(Big) Usage Data in Web Search
A WWW’2013 Tutorial

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

Yoelle Maarek
Yahoo! Labs, Haifa, Israel

Schedule/Outline

• Web Search, a retrospective, pp 3-9 (7 slides - 15m)
• Usage Data pp 10-28, (19 slides, 35m)
  – Query Logs
  – Click Data
  – Other Signals
• Benefits of Usage Data, pp 29-53 (25 slides – 40m)
• 30m Break
• Wisdom of Crowds & the Long Tail, pp 54-68 (14 slides – 30m)
• Risks and Limitations pp 69-78 (10 slides – 20m)
  – Size
  – Personalization
  – Privacy
• Wisdom of "ad-hoc" crowds, pp 79-100 (21 slides – 30m)
• Conclusions, pp 101-103 (3 slides – 10m)
Compare queries and documents as vectors of words

Information Retrieval “invented” by G. Salton at Cornell

- Single (lemmatized?) word as indexing unit, count words
- Vector Space model ($d_i$, document, $q$, query)
- Evaluation on predefined test collections (e.g., 300 docs CACM)

1970-90 Information Retrieval
Looking Back

1970-today
Information Retrieval

Compare queries and documents as vectors of words
Information Retrieval “invented” by Salton at Cornell: Text analysis, count words
Vector Space model \((d_1, \text{document}, q\ \text{query})\)
Evaluation on predefined test collections (e.g., 300 docs CACM)

1994-98
Analyze words in millions of Web pages

Excite, Lycos, Inktomi, Alta-Vista
- HTTP (CERN) and the Web,
- First browser: Mosaic (U of Illinois)
- First crawler by Fuzzy Mauldin (Lycos’ founder)
- Yahoo! founded as a directory service
- Evaluation by TREC (NIST)

Then First Revolution

1998: Analyze links on a very large scale

Google PageRank
- [Brin & Page WWW’1998], “The anatomy of a large-scale hypertextual web search engine”

IBM Hubs and Authorities
- [Kleinberg SIAM’1998], “Authoritative sources in a hyperlinked environment”

- Recommendation of \(T\) by \(S\)
- Evidence that \(S\) and \(T\) are related
- Relation between \(T\) and \(T'\)
First Revolution (cont.)

2000: Analyze the structure

Structure & connectivity
- [Broder et al. WWW 2000] “Graph structure of the Web”

Anchor text is a key signal
- [Eiron & McCurley SIGIR’2003] “Analysis of anchor text for web search”

Then the Second Revolution

2003 Usage Data

AOL query log transcript

Implicit feedback
- Query logs and click through data

Live experiments
- Try out experimental features on x% of users

Today’s Goals

- Understand the benefits of Usage Data
- Understanding its risks and limitations
- Give some directions on how to circumvent them

Disclaimer!!!
Most (not all) of this lecture is high-level as most algorithms and implementation details are still kept as trade secrets by most search engines

Usage Data
Collective Usage Data

- Usage data is the new entry barrier for any search engine. Two key types
  - Query logs
  - Click data

- The bigger it gets, the better mining technologies get

- Collective usage data is key to
  1. Interpreting users’ information needs
  2. Improving ranking and all search artifacts

Queries differ from Documents

Word distribution in queries in documents sorted such that same word on x

2 words way more frequent in query logs than in documents

Queries and Text

Each set with its own power law!

