Applications of Mining Web Queries

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

Joint work with Carlos Hurtado & Marcelo Mendoza (CWR, U. of Chile) and Georges Dupret (Yahoo! Research Latin America)

European Yahoo! Research Lab

- Areas: Web Mining & Web Search
- Hosted by the Barcelona Media Innovation Center
- Synergy with the Web Research Group of UPF
  - Web Mining
  - Web Characterization
  - Web Search

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

(Oct. '05, April '06)

15 languages, 20 countries

World: 6B people
France: 70M visitors p.y.

- 1 million new accounts a day
- 3.4 billion page views per day
- 429 million unique users each month
- 201 million registered users each month

- 20 Pb of storage (20M Gb)
  - US Library of congress every day (28M books, 20TB).
- 10 Tb of data processed per day
- 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?

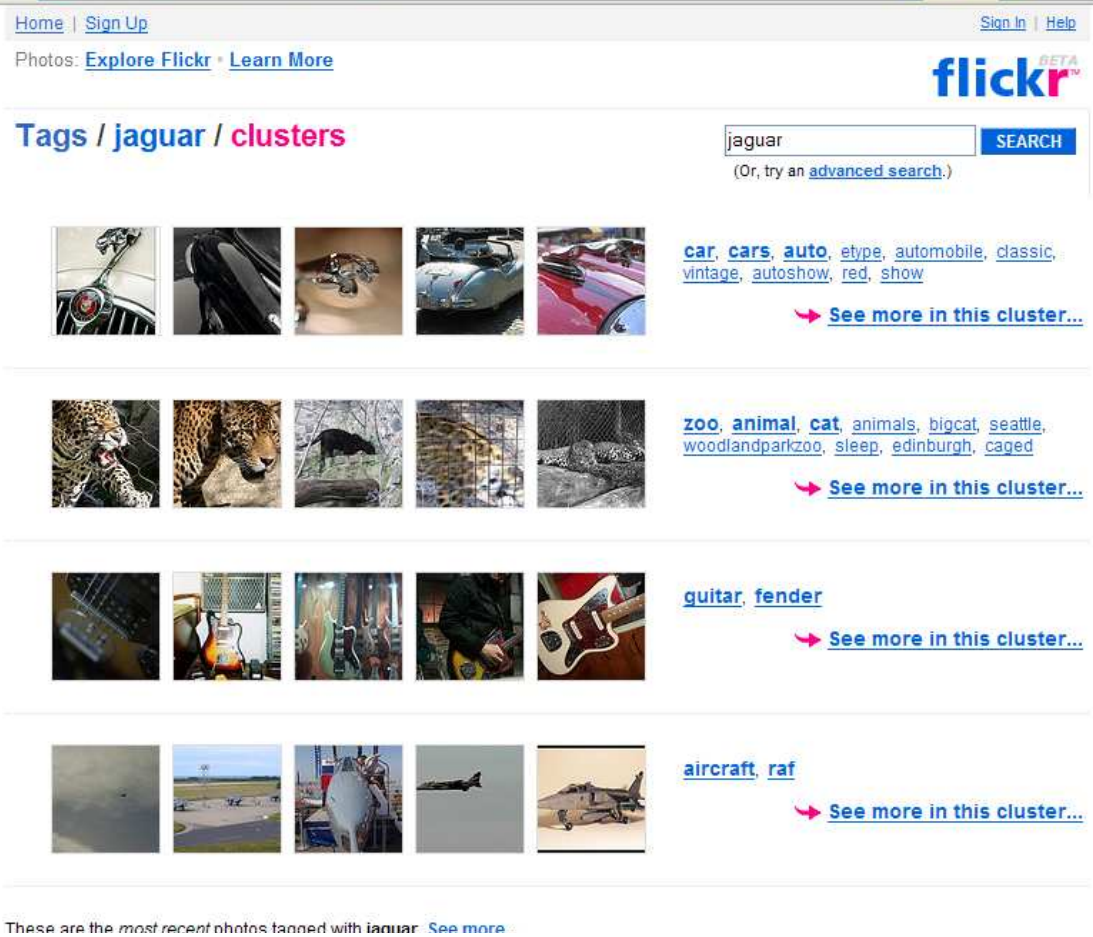
"Information Games"

Observed Data

- Query Logs
  - spelling, synonyms, phrases (named entities), substitutions
- Click-Thru
  - relevance, intent, wording
- Advertising
  - relevance, value, terminology
- Social
  - links, communities, dialogues...

