Web Retrieval: The Role of Users
Ricardo Baeza-Yates, Yahoo! Research, Barcelona, Spain
Yoelle Maarek, Yahoo! Research, Haifa, Israel

ABSTRACT
Web retrieval methods have evolved through three major steps in the last decade or so. They started from standard document-centric IR in the early days of the Web, then made a major step forward by leveraging the structure of the Web, using link analysis techniques. A more recent, no less important step forward, has been to enter the user in this equation in two ways: (1) implicitly, through the analysis of usage data captured by query logs, and session and click information in general, the goal being to improve ranking as well as to measure user's happiness and engagement; (2) explicitly, by offering novel interactive features; the goal here being to better answer users' needs. We cover here the user-related challenges associated with the implicit and explicit uses activities, namely:

Usage data analysis and metrics: It is critical to monitor how users interact with Web retrieval systems, as this implicit relevant feedback aggregated at a large scale can approximate quite accurately the level of success of a given feature. This is the focus of the first part of the tutorial.

User interaction: Web retrieval engines interact with the user at two key stages, each associated with its own challenges: expressing a query and interpreting and using results. After quite some stagnation on the front-end of Web retrieval, we have seen numerous novel interactive features appear in the last 3 to 4 years, as the leading commercial search engines seem to compete for users’ attention. The second part of the tutorial will be dedicated to explicit user interaction. We will introduce novel material (as compared to previous versions of this tutorial.)

An earlier version of this tutorial was given at SIGIR’2010, WSDM’2011 and ECIR’2011
Table of contents

• Session 1: Introduction ..........................................................2-7
  • Web search - retrospective
  • Major game changers

• Session 2: Implicit user aspects ...........................................8-59
  – Mining query logs
  – Monitoring user’s behavior (click models)

• Session 3: Explicit user aspects ...........................................60-96
  – Before the query is issued
  – Search results, traditional, interactive and actionable
  – Query assistance
  – Query introspection

• Session 4: Conclusions..........................................................101-104
  – Current and next challenges

• Bibliography
1970-90
Information Retrieval

Compare queries and documents as vectors of words

- Vector Space model (d, query)
  - Evaluation on predefined test collections (e.g., 300 docs CACM)

1990-96
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)
• **Classic IR**
  - **Corpus:** Fixed document collection
  - **Goal:** Retrieve documents with information content that is relevant to user’s information need

• **Classic relevance**
  - For each query Q and stored document D in a given corpus assume there exists a relevance score $\text{Score}(Q, D)$, where
    
    $\text{Score}(Q, D)$ is averaged over users U and contexts C
  - Optimize $\text{Score}(Q, D)$ as opposed to $\text{Score}(Q, D, U, C)$
  - In other words, usually:
    1. Corpus is *predetermined*
    2. Context is *ignored*
    3. Users are *ignored*

**Bad assumptions in the web context**
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’

Structure and 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

2000 Usage Data

Implicit feedback
Query logs and Click through data

Live experiments
Try out experimental features on % of users

2000 Usage Data

AOL query log transcript

<table>
<thead>
<tr>
<th>Annual</th>
<th>Query</th>
<th>QueryTime</th>
<th>Frequency</th>
<th>ClickRate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>florid bureau of investigations</td>
<td>2000-06-08-09:31:00</td>
<td>1</td>
<td><a href="http://www.fbi.com">http://www.fbi.com</a></td>
</tr>
<tr>
<td>2000</td>
<td>florid bureau of investigations</td>
<td>2000-06-08-09:31:00</td>
<td>1</td>
<td><a href="http://www.fbi.com">http://www.fbi.com</a></td>
</tr>
<tr>
<td>2000</td>
<td>tennessee constitutional city council</td>
<td>2000-06-06-17:17:46</td>
<td>1</td>
<td><a href="http://www.state.tn.us">http://www.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
<tr>
<td>2000</td>
<td>municipal fraud detection internal controls</td>
<td>2000-06-07-21:17:40</td>
<td>5</td>
<td><a href="http://www.tnattorneygeneral.state.tn.us">http://www.tnattorneygeneral.state.tn.us</a></td>
</tr>
</tbody>
</table>

