Query Log Mining
Prof. Ricardo Baeza-Yates, Yahoo! Research, Barcelona, Spain
Dr. Fabrizio Silvestri, ISTI - CNR, Pisa, Italy

What is a Web Search Engine
Real Web Search Engines

History in Search Engines

Alphonse de Lamartine


History Teaches Everything... Even the Future!
What is History?

- Past Queries
- Query Sessions
- Clickthrough Data

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
Tutorial Outline

- Query Logs
- Data Mining Techniques for QL Mining
- Enhancing Efficiency of Search Systems
- Enhancing Effectiveness of Search Systems
- New Directions

Tutorial Outline

- Query Logs
  - The Nature of Queries
  - User Actions
- Data Mining Techniques for QL Mining
- Enhancing Efficiency of Search Systems
- Enhancing Effectiveness of Search Systems
- New Directions
What’s in Query Logs?

The 250 most frequent queried terms in the “famous” AOL query log!
Thanks to http://www.wordle.net for the tagcloud generator

Some Examples!

• AOL's user 2708:
  • revenge tactics
  • the woman's book of revenge
  • dirty tricks for chicks
  • ...
  • locatecell.com
  • what can i do to an old lover for revenge
  • mean revenge tactics
  • death records in hampstead new hampshire
Some Examples

• AOL User 23187425 typed the following queries within a 10 minutes time-span:
  • you come forward 2006-05-07 03:05:19
  • start to stay off 2006-05-07 03:06:04
  • i have had trouble 2006-05-07 03:06:41
  • time to move on 2006-05-07 03:07:16
  • all over with 2006-05-07 03:07:59
  • joe stop that 2006-05-07 03:08:36
  • i can move on 2006-05-07 03:09:32
  • give you my time in person 2006-05-07 03:10:07
  • never find a gain 2006-05-07 03:10:47
  • i want change 2006-05-07 03:11:15
  • know who iam 2006-05-07 03:11:55
  • curse have been broken 2006-05-07 03:12:30
  • told shawn lawn mow burn up 2006-05-07 03:13:50
  • burn up 2006-05-07 03:14:14
  • was his i deal 2006-05-07 03:15:13
  • i would have told him 2006-05-07 03:15:46
  • to kill him too 2006-05-07 03:16:18

I Love Alaska!

• http://www.minimovies.org/documentaires/view/ilovealaska

• “I love Alaska tells the story of one of those AOL users. We get to know a religious middle-aged woman from Houston, Texas, who spends her days at home behind her TV and computer. Her unique style of phrasing combined with her putting her ideas, convictions and obsessions into AOL’s search engine, turn her personal story into a disconcerting novel of sorts.

Over a period of three months, a portrait of a woman emerges who is diligently searching for likeminded souls. The list of her search queries read aloud by a voice-over reads like a revealing character study of a somewhat obese middle-aged lady in her menopause, who is looking for a way to rejuvenate her sex life. In the end, when she cheats on her husband with a man she met online, her life seems to crumble around her. She regrets her deceit, admits to her Internet addiction and dreams of a new life in Alaska.”
Query Logs Analyzed in the Literature

<table>
<thead>
<tr>
<th>Query log name</th>
<th>Public</th>
<th>Period</th>
<th># Queries</th>
<th># Sessions</th>
<th># Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excite '97</td>
<td>Y</td>
<td>Sep '97</td>
<td>1,025,908</td>
<td>211,063</td>
<td>~ 410,360</td>
</tr>
<tr>
<td>Excite '97 (small)</td>
<td>Y</td>
<td>Sep '97</td>
<td>51,473</td>
<td>N.D.</td>
<td>~ 18,113</td>
</tr>
<tr>
<td>Altavista</td>
<td>N</td>
<td>Aug 2nd - Sep 13th '98</td>
<td>993,208,159</td>
<td>285,474,117</td>
<td>N.D.</td>
</tr>
<tr>
<td>Excite '99</td>
<td>Y</td>
<td>Dec '99</td>
<td>1,025,910</td>
<td>325,711</td>
<td>~ 540,000</td>
</tr>
<tr>
<td>Excite '01</td>
<td>Y</td>
<td>May '01</td>
<td>1,025,910</td>
<td>262,025</td>
<td>~ 446,000</td>
</tr>
<tr>
<td>Altavista (public)</td>
<td>Y</td>
<td>Sep '01</td>
<td>7,175,648</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>Tiscali</td>
<td>N</td>
<td>Apr '02</td>
<td>3,278,211</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>TodoBR</td>
<td>Y</td>
<td>Jan - Oct '03</td>
<td>22,589,568</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>TodoCL</td>
<td>N</td>
<td>May - Nov '03</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>AOL (big)</td>
<td>N</td>
<td>Dec 26th '03 - Jan 1st '04</td>
<td>~ 100,000,000</td>
<td>N.D.</td>
<td>~ 50,000,000</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>N</td>
<td>Nov '05 - Nov '06</td>
<td>N.D.</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
<tr>
<td>AOL (small)</td>
<td>Y</td>
<td>Mar 1st - May 31st '06</td>
<td>36,389,567</td>
<td>N.D.</td>
<td>N.D.</td>
</tr>
</tbody>
</table>

Caveat Emptor!