The power of social media

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



Fight Spam

- Adversarial Web Retrieval
  - Text Spam (e.g. Cloaking, Quilt-like pages)
  - Link Spam (e.g. Link Farms)
  - Metadata spam
  - Ad spam (e.g. clicks, bids, etc)

My Motivations for Web Mining

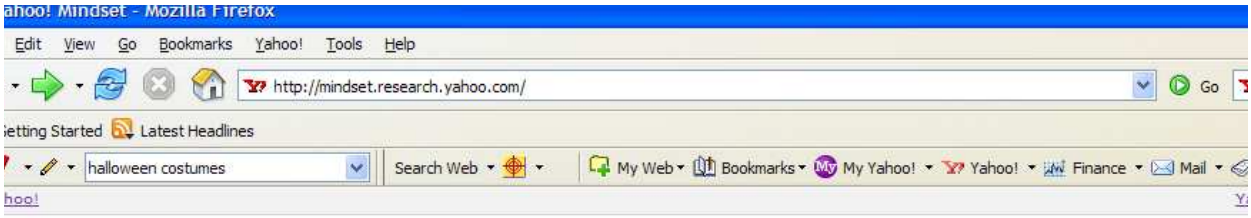
- The Dream of the Semantic Web
  - Hypothesis: Explicit Semantic Information
  - Obstacle: Us
- User Actions: Implicit Semantic Information
  - It's free!
  - Large volume!
  - It's unbiased!
  - Can we capture it?
  - Hypothesis: Queries are the best source

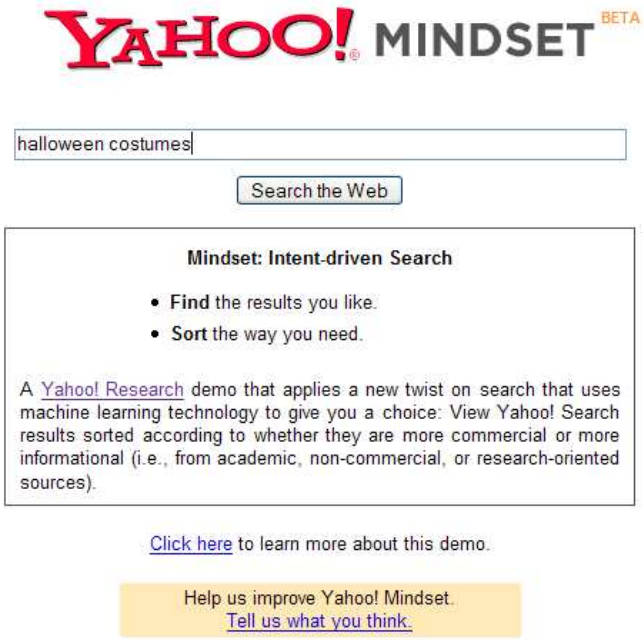
Mining Queries for ...

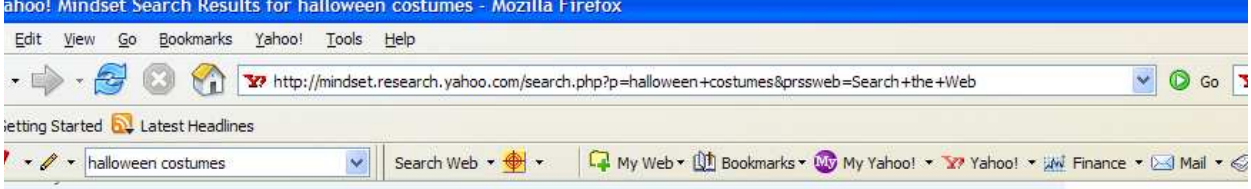
- Improved Web Search
- User Driven Design
  - Information Scent
  - The Web Site that the Users Want
  - The Web Site that You should Have
  - Improve content & structure

