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

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

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

- News Index
  - RSS Feeds
  - Contracted information

Produced data

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

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

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

heterogeneous, large, dangerous.

very high quality & structure, expensive, sparse, safe

high quality, sparse, redundant

homogeneous, high quality, safer, highly structured

Trusted, high quality, sparse

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)
- 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
- 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. **BuyCostumes -- Children's Halloween Costumes & Costume Accessories**
   From Barbie to Winnie the Pooh, for the latest children's Halloween costumes and costume accessories, visit BuyCostumes.com today.

2. **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. And, much more. We also carry a wide variety of costume accessories, costume wigs, costume makeup, Halloween masks, Halloween deco, and much more.
   www.e-halloweencostumes.com

3. **Amazon.com: Halloween Costumes (Singer Sewing Reference Library): Books, Cowles Creative Publishing**
   Halloween Costumes (Singer Sewing Reference Library) (Hardcover...Illegal easy Halloween costumes for kids by Leila Peltz-Scarr...)
   www.amazon.com/exec/obidos/redirect/-/0865733179?vegnance

4. **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

   Halloween Costumes (Singer Sewing Reference Library) (Pb...Illegal easy Halloween costumes for kids by Leila Peltz-Scarr...)
   www.amazon.com/exec/obidos/redirect/-/0865733179?vegnance

6. **Halloween costumes at Pinatas.com**

7. **Halloween - Wikipedia**
   Hypertextual history of the holiday and its traditions. Also includes information about Halloween symbols, cultural history, and religious viewpoints.
   en.wikipedia.org/wiki/Halloween

8. **CBS News | Halloween Costumes Go Upscale | October 5, 2004 10:54:35**
   Halloween Costumes Go Upscale

9. **Halloween Costumes - Space related Halloween Costumes**
   Space-related Halloween Costumes have become popular in the United States, with plenty of Halloween parties every year, with everyone wearing Halloween costumes. Be the hit...
   space.about.com/b/a/2006/07/01/halloween.htm

10. **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/subject/45943/192639

11. **Halloween costumes and supplies at the lowest prices! - Halloween**
    Halloween costumes and supplies at the lowest prices! Halloween decorations, halloween masks, halloween party supplies, child halloween costumes, children's halloween costumes, adult halloween costumes...
    www.halloweencostumes-masks.com

12. **Halloween Costumes and Masks - Store Bought or Homemade Costumes and Masks**
    Store bought or homemade costumes and masks. 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...
Relevance of the Context

- There is no information without context
- Context and hence, content, will be implicit
- Balancing act: information vs. form
  - 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, ...)
<table>
<thead>
<tr>
<th>SEARCH GOAL</th>
<th>DESCRIPTION</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Navigational</td>
<td>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.</td>
<td>aol/a airlines duke university hospital kelly blue book</td>
</tr>
<tr>
<td>2. Informational</td>
<td>My goal is to learn something by reading or viewing web pages.</td>
<td>what is a supercharger 2004 election dates baseball death and injury why are metals shiny color blindness jfk jr help quitting smoking walking with weights pella windows phone card travel amsterdam universities florida newspapers</td>
</tr>
<tr>
<td>2.1 Directed</td>
<td>I want to learn something in particular about my topic.</td>
<td></td>
</tr>
<tr>
<td>2.1.1 Closed</td>
<td>I want to get an answer to a question that has a single, unambiguous answer.</td>
<td></td>
</tr>
<tr>
<td>2.1.2 Open</td>
<td>I want to get an answer to an open-ended question, or one with unconstrained depth.</td>
<td></td>
</tr>
<tr>
<td>2.2 Undirected</td>
<td>I want to learn anything/everything about my topic. A query for topic X might be interpreted as &quot;tell me about X.&quot;</td>
<td></td>
</tr>
<tr>
<td>2.3 Advice</td>
<td>I want to get advice, ideas, suggestions, or instructions.</td>
<td></td>
</tr>
<tr>
<td>2.4 Locate</td>
<td>My goal is to find out whether/where some real world service or product can be obtained.</td>
<td></td>
</tr>
<tr>
<td>2.5 List</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>3. Resource</td>
<td>My goal is to obtain a resource (not information) available on web pages.</td>
<td></td>
</tr>
<tr>
<td>3.1 Download</td>
<td>My goal is to download a resource that must be on my computer or other device to be useful.</td>
<td></td>
</tr>
<tr>
<td>3.2 Entertainment</td>
<td>My goal is to be entertained simply by viewing items available on the result page.</td>
<td></td>
</tr>
<tr>
<td>3.3 Interact</td>
<td>My goal is to interact with a resource using another program/service available on the web site I find.</td>
<td></td>
</tr>
<tr>
<td>3.4 Obtain</td>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