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

[Xiaodong Shi, Technical report, University of Michigan, 2007]
“Social Network Analysis of Web Search Engine Query Logs”
• **2000-to-date: Usage data & live experiments**
  
  - A wealth of *implicit* information
    
    - Query logs and CTR represent “implicit relevant feedback”
      
    
    - Increased traffic generates more and more usage data
    - Wisdom of crowds works (see “did you mean”)
    - Dashboards and 1% AKA bucket experiments
    - Users’ actions give competitive advantage to the engine ranking

  - *Explicit* information from more and more sophisticated users
    
    - Front-end apps and features are the key differentiator
    - Launch and iterate!
Goals of today’s tutorial

• Define and organize concepts you are all familiar with, as users.
• Identify the technical/research challenges
• Some hints on how it works
• Some understanding on the explicit/implicit role of users in Web search

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
Implicit users aspects

Session 2
Collective usage data

• Usage data is the new entry barrier for any search engine

• The more it grows, and the better mining technologies get, the more relevant search results become

• Collective usage data is key to
  1. Interpreting users’ information needs
  2. Improving ranking and all search artifacts
Query logs: mining queries for ...

Improving Web search

- Index layout and caching
- Features for ranking
- Query intention
- Query recommendations
- Matching advertising
• User search modelling

• Discoverability for search engines & people

• User driven Web design
  – The Web site that the users want
  – Improve information scent, content & structure

• Bootstrap pseudo-semantic resources
Web queries

- Cultural and educational diversity
- Short queries & impatient interaction
  - Few queries posed & few answers seen
- Smaller & different vocabulary
- Different user goals
  - Information needs
  - Navigational needs
  - Transactional needs
  See refinement by [Rose & Levinson, WWW 2004] “Understanding user goals in web search”
Who issues head/tail queries?

- Long tail phenomenon
  - Popular hypothesis: Majority of consumers consistently follow the crowd, only eccentrics issue tail queries

- Study of long tail
    - 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.
    - Supporting the tail may boost the head by providing users a one-stop shop for both their mainstream and niche interests.
Queries and text

Word distribution in queries and in documents are different

Queries and text

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

2 words way more frequent in query logs than in documents

Log-log plot – weak correlation bw query & document vocabularies

Term Pairs

[Baeza-Yates et al, SIGIR 2007]
“The Impact of Caching on Search Engines”
Queries and text

Each set with its own power law!
Other implicit signals
How far do people look for results?

68% Of Search Engine Users Click Results Within The First Page of Results

- Only a few: 27%
- The first page: 41%
- The first 2 pages: 17%
- The first 3 pages: 7%
- More than 3 pages: 8%

68% Combined

Source: iprospect.com iProspect Blended Search Results Study – April 2008
“The first three pages of search results now appear to be the last frontier”

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Only a few</td>
<td>27%</td>
<td>23%</td>
<td>24%</td>
<td>16%</td>
</tr>
<tr>
<td>The first page</td>
<td>41%</td>
<td>39%</td>
<td>36%</td>
<td>32%</td>
</tr>
<tr>
<td>The first 2 pages</td>
<td>17%</td>
<td>19%</td>
<td>20%</td>
<td>23%</td>
</tr>
<tr>
<td>The first 3 pages</td>
<td>7%</td>
<td>9%</td>
<td>8%</td>
<td>10%</td>
</tr>
<tr>
<td>More than 3 pages</td>
<td>8%</td>
<td>10%</td>
<td>12%</td>
<td>19%</td>
</tr>
</tbody>
</table>

Source: iprospect.com iProspect Blended Search Results Study – April 2008
“When you [...] don’t find what you are looking for, [when] do you [...] revise your search, or move on to another search engine?”

