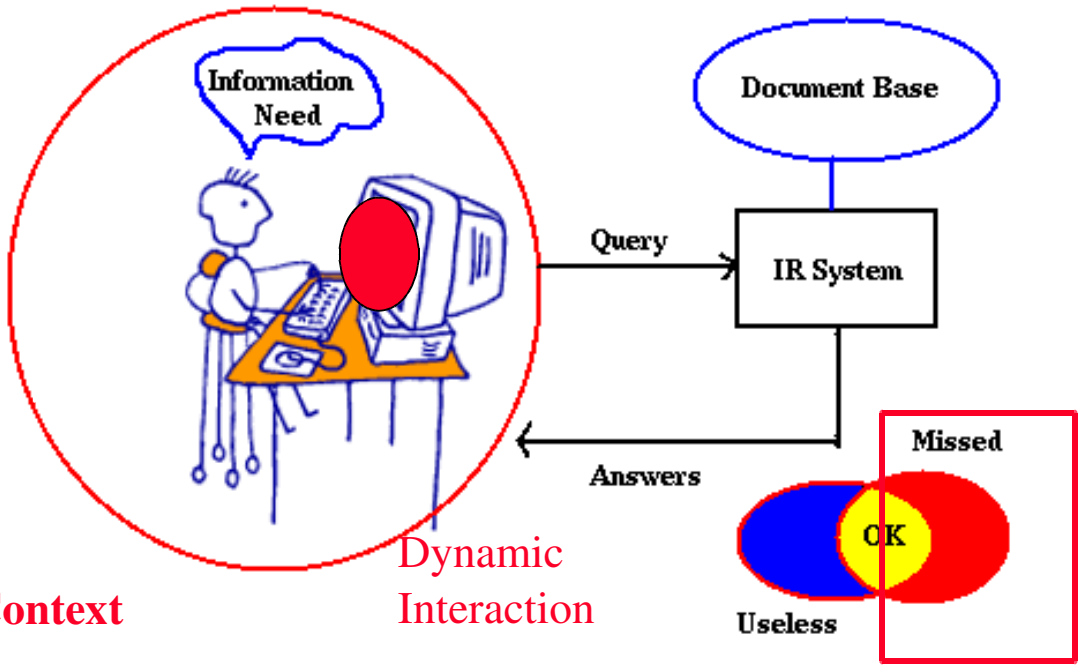
The Different Facets of the Web



The Structure of the Web



Challenges in Current IR Systems

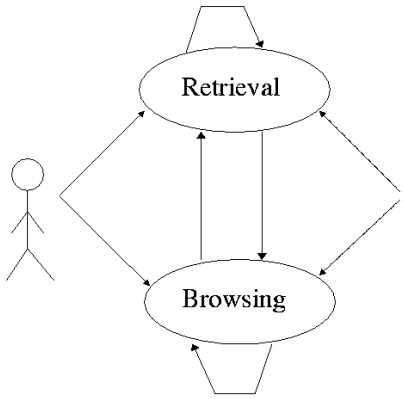


Context

Dynamic Interaction

Interaction

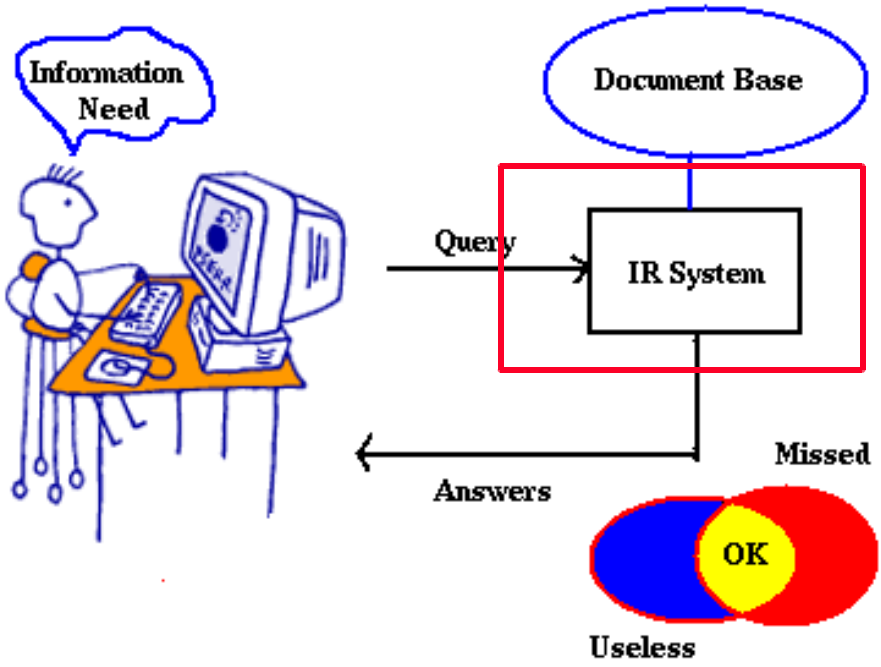
- Inexperienced users
- Dynamic information needs
- Varying task: querying, browsing



- 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

Challenges in Current IR Systems



Web Retrieval

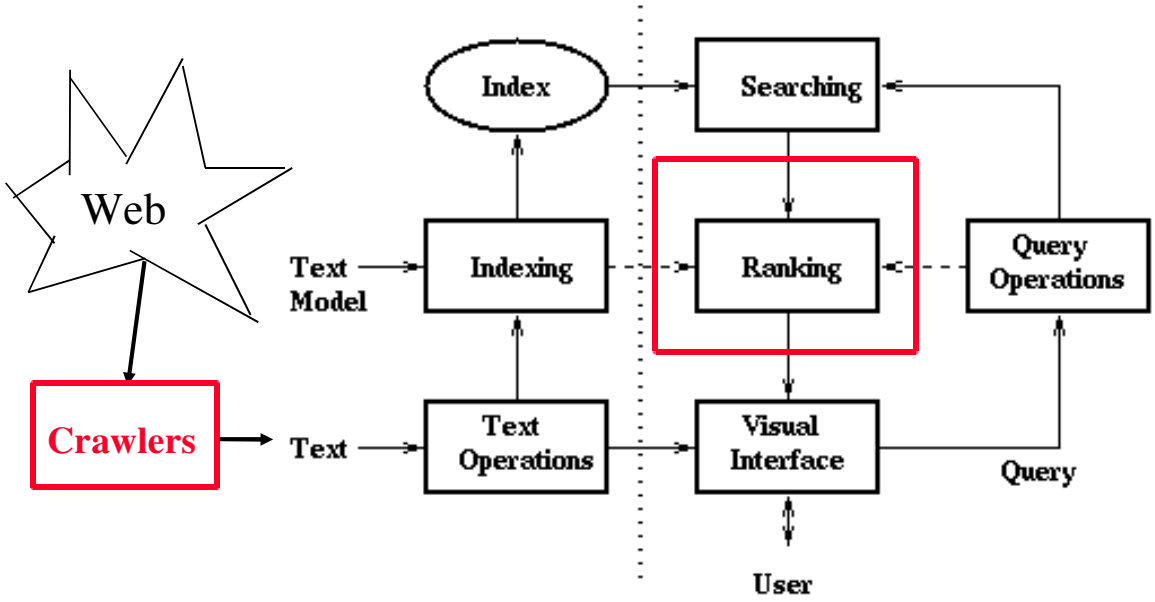
- Centralized Software Architecture
- Hypertext Structure
  - Allows to include link ranking
- On-line Quality Evaluation
- Distributed Data
  - Crawling
- Locally Distributed Index
  - Parallel Indexing
  - Parallel Query Processing
- Business Model based in Advertising
  - E.g. Word based and pay-per-click

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

  Conflict
  - Politeness vs. Usage of Resources

Adversarial IR

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

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

What for?

- The Web as an Object
- User Driven Web Design
- Improving Web Applications
- Social Mining
- .....

The Mining Process

- Gather the data
- Clean, organize and store the data
- Process the data
- Evaluate the quality of your results

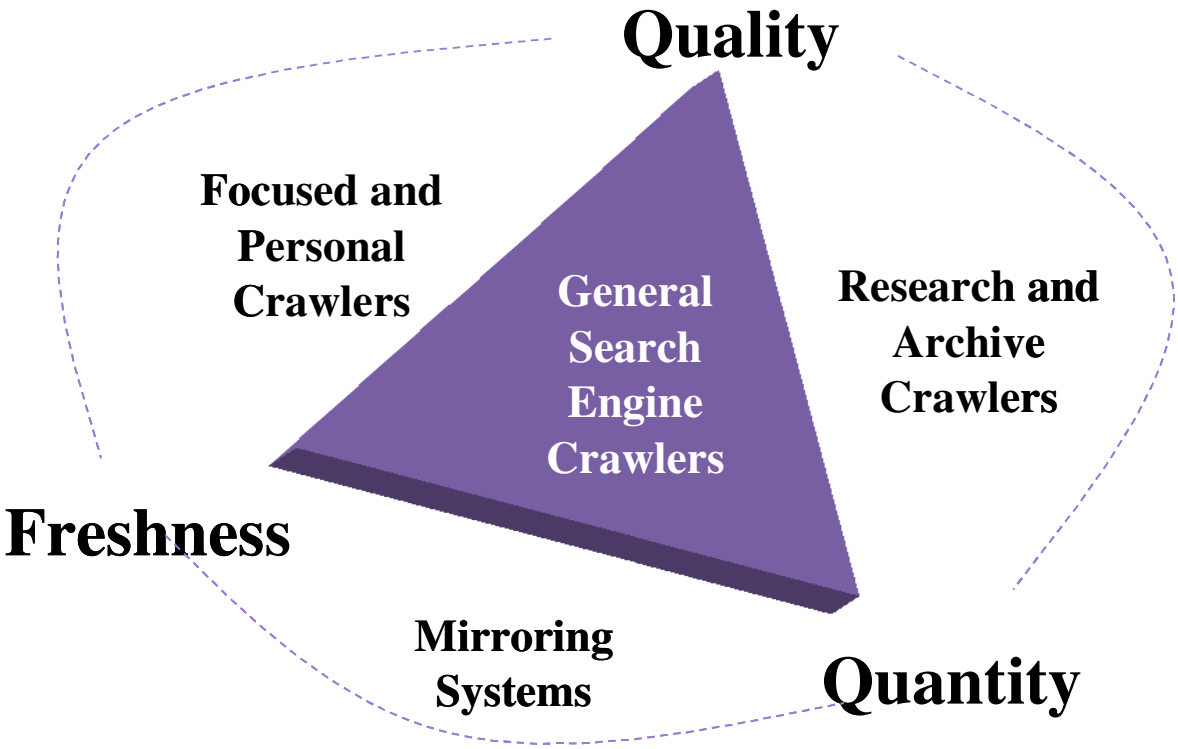
Data Recollection

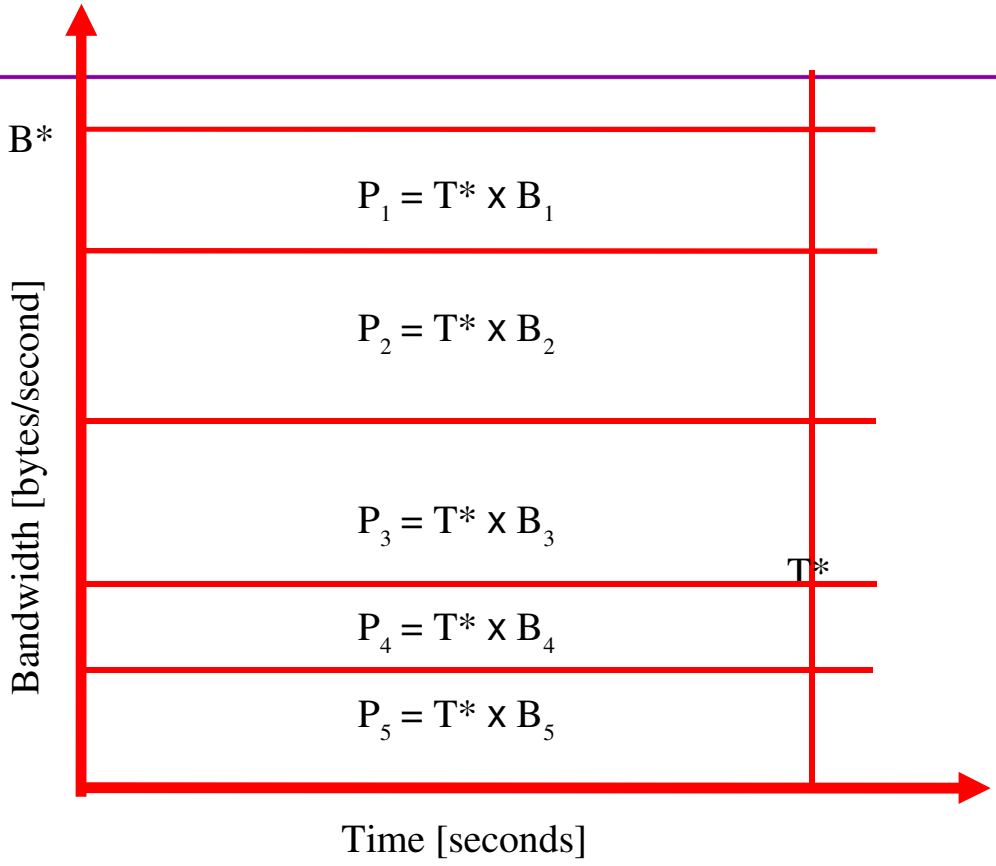
- Content and structure: Crawling
- Usage: Logs
  - Web Server logs
  - Specific Application logs

