From User Actions to Metadata

Ricardo Baeza-Yates

Center for Web Research
www.cwr.cl

CS Dept., Univ. of Chile, Santiago

&

ICREA Research Professor

Technology Dept., Univ. Pompeu Fabra

Barcelona, Spain

ricardo@baeza.cl

Semantic Jokes at Madrid Airport

German Wings Advertising:

Vuelos desde 49 Euros.
No bromeamos. Somos alemanes.

Iberia Advertising:

Punctuality is our aim.

We are Spanish.

Summary

- Motivations
- Context as Metadata
- Web Queries
- User Goals
- Clustering Queries
- Taxonomies from Queries
- Examples

- Joint work with Georges Dupret, Carlos Hurtado & Marcelo Mendoza (CWR, Chile)

Motivations

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

Metadata

- **Normally associated to documents**
- **Is a logical concept**
- **What about users?**
  - Associated to sites?
- **Different types of metadata**
  - Associated to the context of the search
  - Source of topical metadata: the most interesting one!

Philosophical Issues

- **Physical document abstraction: content & metadata**
  - A subtle assumption
  - This asymmetry is not necessary for the storage mechanism
- **A document could be just a set of attributes and values**
  - One of them can be the content
  - For different applications, different attributes will be more important than others
  - The content (& metadata) depends on the application
- **Intrinsically there is no reason to physical metadata**
  - This asymmetry is application driven and it's dynamic

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)

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, etc.**
- ***Goals*****: personalization, localization, better ranking in general, etc.**

Using the Context

Example: *I want information about Santiago*

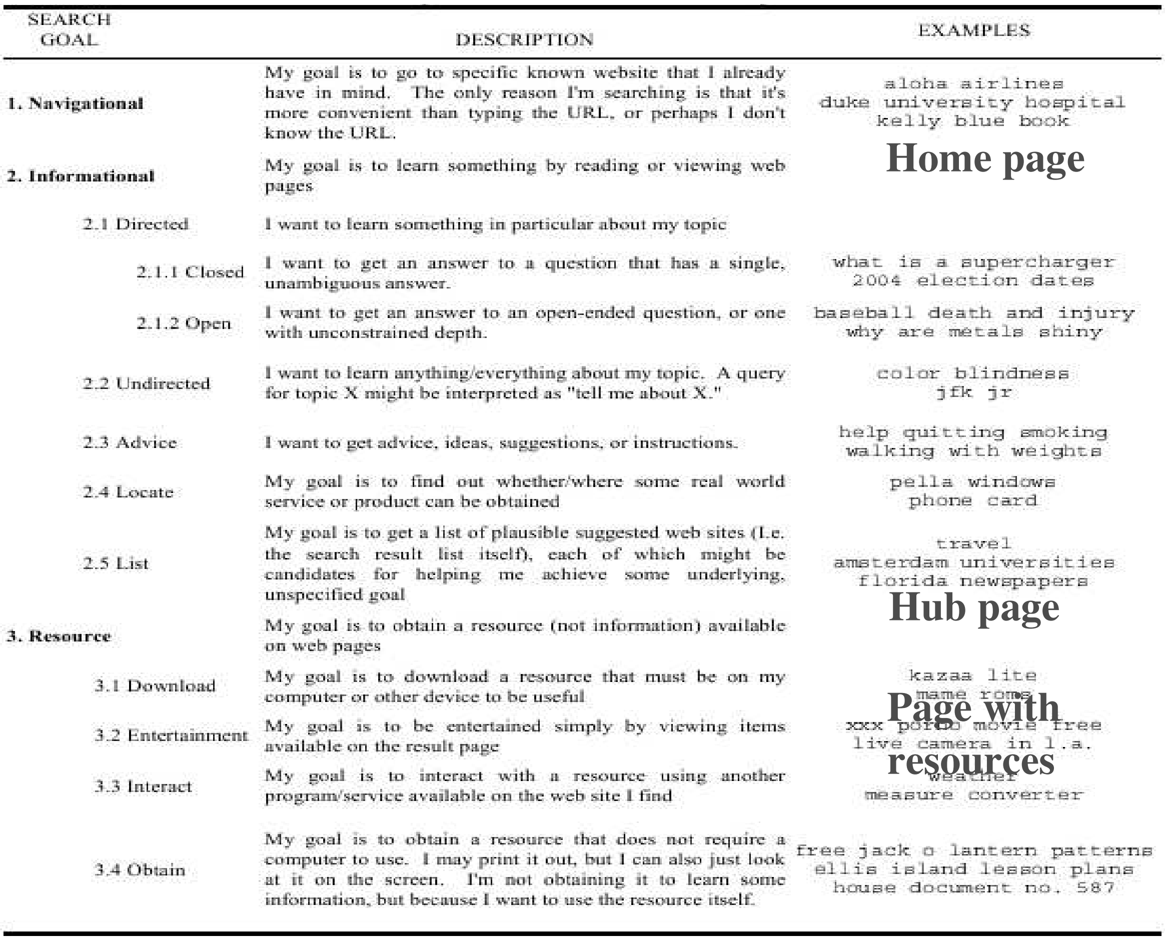
- **Context**
  - Family in Chile
  - Catholic
  - Travelling to Cuba
  - Lives in Argentina
  - Located in Santo Domingo
  - Architect
  - Spanish movies fan
  - Baseball fan
- **Probable Answer**
  - *Santiago de Chile*
  - *Santiago de Compostela*
  - *Santiago de Cuba*
  - *Santiago del Estero*
  - *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, ...)**

