



Queries and Clicks as a Source of Knowledge

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

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



Yahoo! Numbers

(Oct. '05, April '06)

15 languages, 20 countries

- 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

6



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)

7



Observed Data

- Query Logs
 - spelling, synonyms, phrases (named entities), substitutions
 - Click-Thru
 - relevance, intent, wording
 - Advertising
 - relevance, value, terminology
 - Social
 - links, communities, dialogues...
- good quality,
sparse,
power law
- good quality,
sparse,
mostly safe
- Trusted,
high quality,
homogeneous,
structured
- trust?
quality?

8



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”.*

9



The Power of Social Media

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

10

Home | Sign Up | Sign In | Help


Photos: [Explore Flickr](#) • [Learn More](#)

flickr BETA

Tags / [jaguar](#) / clusters


jaguar

(Or, try an [advanced search](#).)




[car](#), [cars](#), [auto](#), [ettype](#), [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](#)



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

12



Mining Queries for ...

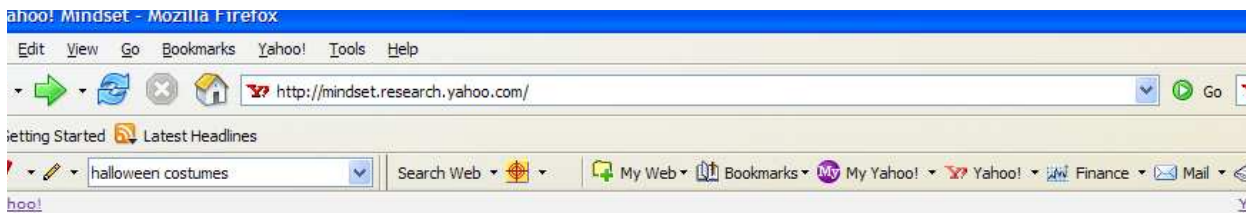
- Improved Web Search
- User Driven Design
 - Information Scent
 - The Web Site that the Users Want
 - The Web Site that You should Have
 - Improve content & structure
- **Bootstrap of pseudo-semantic resources**

13



- 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

14



YAHOO! MINDSET BETA

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.](#)

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, dñer 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
- (44) [HalloweenOnly.com](#)
Costumes, masks, props, and special effects equipment for **Halloween**.
www.halloweenonly.com
 - (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
 - (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
 - (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
 - (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
 - (16) [Halloween Mart](#)

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, dñer 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

- (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
- (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
- (82) [Halloween](#)
... **Halloween** Holiday. **halloween costumes halloween masks halloween decorations halloween recipes halloween crafts halloween ideas. Halloween > halloween costumes, halloween** ... ideas, **halloween crafts** ...
halloween.xuyase.com
- (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
- (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 Costu](#)
[Halloween I](#)
At AnytimeCo
an exclusive s
quality costu
theatrical mal
beards, props
decorations.
www.anytin

[Halloween I](#)
[BuyCostum](#)
BuyCostumes.
Halloween co:
Huge selectio
shopping, gre
and fast ship
costumes at B
[buycostum](#)

[Costumes -](#)
Costumes, Hal
costume wigs,
costume eye
www.bestw

[Halloween I](#)
[More](#)
Starcostumes
extensive line
costumes and
for adults and
wigs, masks, p
Buy online or
www.starco

[Buy a Hallo](#)
Huge selectio
costumes - ev
heros, movie
accessories, p
[halloweenr](#)



Relevance of the Context

- There is no information without context
- Context and hence, content, will be implicit
- Balancing act: information vs. form
- Brown & Digid: *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)

18



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 (IP, registered users), intrusion, will to do it, etc.
- *Other sources*: Web, CV, usage logs, computing environment, ...
- *Goals*: personalization, localization, better ranking in general, etc.

19



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*

20



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
($operating\ system$, $browser$, ...)