- We will show results from published papers.
- No results have been computed for the purpose of this Tutorial.
- No query logs were harmed during the preparation of this tutorial :-)

Some Popular Terms: Excite and Altavista

<table>
<thead>
<tr>
<th>query</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Empty Query</em></td>
<td>2,586</td>
</tr>
<tr>
<td>sex</td>
<td>229</td>
</tr>
<tr>
<td>chat</td>
<td>58</td>
</tr>
<tr>
<td>lucky number generator</td>
<td>56</td>
</tr>
<tr>
<td>p****</td>
<td>55</td>
</tr>
<tr>
<td>porno</td>
<td>55</td>
</tr>
<tr>
<td>b****y</td>
<td>55</td>
</tr>
<tr>
<td>nude beaches</td>
<td>52</td>
</tr>
<tr>
<td>playboy</td>
<td>46</td>
</tr>
<tr>
<td>bondage</td>
<td>46</td>
</tr>
<tr>
<td>porn</td>
<td>45</td>
</tr>
<tr>
<td>rain forest restaurant</td>
<td>40</td>
</tr>
<tr>
<td><em>f</em>**<em>ing</em></td>
<td>40</td>
</tr>
<tr>
<td>crossdressing</td>
<td>39</td>
</tr>
<tr>
<td>crystal methamphetamine</td>
<td>36</td>
</tr>
<tr>
<td>consumer reports</td>
<td>35</td>
</tr>
<tr>
<td>xxx</td>
<td>34</td>
</tr>
<tr>
<td>nude tanya harding</td>
<td>33</td>
</tr>
<tr>
<td>music</td>
<td>33</td>
</tr>
<tr>
<td>sneaker stories</td>
<td>32</td>
</tr>
</tbody>
</table>

(a) Excite.

<table>
<thead>
<tr>
<th>query</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>christmas photos</td>
<td>31,554</td>
</tr>
<tr>
<td>lyrics</td>
<td>15,818</td>
</tr>
<tr>
<td>cracks</td>
<td>12,670</td>
</tr>
<tr>
<td>google</td>
<td>12,210</td>
</tr>
<tr>
<td>gay</td>
<td>10,945</td>
</tr>
<tr>
<td>harry potter</td>
<td>7,933</td>
</tr>
<tr>
<td>wallpapers</td>
<td>7,848</td>
</tr>
<tr>
<td>pornografia</td>
<td>6,893</td>
</tr>
<tr>
<td><em>yahoo com</em></td>
<td>6,753</td>
</tr>
<tr>
<td>juegos</td>
<td>6,550</td>
</tr>
<tr>
<td>lingerie</td>
<td>6,078</td>
</tr>
<tr>
<td>symbiosis logic 53c400a</td>
<td>5,701</td>
</tr>
<tr>
<td>letras de canciones</td>
<td>5,518</td>
</tr>
<tr>
<td>humor</td>
<td>5,400</td>
</tr>
<tr>
<td>pictures</td>
<td>5,293</td>
</tr>
<tr>
<td>preteen</td>
<td>5,137</td>
</tr>
<tr>
<td>hypnosis</td>
<td>4,550</td>
</tr>
<tr>
<td>cpc view registration key</td>
<td>4,553</td>
</tr>
<tr>
<td>sex stories</td>
<td>4,021</td>
</tr>
<tr>
<td>cd cover</td>
<td>4,267</td>
</tr>
</tbody>
</table>

(b) Altavista.


Topic Distribution: Excite and AOL

<table>
<thead>
<tr>
<th>Topic</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertainment or recreation</td>
<td>19.9%</td>
</tr>
<tr>
<td>Sex and pornography</td>
<td>16.8%</td>
</tr>
<tr>
<td>Commerce, travel, employment, or economy</td>
<td>13.3%</td>
</tr>
<tr>
<td>Computers or Internet</td>
<td>12.5%</td>
</tr>
<tr>
<td>Health or sciences</td>
<td>9.5%</td>
</tr>
<tr>
<td>People, places, or things</td>
<td>6.7%</td>
</tr>
<tr>
<td>Society, culture, ethnicity, or religion</td>
<td>5.7%</td>
</tr>
<tr>
<td>Education or humanities</td>
<td>5.6%</td>
</tr>
<tr>
<td>Performing or fine arts</td>
<td>5.4%</td>
</tr>
<tr>
<td>Non-English or unknown</td>
<td>4.1%</td>
</tr>
<tr>
<td>Government</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

Excite

<table>
<thead>
<tr>
<th>Topic</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertainment</td>
<td>13%</td>
</tr>
<tr>
<td>Shopping</td>
<td>13%</td>
</tr>
<tr>
<td>Porn</td>
<td>10%</td>
</tr>
<tr>
<td>Research &amp; learn</td>
<td>9%</td>
</tr>
<tr>
<td>Computing</td>
<td>9%</td>
</tr>
<tr>
<td>Health</td>
<td>5%</td>
</tr>
<tr>
<td>Home</td>
<td>5%</td>
</tr>
<tr>
<td>Travel</td>
<td>5%</td>
</tr>
<tr>
<td>Games</td>
<td>5%</td>
</tr>
<tr>
<td>Personal &amp; Finance</td>
<td>3%</td>
</tr>
<tr>
<td>Sports</td>
<td>3%</td>
</tr>
<tr>
<td>US Sites</td>
<td>3%</td>
</tr>
<tr>
<td>Holidays</td>
<td>1%</td>
</tr>
<tr>
<td>Other</td>
<td>16%</td>
</tr>
</tbody>
</table>

AOL
Long Tail Distribution

URLs ordered by number of clicks

Power-Law In Query Popularity: Altavista

Power-Law In Query Popularity: Excite


Power-Law In Query Popularity: Yahoo!

Query Resubmission


Frequency of Query Submission

Query Statistics: Excite

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>1997</th>
<th>1999</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean terms per query</td>
<td>2.4</td>
<td>2.4</td>
<td>2.6</td>
</tr>
<tr>
<td>Terms per query</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 term</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 terms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3+ terms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean queries per user</td>
<td>2.5</td>
<td>1.9</td>
<td>2.3</td>
</tr>
</tbody>
</table>

In 2008: 2.5 terms per query.