Web Queries

- **Cultural and educational diversity**
- **Short queries**
  - Inherent to users or due to the query language?
- **Short patience**
  - few queries posed & few answers seen
- **Smaller & different vocabulary**
- **Different user goals (Broder, 2002):**
  - Information need
  - Navigational need
  - Transactional need
- **Refined by Rose & Levinson, WWW 2004**



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

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

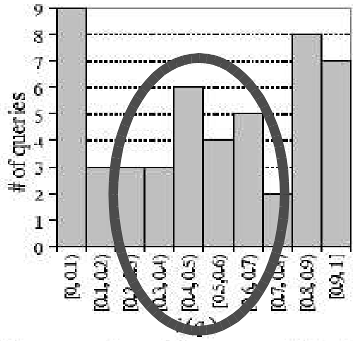


Figure 1: Query distribution along the $i(q)$ axis

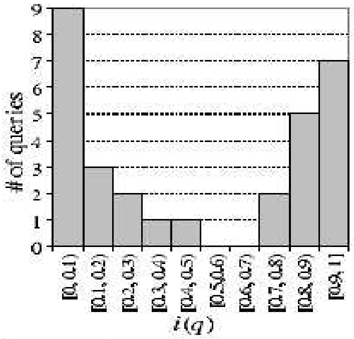


Figure 2: After removing software and person-name queries

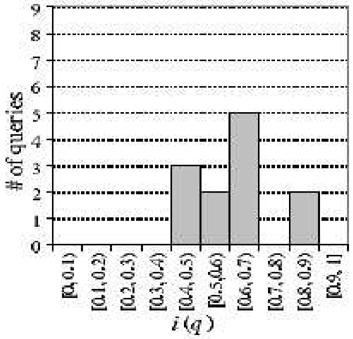


Figure 3: Distribution of the 12 software queries

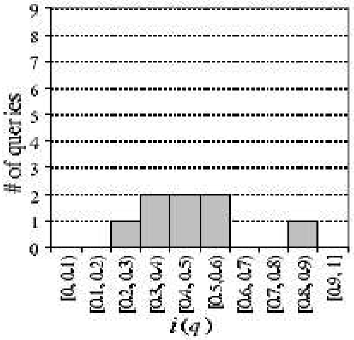
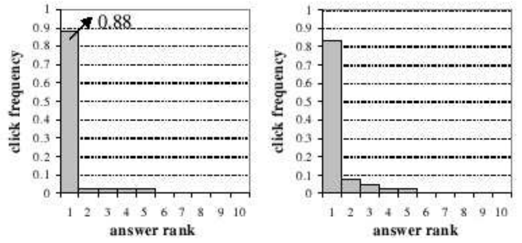