21

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 name rone
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



User Intention

- Kang & Kim, SIGIR 2003
- Liu, Lee & Cho, WWW 2005
- **Features:**
 - Anchor usage rate
 - Query term distribution in home pages
 - Term dependence
- Top 50 CS queries, manual query classification
- *Removed software & person-names, 30 queries left*
- **Not effective: 60%**
- **Features:**
 - Average number of clicks
 - Median of clicks distribution
 - Median of anchor text distr.
- **Drawbacks:**
 - small evaluation
 - a posteriori feature
- **Prediction power: 90%**

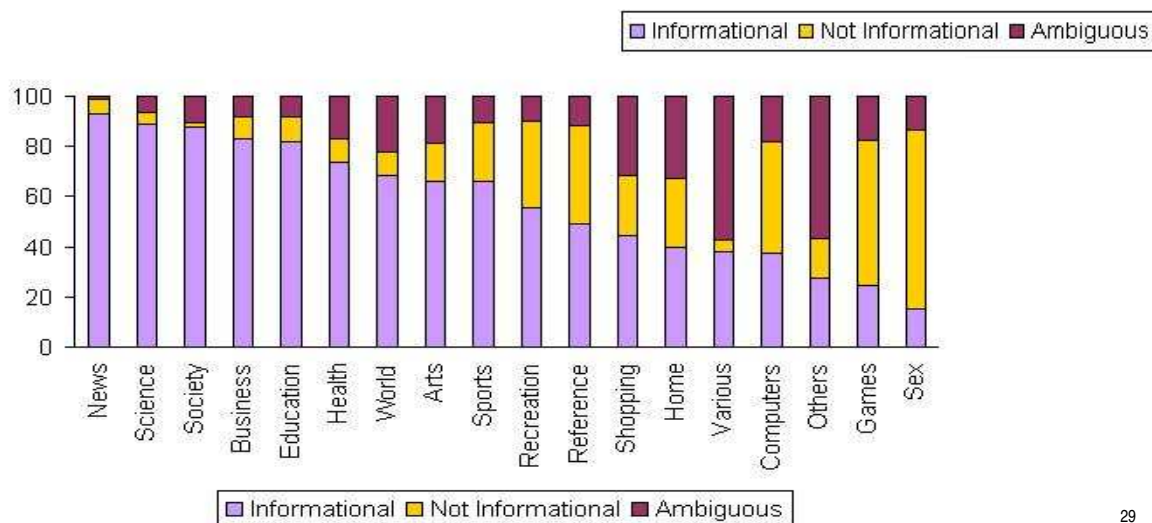
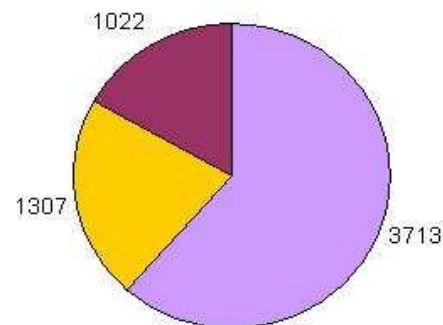


- Manual classification of more than 6,000 popular queries
- Query intention & topic
- Classification & Clustering
- Machine Learning on all the available attributes
 - Baeza-Yates, Calderon & Gonzalez (SPIRE 2006)

28



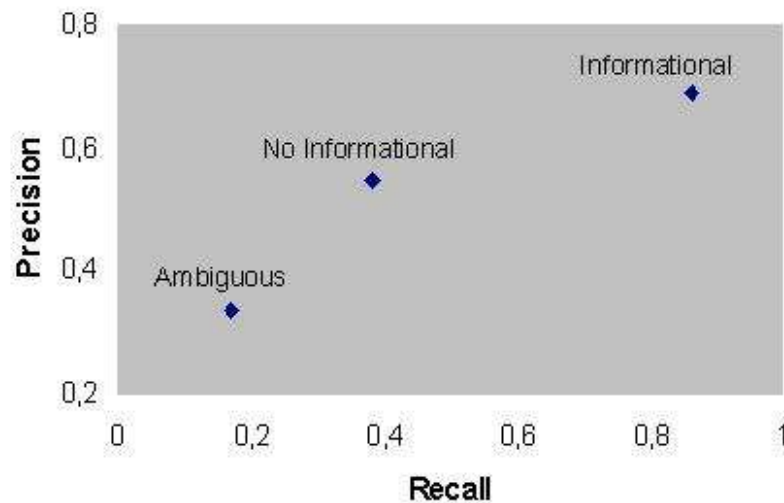
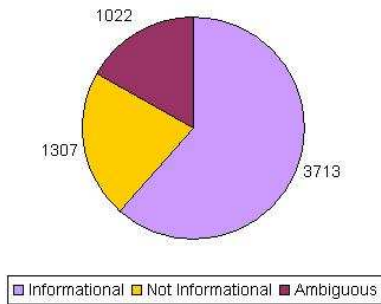
Classified Queries