Web 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







1. (51) BuyCostumes -- Children's **Halloween Costumes** & Costume Accessories
From Barbie to Winnie the Pooh, for the latest children's **Halloween costumes** and costume accessories, visit BuyCosturmes.com today.
www.buycostumes.com/category.aspx?pcatID=childcostumes

2. (23) 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

3. (36) 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

4. (13) **Halloween** Mart
Sells **Halloween costumes** for adults, kids, babies, and pets. Also offers plus size **costumes**, **costumes** for Christmas, Easter, and other holidays, accessories, masks, and makeup.
www.halloweenmart.com

5. (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

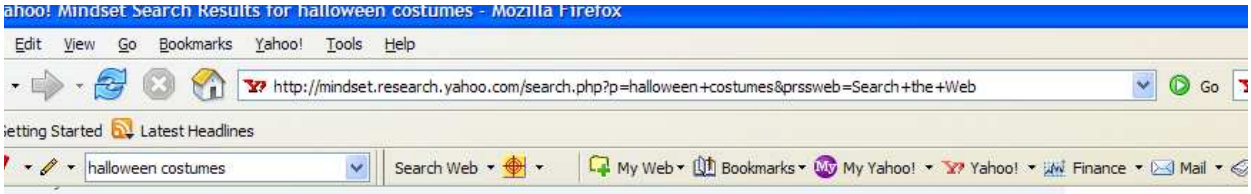
6. (55) **Halloween costumes** at Pinatas.com

Halloween Cos
SoldierCity
SoldierCity is the
shop for Hallowee
including BDUs, fl
fatigue caps, face
www.soldiercit

Halloween Cos
Skeleton's Clo
Featuring over 10
costumes, masks
decorations, all a
prices. At The Ske
have so much fun
www.theskelet

Halloween Cos
Bargain Prices
Shop fast, buy sm
computer software
software products.
from every store
bargain price. Hu
Shopzilla.
www.shopzilla

Costumes Spe
discoverlounge
Costumes, sexy c
halloween costum
costumes and mo
cost and fast ship



1. (95) **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

2. (84) CBS News | **Halloween Costumes** Go Upscale | October 5, 2004 10:54:35
Halloween Costumes Go Upscale
www.cbsnews.com/stories/2004/1...ent/main647447.shtml

3. (39) **Halloween Costumes** - Space related **Halloween Costumes**
... **Halloween Costumes** - Space related **Halloween Costumes**. Here in the United States, the ... plenty of **Halloween** parties this year, with everyone wearing **Halloween costumes**. Be the hit ...
space.about.com/b/a/206745.htm

4. (54) **Halloween Costumes**
messages from 1 to 9 of Discussions relating to The Lounge - **Halloween Costumes** - dewey decimal 007 ... Discussion: **Halloween Costumes**. Date: October 3, 2000 10:34 AM ... Discussion: **Halloween Costumes**. Date: October 3, 2000 1:42 PM ...
www.suite101.com/discussion.cfm/funlight/45943/latest/9

5. (1) **Halloween costumes** and supplies at the lowest prices! - **Halloween**
Halloween ... **halloween**. **halloween costumes**. **halloween** decorations. **halloween** masks. **halloween** party ... **halloween** stuff. child **halloween costumes**. childrens **halloween costumes**. adult **halloween costumes** ...
www.halloween-costumes-masks.com

6. (41) **Halloween Costumes** and Masks - Store Bought or Homemade **Costumes** and Masks
Halloween Costumes and Masks - Store Bought or Homemade **Costumes** and Masks. Celebrate **Halloween** in the proper spirit with a homemade costume and store bought safe adhesive mask. ... Ideas for making your own **Halloween costumes** and buying safe masks ... Also note her feature on Making Your Own Haunted House.) From **Halloween Costumes**, you can dress up as Dr ...