Hourly Topic Distribution

Surprising Topics

- KL-Divergence between observing a query topic u.a.r. and the actual topic observed.

\[
D(p(q|t) \| p(q|c, t)) = \sum_q p(q|t) \log \frac{p(q|t)}{p(q|c, t)}
\]


Tutorial Outline

- Query Logs
- Data Mining Techniques for QL Mining
  - “Classical” DM Tasks
  - New Mining Tasks for Query Logs
- Enhancing Efficiency of Search Systems
- Enhancing Effectiveness of Search Systems
- New Directions
Data Mining

• Many Definitions

• Non-trivial extraction of implicit, previously unknown and potentially useful information from data

• Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Typical DM tasks

• Given a large collection of documents, all concerning sport, build automatically a classifier. Use it to determine all the queries whose topic is sport with high probability

• Subdivide the queries in distinct clusters, so that queries in the same cluster are much more similar to each other than to others in different clusters
What is not DM

• Find all the web search queries that include the phrase “September 11”
• Look at a query log, and select all the queries submitted by the same user ID2378

DM is part of a process

The iterative and exploratory Knowledge Discovery process
Origins of DM

- Traditional techniques may be unsuitable
- Data have huge size, high dimensionality, are heterogeneous and distributed

DM tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other unknown variables

- Description Methods
  - Find human-interpretable patterns that describe the data
DM tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Outlier Detection [Predictive]

Web Mining

- DM techniques applied to Web data
- Web content mining
  - From text, image, in general contents of Web pages
- Web structure mining
  - From hyperlink structure (graph) of Web
- Web usage mining
  - From usage data, like logs
Classification

• Collection of records (training set)
  • One of the attributes is the class
• Find a model for the class attribute
  • A function of the values of other attributes.
• Goal of the model
  • Previously unseen records should be assigned a class (classified) as accurately as possible
  • A test set is used to determine the accuracy of the model

Classification example

<table>
<thead>
<tr>
<th>Qid</th>
<th>Query + snippets</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>soccer, Kaka, ...</td>
<td>Sport</td>
</tr>
<tr>
<td>2</td>
<td>Berlusconi, Sardinia, ...</td>
<td>Politics</td>
</tr>
<tr>
<td>3</td>
<td>tennis, Federer, ...</td>
<td>Sport</td>
</tr>
<tr>
<td>4</td>
<td>Nicolas, Carla, ...</td>
<td>Politics</td>
</tr>
<tr>
<td>5</td>
<td>swimming, Phelps, team, ...</td>
<td>Sport</td>
</tr>
<tr>
<td>6</td>
<td>Obama, president, elect, ...</td>
<td>Politics</td>
</tr>
<tr>
<td>7</td>
<td>hundred, outdoor, race, ...</td>
<td>Sport</td>
</tr>
<tr>
<td>8</td>
<td>beijing, olympic, ...</td>
<td>Sport</td>
</tr>
<tr>
<td>9</td>
<td>Brown, bank, ...</td>
<td>Politics</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Qid</th>
<th>Query + snippets</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pelé, Brasil, ...</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Veltroni, Left, Centre, ...</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Nadal, court, ...</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>President, Carla, ...</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>football, rugby, ...</td>
<td>1</td>
</tr>
</tbody>
</table>
Clustering

• Given
  • a set of records
  • a similarity measure based on the attributes

• Find clusters such that
  • Data points in one cluster are more similar to one another: **MAX** intra-cluster similarity
  • Data points in separate clusters are less similar to one another: **MIN** inter-cluster similarity

Clustering example

• 3D-space: each record is a point corresponding to an \( \mathbb{R}^3 \) vector
• Similarity measure: Euclidean distance
Clustering of text data

• Each document: term vector with word frequency

• Similarity measure:

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>pt</th>
<th>pd</th>
<th>lost</th>
<th>won</th>
<th>game</th>
<th>score</th>
<th>win</th>
<th>tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Document 2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Clustering of queries

• Query are too short text documents

• Expanded representation for the query “apple pie” by using snippet elements [Metzler et al. ECIR07]
Association rules

- Market basket analysis
- extract rules from transactional databases
- each transaction corresponds to a customer basket

<table>
<thead>
<tr>
<th>TID</th>
<th>items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>beer, coke, milk</td>
</tr>
<tr>
<td>2</td>
<td>water, coke, chips, milk</td>
</tr>
<tr>
<td>3</td>
<td>coke, milk</td>
</tr>
<tr>
<td>4</td>
<td>milk, bread</td>
</tr>
<tr>
<td>5</td>
<td>bread, water, coke</td>
</tr>
</tbody>
</table>

- Example of unveiled rule
  coke ➔ milk
  \[\text{sup}=60\%, \text{conf}=75\%\]

Association Rules for Query Expansion

- A user submitting the query “Obama”
- The system suggests to expand the query with the word “President”

<table>
<thead>
<tr>
<th>QueryID</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Elect Obama</td>
</tr>
<tr>
<td>2</td>
<td>President US Obama</td>
</tr>
<tr>
<td>3</td>
<td>Obama President</td>
</tr>
<tr>
<td>4</td>
<td>President Sarcozy</td>
</tr>
<tr>
<td>5</td>
<td>President Obama</td>
</tr>
</tbody>
</table>

- Example of unveiled rule
  Obama ➔ President
  \[\text{sup}=60\%, \text{conf}=75\%\]
Sequential patterns

• For each user (customer), record sequences of associated events
• Temporally ordered sequences of events (e.g., sets of bought objects)
• Find frequent rules that predict strong sequential dependencies among different events
Sequential mining example

- Sessioned query logs
  - A session is a temporal sequence, where each query can be seen as a transaction of words
- Extract frequent features from each query
  - E.g. single words or pairs of words
- Find frequent sequences, and then association rules
- Use the rules for query suggestions
  - (“mortgage loan” → “subprime mortgage”) → “lehman brothers”

Predict continuous variables by regression

- Linear regression: \( Y = a + bX \)
- Model a variable \( Y \) (to be predicted) as a linear function of \( X \)
- Determine coefficients \( a \) and \( b \) on the basis of the training set

<table>
<thead>
<tr>
<th>( X ) years experience</th>
<th>( Y ) salary (in $1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>57</td>
</tr>
<tr>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>13</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>21</td>
<td>90</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>83</td>
</tr>
</tbody>
</table>

\[ Y = 21.7 + 3.7X \]
Regression to learn query ranking

- For a query-document pair \((q; d)\), extract a feature vector \(x = [x_Q; x_D; x_{QD}]\)

  - examples of features extracted from query-document pair \((q; d)\)
    - \(x_Q\): e.g., number of terms in query \(q\)
    - \(x_D\): e.g., the language identity of document \(d\)
    - \(x_{QD}\): e.g., the number of times each term in \(q\) appears in the anchor-texts of document \(d\)