Not in Text
Not in Queries

Normalized Distribution

Counts

Queries

Collection

1e-05
1e-06
1
10
100
1000
10000

Words

Queries and Text

Simplest method: count! and compare

Eg. spelling – from http://www.google.com/jobs/britney.html
Query-flow graph

- Correct
- Specialize
- Generalize
- Parallel Move


Query “research” sessions

- Cut query sessions not by time but by task
- Produces “research sessions”
  - Set of all user activities (queries and clicks) occurring during when trying to satisfy a set of related and complex information needs, eg. travel, education, Medical, etc.
    - Signals: Topical coherence and user’s engagement, impose constraints on R, maximal order sequence
    $$R = \left\{ q_{i_1} , u_{i_1} , t_{i_1} , C_{i_1} \right\} , \ldots , \left\{ q_{i_k} , u_{i_k} , t_{i_k} , C_{i_k} \right\}$$
  - Account for
    - 10% of the search sessions
    - 25% of query volume

[Donato, Bonchi, Chi & Maarek WWW’2010] “Do you want to take notes? Identifying research missions in Yahoo! Search
Clustering queries

• Define relations among queries
  – Common words: sparse set
  – Common clicked URLs: better
  – Natural clusters

• Define distance function among queries

• Using search results to measure the similarity between queries
  – [Sahami and Heilman WWW'2006] “A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets” who propose a method for: “measuring the similarity between short text snippets (even those without any overlapping terms) by leveraging web search results to provide greater context for the short texts”

Clicks, Visits, Pageviews, etc.

• Clicks:
  – The number of times an artifact was clicked by a visitor

• CTR:
  – Click through rate

• Visits:
  – The number of unique sessions initiated by a visitor (Google analytics considers a session to be terminated after 30m of inactivity)

• Pageview:
  – A view on a page by a visitor. New pageviews are counted for each page reload, or return to a page. A unique pageview represents the number of sessions during which a page was viewed one or more times.
Dashboards

Example of dashboard from Google Analytics

More Signals: Non-intrusive Eye-Tracking devices

- Camera by the monitor, facing the user

- Data Representation
  - Static “saccade path”
  - Animated “saccade path”
  - Blind zones maps
  - Heat maps

University of Maryland TOBII system
http://www.rhsmith.umd.edu/behaviorlab/
Heatmap for aggregating eye-tracking data

• "Heat maps" allow to study user’s behavior, e.g.,
  – best add placement,
  – the impact of changing the length of snippets in search result pages, see "What are you looking for? An eye-tracking study of information usage in Web search" [Cutrell and Guan CHI 2007]
  – Etc.

Additional Signals: Cursor movements

Tracked by using an instrumented browser/page/toolbar
• [Guo & Agichtein, SIGIR’2008] “Exploring mouse movements for inferring query intent” (piece of Javascript code in Firefox toolbar)

• Cursor movements at scale
  – [Huang et al., CHI’2012], “User see, user point: gaze and cursor alignment in web search”
  – [Huang et al. SIGIR’2012] “Improving Searcher Models Using Mouse Cursor Activity”
Combination of Signals

- Signals can also be combined, [Feild et al. SIGIR’2010] “Predicting searcher frustration”: study of searchers’ frustration, correlating query log data with sensor information
  - A mental state camera, for six mental states: agreeing, disagreeing, unsure, interested, thinking, and confident.
  - A pressure sensitive mouse
  - A pressure sensitive chair
- Many signals cannot be tracked at a large scale as they require client-side machinery. Reserved to small users' studies (out of scope for big usage data)
- If additional instrumentation is required, value needs to be demonstrated to make it “worth it”

Predicting/Analyzing ClickThrough

- Key question: whether a user examined a specific position [Srikant et al. KDD’2010] “User Browsing Models: Relevance versus Examination”
- Three models
  - Assume examination is independent of the other results for query
    - The examination hypothesis, [Richardson et al WWW’2007] “Predicting clicks: Estimating the click-through rate for new ads”: To be clicked, a result must be both examined and relevant
  - Assume examination depends on the pattern of clicks on prior results
  - Assume examination depends on both the pattern of clicks on prior results, and the relevance of prior results.
Examination depends on other results

- The **cascade hypothesis and model**, [Craswell et al. WSDM'2008] "An experimental comparison of click position-bias models"
  - Hypothesis assumes that users scan each result sequentially without any skips
  - Model further constrains that the user continues examining results until she clicks on a result, and does not examine any additional results after the click