Halloween Cos
SoldierCity
SoldierCity is the
shop for Hallowee
including BDUs, fl
fatigue caps, face
www.soldiercit

Halloween Cos
Skeleton's Clo
Featuring over 10
costumes, masks
decorations, all a
prices. At The Ske
have so much fun
www.theskelet

Halloween Cos
Bargain Prices
Shop fast, buy sm
computer software
software products.
from every store
bargain price. Hu
Shopzilla.
www.shopzilla

Costumes Spe
discoverlounge
Costumes, sexy c
halloween costum
costumes and mo
cost and fast ship

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	**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	**Page with resources** 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

Kang & Kim, SIGIR 2003

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

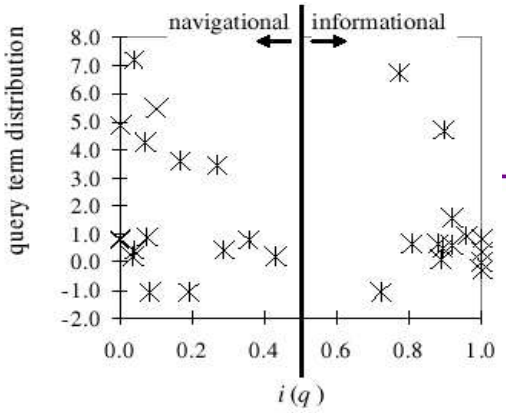


Figure 16: Query term distribution

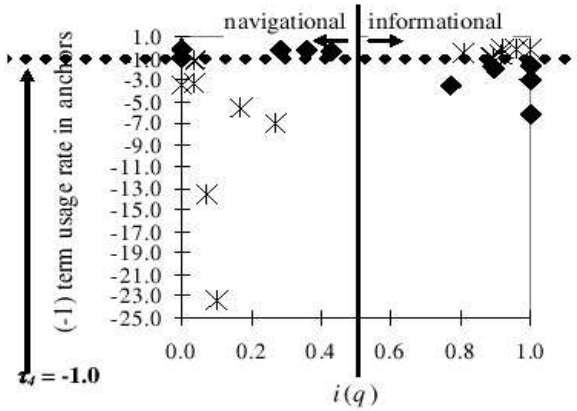


Figure 15: Anchor usage rate