- Assign a numerical grade to each pair \((q; d)\) based on the degree of relevance inferred from the query clicktroughs in the logs

- Use numerical grades as target values for multivariate regression based on the query-document features

Tutorial Outline

- Query Logs
- Data Mining Techniques for QL Mining
- Enhancing Efficiency of Search Systems
  - Caching
  - Index Partitioning and Querying in Distributed IR Systems
- Enhancing Effectiveness of Search Systems
- New Directions
Sketching a Distributed Search Engine

Caching in General

Larger, but slower memory

Smaller, but faster memory

CPU
This is true on an ideal world.
Filtering Effect of Caching


Filtering Effect of Caching

Filtering Effect of Caching

Caching Performance Evaluation

- **Hit-Ratio**: i.e. how many times the cache is useful

- **Query Throughput**: i.e. the number of queries the cache can serve in a second

But... what really impacts on caching performance?

---

Caching for Search Engines Workloads

- Caching Architectures:
  - Two-Level Caching
  - Three-Level Caching

- Caching Policies
  - PDC
  - SDC
  - AC

Two-Level Caching

- Firstly studied in:

- Further analyzed in:
Two-Level Caching

Broker

Result Cache

Posting Cache

IR Core

1st level

2nd level

Three-level Caching

- Adds one level between results and posting lists cache.
- Usually stores frequently occurring pairs of terms.


“Things” to Cache in Search Engines

- Results
  - in answer to a user query
- Posting Lists
  - e.g. for the query “new york” cache the posting lists for term new and for term york
- Partial queries
  - cache subqueries, e.g. for “new york times” cache only “new york”

Cache Replacement Policies

![Diagram showing cache replacement policies](image_url)
Cache Replacement Policies

Cache

Is the object in cache?

Replacement Policy

Traditional Replacement Policies

- LRU
- LFU
- SLRU
- ...

Hit Ratios on Excite

SLRU vs. LRU on Excite
Search Engine Tailored Policies

- PDC
  - Probability Driven Caching
- SDC
  - Static Dynamic Caching
- AC
  - Admission Control

IDEA: design a policy tailored over users' behavior on search pages

With high probability users do not go beyond the first page of results

For some query users browse many result pages.

PDC Priorities

- Priorities are assigned using an approximation of the Markovian SERP request model
- Each SERP different from the first one has a priority computed on historical queries (query log)
  - we cache pages that have follow-up queries more likely to be submitted. Why?

PDC and Prefetching

- in PDC results are organized according to “Fetch Units”
- When SERP i is requested for a query Q, we look up the cache to probe its presence.
- If i is not cached, we request SERP i, i + 1, ..., i + f
- That is we prefetch f SERPs.
- The fetch unit is of size f.


PDC Results

![Chart showing PDC Results]

PDC’s Main Drawback

- Priority Queue Housekeeping Complexity.
  - O(log k) (amortized)
  - LRU is O(1)

PDC’s Main Lessons Learned

- Hit ratio benefits a lot from the use of historical data
- Prefetching helps a lot!
- Differently from previous caching policies, PDC not necessarily caches every submitted queries!!!

Overcoming PDC Complexity

- PDC uses query logs to estimate the likelihood of follow-up queries.
- Why not using query logs to estimate likelihood of resubmitting a query.
- Catching the head of the long tail distribution we might obtain high hit ratios.

That is...

~80% of submitted queries represents the 20% of the unique queries submitted.
But...

Static Dynamic Caching

- SDC (Static Dynamic Caching) adds to the classical static caching schema a dynamically managed section.

- The idea:

  - LRU
  - SLRU
  - PDC
  - ...

SDC and Prefetching

- SDC adopts an “adaptive” prefetching technique:
  - For the first SERP do not prefetch
  - For the follow-up SERPs prefetch \( f \) pages


SDC Hit-Ratios

![SDC Hit-Ratios Chart]

Altavista: hit-ratio vs. \( f_{\text{static}} \) and prefetching factor. Dynamic set policies: LRU, PDC. Size 256,000
SDC’s Main Lessons Learned

• Hit ratio benefits a lot from the use of historical data
• Prefetching helps a lot!
• Static caching alone is not useful, yet...
  • A good combination of a static and a dynamic approach helps a lot!!!


That’s not All Folks!

Admission Control

• An interesting idea of SDC: frequent queries are cached permanently
• AC of Baeza-Yates et al. generalizes the idea by using two dynamically updated sets:
  • A Controlled Cache (CC)
  • An Uncontrolled Cache (UC)
• When a new query arrives an admission policy is applied to steer a query to the CC or to the UC.
• If the query is likely to be seen in the future move it to CC, otherwise send it to UC.


Admission Policy

• Makes use of features, e.g.:
  • Stateful features:
    • $PastF$: the frequency of the query in the (relatively recent) past
  • Stateless features:
    • $LenC$: the length of the query in characters
    • $LenW$: the length of the query in words

Hit-Ratio Results (Past1-5)

Altavista log

Hit ratio (%)

Infinite
LRU, 100k
SDC, 100k
AC, 100k

Frequency threshold

Hit-Ratio Results (LenC- LenW)

UK log

Hit ratio (%)

Infinite
LRU, 500k
SDC, 500k
AC, 500k

Length in characters threshold

UK log

Hit ratio (%)

Infinite
LRU, 500k
SDC, 500k
AC, 500k

Length in words threshold

Caching Posting Lists

- SERP size is fixed
- Posting lists have different lengths.
- Posting list caching techniques adopt policies sensitive to list sizes.

\[ Q_{TFDF} \] Policy

- Idea:
  - suppose you have 10 free slots and 3 postings lists to cache \( l_1, l_2, \) and \( l_3 \). \( l_1 \) appears 10 times and it is long 6 postings, \( l_2 \) and \( l_3 \) appear 6 times each and are long 5 postings.
  - Traditional frequency-only-based policies will choose to cache \( l_1 \) filling up 6 slots and not leaving space for any of the two other lists.
  - \( Q_{TFDF} \) decides to cache \( l_2 \) and \( l_3 \) since they optimize the ratio frequency/size instead of just frequency.