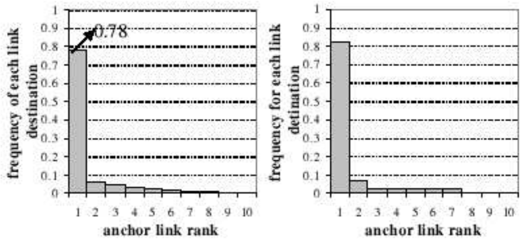
Figure 4: Distribution of the 8 person-name queries

- Click & anchor text distribution



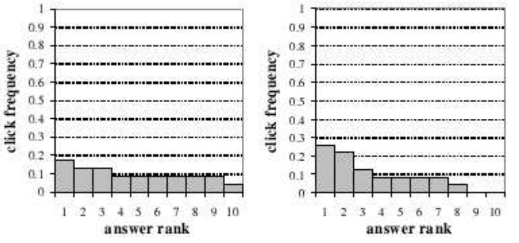
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 5: Click distributions for sample navigational queries



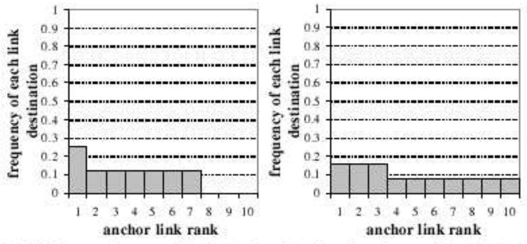
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

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

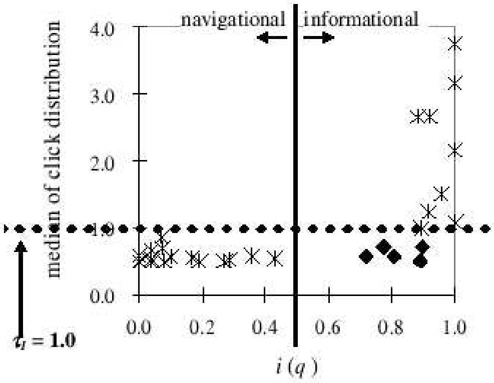


Figure 11: Median of click distribution

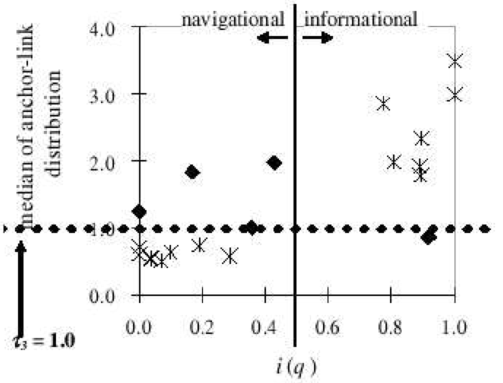


Figure 13: Median of anchor-link distribution

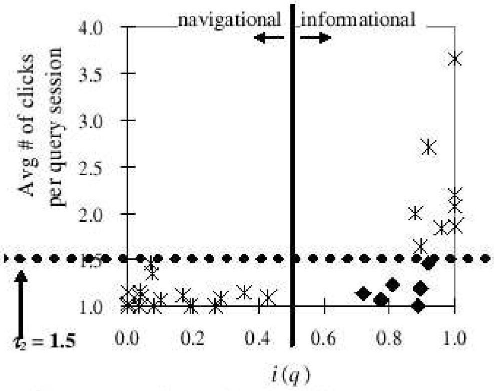


Figure 12: Avg # of clicks per query

Prediction power:

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

- **Drawback: Small evaluation**

Kang & Kim, SIGIR 2003

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

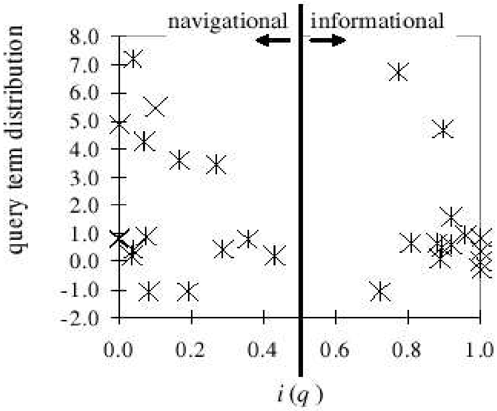


Figure 16: Query term distribution

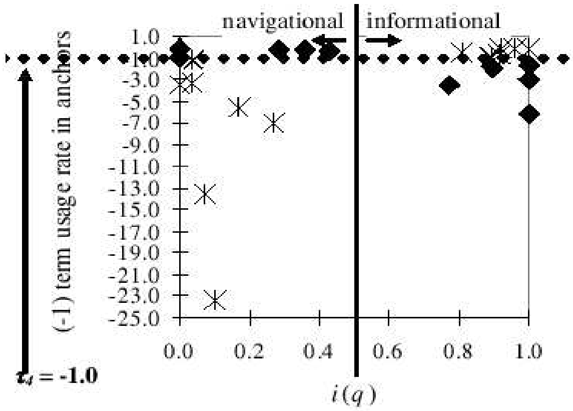


Figure 15: Anchor usage rate

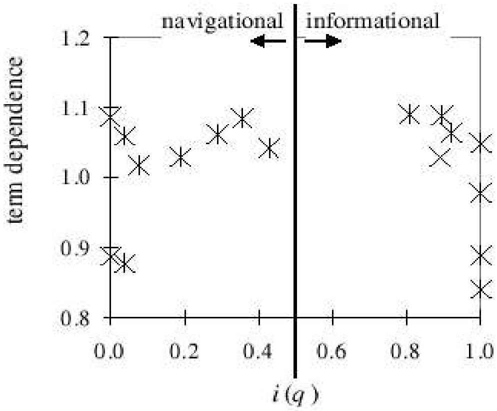
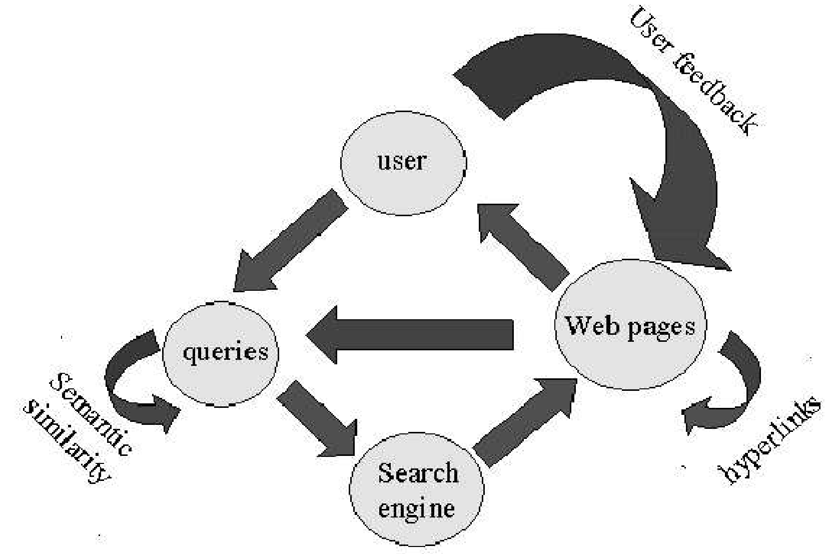