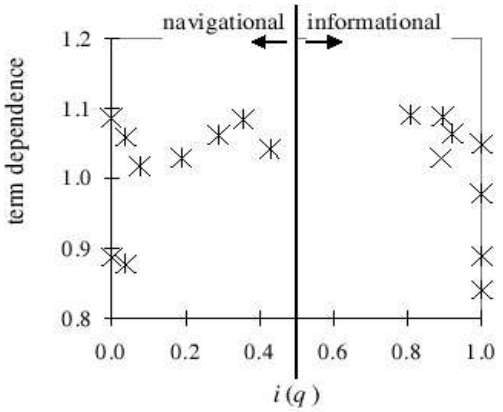


Figure 17: Term dependence

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

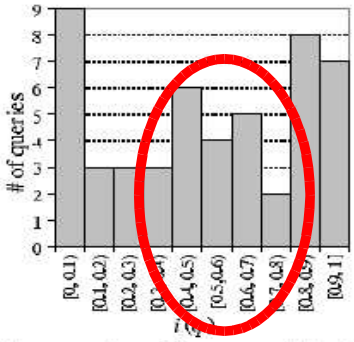


Figure 1: Query distribution along the $i(q)$ axis

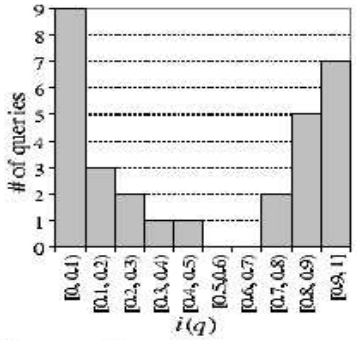


Figure 2: After removing software and person-name queries

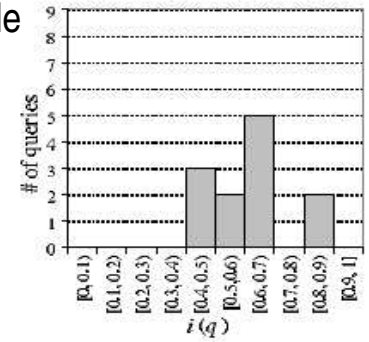


Figure 3: Distribution of the 12 software queries

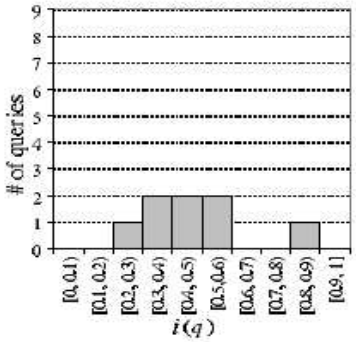
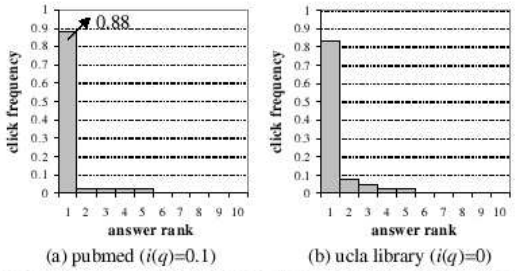


Figure 4: Distribution of the 8 person-name queries

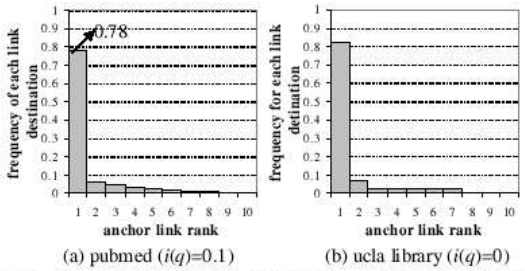
Features

- **Click & anchor text distribution**



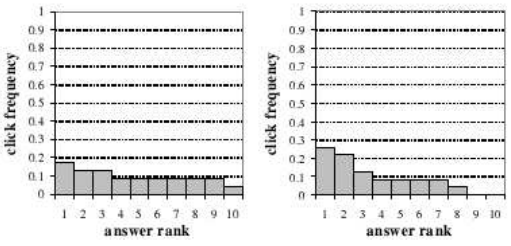
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 5: Click distributions for sample navigational queries



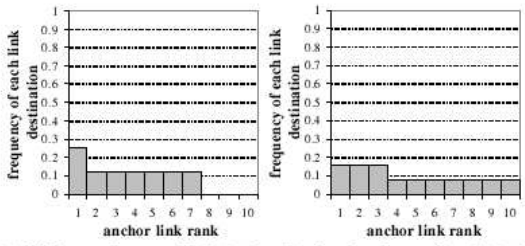
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