29



Results: User Intention

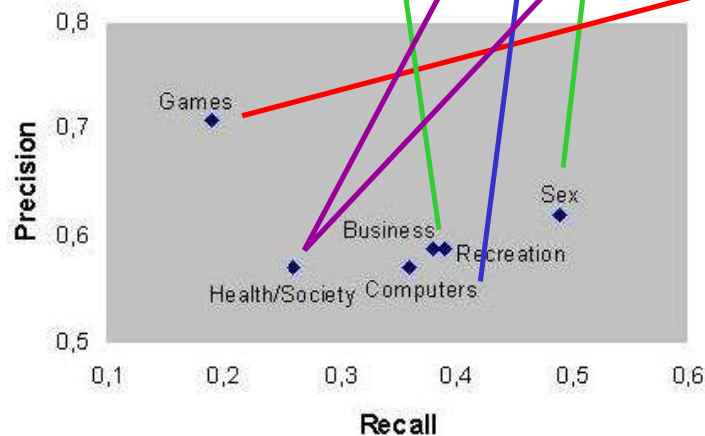
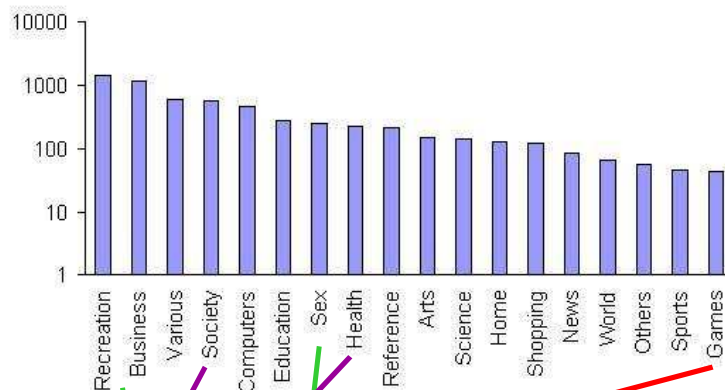