<table>
<thead>
<tr>
<th>After reviewing</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>the first few search results</td>
<td>23%</td>
</tr>
<tr>
<td>the first page of search results</td>
<td>26%</td>
</tr>
<tr>
<td>the first 2 pages of search results</td>
<td>28%</td>
</tr>
<tr>
<td>the first 3 pages of search results</td>
<td>15%</td>
</tr>
<tr>
<td>more than 3 pages of search results</td>
<td>9%</td>
</tr>
</tbody>
</table>

Source: iprospect.com *iProspect Blended Search Results Study – April 2008*
## Same question over time

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>After reviewing the first few search results</td>
<td>23%</td>
<td>16%</td>
<td>23%</td>
<td>14%</td>
</tr>
<tr>
<td>After reviewing the first page of search results</td>
<td>26%</td>
<td>25%</td>
<td>19%</td>
<td>14%</td>
</tr>
<tr>
<td>After reviewing the first 2 pages of search results</td>
<td>28%</td>
<td>27%</td>
<td>26%</td>
<td>28%</td>
</tr>
<tr>
<td>After reviewing the first 3 pages of search results</td>
<td>15%</td>
<td>20%</td>
<td>15%</td>
<td>22%</td>
</tr>
<tr>
<td>After reviewing more than 3 pages of search results</td>
<td>9%</td>
<td>12%</td>
<td>17%</td>
<td>22%</td>
</tr>
</tbody>
</table>

**Source:** iprospect.com *iProspect Blended Search Results Study – April 2008*
Counting queries

Simplest method: count! and compare
Revolutionized spelling correction by departing from the usual dictionary-based model.

- Classic approach was to use edit distances to identify typing mistakes such as letter inversions [Kukich, ACM Computing Surveys 1992] “Techniques for automatically correcting words in text”
- Instead, “Did you mean” learns its spelling corrections simply from usage and a great deal of usage.

Extensive use of query logs analysis

- Frequency
- Transition
- Clicks
- 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 Webbased 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”
• Can we cluster queries well?
• Can we assign user goals to clusters?
Evaluating query similarity

• 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
Using the clusters

- Improved ranking

- Word classification
  - Synonyms & related terms are in the same cluster
  - Homonyms (polysemy) are in different clusters

- Query recommendation (ranking queries!)
  - Real queries, not query expansion

\[
\text{Rank}(q) = \gamma \times \text{Sup}(q, q_{ini}) + (1 - \gamma) \times \text{Clos}(q, q_{ini})
\]
Relating queries

## Qualitative analysis

<table>
<thead>
<tr>
<th></th>
<th>Strength</th>
<th>Sparsity</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td>Medium</td>
<td>High</td>
<td>Polysemy</td>
</tr>
<tr>
<td><strong>Session</strong></td>
<td>Medium</td>
<td>High</td>
<td>Physical sessions</td>
</tr>
<tr>
<td><strong>Click</strong></td>
<td>High</td>
<td>Medium</td>
<td>Click spam</td>
</tr>
<tr>
<td><strong>Link</strong></td>
<td>Weak</td>
<td>Medium</td>
<td>Link spam</td>
</tr>
<tr>
<td><strong>Term</strong></td>
<td>Medium</td>
<td>Low</td>
<td>Term spam</td>
</tr>
</tbody>
</table>
Query log: edge=lexical relation
Query log: edge = subsequent queries (Reversed chronological order)
Query log: edge=common result click
Beyond counting: query flow analysis

- Correct
- Specialize
- Generalize
- Parallel Move

Mining query logs + clicks for "research sessions"

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

- **Research missions**
  - Defined as a “set of related and complex information needs”
    - Travel needs
    - Education needs
    - Medical needs

- **Research missions are conducted during research sessions**
  - Defined as a set of all “user activities (queries and clicks) occurring during a research mission
    - May contained queries from distinct sessions
    - Key signals: topical coherence and user’s engagement
  - Account for
    - 10% of the search sessions
    - 25% of query volume
A research session \( R \) is a maximal order sequence