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

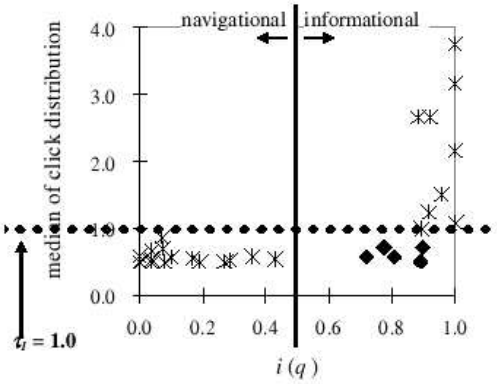


Figure 11: Median of click distribution

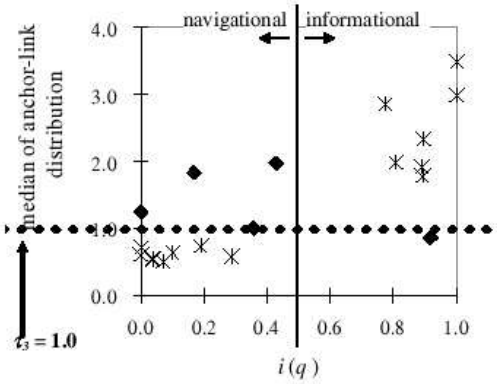


Figure 13: Median of anchor-link distribution

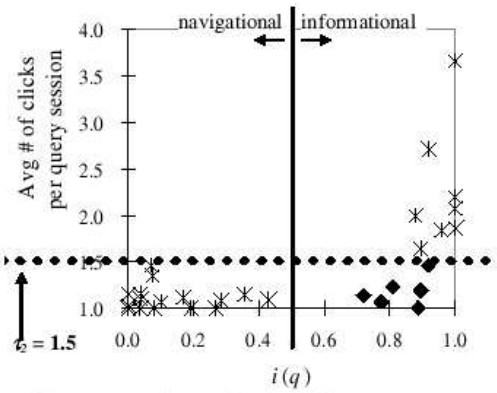


Figure 12: Avg # of clicks per query

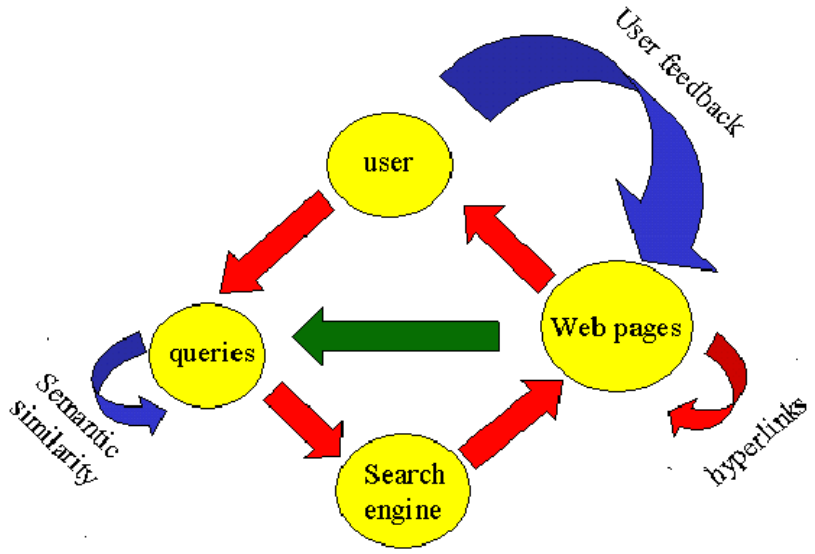
Prediction power:

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

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

Clustering Queries

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



Our Approach

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

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

- Recommend a better query (precise goal)
  - Query ranking

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

- 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

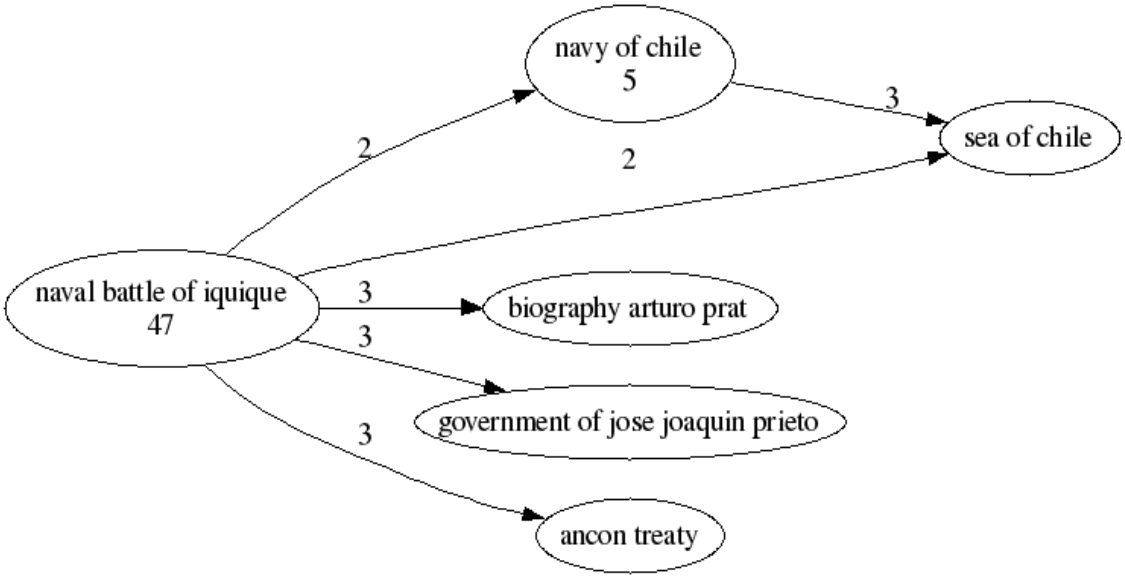
- Improved ranking
- Word classification
  - Synonyms are in the same cluster
  - Homonyms (polysemy) are in different clusters
- Query recommendation
  - Real queries, not query expansion