- Results:
  - Traditional static caching has a hit ratio of 10
  - \( Q_{TFDF} \) static policy has a hit ratio of 12
Q_{TFDF} Results

Caching posting lists -- UK dataset

SDC-like Q_{TFDF}

Adding dynamic cache for caching posting lists
Not Only Caching

- Improve efficiency using query logs can also be done by:
  - query routing
  - data/index partitioning

Sketching a Distributed Search Engine
Index Partitioning

Term Partitioning Systems

- Query routing is “trivial”...
- Whenever $Q=(t_1, t_2, \ldots, t_n)$ is received route it to the servers containing those terms.
- But..
  - *not scalable* (indexing is $n \log n$, term partitioning needs reindexing the entire collection from scratch)
  - *load unbalancing* among IR cores
Why Studying Term Partitioned Systems?

• In principle:
  • less IR Cores queried
  • less operation performed
• Briefly...
  • More available capacity!

Term Partitioning
(Random Partitioning)

IR Core 1

T_{10} \ldots T_{12}

IR Core 2

T_{1} \ldots T_{3}

IR Core 3

T_{1} \ldots T_{9}

IR Core 4

T_{2} \ldots T_{8}
Pipelined Term Part.
(Random Partitioning)

How can we...

- Balance the load?
- Better exploit resources?
- In light of... The Power Law!!!!
Two Approaches in Literature


The Idea...

- If terms co-occur frequently in past queries...
  - pack them together up in the same IR Core.
- but...
  - Power law prevents load balancing...
    - Knapsack problem can help in balancing the load.
    - Fit in partitions terms with weight \( L_t = Q_t \times B_t \)
      \( (Q_t \) occurrences of \( t \) in the query log, \( B_t \) length of \( t \)’s postings list)
Load Balancing

T₂ T₄ T₃  T₂ T₄ T₃  T₂ T₄ T₃  T₂ T₄ T₃

Query Broker


Load Balancing + Replication

T₂ T₄ T₃  T₂ T₄ T₃  T₂ T₄ T₃  T₂ T₄ T₃

Query Broker

IR Core 1  IR Core 2  IR Core 3  IR Core 4

T₂  T₄  T₃  Whatever

Load Balancing Results

- On .GOV2 collection. Queries adapted from WT10G.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1.45</td>
<td>1.44</td>
<td>1.46</td>
<td>1.50</td>
<td>1.48</td>
<td>1.47</td>
</tr>
<tr>
<td>Using $f_t$</td>
<td>1.43</td>
<td>1.20</td>
<td>1.23</td>
<td>1.40</td>
<td>1.42</td>
<td>1.34</td>
</tr>
<tr>
<td>Past $L_t$</td>
<td>1.14</td>
<td>1.26</td>
<td>1.23</td>
<td>1.19</td>
<td>1.20</td>
<td>1.20</td>
</tr>
<tr>
<td>Current $L_t$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>


Load Balancing + Replication Results

- On .GOV2 collection. Queries adapted from WT10G.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicate 1</td>
<td>1.26</td>
<td>1.20</td>
<td>1.10</td>
<td>1.17</td>
<td>1.11</td>
<td>1.17</td>
</tr>
<tr>
<td>Duplicate 10</td>
<td>1.06</td>
<td>1.29</td>
<td>1.17</td>
<td>1.18</td>
<td>1.16</td>
<td>1.17</td>
</tr>
<tr>
<td>Duplicate 100</td>
<td>1.09</td>
<td>1.14</td>
<td>1.10</td>
<td>1.13</td>
<td>1.15</td>
<td>1.12</td>
</tr>
<tr>
<td>Duplicate 1000</td>
<td>1.08</td>
<td>1.09</td>
<td>1.07</td>
<td>1.19</td>
<td>1.09</td>
<td>1.10</td>
</tr>
<tr>
<td>Multi-replicate</td>
<td>1.05</td>
<td>1.12</td>
<td>1.09</td>
<td>1.16</td>
<td>1.12</td>
<td>1.11</td>
</tr>
</tbody>
</table>

But...

- Query Throughput.
- On .GOV2 collection. Queries from a real life MSN query log

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Batch 2</th>
<th>Batch 3</th>
<th>Batch 4</th>
<th>Batch 5</th>
<th>Batch 6</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hashed</td>
<td>1.82</td>
<td>1.82</td>
<td>1.86</td>
<td>1.84</td>
<td>1.83</td>
<td>1.83</td>
</tr>
<tr>
<td>Duplicate 100</td>
<td>2.21</td>
<td>2.14</td>
<td>2.25</td>
<td>2.17</td>
<td>2.20</td>
<td>2.19</td>
</tr>
<tr>
<td>Doc-distributed</td>
<td>2.21</td>
<td>2.25</td>
<td>2.24</td>
<td>2.31</td>
<td>2.27</td>
<td>2.26</td>
</tr>
</tbody>
</table>


Adding a Dimension

- Number of IR Cores used by each query
- Basically, instead of sending lists around try to keep many query processing local to a node.
  - More load unbalance
- Trade-off: The Term-Assignment Problem

The Term-Assignment Problem

Use a Frequent Itemsets Mining Algorithm to find pairs of co-occurring terms. Then, optimize Omega not only considering single terms but also term-sets.


---

Number of Queried IR Cores (Servers)

<table>
<thead>
<tr>
<th>Servers</th>
<th>Baseline Cases</th>
<th>Term Assignment $\alpha = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>random</td>
<td>bin packing</td>
</tr>
<tr>
<td>1</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>&gt; 3</td>
<td>24</td>
<td>25</td>
</tr>
</tbody>
</table>

Impact of Replications on the # of IR Cores

<table>
<thead>
<tr>
<th>Servers</th>
<th>Replication Factors</th>
<th>0.0001</th>
<th>0.0005</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bin pack.</td>
<td>term ass.</td>
<td>bin pack.</td>
<td>term ass.</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td>54</td>
<td>56</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>22</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>&gt; 3</td>
<td>15</td>
<td>14</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

The Overall Picture

With an average of (slightly more than) two IR Cores per query, the load is only slightly unbalanced

Collection Selection

- Traditionally used in Federated Distributed IR systems to reduce the number of queried servers.
- Rarely (???) used in Web Search Engine systems
- see Google’s MICRO paper on their architecture.
To Select or Not To Select?