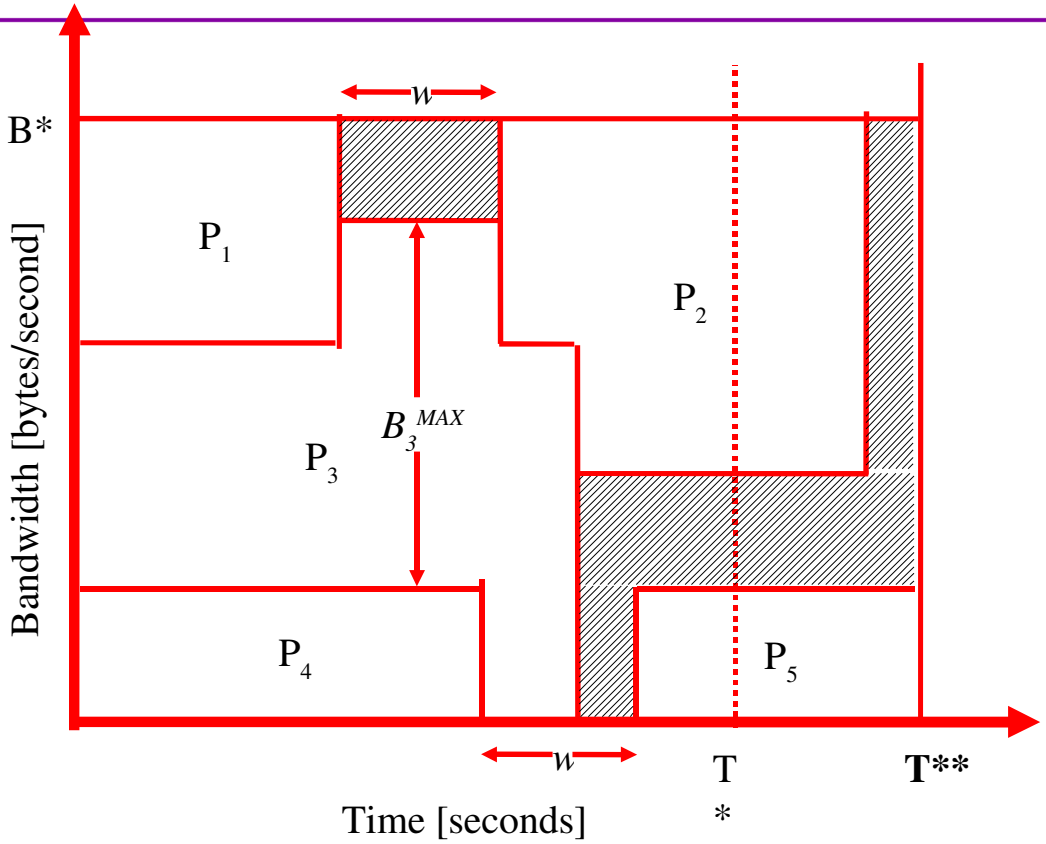
Crawling

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

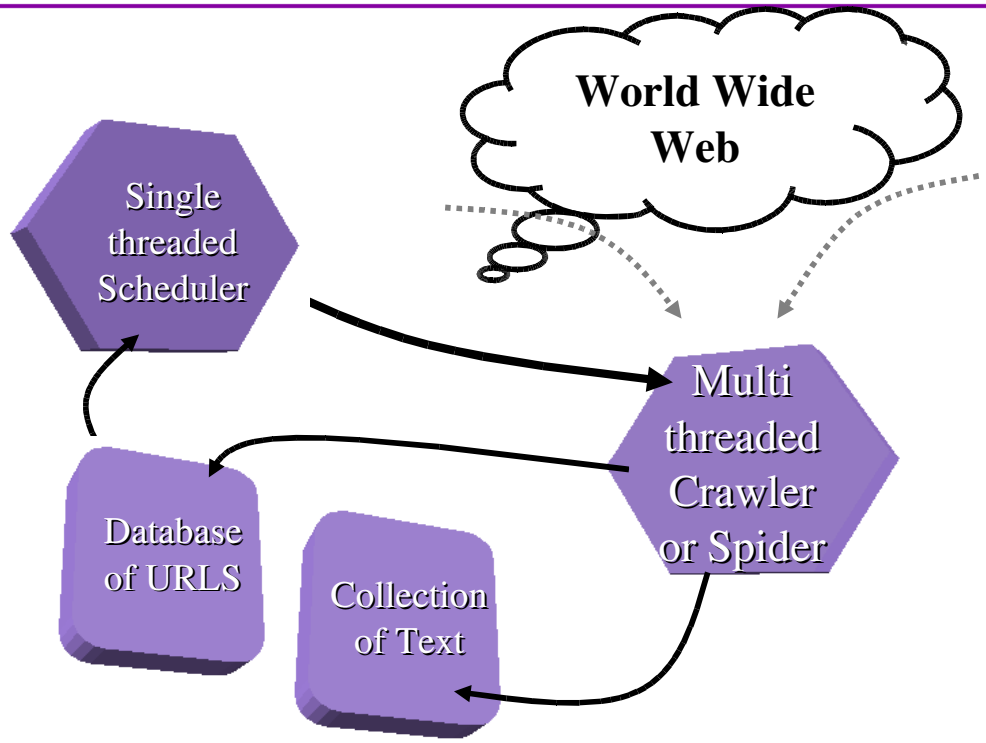
Crawling Goals

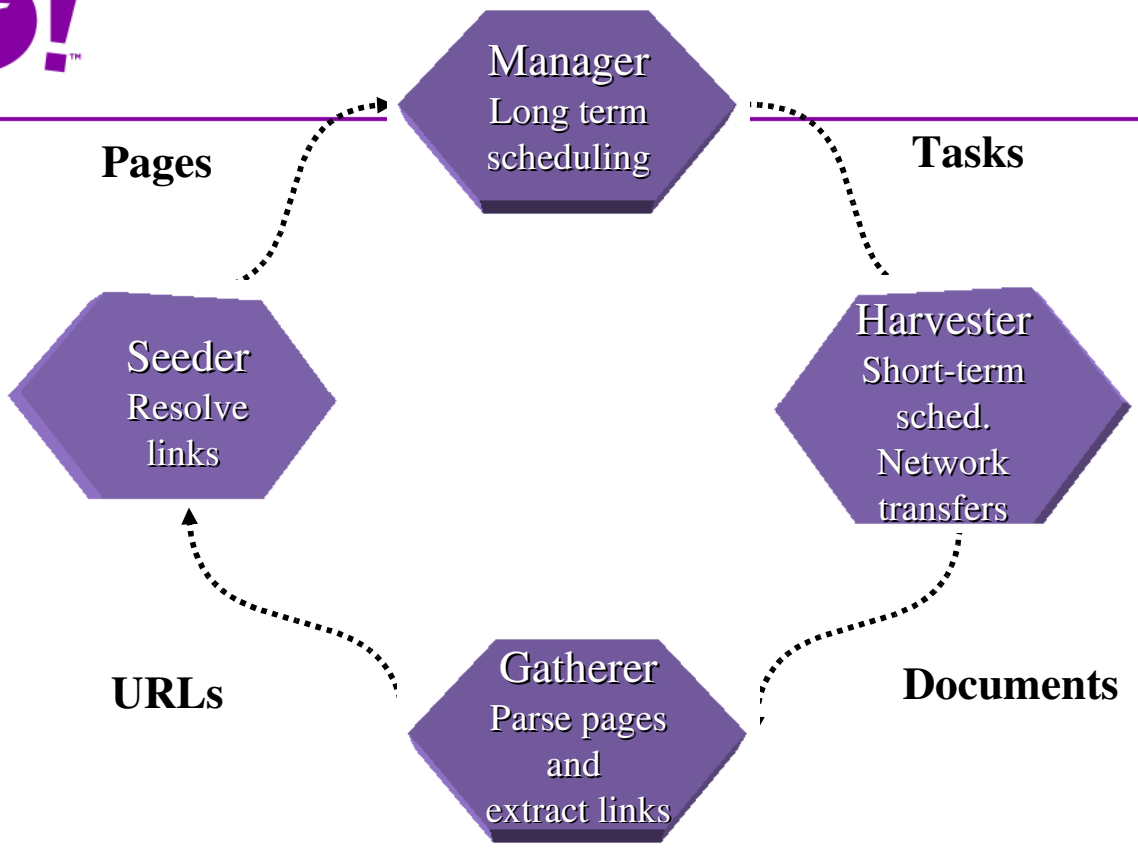



B*
$P_1 = T^* \times B_1$
$P_2 = T^* \times B_2$
$P_3 = T^* \times B_3$
T*
$P_4 = T^* \times B_4$
$P_5 = T^* \times B_5$
Bandwidth [bytes/second]
Time [seconds]