Figure 17: Term dependence

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

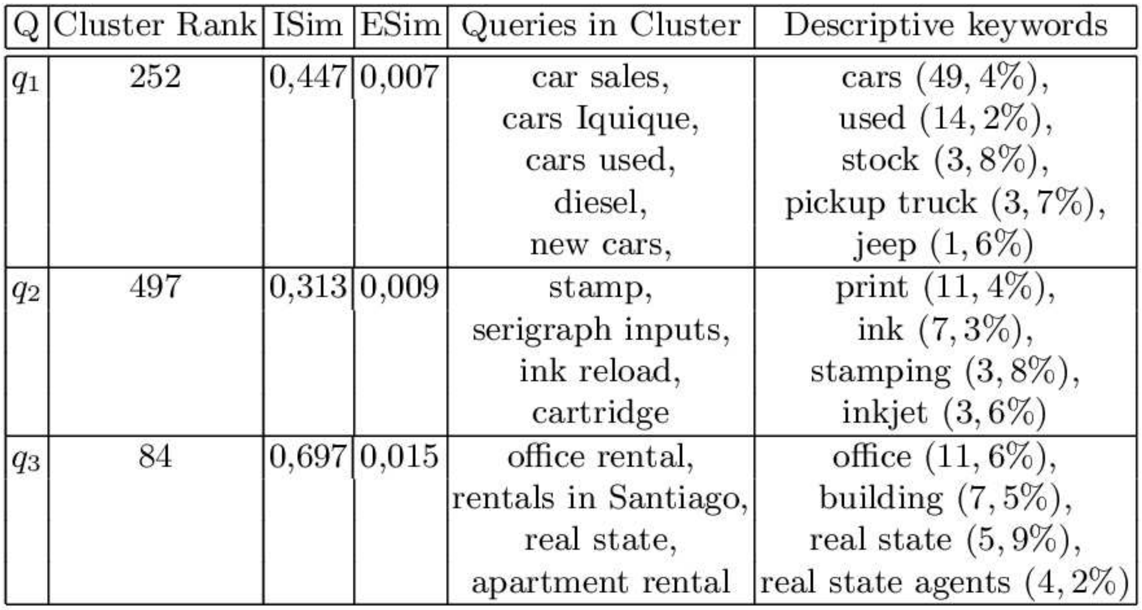
$$\boldsymbol{q}[i] = \sum_{URLu} \frac{\mathtt{Pop}(q,u) \times \mathtt{Tf}(t_i,u)}{\max_t \mathtt{Tf}(t,u)}$$

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

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

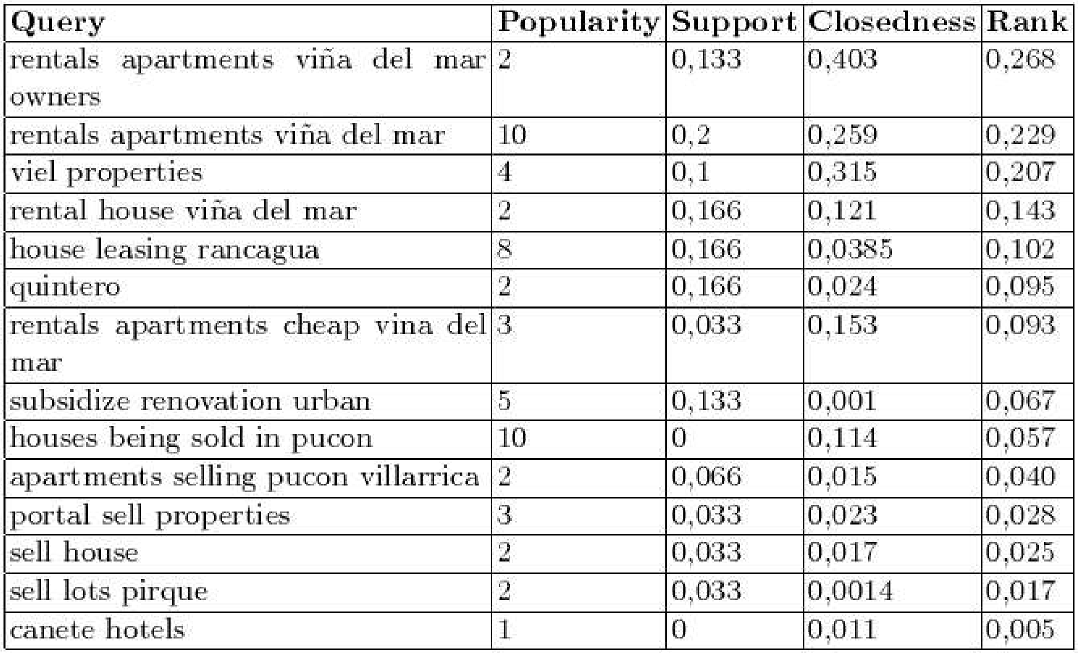
- **Pseudo-taxonomies for queries**
  - Clusters dendogram
  - Real language (slang?) of the Web
  - Can be used for classification purposes