• Pros
  • Reduced Load on IR Cores
  • Potentially Eliminates Noise due to the presence of non relevant documents w.r.t. a query

• Cons
  • Load Imbalance
  • Reduced Precision

The Curse of Reduced Precision

• The reduction in precision is an issue that have to be taken into serious consideration.

• “Luckily”, precision in collection selection architectures can be enhanced by using:
  • Ad-hoc partitioning strategies
  • Collection Prioritization
  • Incremental Caching
Query-Vector Document Model

- It is a vector-space like model.
- Documents are represented by queries they answer.
- Example.
  Query “munch” is answered by documents d1, d2, d7, d3, d4. “munch” is represented by “1 2 3 4 7”.

Possible Refinements:
- Consider Ranking Scores
- Consider Clicks


QV-Based Partitioning

The co-clustering implementation is derived from the paper:
Inderjit S. Dhillon, Subramanyam Mallela, Dharmendra S. Modha: Information-theoretic co-clustering. KDD 2003: 89-98
It is available at the following address
http://hpc.isti.cnr.it/~diego/phd/fastcluster.tgz
QV-Based Collection Selection

Comparison with SoA-Collection Selection

- CORI
- does not make use of usage information.
- It exploits simple statistics on collections’ vocabulary.
- At the Query Broker side, information are stored using quite a lot of memory.
- QV representation for collection selection is about 19% of the size needed by CORI metadata.

James P. Callan, Zhihong Lu, W. Bruce Croft. Searching Distributed Collections with Inference Networks. SIGIR 1995.
Comparison in Terms of Precise Results

<table>
<thead>
<tr>
<th>Intersection (%) at 10</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>OVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>5</td>
<td>11</td>
<td>25</td>
<td>50</td>
<td>93</td>
<td>100</td>
</tr>
<tr>
<td>QV-based Improvement</td>
<td>34</td>
<td>45</td>
<td>58</td>
<td>76</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>6.8X</td>
<td>4.1X</td>
<td>2.3X</td>
<td>1.5X</td>
<td>&gt;1X</td>
<td>-</td>
</tr>
</tbody>
</table>


Load Balancing...Again!

![Peak Load on Each IR Core](image_url)
Guess Why...

Queries here are always sent to the same IR Cores

Collection Prioritization

Diego Puppin, Raffaele Perego, Fabrizio Silvestri, Ricardo Baeza-Yates. **Tuning the Capacity of Search Engines: Load-driven Routing and Incremental Caching to Reduce and Balance the Load.**

How IR Cores Take Decisions

- **load-driven basic**<sub><L></sub>
  - Accept queries only if the load is below \((1 - \text{rank}/\text{NServers}) \times \text{L}\). E.g. the second best server accepts the query only if its load is below 90% of \text{L}.

- **load-driven boost**<sub><L, T></sub>
  - It is the same as the basic strategy except that the first \text{T} servers’ load threshold is always \text{L} and then it starts to decrease in a linear fashion as in the basic case.


---

Load... Balanced!

### What about Precision?

<table>
<thead>
<tr>
<th></th>
<th>FIXED 4</th>
<th>BASIC&lt;25&gt;</th>
<th>BOOST&lt;4, 25&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>34</td>
<td>56</td>
<td>71</td>
</tr>
<tr>
<td>10</td>
<td>34</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>20</td>
<td>34</td>
<td>68</td>
<td>70</td>
</tr>
</tbody>
</table>

Intersection (%) at

### Competitive Similarity

<table>
<thead>
<tr>
<th></th>
<th>FIXED 4</th>
<th>BASIC&lt;25&gt;</th>
<th>BOOST&lt;4, 25&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.88</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>10</td>
<td>0.87</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>20</td>
<td>0.85</td>
<td>0.89</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Incremental Caching

Does it Work?

<table>
<thead>
<tr>
<th></th>
<th>BASIC&lt;25&gt;</th>
<th>BOOST&lt;4, 25&gt;</th>
<th>BOOST&lt;4, 25&gt; + INC</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>10</td>
<td>0.90</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>20</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Tutorial Outline

- Query Logs
- Data Mining Techniques for QL Mining
- Enhancing Efficiency of Search Systems
- Enhancing Effectiveness of Search Systems
  - Query intention
  - Query recommendations
  - Spelling Correction
  - User Driven Web Design
  - Extracting Semantics from Queries
- New Directions

More on Query Characteristics
**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”.

---

**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!)
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

User Needs (Broder 2002)

- **Informational** – want to learn about something (~40% / 65%)
  - Low hemoglobin
- **Navigational** – want to go to that page (~25% / 15%)
  - Warsaw Airport
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
  - Access a service
  - Barcelona weather
  - Downloads
  - Mars surface images
  - Shop
  - Canon S410
- Gray areas
  - Find a good hub
  - Car rental Poland
  - Exploratory search “see what’s there”
Queries and Text

Term Pairs

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

- 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
  - 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 (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.