30

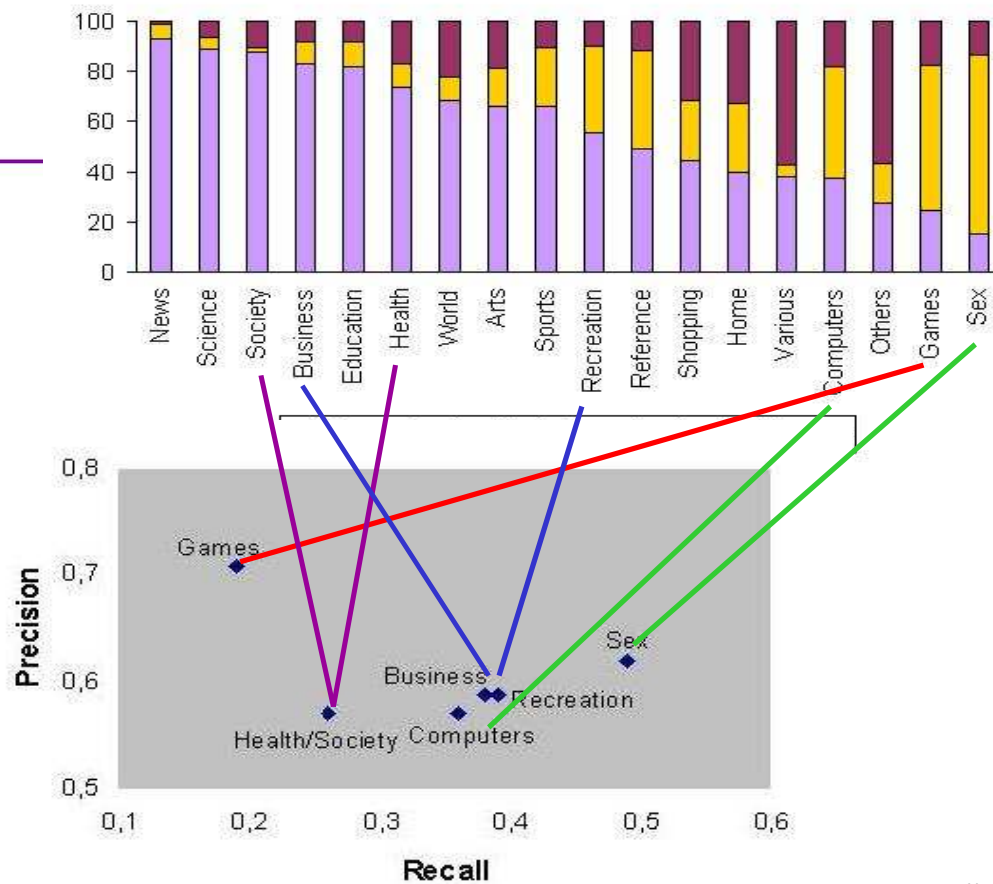


Results: Topic

- Volume wise the results are different



31



32

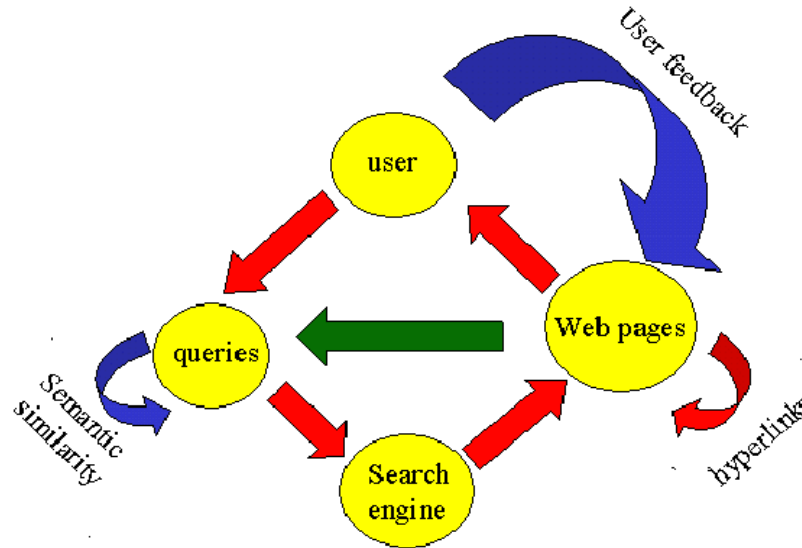


Clustering Queries

- Define relations among queries
 - Common words: sparse set, polysemy problems
 - 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)

33

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



34

Y! Key Idea

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

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

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

35



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%)