Kang & Kim, SIGIR 2003

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

**Figure 15: Anchor usage rate**

**Figure 16: Query term distribution**

**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

Figure 5: Click distributions for sample navigational queries
Figure 6: Click distributions for sample informational queries
Figure 7: Anchor-link distributions for sample navigational queries
Figure 8: Anchor-link distributions for sample informational queries
Prediction power:
- Single features: 80%
- Mixed features: 90%
- Drawbacks: Small evaluation, a posteriori feature

![Figure 11: Median of click distribution](image)
![Figure 13: Median of anchor-link distribution](image)

![Figure 12: Avg # of clicks per query](image)

---

**Y! 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
    \[ q[i] = \sum_{URL_u} \frac{Pop(q, u) \times Tf(t_i, u)}{\max_t Tf(t, u)} \]

- Recommend a better query (precise goal)
  - Query ranking
    \[ \text{Rank}(q) = \gamma \times \text{Sup}(q, q_{ini}) + (1 - \gamma) \times \text{Clos}(q) \]

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

Clusters Examples

<table>
<thead>
<tr>
<th>Q</th>
<th>Cluster Rank</th>
<th>ISim</th>
<th>ESim</th>
<th>Queries in Cluster</th>
<th>Descriptive keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1</td>
<td>252</td>
<td>0.447</td>
<td>0.007</td>
<td>car sales, cars Iquique, cars used, diesel, new cars</td>
<td>cars (49.4%), used (14.2%), stock (3.8%), pickup truck (3.7%), jeep (1.6%)</td>
</tr>
<tr>
<td>q2</td>
<td>497</td>
<td>0.313</td>
<td>0.009</td>
<td>stamp, serigraph inputs, ink reload, cartridge</td>
<td>print (11.4%), ink (7.3%), stamping (3.8%), inkjet (3.6%)</td>
</tr>
<tr>
<td>q3</td>
<td>84</td>
<td>0.697</td>
<td>0.015</td>
<td>office rental, rentals in Santiago, real state, apartment rental</td>
<td>office (11.6%), building (7.5%), real state (5.9%), real state agents (4.2%)</td>
</tr>
</tbody>
</table>
• Improved ranking

• Word classification
  – Synonyms are in the same cluster
  – Homonyms (polysemy) are in different clusters

• Query recommendation
  – Real queries, not query expansion

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

- navy of chile
- sea of chile
- naval battle of iquique
- biography arturo prat
- government of jose joaquin prieto
- ancon treaty
• Infer topics from queries that imply documents

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) business:finances:banks</td>
<td>negocios:finanzas:banco</td>
</tr>
<tr>
<td>(2) society:law:norm:codes</td>
<td>sociedad:derecho:normas:códigos</td>
</tr>
<tr>
<td>(3) business:building-industry:builders</td>
<td>negocios:construcción:constructoras</td>
</tr>
<tr>
<td>(4) business:environment:engineering</td>
<td>negocios:medio-ambiente:ingeniería</td>
</tr>
<tr>
<td>(5) business:gifts:flowers</td>
<td>negocios:compras:regalos:flor</td>
</tr>
<tr>
<td>(6) society:history</td>
<td>sociedad:historia</td>
</tr>
<tr>
<td>(7) leisure:sports:motorcycling</td>
<td>tiempo libre:deportes:motociclismo</td>
</tr>
<tr>
<td>(8) business:informatics:supp</td>
<td>negocios:informática:soporte</td>
</tr>
<tr>
<td>(9) leisure:gastronomy:drinks:</td>
<td>tiempo libre:gastronomía:bebidas:vino</td>
</tr>
<tr>
<td>(10) business:foreign trade:</td>
<td>negocios:comercio exterior:zonas francas</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th>Number of Docs</th>
<th>Relevant</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>100</td>
<td>83</td>
<td>83%</td>
<td>71%</td>
</tr>
<tr>
<td>$H$</td>
<td>100</td>
<td>76</td>
<td>76%</td>
<td>65%</td>
</tr>
<tr>
<td>$H \cap A$</td>
<td>48</td>
<td>43</td>
<td>93%</td>
<td>37%</td>
</tr>
<tr>
<td>$H - A$</td>
<td>52</td>
<td>33</td>
<td>63%</td>
<td>28%</td>
</tr>
<tr>
<td>$A - H$</td>
<td>52</td>
<td>40</td>
<td>77%</td>
<td>34%</td>
</tr>
</tbody>
</table>

• Quality of answers

[Graph showing precision (%) vs. number of results]
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