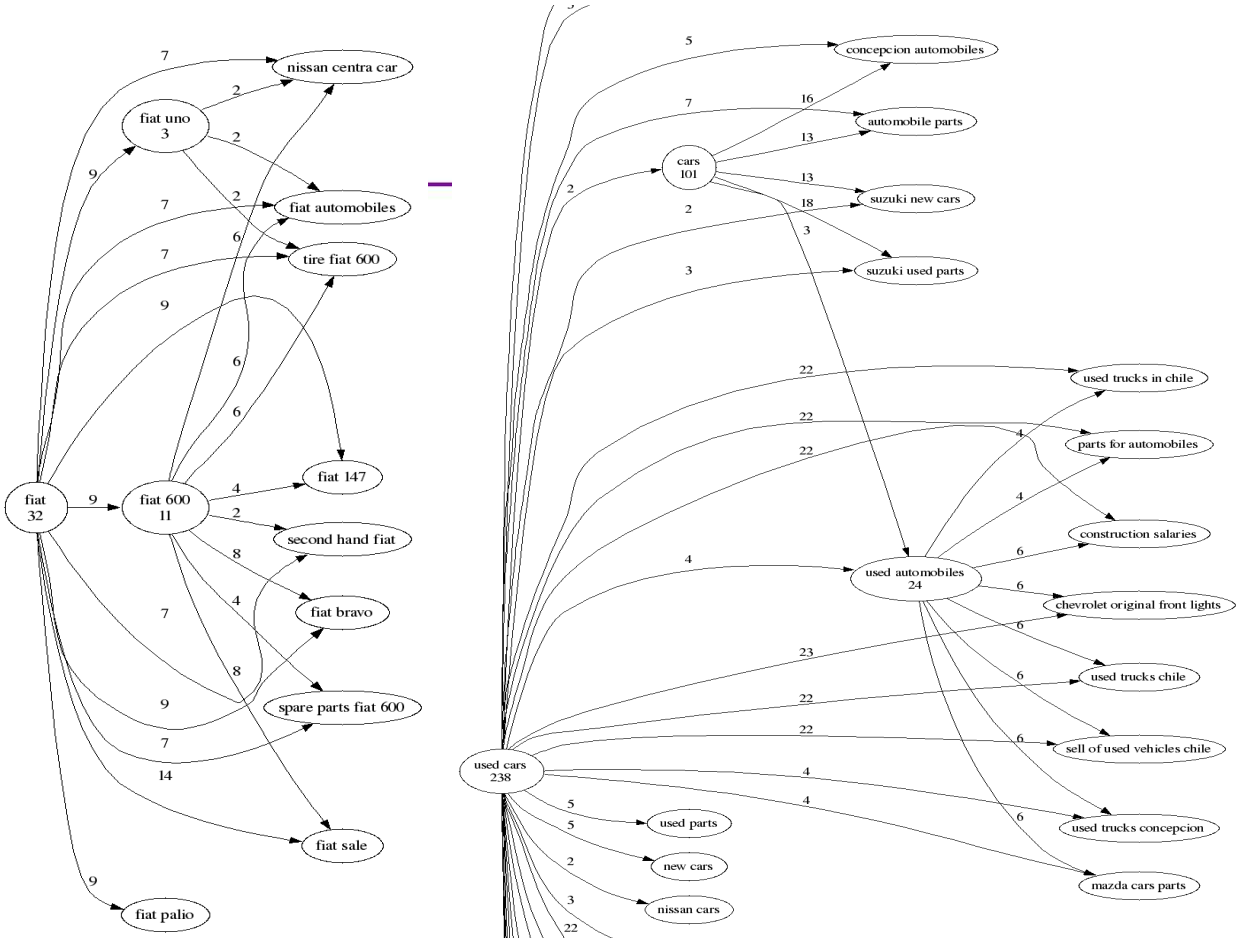
Query Recommendation

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 Query Recommendation

- Query dominance based on clicked pages





Taxonomies

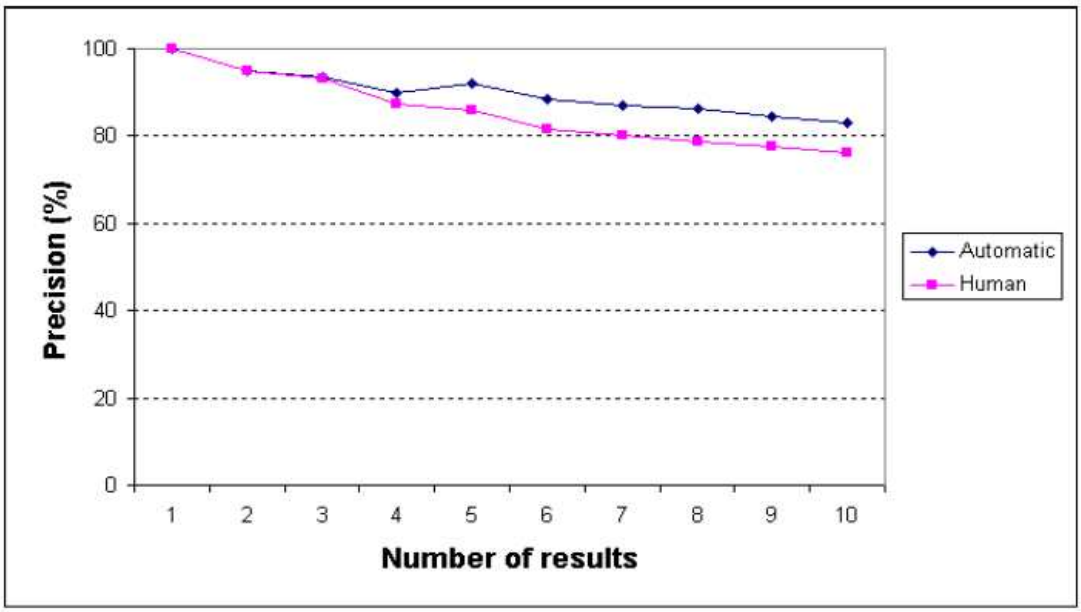
- Infer topics from queries that imply documents

	English	Spanish
(1)	*business:finances:banks*	*negocios:finanzas:bancos*
(2)	*society:law:norm:codes*	*sociedad:derecho:normas:códigos*
(3)	*business:building-industry:builders*	*negocios:construcción:constructoras*
(4)	*business:environment:engineering*	*negocios:medio-ambiente:ingeniería*
(5)	*business:sales:gifts:flowers*	*negocios:compras:regalos:flores*
(6)	*society:history*	*sociedad:historia*
(7)	*leisure:sports:motorcycling*	*tiempo libre:deportes:motociclismo*
(8)	*business:informatics:support*	*negocios:informática:soporte*
(9)	*leisure:gastronomy:drinks:wine*	*tiempo libre:gastronomía:bebidas:vinos*
(10)	*business:foreign trade:customs duty*	*negocios:comercio exterior:zonas francas*

Set	Number of Docs.	Relevant	Precision	Recall
A	100	83	83%	71%
H	100	76	76%	65%
$H \cap A$	48	43	93%	37%
$H - A$	52	33	63%	28%
$A - H$	52	40	77 %	34%

Taxonomies

- Quality of answers



Ongoing Work

- Build baseline set to evaluate quality of clusters
- Predict user goal + query recommendation
- Better queries have more precise goals
- 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

Questions?

Applications of Mining Web Queries

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