36



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

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

37



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

38



Building Taxonomies

- Infer topics from queries that imply documents

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

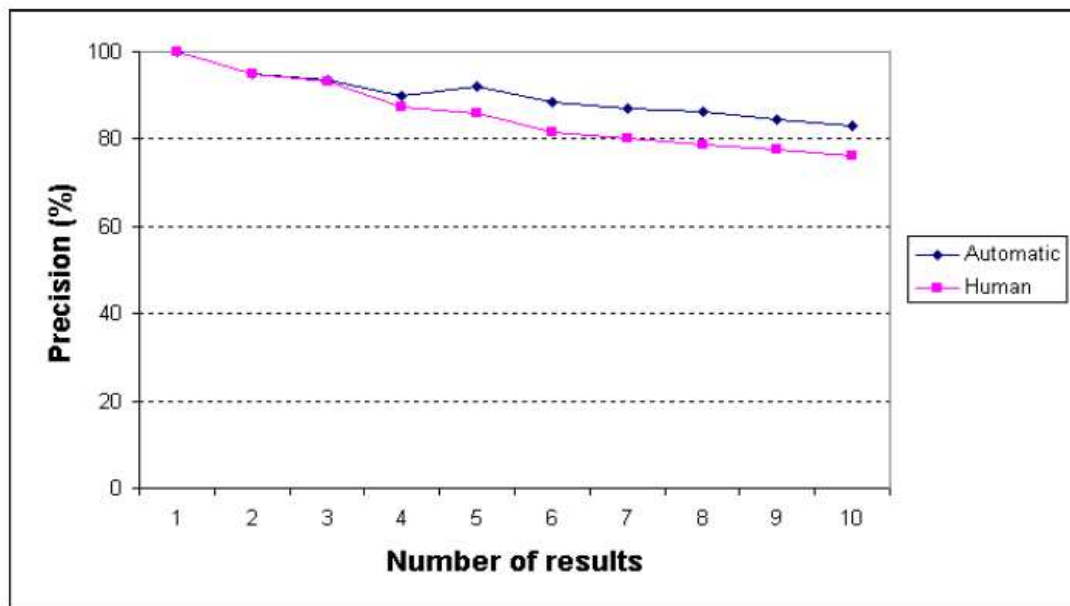
Set	Number of Docs.	Relevant	Precision	Recall
A	100	83	83%	71%
H	100	76	76%	65%
$H \cap A$	48	43	93%	37%
$H - A$	52	33	63%	28%
$A - H$	52	40	77 %	34%

40



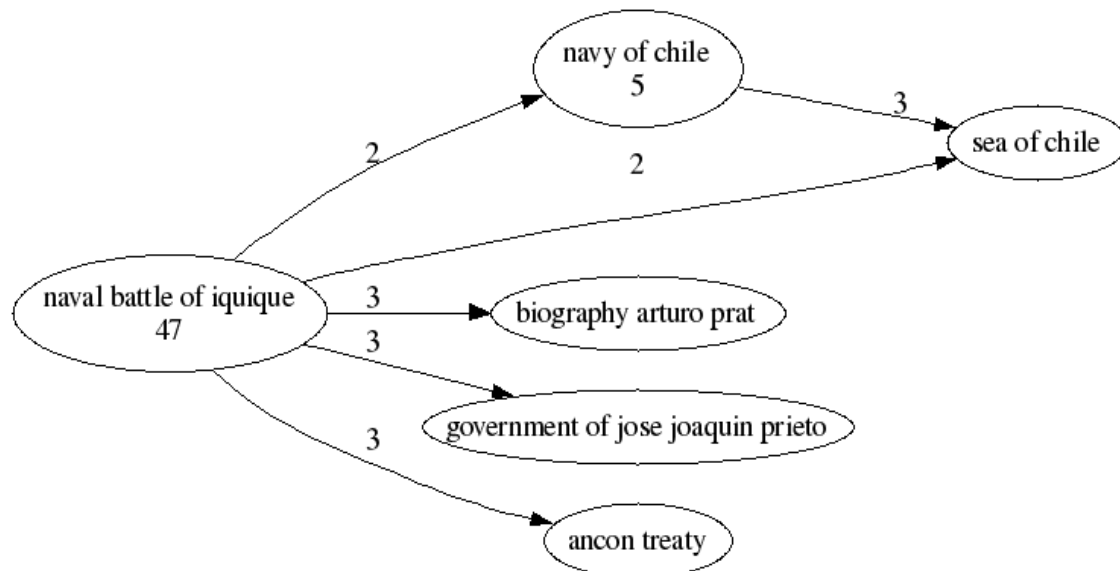
Results better than humans!