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

Where, for given thresholds \( s_\theta, t_\theta \) and \( c_\theta \), we have

\[
u_{i_1} = \Lambda = u_{i_k} = u \in U \quad \text{and} \quad t_{i_1} \leq \Lambda \leq t_{i_k} \leq \tau
\]

\[
\forall l, j \in \{i_1, ..., i_k\}: s(f(q_l), f(q_j)) \geq s_\vartheta
\]

\[
|R| = k \geq k_\vartheta
\]

\[
\sum_{j=1}^{k} |C_{i_j}| \geq c_\vartheta
\]
• **Features extractor:** computes ~30 features from queries stream
  - Textual features:
    • size of the intersection on character-level 3-grams
    • cosine similarity computed on sets of stemmed words
  - Session features: number of clicks and queries since the beginning of the session
  - Time related features: interval between two queries

• **Research detector:** boosted decision tree that decides whether two queries q1 and q2 belong to the same research task
  • q1 = house buying in Palo Alto q2 = Real estate Bay Area q3 = pizza hut
  • research(q1, q2) = 1 research(q2, q3) = 0

• **Mission boundaries detector:** boosted decision tree that decides if two queries q1 and q2 are topically coherent and thus susceptible to be part of the same mission.
An application: Search Pad*
More details in Session 3

Appears when research session is identified

*removed on April 20–web search interface is a moving ground!
Beyond search: implicit folksonomy?
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)

Experimental evaluation

[Baeza-Yates & Tiberi KDD’ 2007]
“Extracting Semantic Relations from Query Logs”
[Francisco et al, SPIRE’2010] “Mining Large Query Induced Graphs towards a Hierarchical Query Folksonomy”
Hierarchical clustering

[Francisco et al., SPIRE’2010] “Mining Large Query Induced Graphs towards a Hierarchical Query Folksonomy”
Some open issues

- Data Volume versus Better Algorithms

- Explicit versus implicit social networks
  - Any fundamental similarities?

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

- User aggregation versus personalization
  - Optimize common tasks
  - Move away from privacy issues
Ultimate evaluation – on real traffic

- Examine users
- Conduct experiments
- Launch (by increments)
• Key question: whether a user examined a specific position

• 3 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

  - 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

  - 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 and 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
• 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 clickthrough 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

• Srikant et al. examined all these models and proved that: “relevance of the result for that query instance is strongly correlated with clicks on other results and is responsible for a substantial portion of the changes in conditioned CTR”
Observe users in users-studies

- Eyetracking, Mouse tracking
- [Cutrell et al., CHI’2007] “What are you looking for?: an eye-tracking study of information usage in web search”
- Etc.

Enroll Beta-testers

- Limited in scope
- Has a cost

Experimentation models
Experiment and launch

- Bucket AKA 1% experiment AKA A/B testing

- A key advantage of large scale – our testers are real users
  - Set aside a random (or nor so random) % of users identified by cookies, i.e. a fraction of real traffic, and deploy the chosen feature just for them for a week or two
    - Measure multiple metrics during this experiment
    - The key is in interpreting the metrics and making the smart choice, a small decrease in revenue might be acceptable if user’s happiness is increased for instance

- For major features, launch in increments to verify that performance is maintained

- Remember to keep “hold back” experiments to keep improving
Explicit user aspects

Session 3
Very little differences between major search engines
A rectangle – text box for your queries

Other forms of rectangles?
• Embedded in a portal
• Always here in a toolbar
• Ultimate rectangle: omnibox
Omniboxes

- **Ultimate version in Google chrome**
  - Merges the functionality of the address and the search boxes.
  - Browser decides whether the user’s intent is navigation or search

- **Firefox since release 2.0**
  - When entering a word entered in the address bar triggers Google’s “I feel lucky”
  - Customizable feature: By modifying the `keyword.URL` property (accessed by typing `about:config` in the address bar)
A few facts

- There is no standard Web query language – semantics are not agreed upon, engines reserve the right to change their interpretation
- Stemming?
- Conjunction or not?
- Query rewrite?

Free-text query format won!