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 $(4, 2\%)$

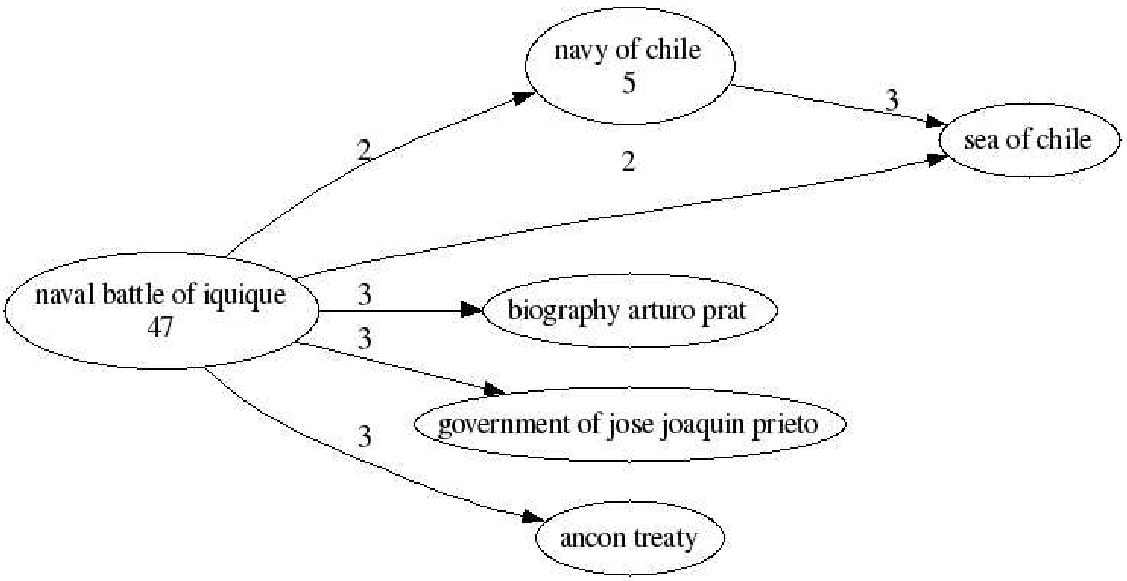
Query Recommendation

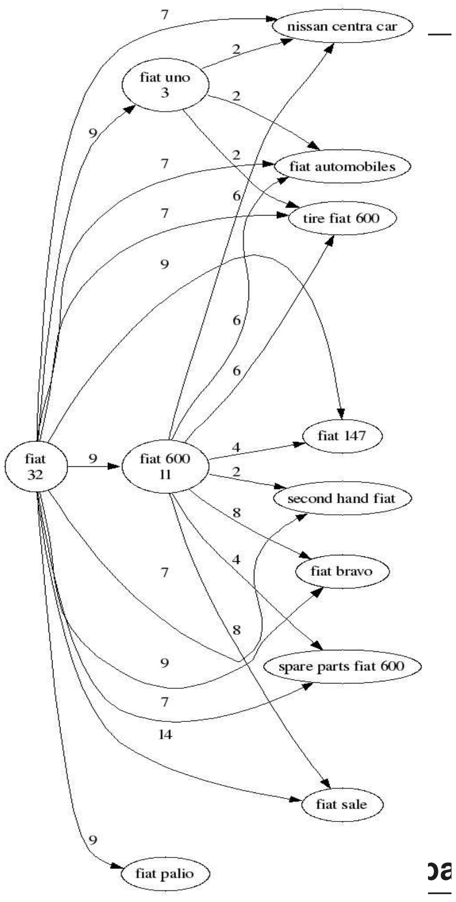


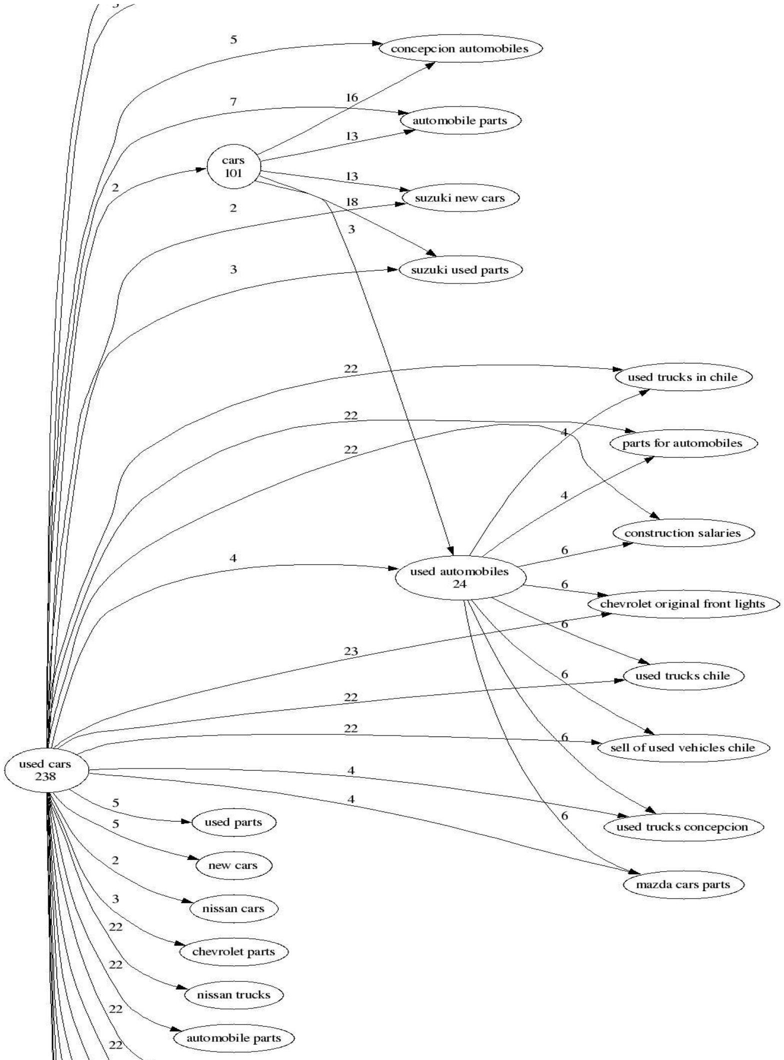
Query	Popularity	Support	Closedness	Rank
rentals apartments viña del mar owners	2	0,133	0,403	0,268
rentals apartments viña del mar	10	0,2	0,259	0,229
viel properties	4	0,1	0,315	0,207
rental house viña del mar	2	0,166	0,121	0,143
house leasing rancagua	8	0,166	0,0385	0,102
quintero	2	0,166	0,024	0,095
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**

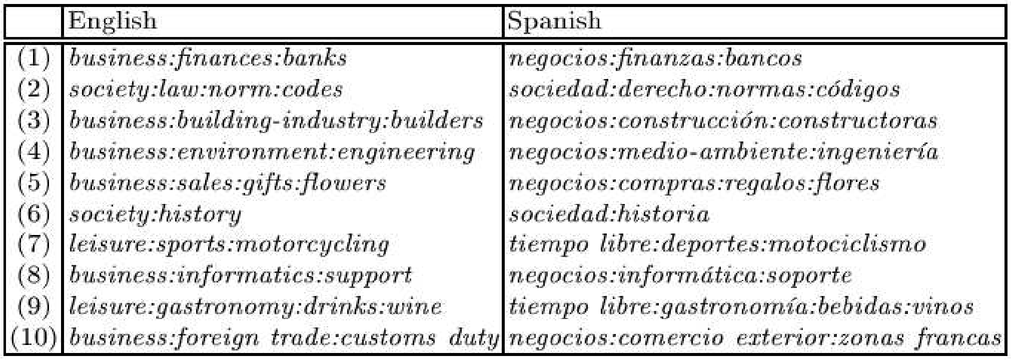




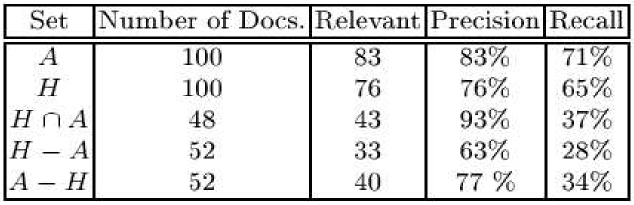


Taxonomies

- Infer topics from queries that imply documents



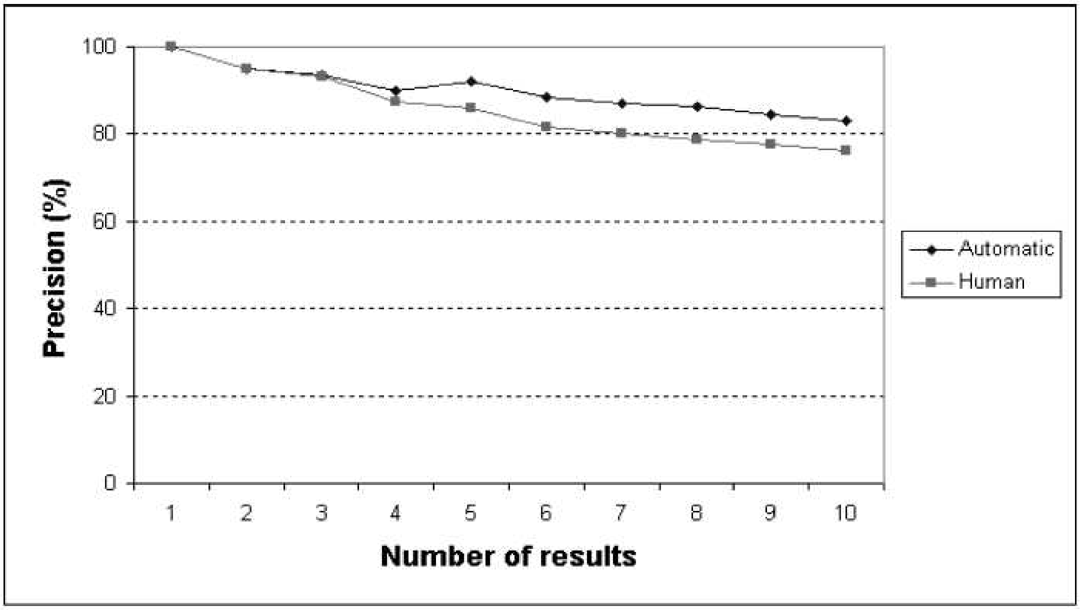
	English	Spanish
(1)	*business:finances:banks*	*negocios:finanzas:bancos*
(2)	*society:law:norm:codes*	*sociedad:derecho:normas:códigos*
(3)	*business:building-industry:builders*	*negocios:construcción:constructoras*
(4)	*business:environment:engineering*	*negocios:medio-ambiente:ingeniería*
(5)	*business:sales:gifts:flowers*	*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, comments, …?

Ricardo Baeza-Yates
ricardo@baeza.cl

Advertisements:

SPIRE 2005, November, Buenos Aires, Argentina
SPIRE 2006, October, Barcelona, Spain