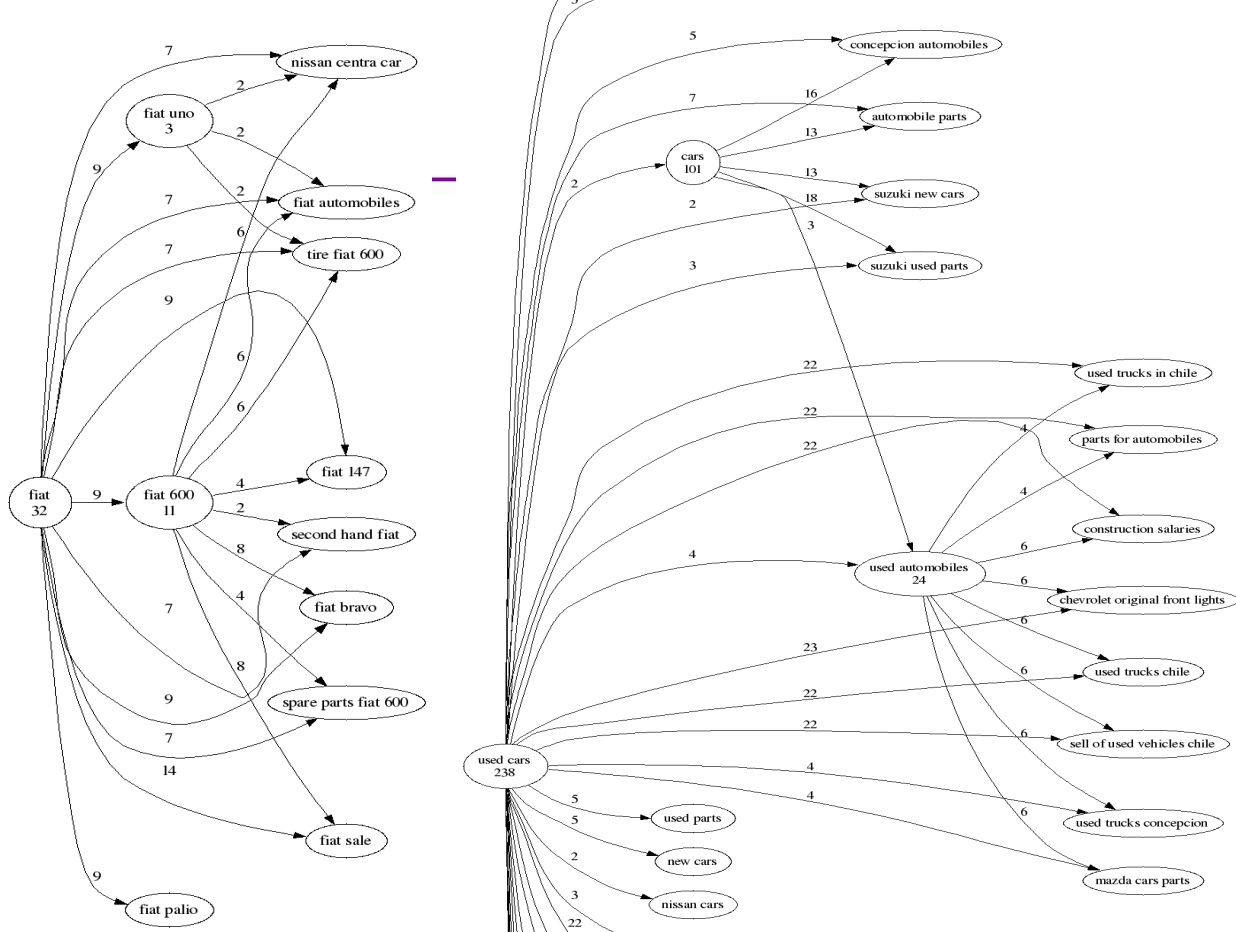
- Quality of classification maintenance



Simple Related Terms

- Query dominance based on clicked pages





Y! Final Remarks

- Many potential uses of the wisdom of people
- Same ideas can be applied to digital libraries
 - Usage logs in DLs
 - Queries as surrogate keywords for documents
 - Practical pseudo-taxonomies?
 - Focus on what people really need!