- The **dependent click model**, [Guo et al, WSDM'2009] "Efficient multiple-click models in web search"
  - Generalizes the cascade model to instances with multiple clicks

- The **user browsing model**, [Dupret et al, SIGIR 2008] "A user browsing model to predict search engine click data from past observations."
  - In contrast, allows users to stop browsing the current results and instead reformulate the query

Examination depends on prior clicks & prior relevance

- The **click-chain model** [Guo et al, WWW’2009] "Click chain model in web Search"
  - If a user clicks on the previous result, the probability that they go on to examine more results ranges between 2 values depending on the relevance of the previous result

- The **general click model** [Zhu et al, WSDM’2010] “A novel click model and its applications to online advertising”
  - Treats all relevance and examination effects in the model as random variables

- All the above models relate to “perceived” relevance – whether the user considers the result relevant before he clicks on the result

- There also exist “post-clicks” models
Post-clicks models

- The **dynamic Bayesian model** [Chapelle et al., WWW'2009] “A dynamic Bayesian network click model for web search ranking”
  - Uses the “user satisfaction” (post-click relevance) of the preceding click to predict whether the user will continue examining additional results

- The **session utility model** [Dupret et al., WSDM'2010] “A model to estimate intrinsic document relevance from the click-through logs of a web search engine.”
  - Proposes a user browsing model based on the “intrinsic” (post-click) relevance of the sequence of clicked results in a user session

In summary

Srikant et al. examined all these models and proved that:

- “relevance of the result for [a given] query instance”
- is strongly correlated with clicks on other results and
- is responsible for a substantial portion of the changes in conditioned CTR”
Benefits of Usage Data

Did you Mean?

• Depart from the usual dictionary-based model
  – Classic approach was to use edit distances to identify typing mistakes such as letter inversions for instance, see [Kukich, ACM Computing Surveys 1992] “Techniques for automatically correcting words in text”
• Extensive use of query logs analysis
  – Frequency
  – Transition
  – Clicks
Dynamic Query Suggestion

- Query suggestions appear as you type in rectangle
- Some history
  - Google labs in 2004
  - Google toolbar in 2006
  - yahoo.com, search.yahoo.com in 2007
  - youtube.com and google.com in 2008

Don’t Stop at Query Completion
Dynamic Query Suggestions

• Key difference between dynamic query suggestions and query assistance?
  – Query suggestions take as input a prefix as opposed to a full query

• Challenges
  – Limited input – prefix only
  – Latency
  – Freshness
  – Locality
  – Diversity (danger of rich gets richer syndrome)
  – Serendipity
  – etc.

The “Voice of the Search Engine”???
Once a query is issued, the users’ needs (informational, navigational as well transactional) can be either:

- **satisfied**
  - Users get their answer immediately from a onebox result, such as calculator, weather, sports results, etc.)
  - Almost immediately after they click on one or a few of the top results

- **partially satisfied**
  - Users have undertaken a “research task”, no single Web page holds all the needed information.
  - Needs susceptible to trigger research tasks: travel needs, homework, education needs, or health information

- **not satisfied at all**
  - Users did not formulate their query well or
  - Relevant content simply does not exist

**Help Users Reformulate their Queries**

**Related Queries**

1. Content-aware approaches: Use SERP or target pages to measure query similarity
2. Content-ignorant approaches: Use Clicks
3. Query-flow approaches: Monitor the users’ sequential search behavior to better understand query intent
Sitelinks/Quick Links

- "Navigational shortcuts [...] are displayed below the website homepage on a search results page and let users directly jump to selected points inside the website"
- Not trivial as the goal is to maximize the benefits for a majority of users, while showing only relevant links in a limited real estate

[Chakrabati, Kumar and Pundera, WWW'2009] "Quicklink selection for navigational query results"

Identifying “research” sessions on the fly

- A research session R is a maximal order sequence

\[ R = \left\langle \langle q_{i_1}, u_{i_1}, t_{i_1}, C_{i_1} \rangle, \ldots, \langle q_{i_k}, u_{i_k}, t_{i_k}, C_{i_k} \rangle \right\rangle \]

Where, for given thresholds \( s_\theta, t_\theta \) and \( C_\theta \), all \((u_i)\)'s refer to the same user and

\[ t_{i_1} \leq \ldots \leq t_{i_k} \leq \tau \]
\[ \forall i, j \in \{i_1, \ldots, i_k\} : s(f(q_i), f(q_j)) \geq s_\theta \]
\[ |R| = k \geq k_\theta \]
\[ \sum_{j=1}^{k} |C_{i_j}| \geq C_\theta \]
Exploiting query research sessions: Search Pad*

Appears when research session is identified

Discontinued!

How to Measure when there is No Click?

- Oneboxes, Direct Displays, Direct Answer on the results page
  - weather, stocks, movie or train schedules, sports results,
- Package tracking (Fedex/UPS), etc.
Who would click on that result?

Analyze Change in User’s Behavior

- Find users who are “tenacious”
  - Reformulate/click, do not let go
  - Measure their abandonment

- Session representation
  - Queries and clicks
  - XQCQX for “start, query, click, query, stop”
  - Tenacity = \( \frac{(XQQ+XQC)}{(XQQ+XQC+XQX)} \)

- Generalize: Analyze behavior of “pitbulls”, “poodles”, “serial shoppers”, kids etc.?

[C. Castillo, A. Gionis, R. Lempel and Y. Maarek, Industrial Track, SIGIR’2010] “When no clicks are good news”
Explicit Usage Data
Relevance feedback Revisited

Google Searchwiki
• User's feedback
  – Comment
  – Promote
  – Remove

Discontinued!

To “Stars in Search”

http://googleblog.blogspot.com/2010/03/stars-make-search-more-personal.html

Discontinued!
To +1 in Social Search

People in my circles

Bing Social Search –Facebook integration

“Social annotations: utility and prediction modeling” [Pantel, Gamon, Alonso & Haas, SIGIR 2012]

http://www.bing.com/community/site_blogs/search/archive/2012/05/15/start-doing-more-now-try-the-new-bing-today.aspx
Bing Social Search – Friends and experts (twitter)

Bing Social Search – foursquare integration
Query Trends Exposed to Users

- **Time series of search trends, based on “query shares” of query term $q$ at time $t_i$ in geo location $geo_j$ where**

  $$\text{query share} = \frac{\text{number of queries for } q}{\text{number of queries} (t_i, geo)}$$

- **Google trends**
  - “Google Trends provides an index of the volume of Google queries by geographic location and category”
  - [Varian and Choi 2009, in Google Search Blog]
    “Predicting the Present with Google Trends”

- **Google Insight for Search**
  - Uses the same data but geared to researchers and advertisers.
  - [Shimshoni et al. 2009 on Google Search Blog]
    “On the Predictability of Search Trends”
Yahoo! Clues clues.yahoo.com

Consider other facets
- Searches over time
- By demographic
- By Income
- By Location
- Search Flow
- Related Search

By Demographic and By Income

Combine Yahoo! search query log with:
- profile information provided by Yahoo! (birth year, gender, ZIP code)
- US-census information aggregated by ZIP code
Annotate each query with the average per-capita income in the ZIP code area


[Serdyukov, Murdock & van Zwol SIGIR’2009] “Placing Flickr Photos on a Map”
Ultimate Evaluation
Controlled Experiments on Real Traffic

- AKA Bucket or A/B testing
- Launch feature exclusively to small % of users for n weeks
- Measure multiple metrics
- Launch in increments to verify that performance is maintained
- Keep “hold back” experiments

[Andy Beal, Market Pilgrim 2005]
“Google Space project at London’s Heathrow airport and lab mice in a maze?”

[Kohavi et al, Data Mining Knowledge Discovery 2009]
“Controlled experiments on the web: survey and practical guide”

[Tang et al. KDD 2010]
“Overlapping experiment infrastructure: More, better, faster experimentation”

[Deng et al, WSDM 2013] “Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data”

Wisdom of Crowds and the Long Tail
The Wisdom of Crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
  - “Under the right circumstances, groups are remarkably intelligent”
- Importance of diversity, independence and decentralization

Aggregating data

“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.”

---

Flickr: Clustering Pictures
Flickr: Geo-tagged pictures

Wisdom in Web Search

<table>
<thead>
<tr>
<th>Generation</th>
<th>Technology</th>
<th>Wisdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>First: 1994-98</td>
<td>Classical IR</td>
<td>Writers</td>
</tr>
</tbody>
</table>
| Second: 1997-2003 | +Link Analysis  
                      | +Anchor text  
                      | +Click-through voting | +Webmasters 
                      | +Readers        |
| Third: 2003-2010 | +Usage data mining               | Everyone     |
| Fourth: 2008-??   | +Query intent detection          | Everyone     
                      | +Learning to rank     |
The Wisdom of Crowds

• Popularity
• Diversity
• Quality
• Coverage
The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

Long tail
Heavy tail
Most measures in the Web follow a power law

Example: Click Distribution

User interaction is a power law!
Heavy tail of user interests

- Many queries, each asked very few times, make up a large fraction of all queries
  - Movies watched, blogs read, words used, …

One explanation

<table>
<thead>
<tr>
<th>People</th>
<th>Interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal people</td>
<td>Weirdos</td>
</tr>
</tbody>
</table>
Heavy tail of user interests

Many queries, each asked very few times, make up a large fraction of all queries

Applies to word usage, web page access, …

We are all partially eclectic

Who is in the Long Tail?

• Popular hypothesis:
  – Majority of consumers consistently follow the crowd, only eccentrics issue tail queries

• Wrong!
  – Extensive study on user preferences for movies, music, Web browsing and Web search
  – Everyone is a bit eccentric, consuming both popular and specialty products, e.g., most people have their own 80-20.
  – Supporting the tail (products or queries) goes beyond direct revenues to second-order gains associated with
    • increased consumer satisfaction
    • repeat patronage.
Risks and Limitations

Conflicting Factors

• Size
  – Web Search requires usage data over larger and larger populations over longer and longer periods of time
• Personalization
  – Size is not there on tail (sparse data), yet tail is critical
  – Personalization exposes private facets
• Privacy
  – Long-term logs endanger privacy
Risks of size: When the crowd dominates

• Risks of “wisdom of crowds”
• What about long tail?
  – See (obsolete now) “shwarzneger” example

Risks of size: endanger the Long Tail

• Neglecting the long tail impacts and upsets everyone!
• Supporting the tail may boost the head by providing users a one-stop shop for both their mainstream and niche interests.
Personalization “Facets”

Successful “facets”:
• Language (not always)
• Geo

http://googleblog.blogspot.com/2009/03/local-flavor-for-google-suggest.html

– Challenge – decide when to trigger frequency or freshness vs geo

• Semantic facets based on previous queries

Social Personalization: personal results (discontinued?)

barcelona
Social Personalization

Risks of Personalization

• The Filter “Bubble”, Eli Pariser
• Avoid the Poor get Poorer Syndrome

• Solutions:
• Diversity
• Serendipity
• Explore & Exploit
Risks of Privacy:
AOL Query Logs Release Incident

A Face Is Exposed for AOL Searcher No. 4417749,
By MICHAEL BARBARO and TOM ZELLER Jr,

• No. 4417749 conducted hundreds of searches over a
  three-month period on topics ranging from “numb fingers” to “60 single men” to “dog
  that urinates on everything.”
• Other queries: “landscapers in Lilburn, Ga,” several people with the last name Arnold
  and “homes sold in shadow lake subdivision gwinnett county georgia.”
• Data trail led to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga.,
  frequently researches her friends’ medical ailments and loves her three dogs.

Risks of Privacy

• (ZIP code, date of birth, gender) is enough to identify 87% of US citizens using public DB
  (Sweeney, 2001)
• K-anonymity
  – Suppress or generalize attributes until each entry is identical to at least k-1 other entries
• Federal Trade Commission in US: Privacy policies should “address the collection of data itself and not
  just how the data is used”, Dec 2010.
• Data Protection Directive in EU
Risks of Privacy: Query Logs

• Profile: [Jones, Kumar, Pang, Tompkins, CIKM 2007]
  – Gender: 84%
  – Age (±10): 79%
  – Location (ZIP3): 35%

• Vanity Queries: [Jones et al, CIKM 2008]
  – Partial name: 8.9%
  – Complete: 1.2%

• Survey:
  – [Cooper ACM TWEB 2008]: A Survey of query log privacy-enhancing techniques from a policy perspective. A good anonymization is still an open problem

Wisdom of Ad-hoc Crowds
Why the Heavy Tail Matters?

- Not only for e-commerce, but because we are all there
- Personalization vs. Contextualization

User interaction is another long tail

**Interests**

**People**

Contextualization

- Context:
  - Local: geography, language, …
  - Person: do we know the user? enough data?
- Social
- Task
- Personalization: Data volume vs. privacy
- Contextualization: Small ad-hoc crowds

- What is the right interface?
How to Circumvent these Limitations?

- Wisdom of “ad-hoc” crowds?
  - Aggregate data in the “right way”
  - When data is sparse
    - Aggregate users around same intent, same task, same facet
    - Change granularity “ad hoc”

- Middle age men
- Fans of Messi

Example: Mining Geo/time Data

- Optimal Touristic Paths from Flickr

- Good for tourists and locals

De Choudhury et al, HT 2010
Another example: Regions from Pictures

Search Example: Query Intent Prediction

- Users have complex interests that go beyond the traditional informational queries
- Relevant for Web-search engines:
  - Intent-aware ranking
  - Intent-aware result page
  - Identify and provide accurate results for new types of queries
- Complements other strategies to improve the quality of the search results:
  - Diversification provides a list results that cover different interpretations of ambiguous queries
  - Hence, an important step to provide the right results is to identify the query type
- Query type is the simplest intent
A Web Query: the Tip of the Iceberg

Dimensions of the User’s Intent

Meta-Dimension
- **News.** University hunger’s strike
- **Business.** Sale of furniture
- **Reference.** Contraceptive method
- **Community.** Your favorite singer
A list of 18 topics built from the first level of categories offered by ODP, Yahoo!, and Wikipedia.

- **News** - Reference

The primary need of the user:
- **Informational**. What is thermodynamics?
- **Not-Informational**. Notebook sales
- **Both**. Airbrush
Dimensions of the User’s Intent

Is the query aimed to:
- Obtain an informational resource? *street map of Barcelona*
- Perform an action? *Renting an apartment*

Dimensions of the User’s Intent

How specialized is a query:
- **Specific.** *Autonomous Univ. of Bucaramanga.*
- **Medium.** *Private universities*
- **Broad.** *Universities*
Dimensions of the User’s Intent

Does the query contains **polysemic** words?

*wood, dove*

Is the query designed to retrieve **authoritative** and trusted answers?

*Hospitals in Goteborg*
Dimensions of the User’s Intent

Does the query includes a geographic location? Hotels in Montevideo
Also, trying to find a location that is not explicitly mentioned? Nearest hotel

Dimensions of the User’s Intent

Could the results be influenced by time?
U.S. President vs. U.S. President 2010
Dimensions of the User’s Intent

<table>
<thead>
<tr>
<th>Facet</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genre</td>
<td>provides a generic context to the user's query intent, and can be thought as a meta-facet.</td>
</tr>
<tr>
<td>Topic</td>
<td>a list of topics built from the first level of categories offered by ODP, Yahoo!, and Wikipedia.</td>
</tr>
<tr>
<td>Task</td>
<td>this facet is related with the type of resource associated with the query.</td>
</tr>
<tr>
<td>Objective</td>
<td>represents if the query is aimed to do some action or to obtain a resource.</td>
</tr>
<tr>
<td>Specificity</td>
<td>this facet describes how specialized is a query.</td>
</tr>
<tr>
<td>Scope</td>
<td>the scope aims at capturing whether the query contains polysemic words or not.</td>
</tr>
<tr>
<td>Authority Sensitivity</td>
<td>through this facet it is possible to determining whether the query is designed to retrieve authoritative and trusted answers</td>
</tr>
<tr>
<td>Spatial Sensitivity</td>
<td>this facet reflects the interest of the user to get a resource related to an explicit spatial location.</td>
</tr>
<tr>
<td>Time Sensitivity</td>
<td>this facet captures the fact that some queries require different results when posed at different times.</td>
</tr>
</tbody>
</table>

The Problem

Web Query  Classifier  User Intent

News Corp Phone Hacking Scandal

Training: queries & multi-dimensional user intent!!

<table>
<thead>
<tr>
<th>Genre</th>
<th>News</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>Finance</td>
</tr>
<tr>
<td>Task</td>
<td>Informational</td>
</tr>
<tr>
<td>Objective</td>
<td>Resource</td>
</tr>
<tr>
<td>Scope</td>
<td>Unique</td>
</tr>
<tr>
<td>Specificity</td>
<td>Specific</td>
</tr>
<tr>
<td>Authority Sensitivity</td>
<td>Yes</td>
</tr>
<tr>
<td>Spatial Sensitivity</td>
<td>No</td>
</tr>
<tr>
<td>Time Sensitivity</td>
<td>Yes</td>
</tr>
</tbody>
</table>
One Facet Prediction

- Most of the work in query intent identification is based on the analysis of only one possible facet of the query.
- The most common of these facets are the topic category and the type of resource associated with the query: informational, navigational, transactional.
- Other approaches to identify query intent consider facets like:
  - Geographic locality
  - Time sensitivity
  - Ambiguity
  - Specificity
- Usually automatic classification of query intent is based on the information of individual facets.

Multiple Facet Prediction

- Better is to classify query intent in more than one facet:
  - [Baeza-Yates et al., SPIRE'06]: Automatic classification of queries into topic and type of resource (informational, not informational and ambiguous).
  - [Nguyen and Kan, WWW'07]: Query log analysis of four facets (ambiguity, authority sensitivity, temporal sensitivity, spatial sensitivity). They reported automatic classification results only for the authority sensitivity facet.
  - [Calderon et al. SIGIR 2011]: Multi-facet prediction using tree structured distributions, improving it with Wordnet. Includes all facets.
  - [Gonzalez & Baeza-Yates, SPIRE 2011]: All the facets using multi-level SVM classification and analyzing their interdependency.

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic/homepage [17]\textsuperscript{1}</td>
<td>200</td>
<td>91%</td>
</tr>
<tr>
<td><em>FastQ (Task)</em></td>
<td>5249</td>
<td>76%</td>
</tr>
<tr>
<td>Rule-based hierarchy [15]</td>
<td>1500000</td>
<td>74%</td>
</tr>
<tr>
<td>Query-log based [19]</td>
<td>50</td>
<td>54%</td>
</tr>
</tbody>
</table>
Challenges

• Facets focus in
  • Completeness
  • Coverage
  • Simplicity

• Sparse training data
• Semantic relationships between query words
• Interdependencies between facets of user intent
• Facet dependent ranking (ad-hoc crowd)

Conclusions
Plenty of Open Problems

- Algorithm scalability
- Dependency on training data
- Privacy Issues

- More on query intent detection
  - Quality
  - Online

[Baeza-Yates & Maarek, SSDBM 2012]

QUESTIONS?