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
Session: ( (t, q), (t, URL, dt) )

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>Navigational</td>
<td>My goal is to go to a 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>aloha airlines data university hospital kelly blue book</td>
</tr>
<tr>
<td>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</td>
</tr>
<tr>
<td>Directed</td>
<td>I want to learn something in particular about my topic</td>
<td>help quitting smoking walking with weights</td>
</tr>
<tr>
<td>Closed</td>
<td>I want to get an answer to a question that has a single, unambiguous answer.</td>
<td>pella windows phone card</td>
</tr>
<tr>
<td>Open</td>
<td>I want to get an answer to an open-ended question, or one with unconstrained depth.</td>
<td>travel amsterdam universities florida newspapers</td>
</tr>
<tr>
<td>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>kasaa lite french loans xxx point movis free live camera in i.a. weather measure converter</td>
</tr>
<tr>
<td>Advice</td>
<td>I want to get advice, ideas, suggestions, or instructions.</td>
<td></td>
</tr>
<tr>
<td>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>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>Resource</td>
<td>My goal is to obtain a resource (not information) available on web pages</td>
<td></td>
</tr>
<tr>
<td>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>Entertainment</td>
<td>My goal is to be entertained simply by viewing items available on the result page</td>
<td></td>
</tr>
<tr>
<td>Interact</td>
<td>My goal is to interact with a resource using another program, perhaps to edit the web site I find.</td>
<td></td>
</tr>
<tr>
<td>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>

Rose & Levinson, WWW 2004
Query Intention

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

Features

- Click & anchor text distribution

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

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

Query Intention Revisited

- 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
Geography and Query Intent

Location 1: query location
“Pizza Amherst, MA”
query1

Distance 1:
home–query intent

Location 2: Home address
IP address / profile zip

Distance 2:
Reformulation
distance
“Pizza Northampton”
query2

Location 3: query location

Query Recommendation
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)

Our Approach

- 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
    \[ q[i] = \sum_{URL_u} \frac{Pop(q,u) \times Tf(t_i,u)}{\max_t Tf(t,u)} \]
  - Unbias the effect of the rank and the interface in the clicks
- 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%),</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>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%),</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>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%),</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>real state agents (4, 2%)</td>
</tr>
</tbody>
</table>
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

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

---

**Query Recommendation**

<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 owners</td>
<td>2</td>
<td>0.133</td>
<td>0.403</td>
<td>0.268</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 vina del mar</td>
<td>3</td>
<td>0.033</td>
<td>0.153</td>
<td>0.093</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>
Noisy Channel Model

Platonic concept of query

Typing quickly
Distracted
Forgot how to spell

Typos/spelling errors

Correct Spelling

Reconstruct original query by “reversing this process”

Modeling Errors

\[ P(q_{correct} | q_{error}) = p(q_{error} | q_{correct}) p(q_{correct}) \]

Error model

Language Model

Character level: \( p(m|n) \) \( p(s|z) \) etc

Mine web data sources for these probabilities

Query level: \( p(\text{“sigir 2008”}) \), \( p(\text{“sigir iraq”}) \)…
Correct Spelling More Common than Misspelling in Query Logs

[Cucerzan and Brill, 2004]

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>albert einstein</td>
<td>4834</td>
</tr>
<tr>
<td>albert einsten</td>
<td>525</td>
</tr>
<tr>
<td>albert einstein</td>
<td>149</td>
</tr>
<tr>
<td>albert einsten</td>
<td>27</td>
</tr>
<tr>
<td>albert einsteins</td>
<td>25</td>
</tr>
<tr>
<td>albert einstain</td>
<td>11</td>
</tr>
<tr>
<td>albert einstein</td>
<td>10</td>
</tr>
<tr>
<td>albert einstein</td>
<td>9</td>
</tr>
<tr>
<td>albeart einstein</td>
<td>6</td>
</tr>
<tr>
<td>aolbert einstein</td>
<td>6</td>
</tr>
<tr>
<td>alber einstein</td>
<td>4</td>
</tr>
<tr>
<td>albert einseint</td>
<td>3</td>
</tr>
<tr>
<td>albert einsteins</td>
<td>3</td>
</tr>
<tr>
<td>albert einsterin</td>
<td>3</td>
</tr>
<tr>
<td>albert einstein</td>
<td>3</td>
</tr>
<tr>
<td>albert einstein</td>
<td>3</td>
</tr>
<tr>
<td>alber einstein</td>
<td>3</td>
</tr>
</tbody>
</table>

Good and bad spellings point to same page

excite.com

- excitement
- exits

[Craswell et al 2001]
## Reformulations from Bad to Good Spellings

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-rewrite</td>
<td>mic amps  →  create taxi</td>
<td>53.2%</td>
</tr>
<tr>
<td>insertions</td>
<td>game codes  →  video game codes</td>
<td>9.1%</td>
</tr>
<tr>
<td>substitutions</td>
<td>john wayne bust  →  john wayne statue</td>
<td>8.7%</td>
</tr>
<tr>
<td>deletions</td>
<td>skateboarding pics  →  skateboarding</td>
<td>5.0%</td>
</tr>
<tr>
<td>spell correction</td>
<td>real eastate  →  real estate</td>
<td>7.0%</td>
</tr>
<tr>
<td>mixture</td>
<td>huston's restaurant  →  houston's</td>
<td>6.2%</td>
</tr>
<tr>
<td>specialization</td>
<td>jobs  →  marine employment</td>
<td>4.6%</td>
</tr>
<tr>
<td>generalization</td>
<td>gm rebates  →  show me all the current auto rebates</td>
<td>3.2%</td>
</tr>
<tr>
<td>other</td>
<td>thansgiving  →  dia de acconde gracias</td>
<td>2.4%</td>
</tr>
</tbody>
</table>

[Jones & Fain, 2003]

## User Driven Web Design
A Causal Model for Web Design

Ubiquity for Crawlers
Ubiquity for People

Using Common Sense

Causality: The five steps in a Web site

- Findability
- Visibility and accessibility
- Content
- Appearance
- Fidelity

Information Architecture

Ubiquity

Usability
How to Improve Web Design?