- Queries are whatever users enter
  - Controlled vocabulary approach is dead
  - Spelling mistakes are acceptable
- Difference with natural language?
- Difference between a query and a question?
Common query operators

<table>
<thead>
<tr>
<th>Operator Syntax</th>
<th>Details</th>
<th>Google</th>
<th>Yahoo! Search</th>
<th>Bing</th>
<th>Ask</th>
</tr>
</thead>
<tbody>
<tr>
<td>“:” double quotes surrounding a string</td>
<td>Phrase search</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>+ preceded by a space, operates on the term/phrase that immediately follows</td>
<td>This operator ensures that the associated term is included “as is” in the results</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>– preceded by a space, operates on the term/phrase that immediately follows, Bing uses NOT as well</td>
<td>This operator ensures that the associated terms do not appear in any result</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>OR (as well as l) operates on preceding and succeeding terms or phrases</td>
<td>Equivalent to a Boolean OR</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>site: Followed by a site name</td>
<td>Returns results from the specific site only</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>hostname: Followed by a host name</td>
<td>Returns results from the specific host only</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>url: Followed by a URL</td>
<td>Checks that the following url exists in the engine index</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>inurl: Followed by a term</td>
<td>Returns results whose URL contains the specified term</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>intitle: Followed by a term</td>
<td>Returns results whose title contains the specific term</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>inlink://inanchor: Followed by a term</td>
<td>Returns results that contain the specific term in their link or anchor metadata</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 2.3: Common query operators

Not so common operators

- **Google**
  - * should be treated “as a placeholder for any unknown term”

- **Yahoo**
  - `link: <url>` returns documents that link to a specific url, a feature provided by the Yahoo! Site Explorer tool.
  - Other shortcuts retrieved by `!list` in the search rectangle include `!news, !flickr, !wiki, !map`, etc.

- **Bing**
  - `AND/` interestingly enough, Bing provides this operator even if it claims that “by default all searches are AND searches”.
  - `( )` used to group words in conjunction with other operators (−/+).
  - `filetype:, contains:, ip:, feed:, prefer:` etc.

- **Ask**
  - `afterdate:, beforedate:` followed by a date (yyyymmdd)
  - `betweendate:` followed by 2 dates separated by a .
  - `last:` followed by a given time period, among `{week, 2weeks, month, 6months, year, 2years}`
• 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
Y! Search assist:
Don’t stop at query completion

Yahoo! Search suggestions:
- hong kong disneyland
- hong kong hotels
- yahoo hong kong
- hong kong airport
- hong kong map

Explore related concepts:
- singapore
- taiwan
- macau
- china
- japan
- shanghai
- tokyo
- beijing
Even ads

Hilton hotels

Hotels by Hilton - Hotel Reservations, Deals, and Room Rates
www1.hilton.com/
hilton hotels corporation
hilton hotels careers
hilton hotels hawaii
hilton hotels new york
hilton hotels chicago
hilton hotels washington dc
hilton hotels jobs
hilton hotels florida
hilton hotels corp
hilton hotels london

Hilton.com Official Site

Google Search I'm Feeling Lucky
Two new features enhance search beyond the results page, 12/11/2009 10:16:00 AM
..and always more
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)
  – etc.
### Parallel with regular search

<table>
<thead>
<tr>
<th>Effectiveness</th>
<th>Query x Docs</th>
<th>Prefix x Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency</td>
<td>~1s</td>
<td>Typing pace</td>
</tr>
<tr>
<td>Automatic spelling correction</td>
<td>Well spelled/formed queries</td>
<td>Well spelled/formed queries</td>
</tr>
<tr>
<td>Filtering</td>
<td>Inappropriate documents</td>
<td>Inappropriate queries</td>
</tr>
<tr>
<td>Freshness</td>
<td>Fresh documents (news)</td>
<td>Fresh queries</td>
</tr>
<tr>
<td>Personalization</td>
<td>Personalized ranking</td>
<td>Favorite queries</td>
</tr>
<tr>
<td>Local</td>
<td>Promote local results</td>
<td>Promote local queries</td>
</tr>
<tr>
<td>Universal</td>
<td>Merge results from different sources</td>
<td>Merge queries from different sources (fresh vs local?)</td>
</tr>
</tbody>
</table>
The “Voice of the search engine”???
Not only with people names ...
SERP – basic layout
Traditional organic results

- Regular result
  - Title
  - Snippet computed on the fly, query dependent, highlights the right keywords
  - URL

- Where do we get titles from?
- How to build snippets?
  - Identify shortest window
  - Use information from parents’ page for inferring titles/snippets?

[Amitay & Paris, CIKM’2000]
"Automatically Summarising Web Sites, Is There A Way Around It?"
More about results

- Results from a same site
  - Earlier in the game: indents
  - Now navigational shortcuts aka 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.
- Will keep improving since the best links are inferred from learning users’ behavior via clicks and toolbar data

[Chakrabati, Kumar & Pundera, WWW’2009] “Quicklink selection for navigational query results”
Oneboxes/Direct Display

- **What are these?**
  - Very specific results (rather than links), answering very precise queries with a unique answer
  - Displayed above regular Web results, due to their high relevance, and in a slightly different format.

- **How does it work**
  - Hack! Triggered by specific terms in the user’s query that indicate a clear intent
  - Ultimate search result experience but only in very specific cases.

- **Examples:**
  - Weather, Movie schedules, ??
Structured display by agreement

Publishers and search engines agree on a preferred display format

- **Yahoo!**: Search Monkey in 2007
- **Google**: Rich snippets launched in 2009
  - [http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=99170](http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=99170)
- **Also on Bing**
Interactive layouts

• Bing “More on this page”
• Google Instant Preview
• Yahoo browse in place

Beijing - Visitor Guide
tavel.yahoo.com

Current Local Time
7:51 pm (CST)

Weather Forecast
Today  Mostly Cloudy  90°F | 75°F
Tomorrow  PM Thunderstorms  90°F | 74°F
Saturday  Thunderstorms  87°F | 71°F
Extended forecast on Yahoo Weather
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.

We focus here on research tasks and reformulation tools
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
Dynamic results as you type

- Google Instant
  - Leverages Google Auto-completion
  - One dimensional
- Yahoo! DirectSearch
  - Leverages SearchAssist
  - Two dimensional
• Cached
• Translate
  \{ most
• Similar
• Demoted features
  – Searchwiki>Stars in Search (Google)
  – Notepad (Google) Search Pad (Yahoo! Search)
• Recent features
  – Google +1
  – Yahoo! Apps

PS: we discuss here only actionable “results” as opposed to actionable web pages through services offered by toolbars or browsers (e.g., sidewiki out of scope)
User’s feedback
- Comment
- Promote
- Remove

Haifa - Wikipedia, the free encyclopedia
Haifa, built on the slopes of Mount Carmel, has a history dating back to .... Haifa is Israel's third-largest city, consisting of 103000 households. ...
Etymology - Early history - Modern history - Demographics
en.wikipedia.org/wiki/Haifa - Cached - Similar - ☰ ☯ ☩

Haifa Wehbe - Wikipedia, the free encyclopedia
Haifa Wehbe was born inMahrouna, a small Shiite farming town in Southern ... Haifa Wehbe has also performed with artists from outside the Arab world. ...
en.wikipedia.org/wiki/Haifa_Wehbe - Cached - Similar - ☰ ☯ ☩
http://googleblog.blogspot.com/2010/03/stars-make-search-more-personal.html
From *personal stars to public Google +1*

- Results annotated with +1 button
- “Click +1 to publicly give something your stamp of approval”
- Visible in the search results of your contacts by default

See [http://www.google.com/+1/button/](http://www.google.com/+1/button/)
• QuickApps displays applications relative to the word or phrase that you searched on in the search box

• Examples
  – Sketch a search for query “restaurants 94109”
  – OpenTable for “evvia palo alto”
  – Restaurant Comparison for “tamarine palo alto”
Beyond explicit search – explicit introspection

• Two examples
  • Google trends
  • Yahoo! Clues
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

\[
\text{query share} = \left( \frac{\text{number of queries for } q}{\text{number of queries}} \right)_{(t_i, geo_j)}
\]

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

- Google Insight for Search
  - Uses the same data but geared to researchers and advertisers.
Consider other facets

- Searches over time
- **By demographic**
- By Location
- **Search Flow**
- Related Search
Combine Yahoo! search query log with:

- profile information provided by Yahoo! (28 million of its users, birth year, gender, ZIP code)
- US-census information aggregated by ZIP code

Annotate each query with ZIP code area


Location based models from geocoded queries
[Serdyukov, Murdock & van Zwol SIGIR’2009] “Placing Flickr Photos on a Map”
Comparing demographics on 2 queries

- angelina jolie
- brad pitt

**By Demo**
- Men under age 24 who searched for brad pitt also looked at:
  - george clooney
  - bruce willis
  - johnny depp

- 9% 25 to 34
- 10% 35 to 44
- 12% 45 to 54
- 13% 55 to 64
- 10% Above 65

- Women 8% 25 to 34
- Men 12% 35 to 44
- 10% 45 to 54
- 10% 55 to 64
- 10% Above 65

**By Location**

- Data © 2010 NAVTEQ

- Data © 2010 NAVTEQ
Query graph flow

- From implicit to explicit – showing the search flow to users

Most Popular Search Queries

1. selena gomez
   - Will Selena Gomez be the next Charlotte York in 'S...?', TheCelebrityCafe.com
   - Selena Gomez And Leighton Meester Talk Monte Carlo..., omg!
   - Selena Gomez: 'Rock of Ages' Gall, Just Jared Jr.

2. iphone 5
   - Link: The iPhone Turns Four: How It Has Changed Us..., Forbes
   - iPhone birthday: How Apple forever changed the sma..., The Christian Science Monitor
   - iPhone 5, iPad 3 launching this October?, ZDNet

3. ryan dunn
   - Ryan Dunn honored by 'Jackass' crew with a tribute..., KABC-TV Los Angeles
   - Ryan Dunn honored with a tribute video - Watch, KABC-TV Los Angeles
   - Ryan Dunn Death Site Vandalsized Recap, antiMUSIC

4. megan fox
   - Megan Fox's rep confirms Shia LaBeouf hookup, TheCelebrityCafe.com
   - Rep confirms Megan Fox, Shia LaBeouf fling, TheCelebrityCafe.com
   - Fox's rep confirms on set tryst with LaBeouf, Hollywood.com

5. justin bieber
   - Justin Bieber Vs 'Douch Bag' Recap, antiMUSIC
   - Justin Bieber's dive-sized ego tempered by Selena..., Chicago Sun-Times
   - Justin Bieber Gets inspirational In Google Chrome..., Billboard
Conclusion and future directions

Session 4
Conclusions

• We still *don’t understand well information needs*
  – Will we ever?
  – Queries are just an approximation and brain electrodes won’t work

• Users crave for
  – more interactive features
  – more “digest” features – show me answers not links
  – more personalization but are scared of privacy infringements
We are far from being done with innovation in search engines.

**Large scale usage data is key BUT**

<table>
<thead>
<tr>
<th>Usage data at a very large scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>over larger and larger populations</td>
</tr>
<tr>
<td>over longer and longer periods of time</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personalization</th>
<th>Privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>More data via larger communities, makes data less personalized</td>
<td></td>
</tr>
<tr>
<td><em>wisdom of crowds does not work well on small corpora</em></td>
<td></td>
</tr>
<tr>
<td>Over personalization endangers privacy</td>
<td></td>
</tr>
<tr>
<td>Long-term logs endanger privacy</td>
<td></td>
</tr>
</tbody>
</table>
The New frontiers

- Front-end and user experience
  - The most probable reason for users to switch between quasi-equivalent engines is a better user experience

- Depart from the rectangle/ranked list paradigm
  - Get rid of queries? Implicit search
    - Content delivery is one flavor
    - But in general, why should we even have to formulate a query?