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

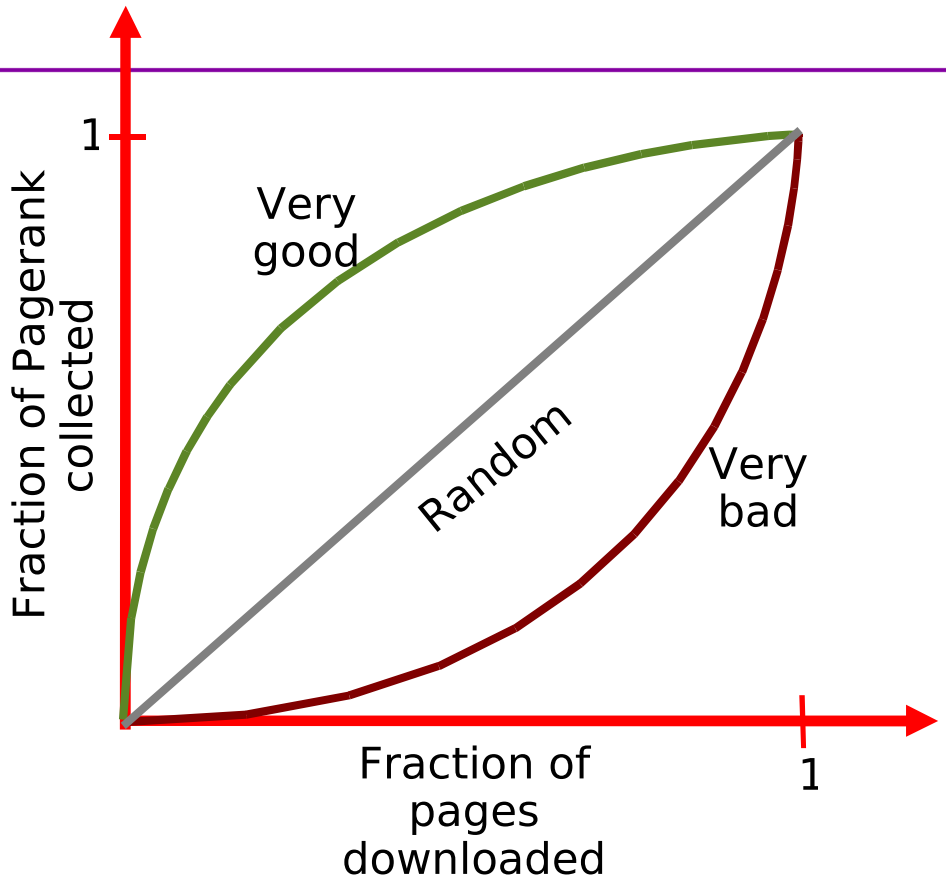
Software Architecture



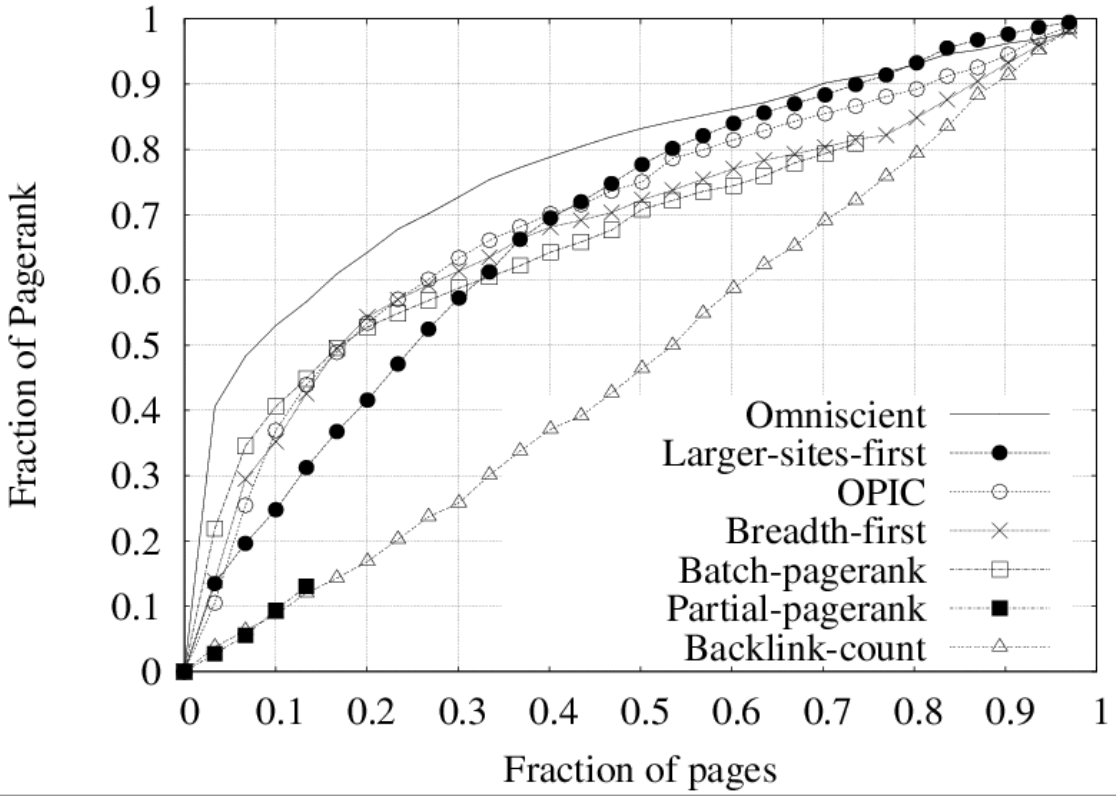


Crawling Heuristics

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

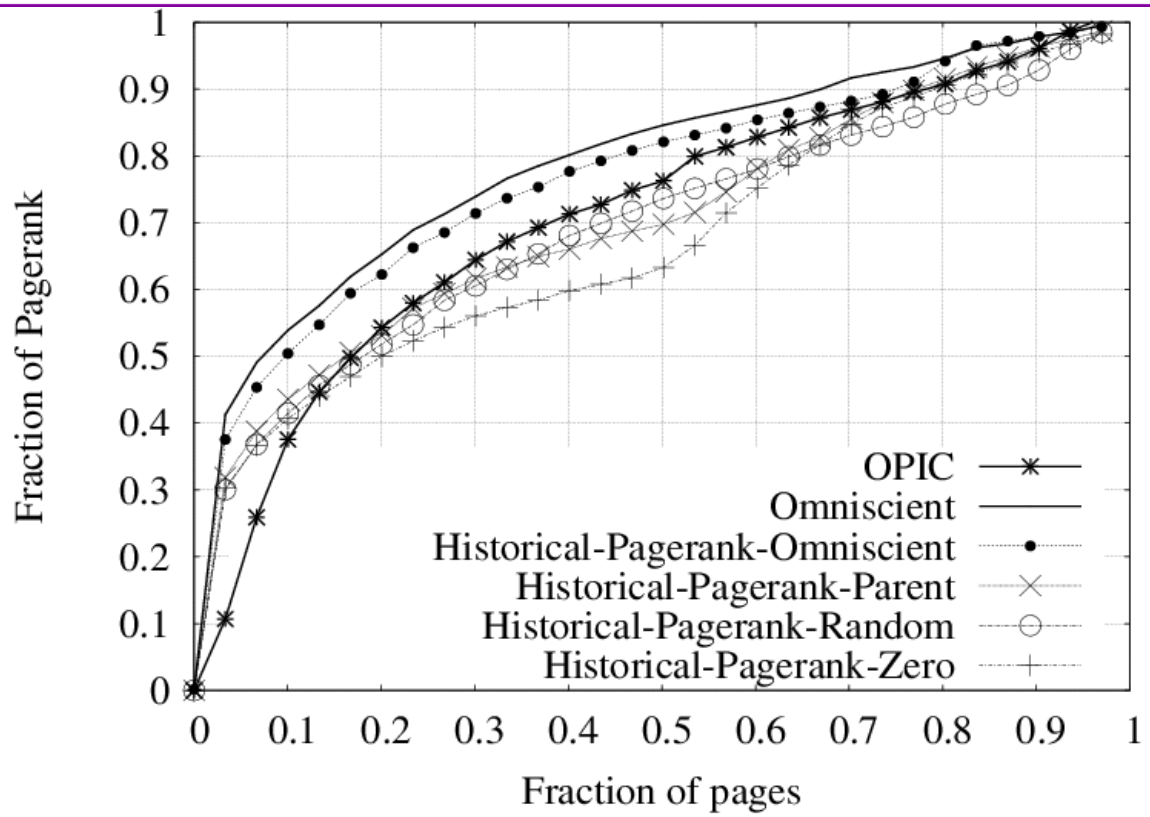


No Historical Information

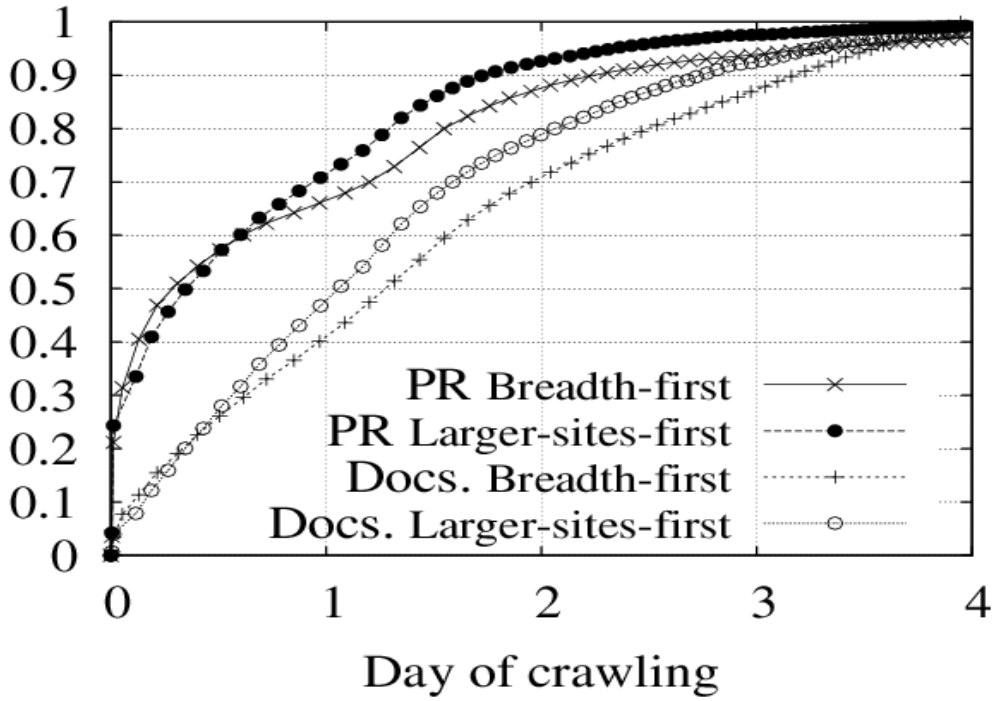


Baeza-Yates, Castillo, Marin & Rodriguez, WWW2005

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

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

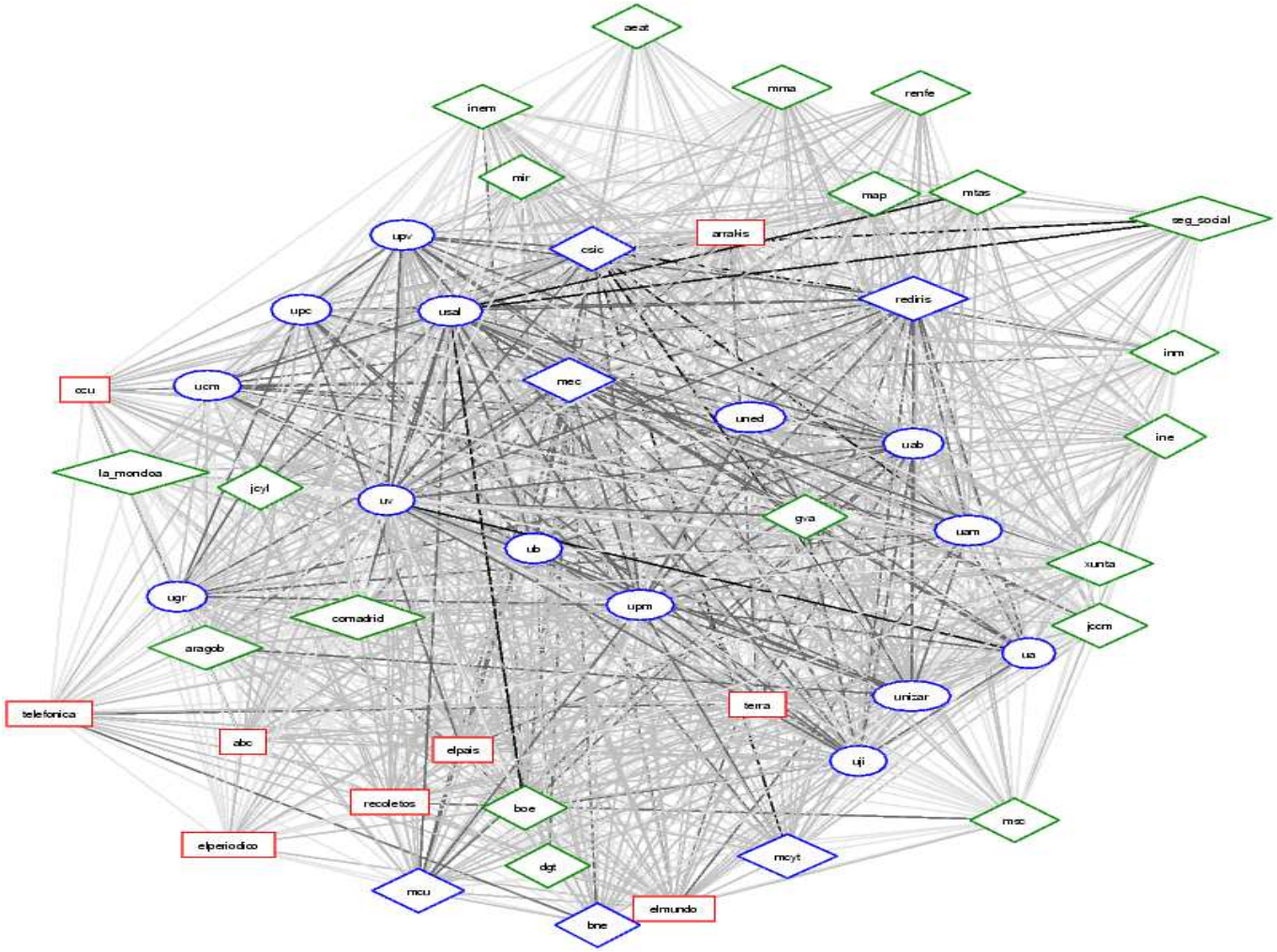
- 20 Pb of storage (20M Gb)
  - – US Library of congress every day (28M books, 20TB)
- **12 Tb of data generated per day**
- 7 billion song ratings
- 2 billion photos stored
- 2 billion Mail+Messenger sent per day

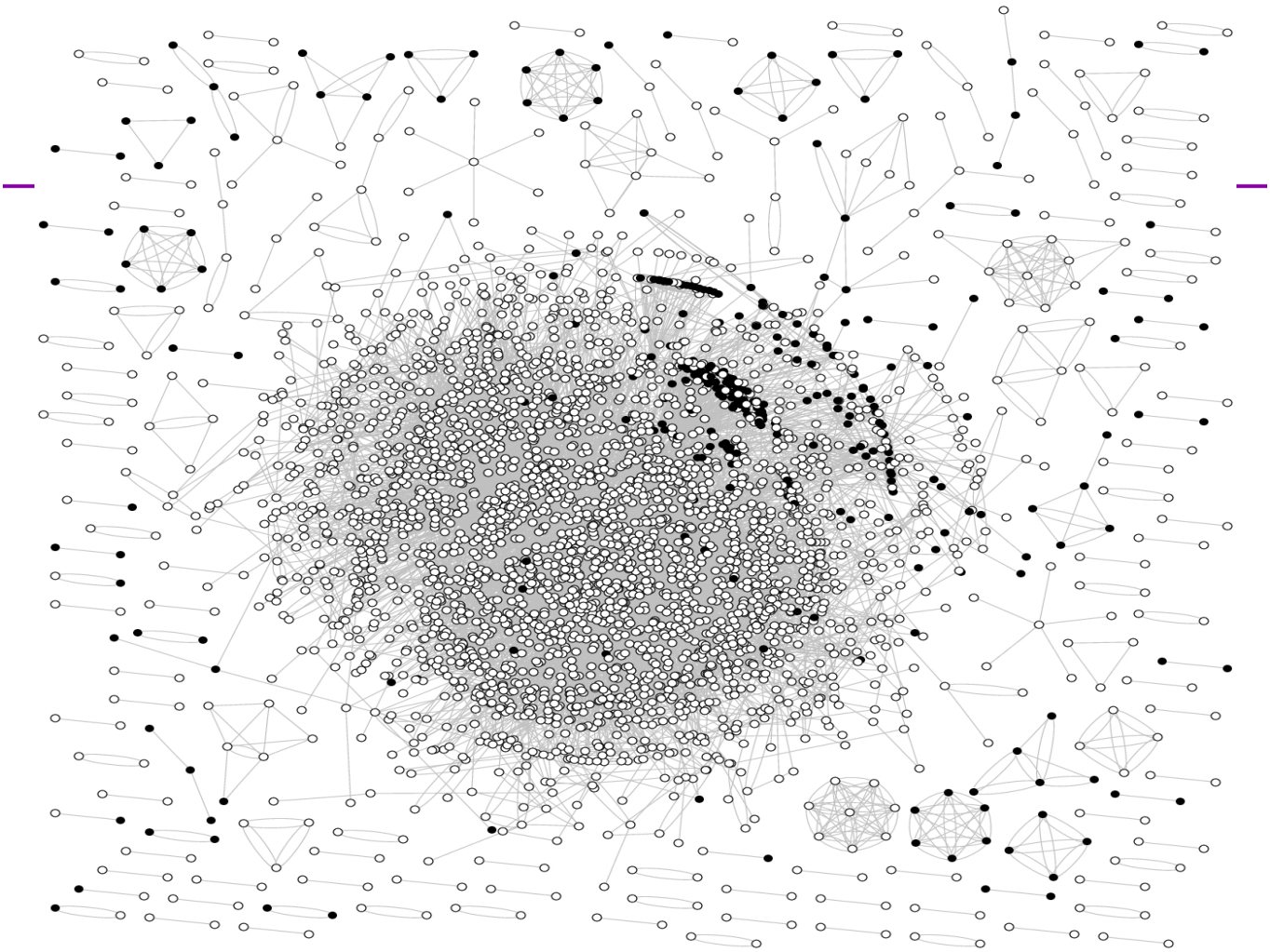
Web Characterization

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

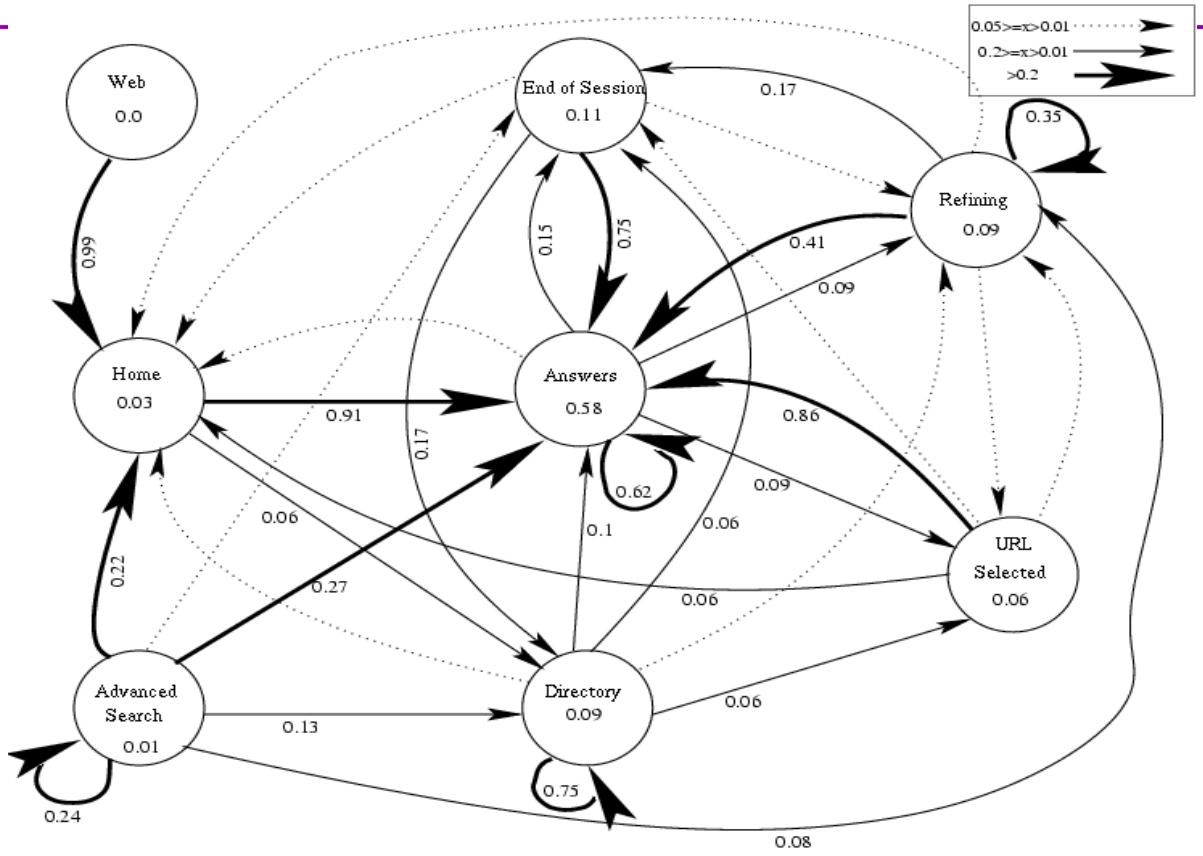
A Few Examples

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

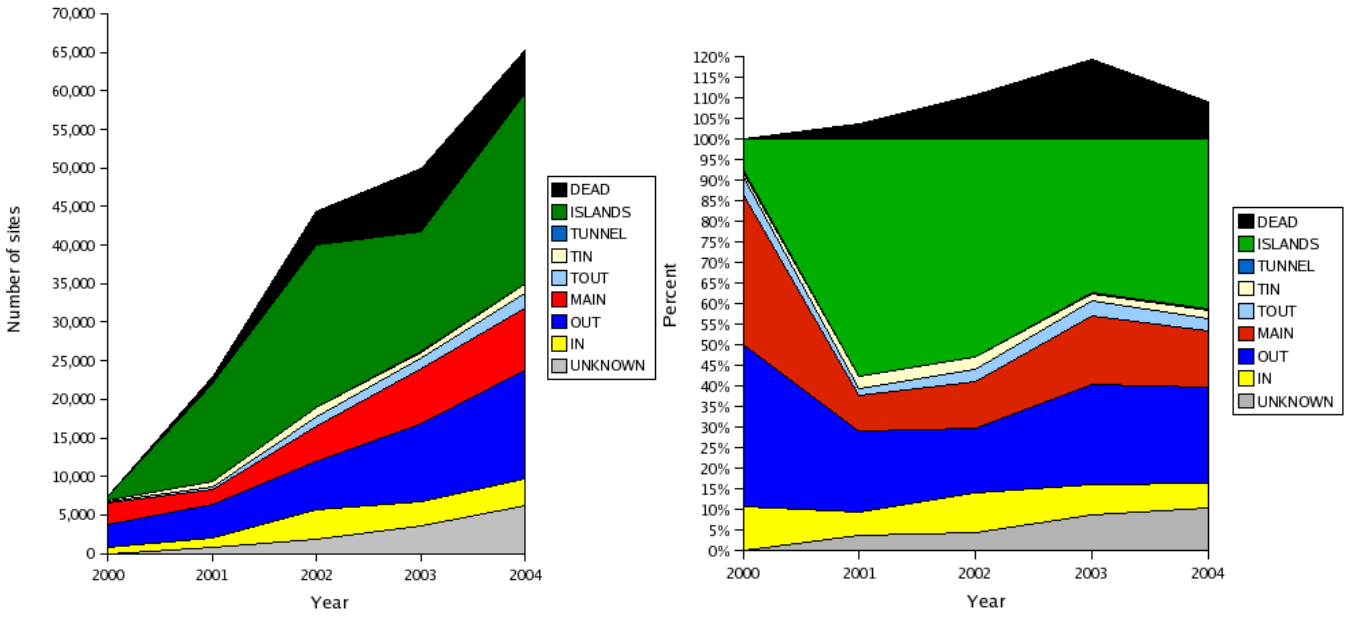

aeat
inem
mma
renfe
mir
map
mtas
seg_social
upv
csic
anaitis
upc
usal
rediris
inm
ocu
ucm
mec
uned
uab
ine
la_mondoa
jcyl
uv
gva
uam
xunta
ub
ugr
upm
comadrid
jccm
aragob
ua
telefonica
terra
unizar
abc
elpais
uji
recoletos
boe
msc
elperiodico
dgt
mcyt
mcu
elmundo
bne



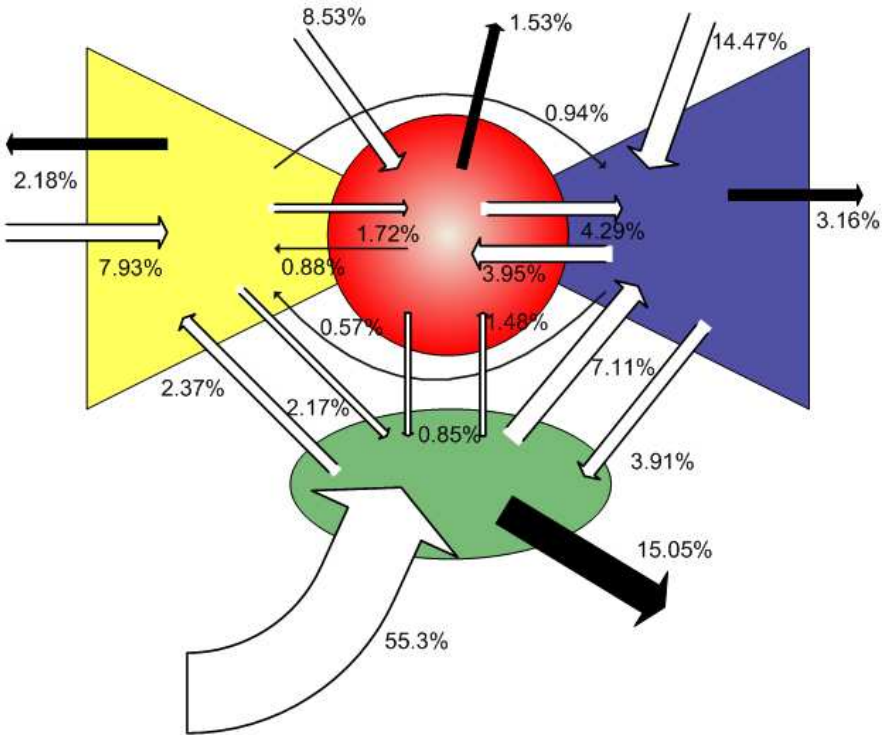
User Modeling



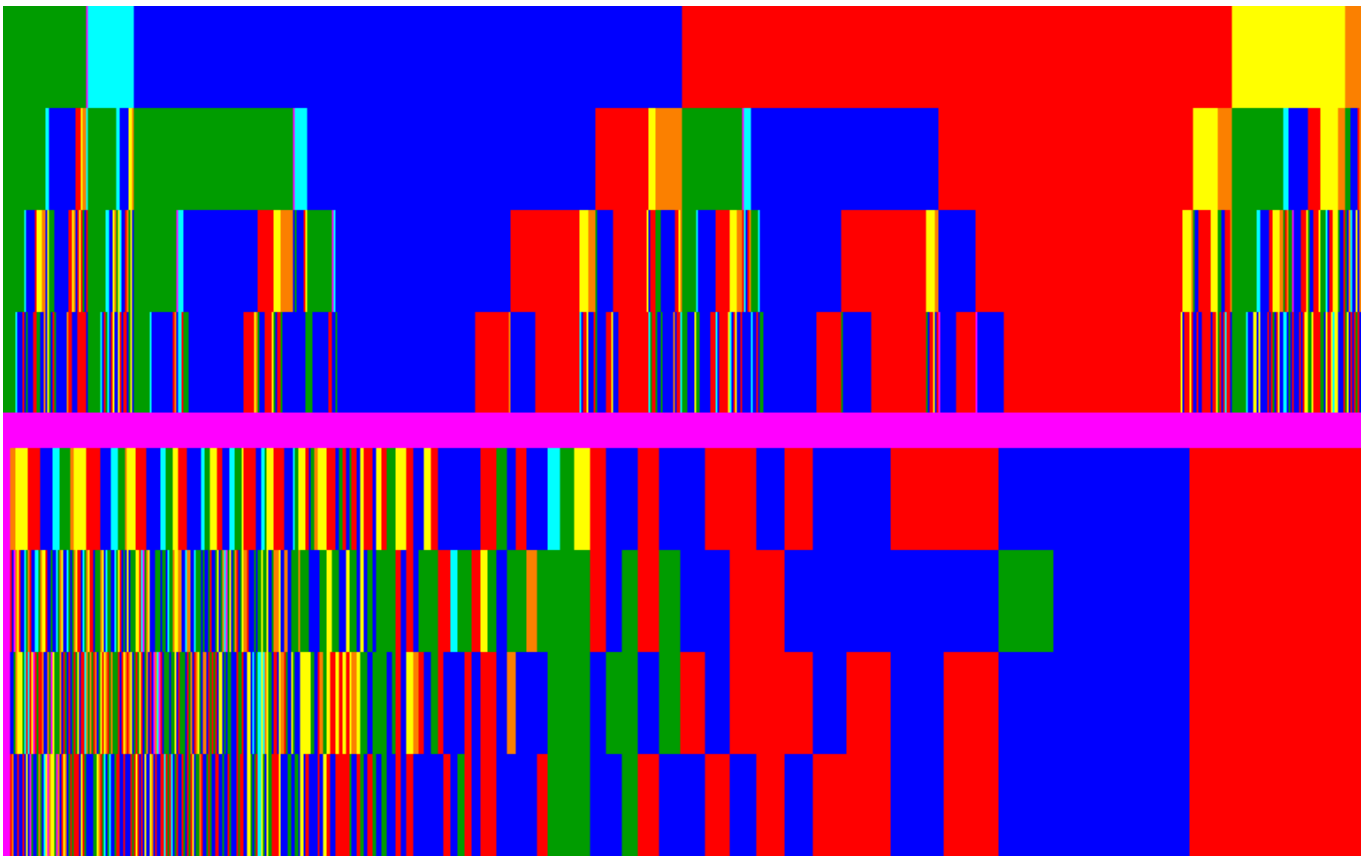
Size Evolution



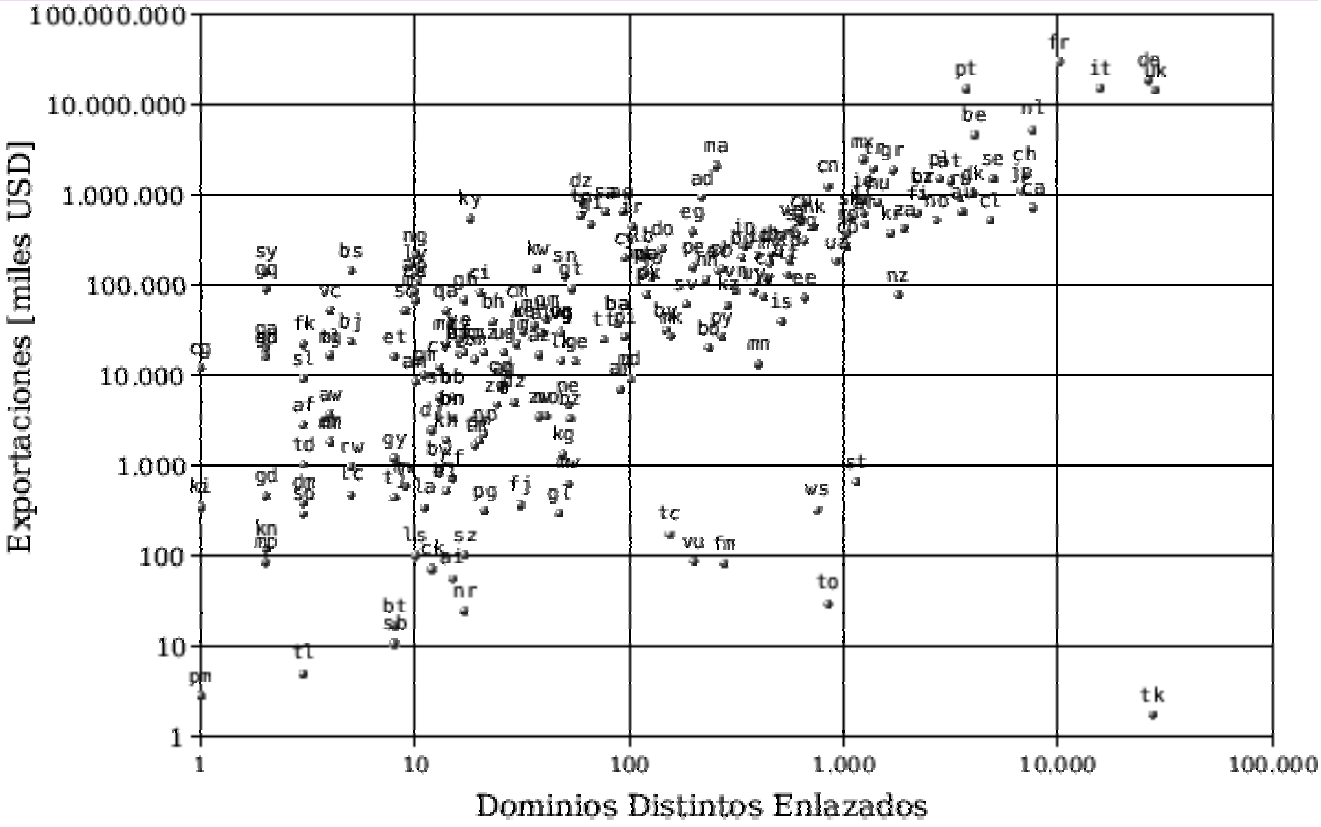
Structure Macro Dynamics



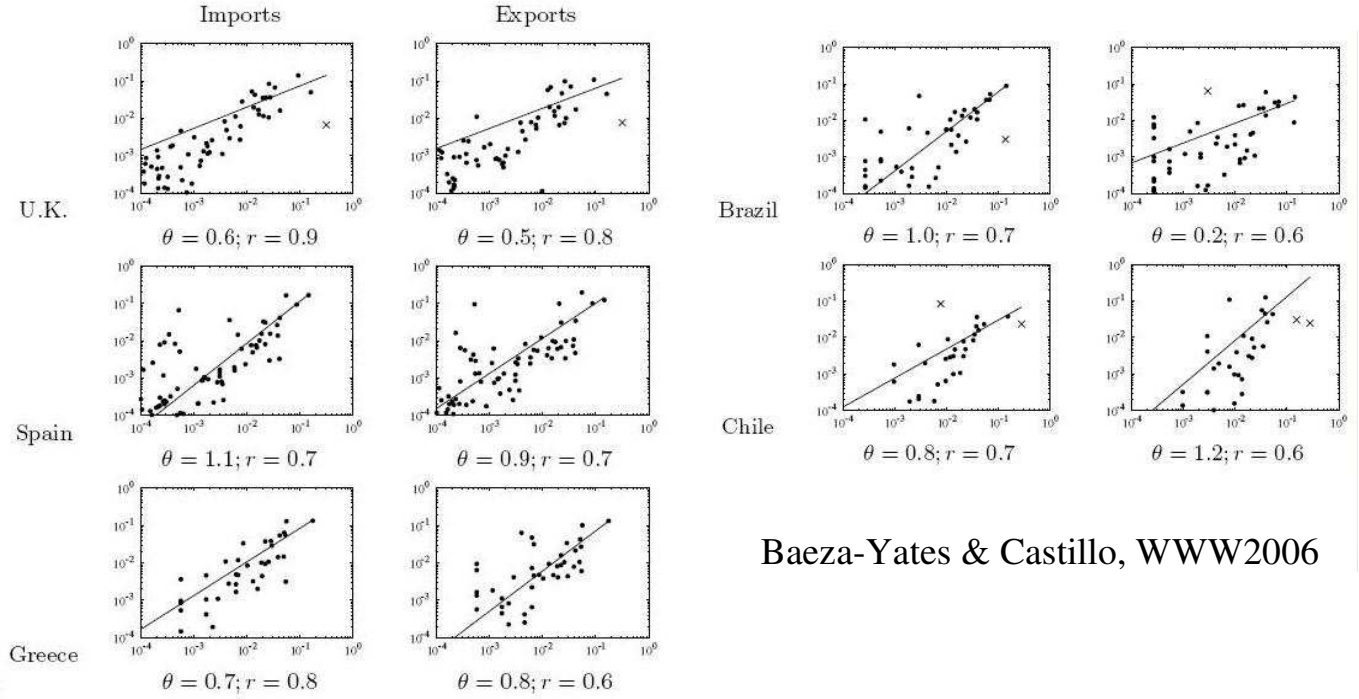
Structure Micro Dynamics



Mirror of the Society



Exports/Imports vs. Domain Links



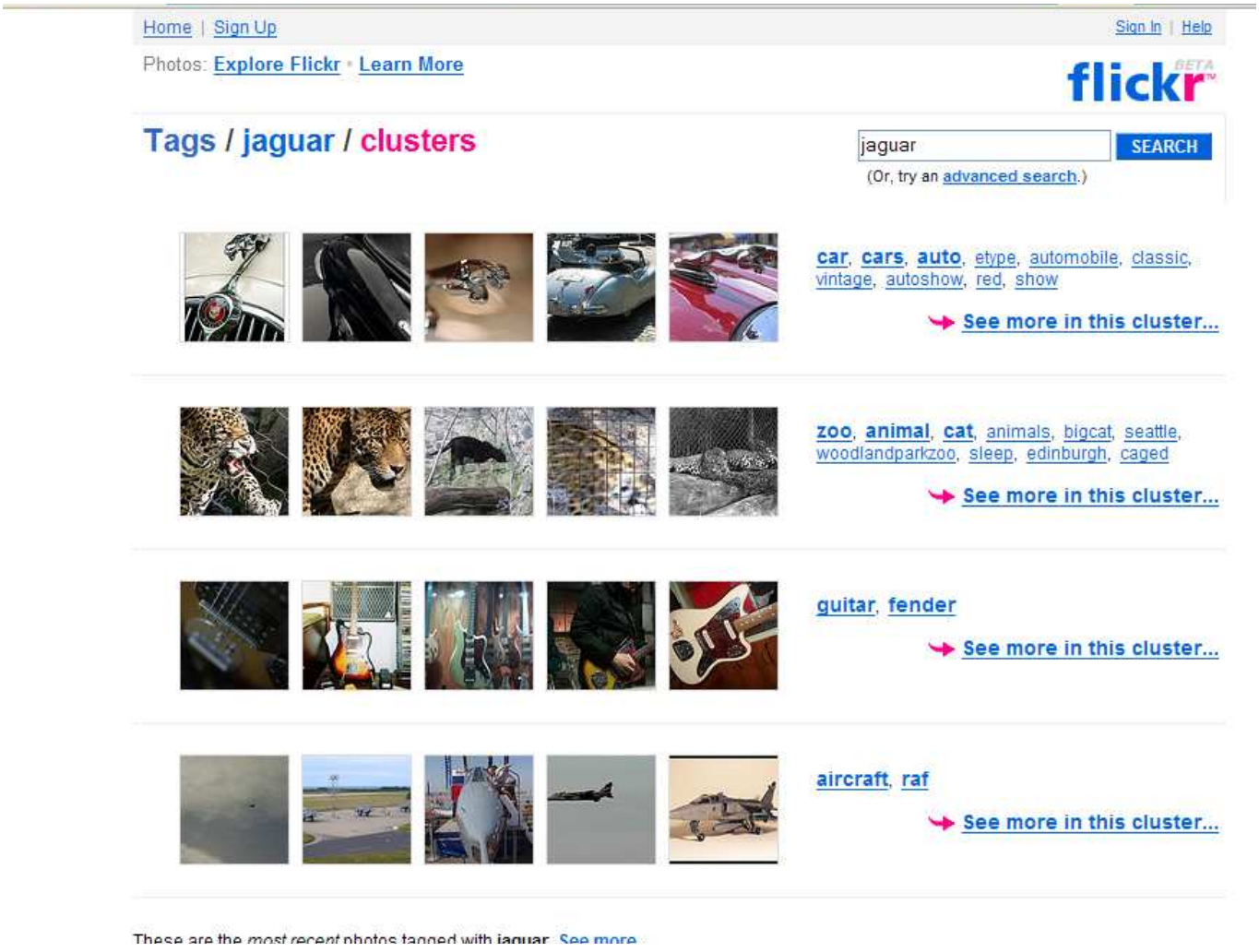
Baeza-Yates & Castillo, WWW2006

The Wisdom of Crowds

- James Surowiecki, a ***New Yorker*** columnist, published this book in 2004
  - **"Under the right circumstances, groups are remarkably intelligent"**
- Importance of diversity, independence and decentralization

Aggregating data

"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".

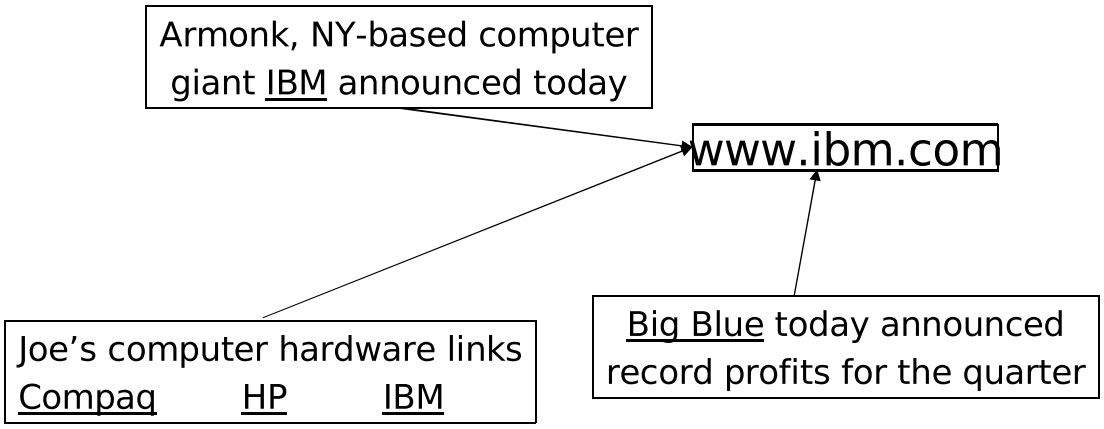


The Power of Social Media

- 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?

Anchor Text

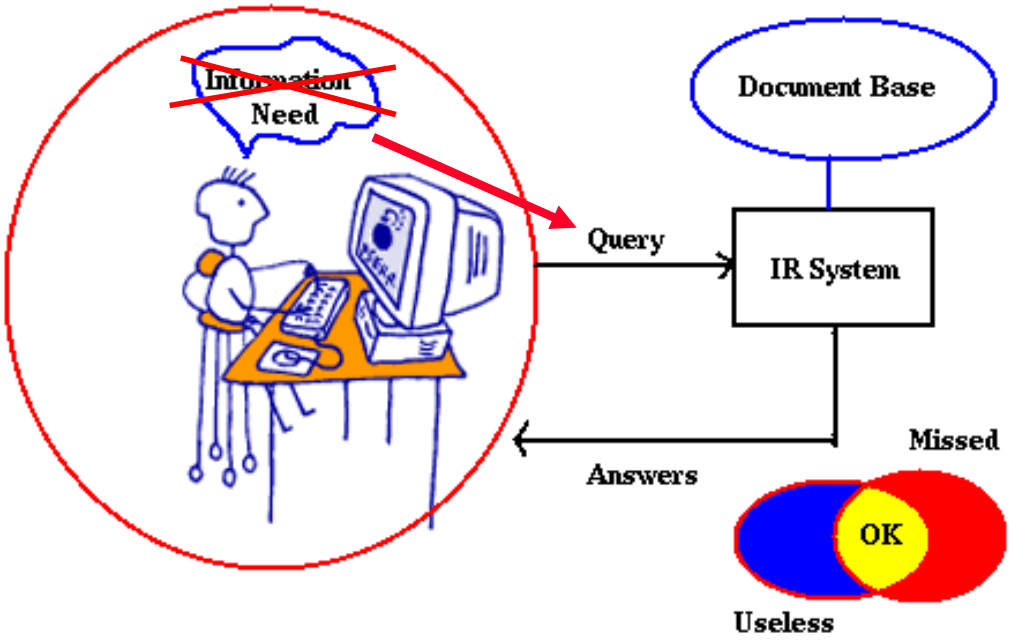
- The wisdom of anchor text:
  - when indexing a document *D*, include **anchor text** from links pointing to *D*



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!)

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%)

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

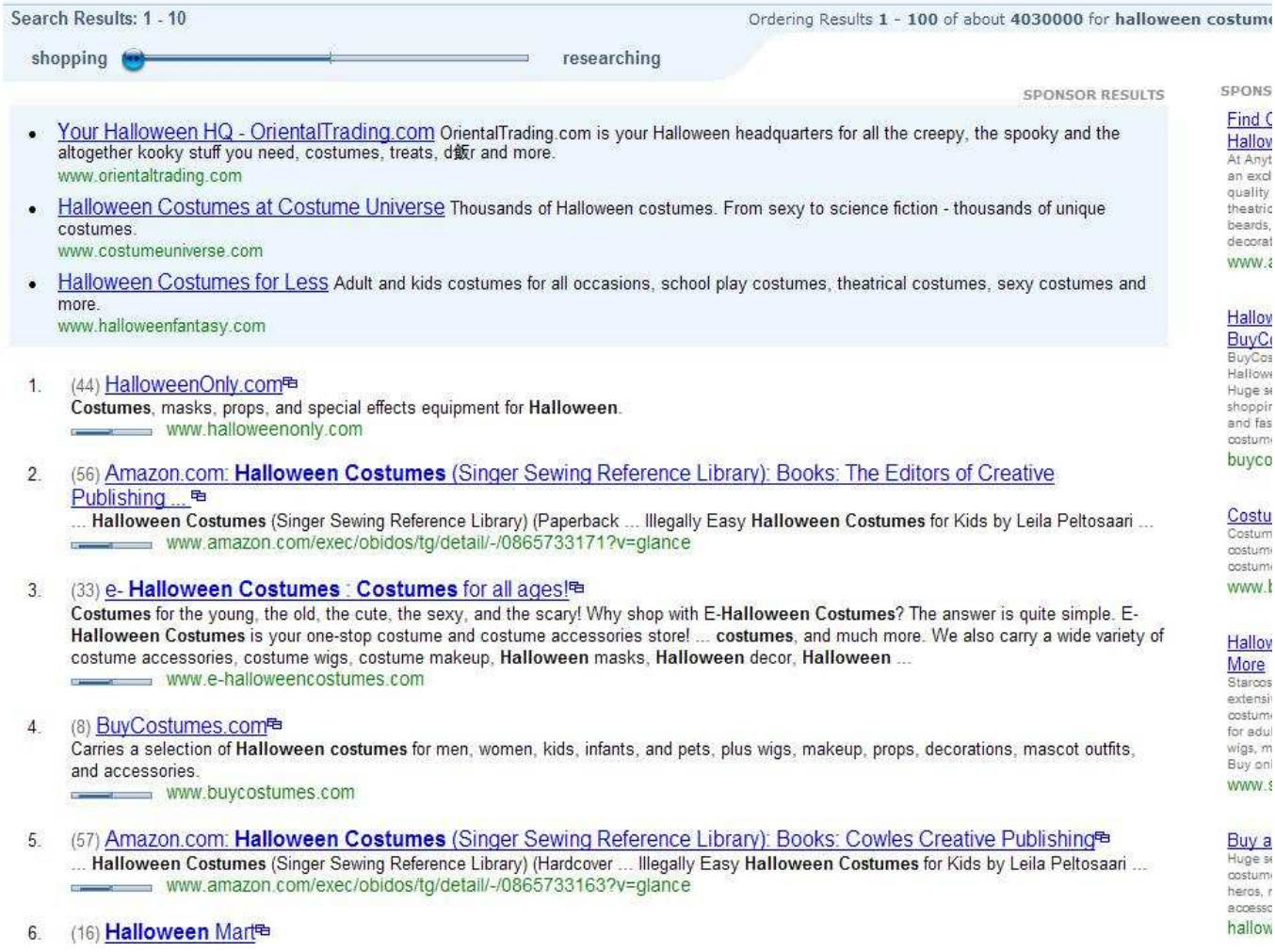
      `Bangalore weather`
    - Downloads

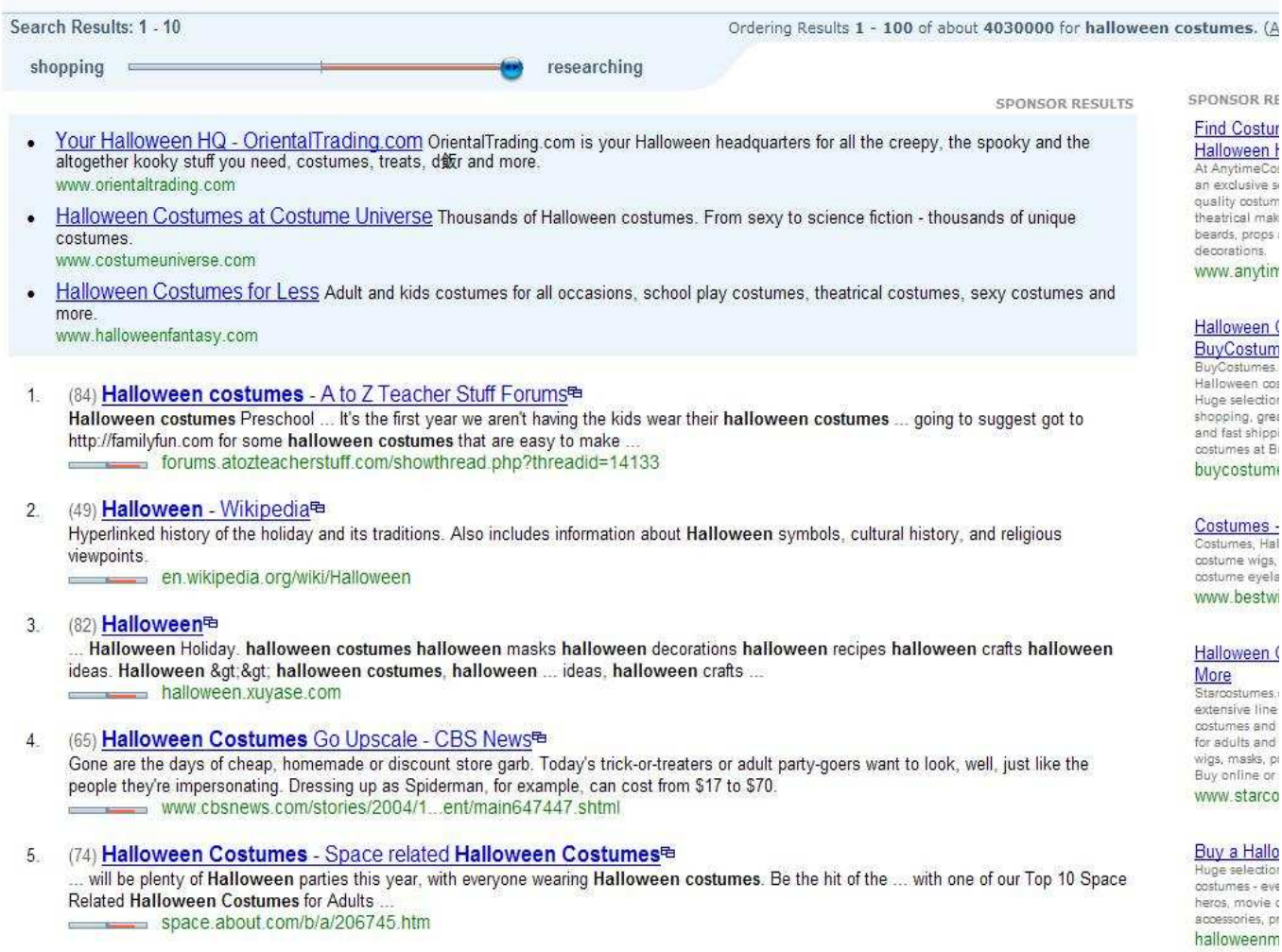
      `Mars surface images`
    - Shop

      `Digital camera`
  - Gray areas

    `Car rental Goa`
    - Find a good hub
    - Exploratory search "see what's there"


rted Latest Headlines
halloween costumes
Search Web
My Web
Bookmarks
My Yahoo!
Yahoo!
Finance
Mail
News
Yahoo! Re
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.
Privacy Policy - Terms of Service - Copyright/IP Policy - Submit Your Site
Copyright © 2005 Yahoo! Inc. All rights reserved.

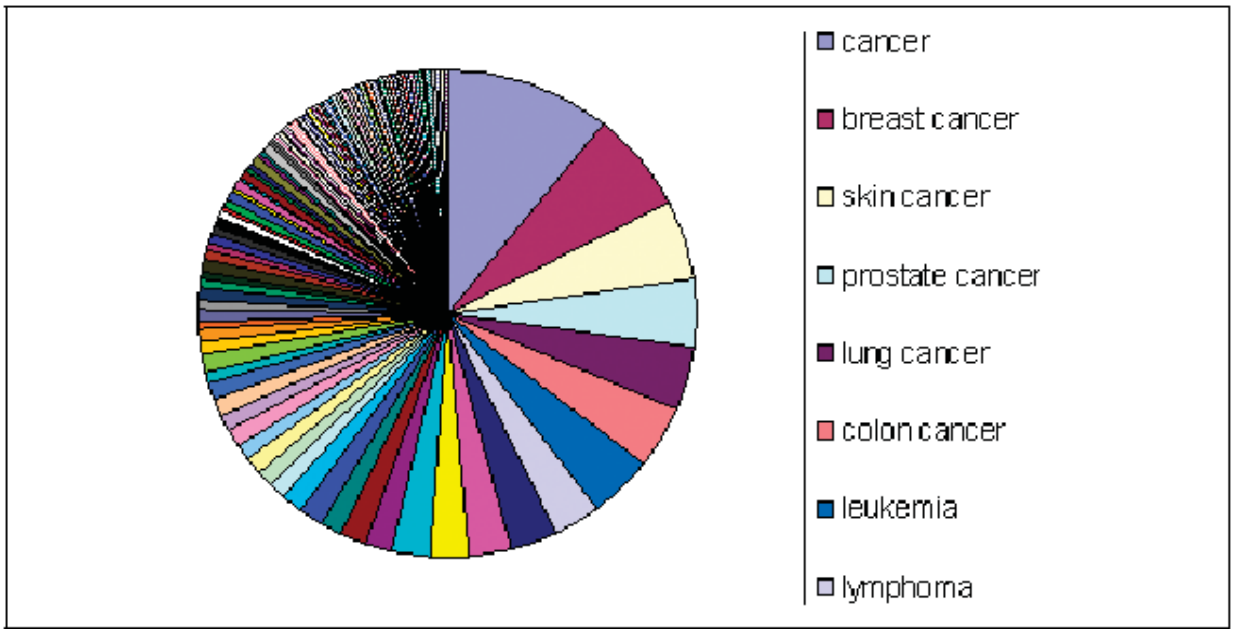

Search Results: 1 - 10
Ordering Results 1 - 100 of about 4030000 for halloween costume
shopping
researching
SPONSOR RESULTS
Your Halloween HQ - OrientalTrading.com OrientalTrading.com is your Halloween headquarters for all the creepy, the spooky and the altogether kooky stuff you need, costumes, treats, 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
SPONS
Find
Hallov
At Anyt
an excl
quality
theatric
beards,
decorat
www.a
Hallov
BuyCo
BuyCos
Hallowe
Huge se
shoppir
and fas
costum
buyco
Costu
Costum
costum
costum
www.b
Hallov
More
Starcos
extensi
costum
for adu
wigs, m
Buy onl
www.s
Buy a
Huge se
costum
heros, r
accesso
hallow



Mining Queries for ...

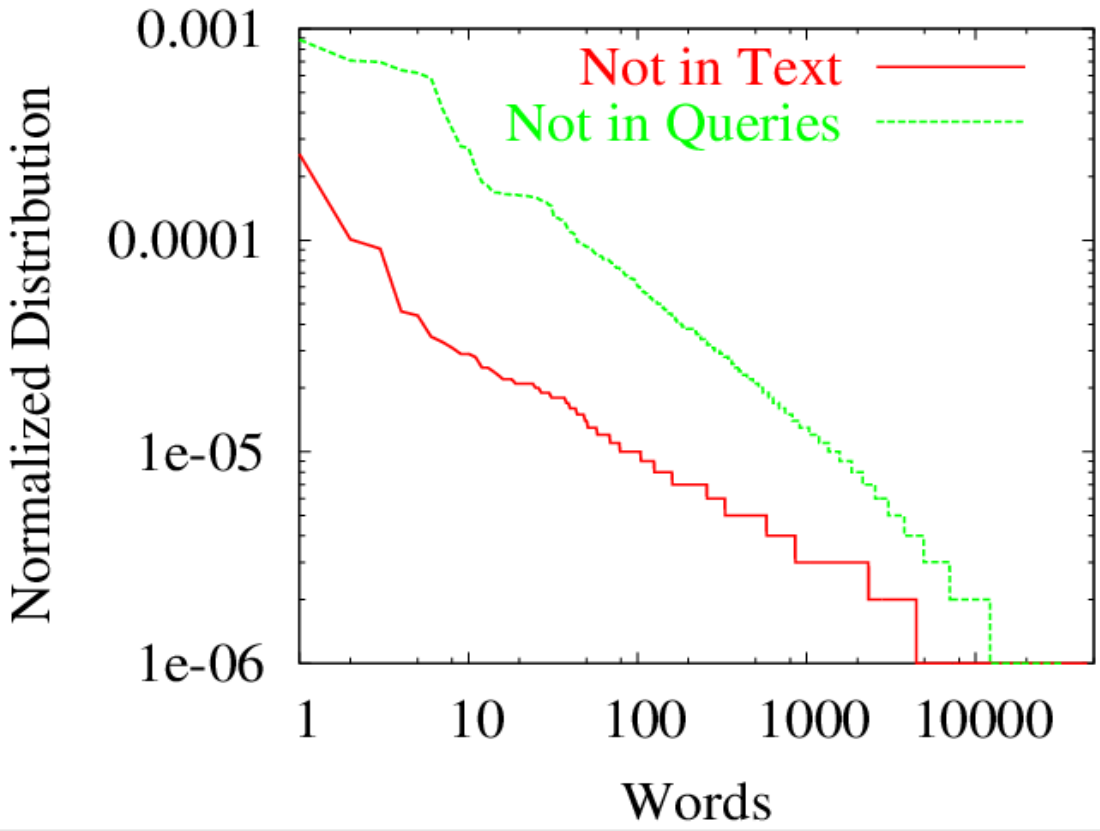
- Improved Web Search: index layout, ranking
- User Driven Design
  - The Web Site that the Users Want
  - The Web Site that You should Have
  - Improve content & structure
- Bootstrap of pseudo-semantic resources

Query Distribution



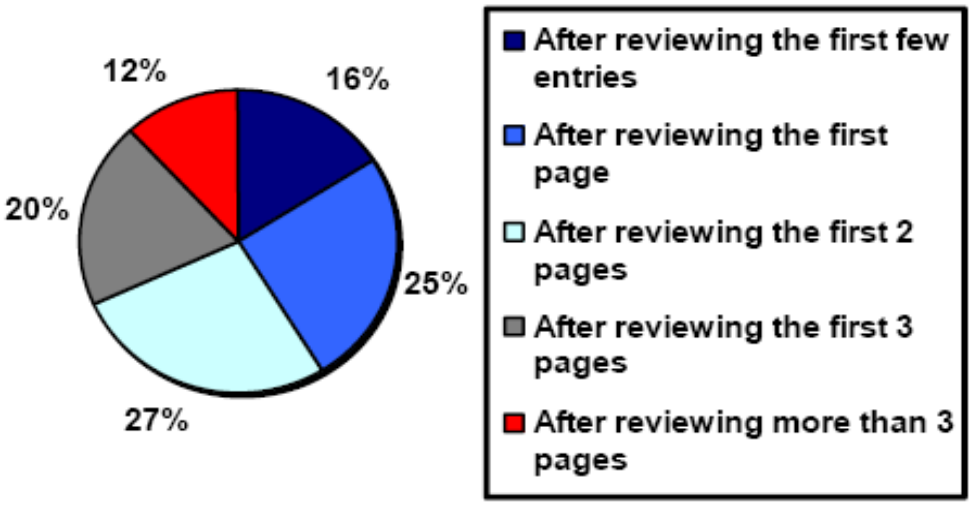
Power law: few popular broad queries, many rare specific queries

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

MP3

games

cars

famous person

pictures

- What is the goal?

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.

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	**Home page**
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	**Hub page**
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

Page with resources

Rose & Levinson 2004

Y! 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*

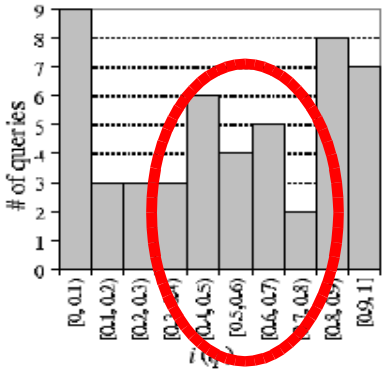


Figure 1: Query distribution along the $i(q)$ axis

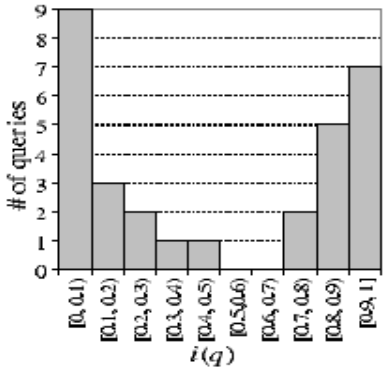


Figure 2: After removing software and person-name queries

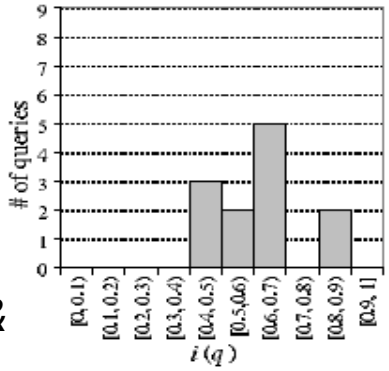


Figure 3: Distribution of the 12 software queries

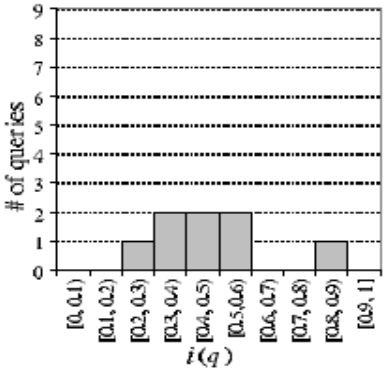
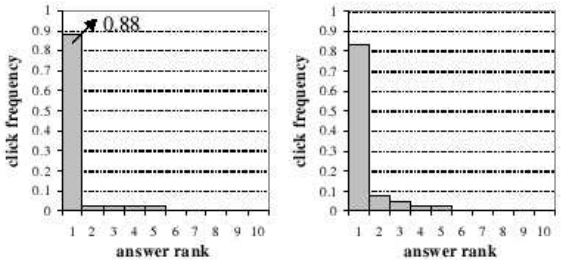


Figure 4: Distribution of the 8 person-name queries

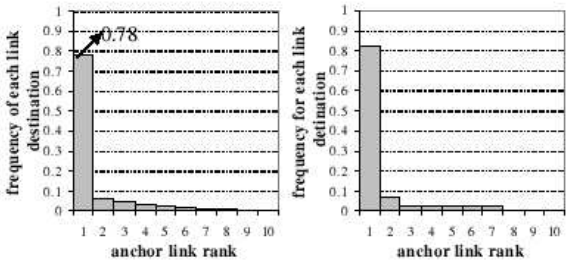
Y! Features

- **Click & anchor text distribution**



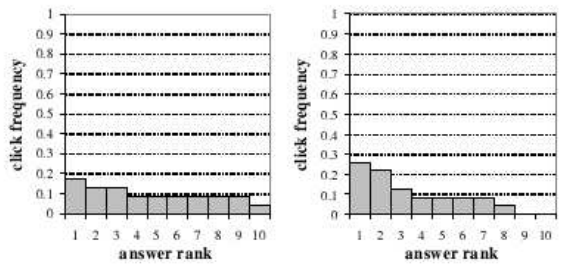
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 5: Click distributions for sample navigational queries



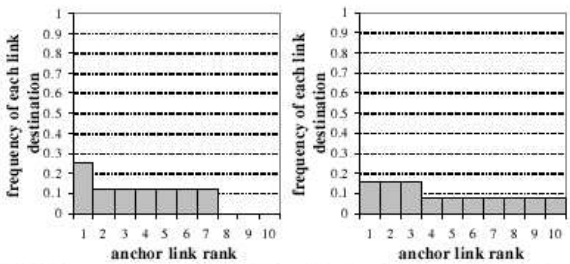
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

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

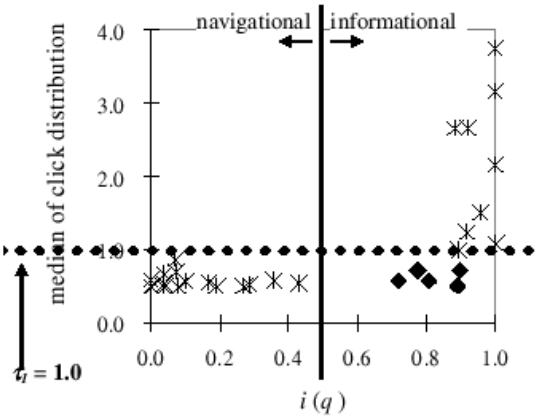


Figure 11: Median of click distribution

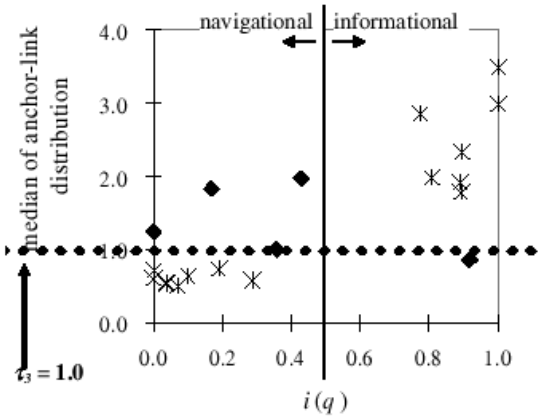


Figure 13: Median of anchor-link distribution

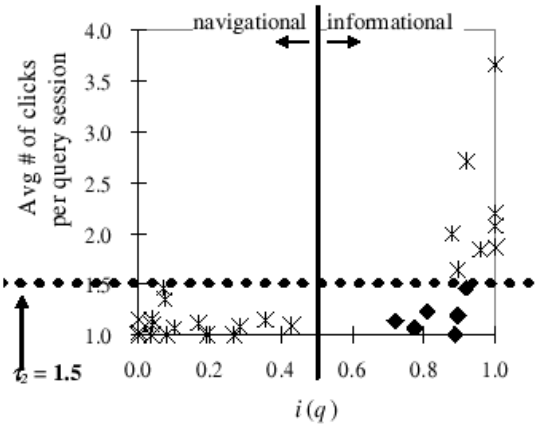


Figure 12: Avg # of clicks per query

Prediction power:

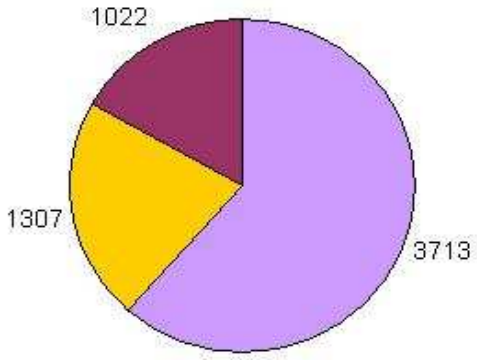
- Single features: 80%
- Mixed features: 90%

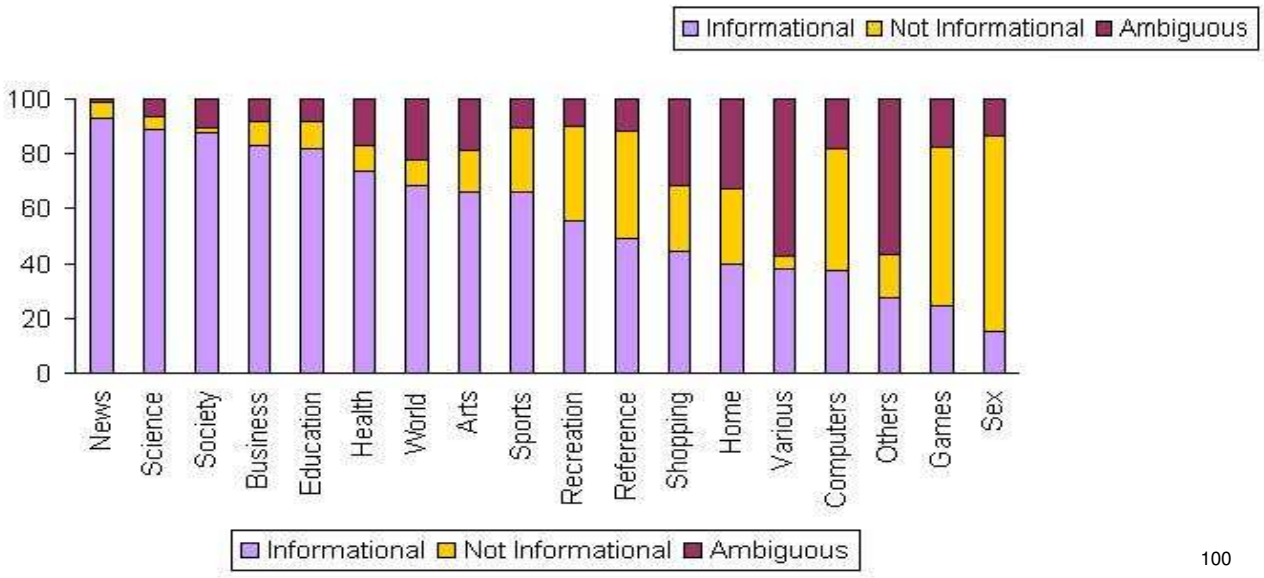
- Drawbacks: Small evaluation, a posteriori feature

User Intention

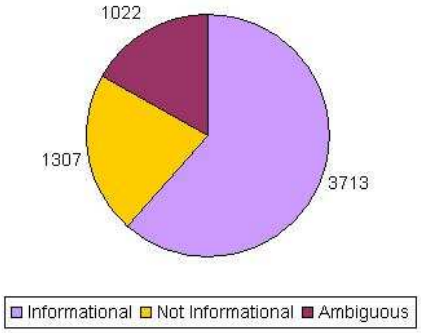
- Manual classification of more than 6,000 popular queries
- Query Intention & topic
- Classification & Clustering
- Machine Learning on all the available attributes
- Baeza-Yates, Calderon & Gonzalez (SPIRE 2006)

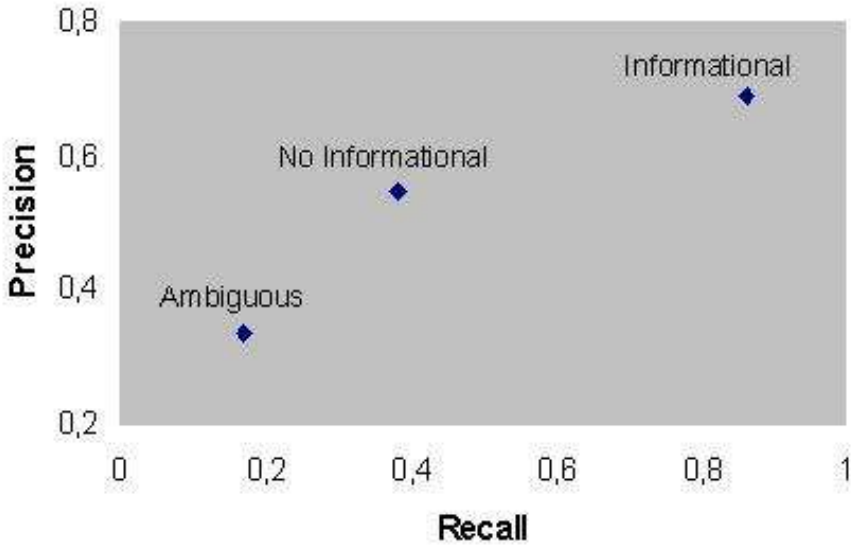
Classified Queries





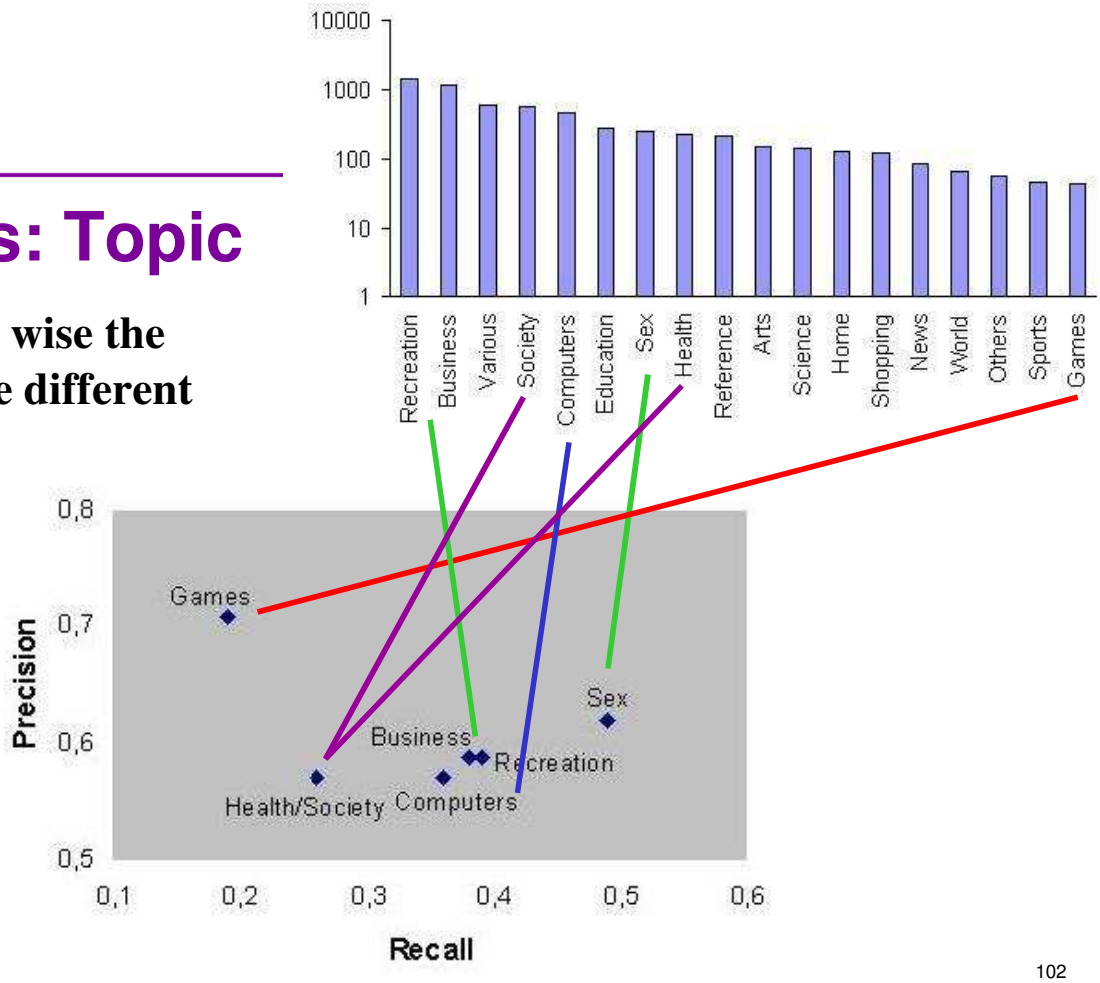
Results: User Intention

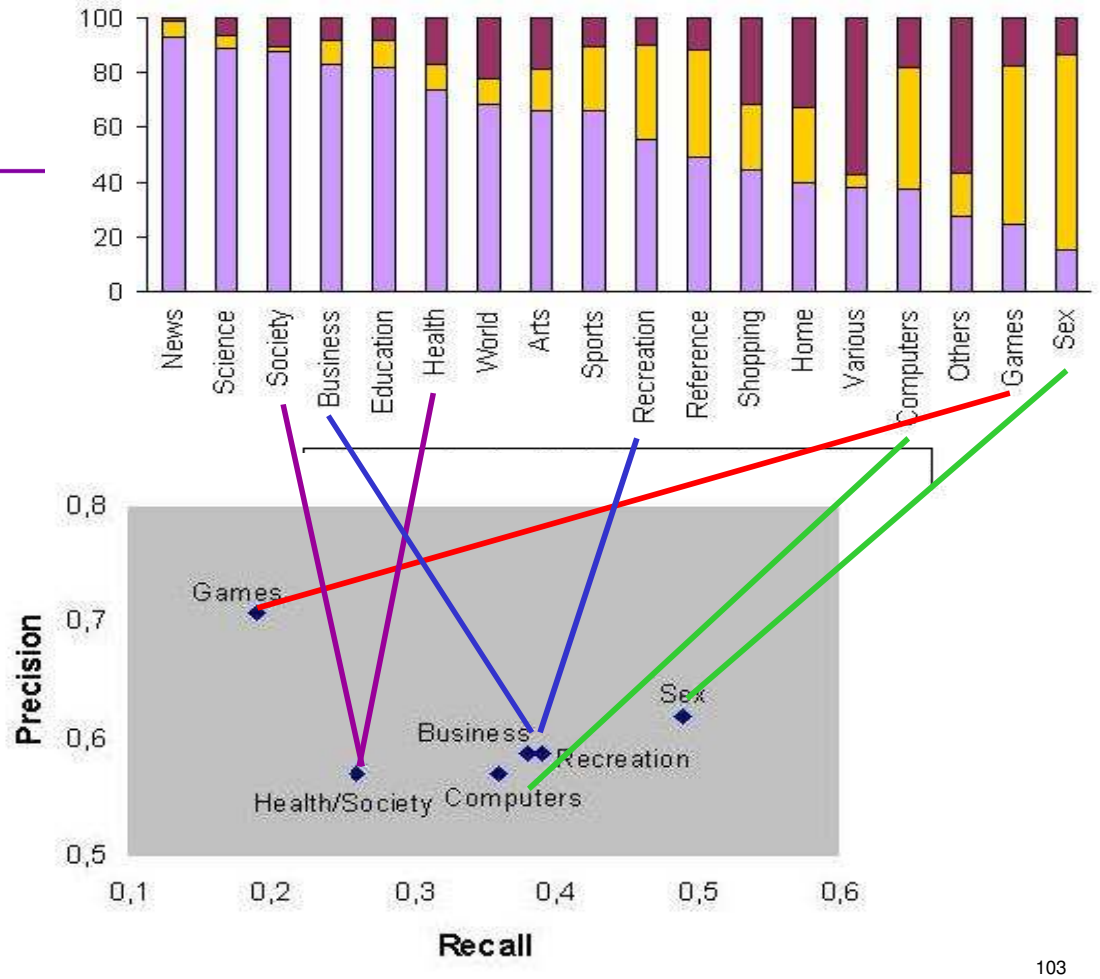




Results: Topic

- **Volume wise the results are different**



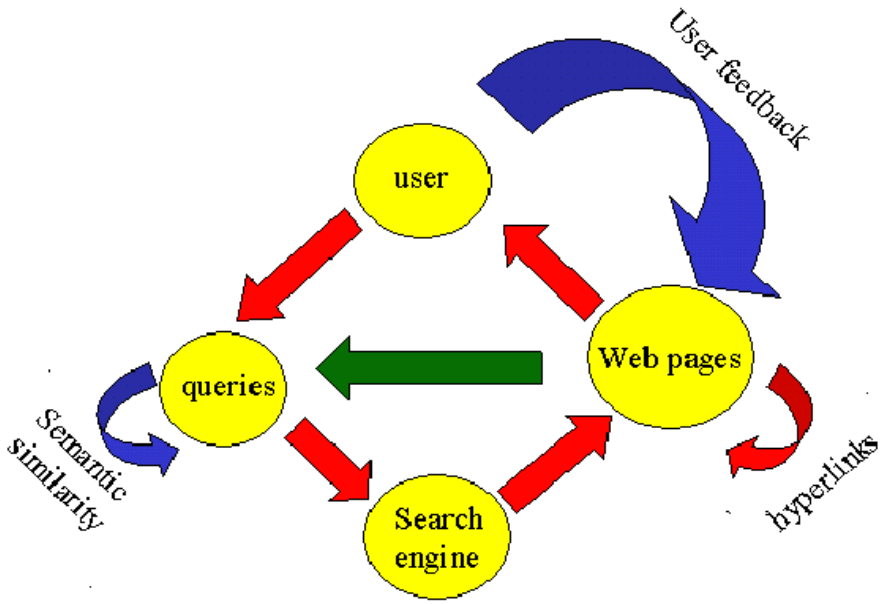


Clustering Queries

- 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

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

- 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

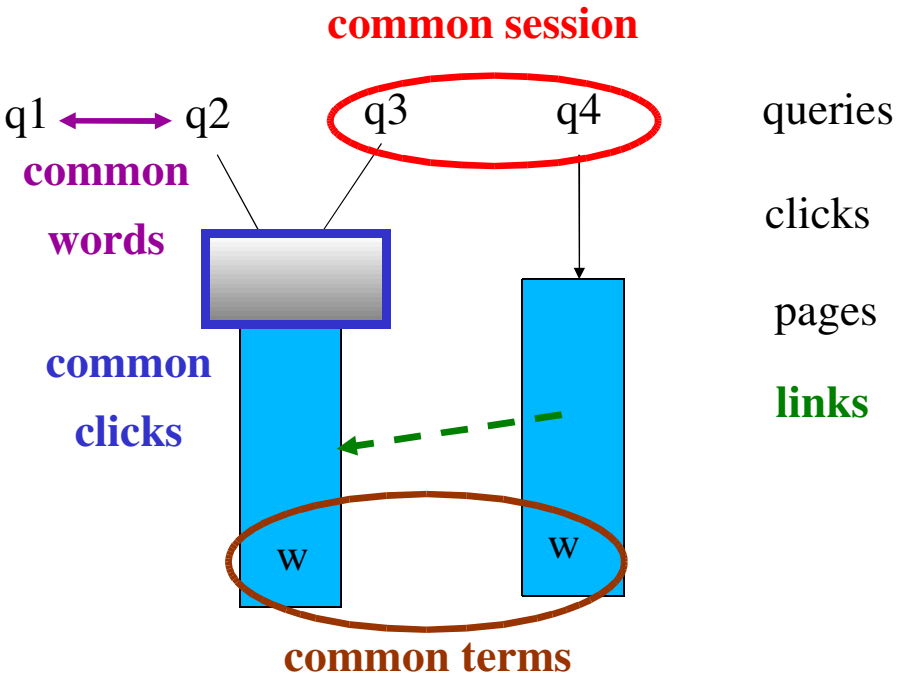
Journal of ASIST 2007

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

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

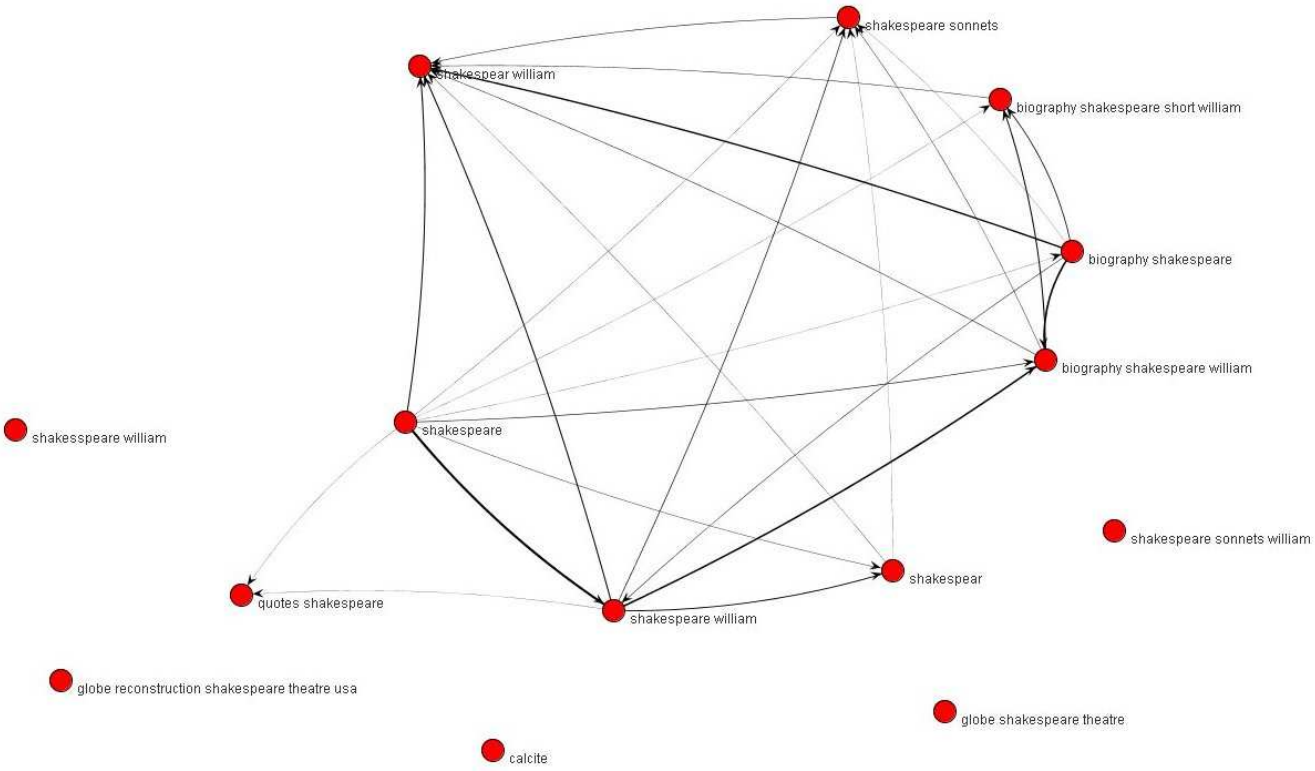
Relating Queries (Baeza-Yates, 2007)



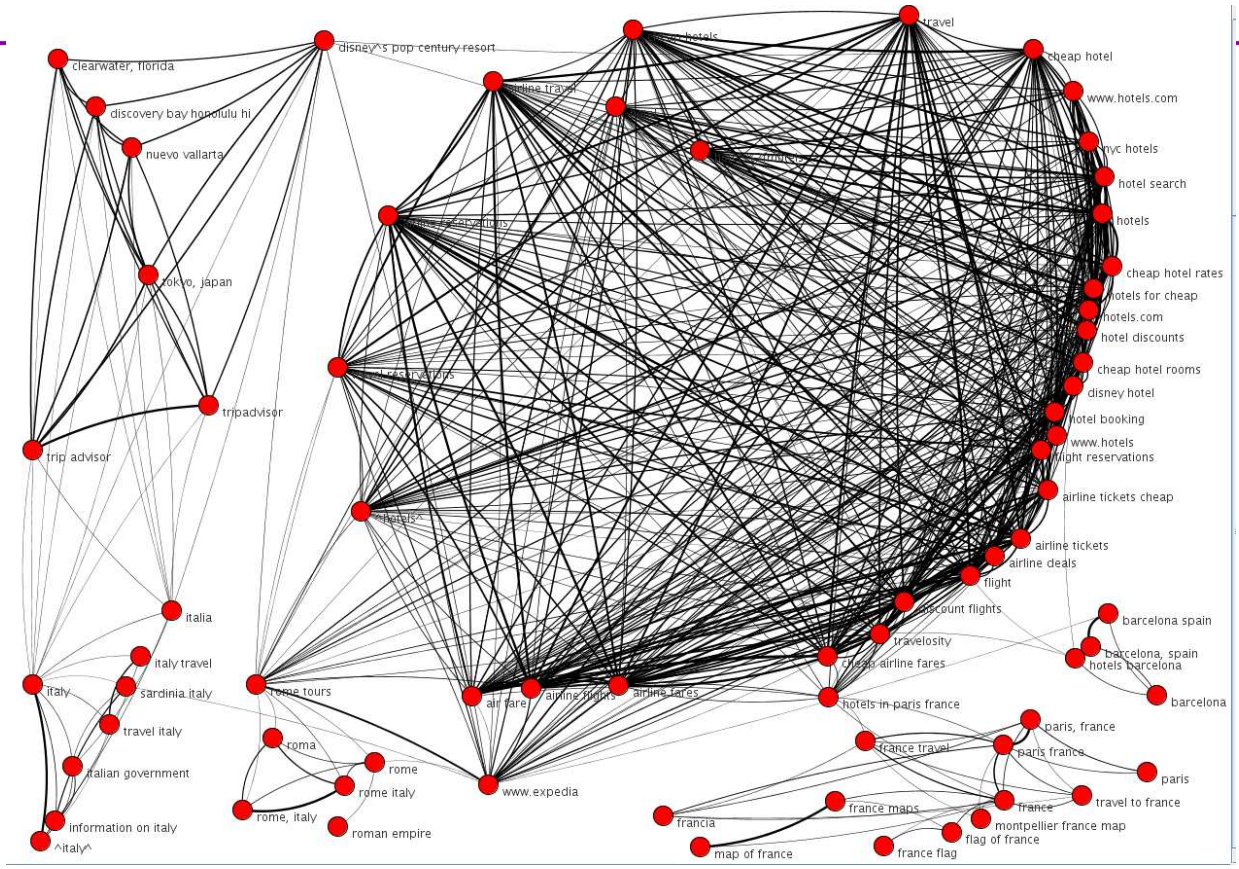
Qualitative Analysis

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



Click Graph



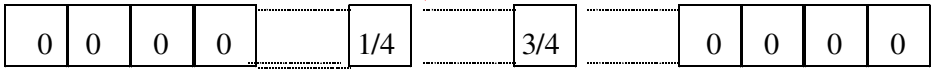
Formal Definition

- 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

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

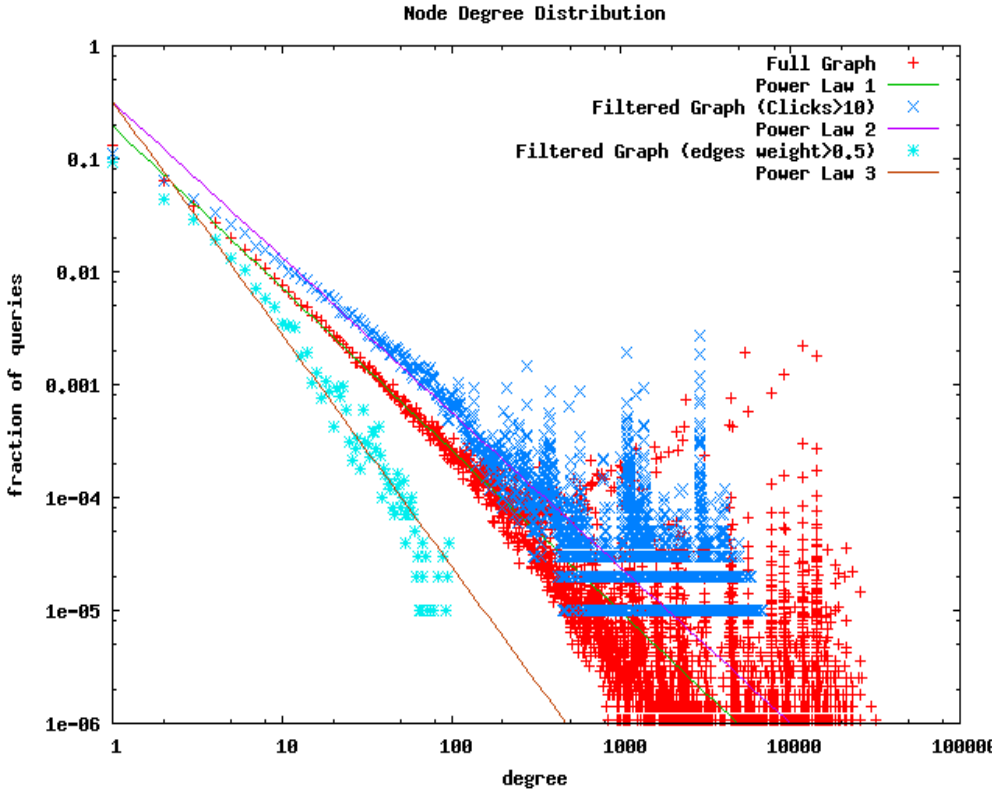
URL based Vector Space

- 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)

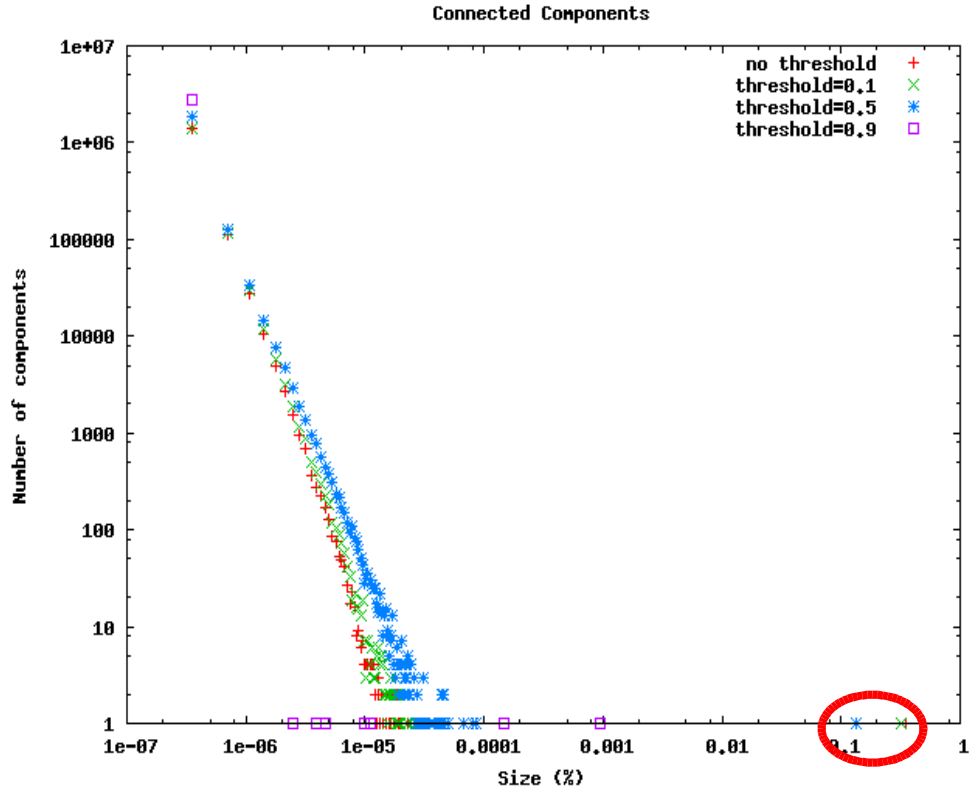


"Complex networks"

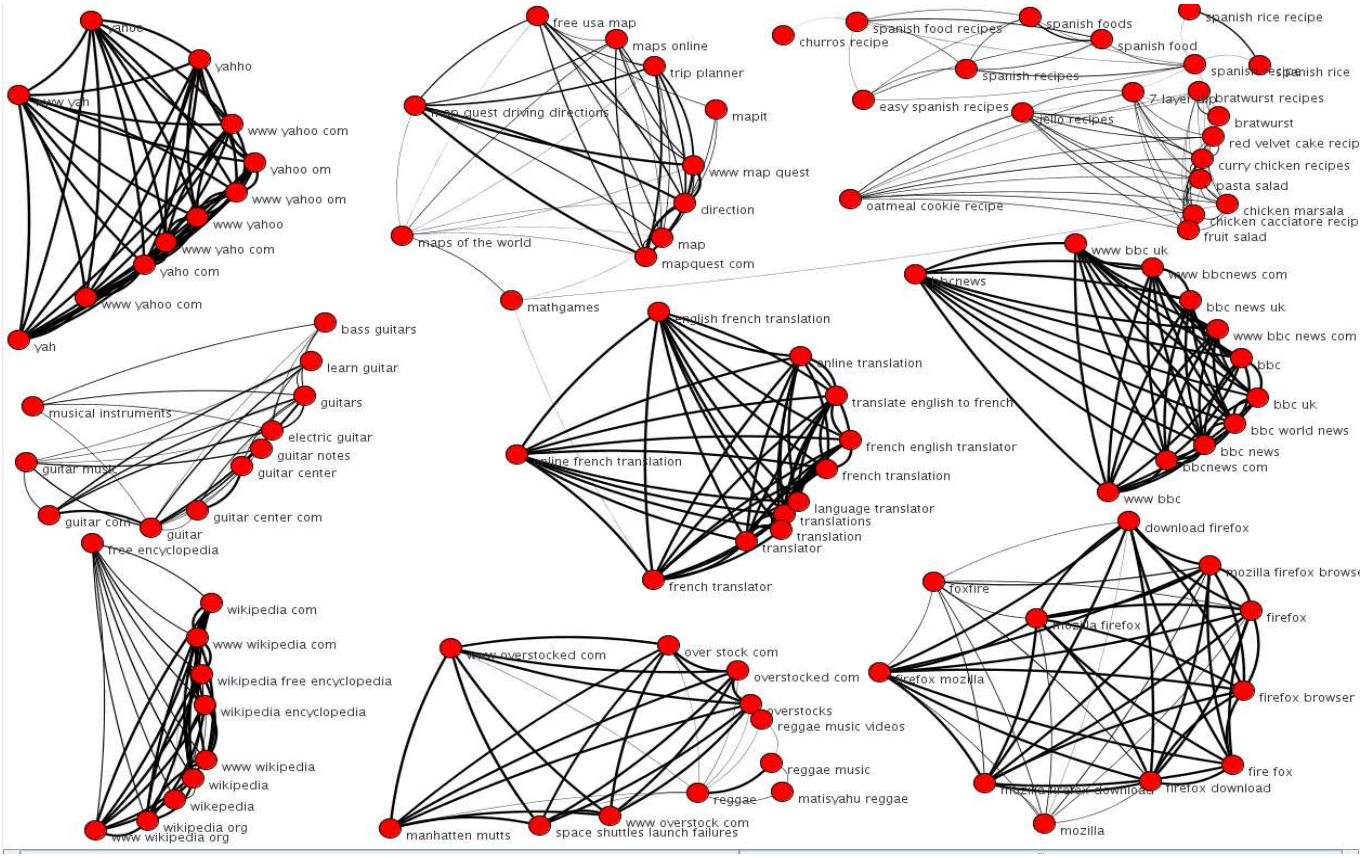
Node Degree Distribution



Connected Components



Implicit Folksonomy?



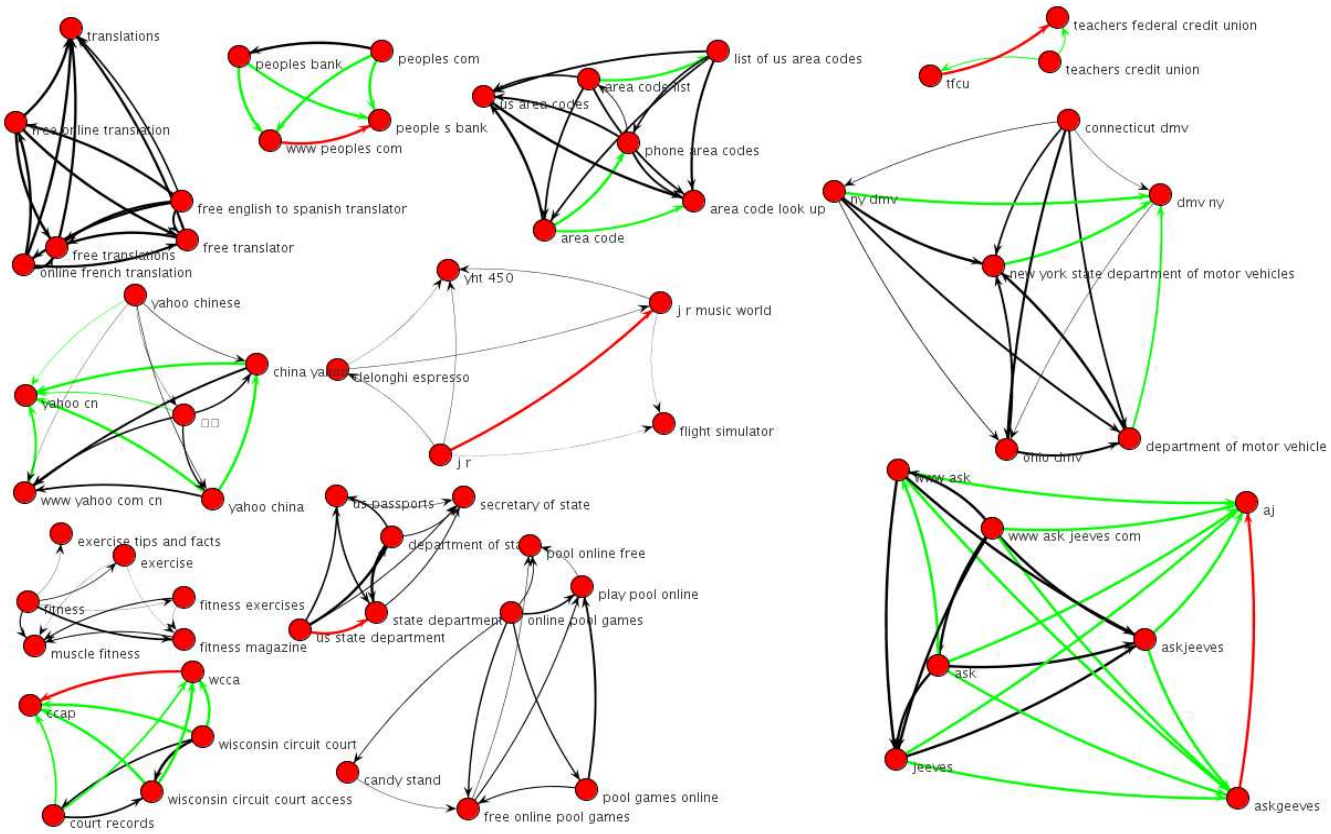
Set Relations and Graph Mining

- 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 }*

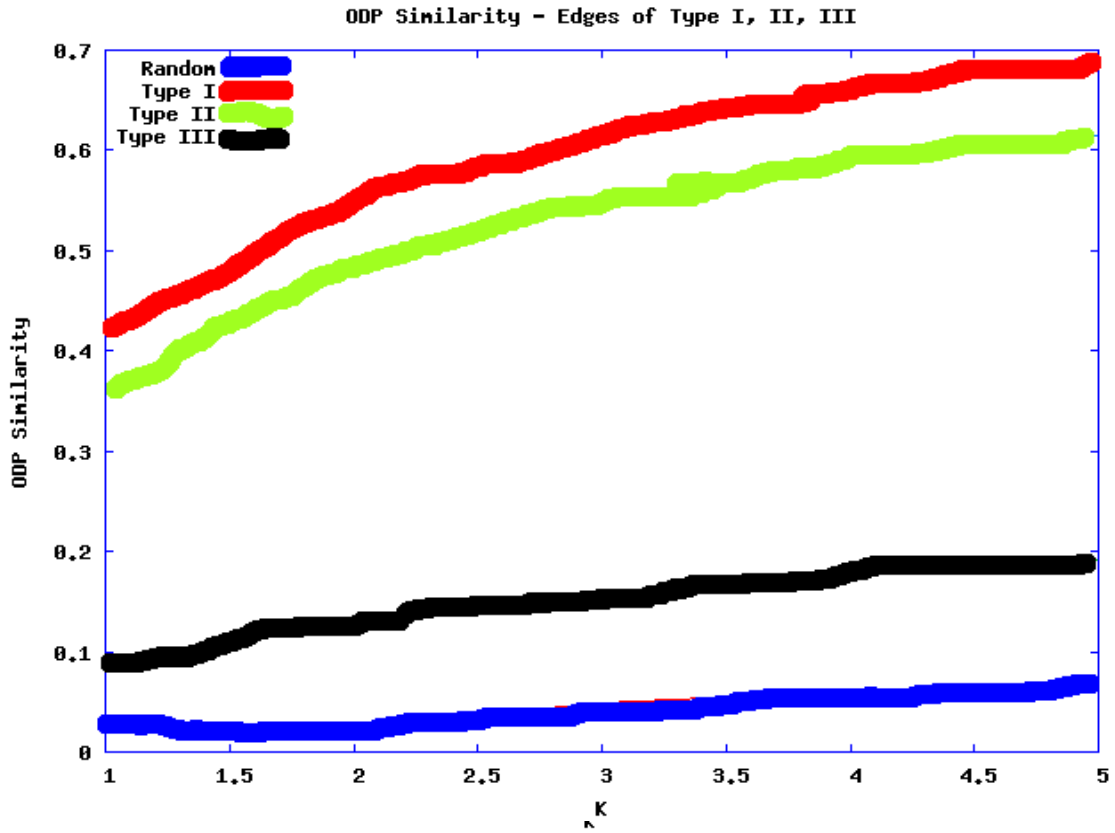
ODP Similarity

- 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

- Explicit vs. implicit social networks
  - 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

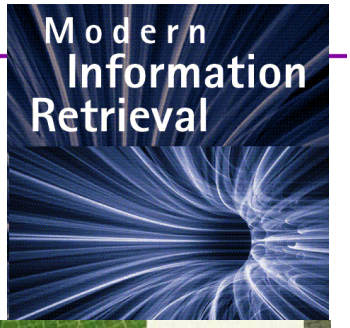
Epilogue

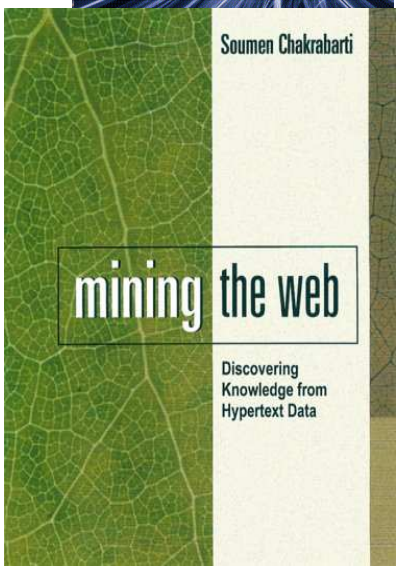
- The Web is scientifically young
- The Web is intellectually diverse
- The technology mirrors the economic, legal and sociological reality
- Web Mining: large potential for many applications
  - A fast prototyping platform is needed
- Plenty of open problems

The Future: Web 3.0?

- We are at Web 2.0 beta
- People wants to get tasks done
  - Where I do go for a original holiday with 1,000 US$?
- Take in account the context of the task

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/