- User-driven Web design
- Web Mining: content, structure, usage
- Analysis of navigation logs
  - Improve organization
- Analysis of query logs (local site search)
  - Improve content and semantics of your text
- Example: Yahoo!
Navigation Mining:

Site Restructuring
Query Mining: Information Scent

External Queries

Internal Queries

A B C D E F

Patent pending, University of Chile

Content & Structure Mining

• Correlate Text Content and Link Structure
  – Text Clustering Baeza-Yates & Poblete, 2005
  – Link Analysis

• Tool for Query, Content and Structure Mining
  – Example of Search Analytics
  – Forthcoming book by Lou Rosenfeld
Extracting Semantics

Relating Queries (Baeza-Yates, 2007)
### Qualitative Analysis

<table>
<thead>
<tr>
<th>Graph</th>
<th>Strength</th>
<th>Sparsity</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Medium</td>
<td>High</td>
<td>Polysemy</td>
</tr>
<tr>
<td>Session</td>
<td>Medium</td>
<td>High</td>
<td>Physical sessions</td>
</tr>
<tr>
<td>Click</td>
<td>High</td>
<td>Medium</td>
<td>Click spam</td>
</tr>
<tr>
<td>Link</td>
<td>Weak</td>
<td>Medium</td>
<td>Link spam</td>
</tr>
<tr>
<td>Term</td>
<td>Medium</td>
<td>Low</td>
<td>Term spam</td>
</tr>
</tbody>
</table>

### Data Cleaning Example

- “Physical” sessions are used
  - Noise in the data
  - Topical/task oriented sessions
- Iterative cleaning procedure
  - Use clusters to detect logical sessions
  - Recluster the clicked pages
  - Repeat until there are no more changes
Clicks: URL Cover Graph

Click Distribution

Data per user is a power law
Set Relations

- 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

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 \( c_1, \ldots, c_k \) and \( c'_1, \ldots, c'_k \) be the top \( k \) categories for two queries. Define the similarity (@\( k \)) between the two queries as \( \max\{ \text{sim}(c_i, c'_j) \mid i, j = 1, \ldots, K \} \)

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)
Some Open Issues

- Implicit social network
  - Any fundamental similarities?

- How to evaluate with partial knowledge?
  - Data volume amplifies the problem

- User aggregation vs. personalization
  - Optimize common tasks: help more people
  - Move away from privacy issues
Tutorial Outline

- Query Logs
- Data Mining Techniques for QL Mining
- Enhancing Efficiency of Search Systems
- Enhancing Effectiveness of Search Systems
- New Directions
  - Eye tracking
  - Web Search Advertisement
  - Extracting Human Activities

Eye Tracking

Why Eye Tracking?

- Understand how searchers evaluate online search results
- Enhanced interface design
- More accurate interpretation of implicit feedback (eg, clickthrough data)
- More targeted metrics for evaluating retrieval performance


Research Problems

- Time to select a result
- Number of snippets looked at before clicking
- Do users click a result but also look at successive snippets
- What is the order snippets are looked at?
- Is the whole snippet read?

Indices to Look at

- Fixations
  - ~200-300ms; information is acquired
- Saccades
  - extremely rapid movements between fixations
- Pupil dilation
  - size of pupil indicates interest, arousal


Indices to Look at

Aggregate eye-tracking graphs depict viewing intensity in key regions

“Scanpath” output depicts pattern of movement throughout screen. Black markers represent fixations.
Search Tasks

- Who discovered the first modern antibiotic?
- Find the homepage of Emeril - the chef who has a TV cooking program.
- What actor starred as the main character in the original 'Time Machine' movie?
- Find the page displaying the routemap for Greyhound buses.
- You are excited to cast your vote in the democratic presidential primary - when can you do so in NY?
- Find the homepage of Michael Jordan, the statistician.
- Where is the tallest mountain in NY located?
- Find the homepage for graduate housing at Carnegie Mellon University.
- A friend told you that Mr. Cornell used to live close to campus - between University and Stewart Aves - does anyone live in his house now; if so who?
- Find the homepage of the 1,000 Acres Dude Ranch.


Time to Decide

Time Spent on Snippets


Time Spent on each Result Part

Lesson Learned

- Document selected in under 5 seconds
- Users click on the first promising link they see
- Results viewed linearly
- Top 2 results most likely to be viewed
- Users rather reformulate query than scroll
- Task type and difficulty affect viewing behavior
- Presentation of results affects selection


Impact on Search Engines
Web Search Advertisement

- Bidding process occurs beforehand
  - Advertisers bid to be displayed by keywords of their choice in auction-like process
- User types a keyword into search engine
- Advertisements displayed next to search results
  - The higher the advertisers bid, the higher the placement on the page

Impact of Past Queries

- History can make revenues higher ;-) 
- ehr... to improve precision!
- How?
  - Better ad matching: selecting most promising advertisers
  - Better query classification
  - Decide when to show advertisements
Better Query Classification

• Hard job
• queries are very short
• What does huawey e220 means?
• For more information see:

When it is Better to Show Ads?

• Motivation
  • Repeatedly showing non-relevant ads can have
detrimental long-term effects
  • Want to be able to automatically decide when to advertise or not to, by analyzing individual ads or a set of ads

The Idea...

• Extract a set of features from a query that will make the engine decide whether showing an ads or not

• Use machine learning approaches to build a scoring function for query,ad pairs suggesting whether an ad should be showed or not

Extracting Human Activities

• Remember the “I Love Alaska” example of the introduction?

• People look for “stuff” on the Web to accomplish a given task

• Often a task is split into subtask
  • each of which is accomplished through a search process
An Example from the AOL Log

• The Birthday Party Planning
• Usually involves sequences of the following

Supplies  Games  Cake
Location  Gifts  Theme Party

Supplies

• personalized birthday party supplies
• paper birthday plates
• birthday party plates

• Submitted by the same user
Cake

- 2263543 casino cake for grandma's birthday
- 2263543 casino cake for grandma's birthday brooklyn new york
- 2722 custom birthday cakes atlanta ga
- 426204 birthday cakes
- 426204 cake pans for birthday
- 426204 over the hill birthday cake
- 426204 horse birthday cake
- 426204 publix cakes

Theme Party

- 71845 12th birthday-party
- 2263543 birthday party ideas for senior citizen
- 3510760 pony rides birthday party
- 3510760 pony rides birthday party long island
- 3510760 guitar playing and sing alongs for child's birthday long island
- 2722 preteen birthday party
- 2722 shimmering butterfly party
Aim...

- Finding arrows...

The End

- Thank you for your attention!
- Hope you will get interested in the topics discussed.
- For more information: