An earlier version of this tutorial was given at SIGIR'2012 and WSDM'2013

(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)

Web Search
Quick Retrospective

YAHOO!

Looking Back

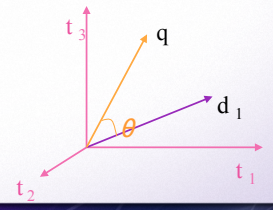
4

1970-90
Information Retrieval

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_1 document, q query)
- Evaluation on predefined test collections (e.g., 300 docs CACM)



4

Looking Back

5

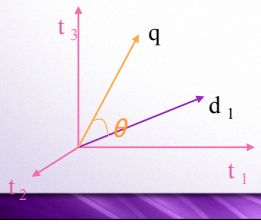
1970-today
Information Retrieval

Compare queries and documents as vectors of words

Information Retrieval "invented" by Salton at Cornell: Text analysis, count words

Vector Space model (d_1 document, q 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)

5

Then First Revolution

6

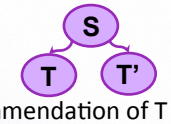
1998: Analyze links on a very large scale

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"



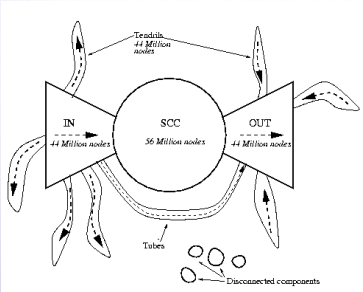
- Recommendation of T by S
- Evidence that S and T are related
- Relation between T and T'

6

First Revolution (cont.)

7

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*"

7

Then the Second Revolution

2003 Usage Data

AOL query log transcript

AnonID	Query	QueryTime	ItemRank	ClickURL
100218	tennessee department of transportation	2006-03-01 11:08:30	1	http://www.tdot.state.tn.us
100218	tennessee federal court	2006-03-01 11:53:44	1	http://www.constructionweblinks.com
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	2	http://www.tennessee.gov
100218	state of tennessee emergency communications board	2006-03-01 12:56:18	1	http://www.tennessee.gov
100218	dixie youth softball	2006-03-02 10:36:48	2	http://www.dixie.org
100218	cdwg	2006-03-03 14:29:07	1	http://www.cdwg.com
100218	cdwg scam cdwge	2006-03-03 14:30:11		
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	2	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	escambia county sheriff's department	2006-03-07 09:26:51	1	http://www.escambiaso.com
100218	pensacola police department	2006-03-07 09:34:28	1	http://www.pensacolapolice.com
100218	memphis pd	2006-03-07 09:42:33	1	http://www.memphispolice.org
100218	nashville metro pd	2006-03-07 09:44:43	1	http://www.police.nashville.org
100218	florida highway patrol	2006-03-07 09:48:35	1	http://www.flhp.state.fl.us
100218	tennessee highway patrol	2006-03-07 09:49:52	1	http://www.state.tn.us
100218	florida bureau of investigations	2006-03-07 09:51:08	2	http://www.fdle.com
100218	florida bureau of investigations	2006-03-07 09:51:08	1	http://www.flhp.state.fl.us
100218	government finance officers association	2006-03-07 21:16:11		
100218	state of tennessee controllers manual	2006-03-07 21:17:12		
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	3	http://www.comptroller.state.tn.us
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	4	http://www.tbr.state.tn.us
100218	state of tennessee audit controllers manual	2006-03-07 21:17:40	9	http://audit.tennessee.edu
100218	internal controls for municipalities under 10 000	2006-03-07 21:38:04	1	http://www.nyscpa.org
100218	internal controls for municipalities under 10 000	2006-03-07 21:38:04	4	http://www.masidor.com
100218	municipality fraud detection techniques	2006-03-07 21:41:40		
100218	municipal fraud audit detection internal controls	2006-03-07 21:43:15		
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	1	http://www.whitehouse.gov
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	4	http://www.nlsgc.org
100218	internal fraud controls for municipalities cities towns local government	2006-03-07 21:45:13	7	http://www.sao.state.ut.us
100218	evaluating internal controls a local government managers guide	2006-03-07 21:51:18	5	http://www.allbusiness.com

[Xiaodong Shi, U. Of Michigan, 2007]
"Social Network Analysis of Web Search Engine Query Logs"

Implicit feedback

- Query logs and click through data

Live experiments

- Try out experimental features on x% of users

8

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

9

Usage Data

YAHOO!

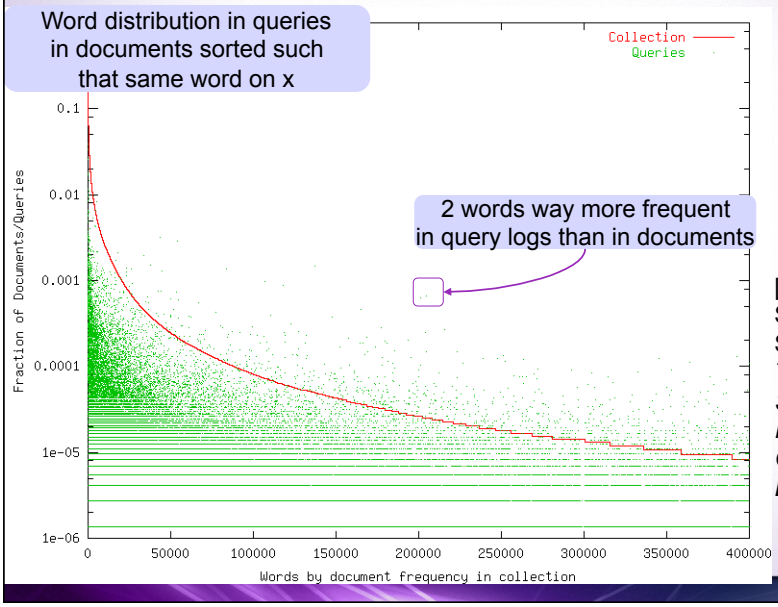
Collective Usage Data

- 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

11

Queries differ from Documents

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



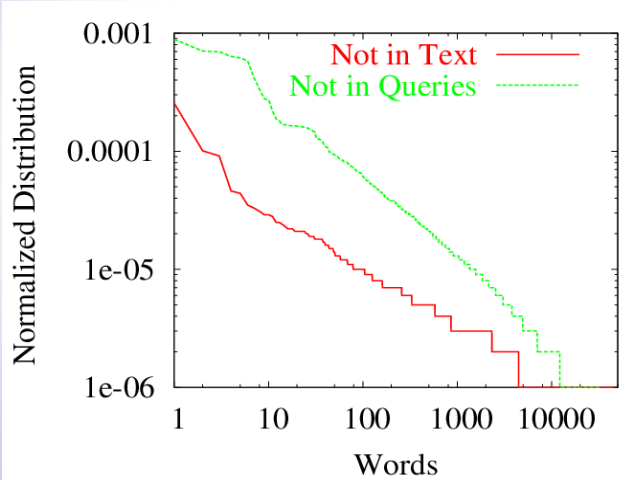
2 words way more frequent in query logs than in documents

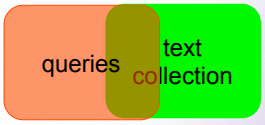
[Baeza-Yates & Saint-Jean, SPIRE'2003] *"A Three Level Search Engine Index based in Query Log Distribution"*

12

Queries and Text

Each set with its own power law!





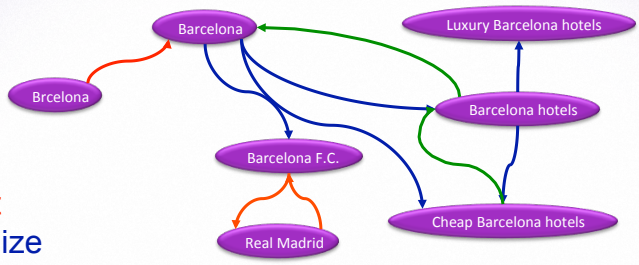
Counting queries

Simplest method: count! and compare

Eg. spelling – from http://www.google.com/jobs/britney.html

488941 britney spears	20 britney spears	9 brinttany spears	5 brney spears	3 britiy spears	2 brizzeny spears
40134 brittany spea	britany spears	9 britanay spears	5 broitney spears	3 britweny spears	2 brittany spears
36315 britney spea	britny spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 brittany spears
24942 brittany spear	briteny spears	9 britn spears	5 bruteny spears	3 britneby spears	2 brittney spears
7331 britny spears	britteny spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spear	briney spears	9 britneyn spears	5 bzrittney spears	3 britnesy spears	2 britane spears
2696 britteny spea	brittny spears	9 britzney spears	5 gzitney spears	3 britnetty spears	2 britaneny spears
1807 briney spears	brintey spears	9 brtiny spears	5 sgritney spears	3 britnez spears	2 britania spears
1635 brittny spear	britanny spears	9 brzitney spears	4 bittny spears	3 britneyxxx spears	2 britann spears
1479 brintey spear	britiny spears	9 brtny spears	4 bnzitney spears	3 britnity spears	2 britanna spears
1479 britanny spea	britney spears	9 brytny spears	4 brandy spears	3 britntey spears	2 britannie spears
1336 britiny spear	britnet spears	9 zbitney spears	4 bzbritney spears	3 britnyey spears	2 britannn spears
1211 britnet spear	britneyspears	8 birtiny spears	4 breatiny spears	3 brittzeny spears	2 britannu spears
1096 britiney spea	britney spears	8 bithney spears	4 breetney spears	3 brittneey spears	2 britanyl spears
991 britaney spea	britnet spears	8 brattany spears	4 bretiney spears	3 brittnney spears	2 britanyt spears
991 britnay spear	britiny spears	8 breitny spears	4 brfitney spears	3 brittnyey spears	2 briteeny spears
811 brithney spea	britiney spears	8 breteny spears	4 briattany spears	3 brityen spears	2 britenany spears
811 britiney spear	britaney spears	8 brightny spears	4 brieteny spears	3 briytney spears	2 britenet spears
664 britney spear	britnay spears	8 brintay spears	4 brietty spears	2 britney spears	2 briteniy spears
664 brintney spea	brithney spears	8 brinttey spears	4 briitny spears	3 broteny spears	2 britenys spears
664 briteney spea	brtiney spears	8 briotney spears	4 briittany spears	3 brtaney spears	2 britianey spears
601 bitney spears	birtney spears	8 britanys spears	4 brinie spears	3 brtiiany spears	2 britin spears
601 brinty spears	brintey spears	8 britley spears	4 brinteney spears	3 brtinay spears	2 britinaky spears
544 brittaney spe	britney spears	8 brineyb spears	4 brintne spears	3 brtinney spears	2 britny spears
544 brittnay spea	brintney spears	8 britnrey spears	4 britaby spears	3 brtitany spears	2 britnaney spears
364 britey spears	briteney spears	8 brintty spears	4 britaey spears	3 brtiteny spears	2 britnat spears
364 brittiny spea	bitney spears	8 brittnaz spears	4 britainey spears	3 brtnat spears	2 britnkey spears
329 brtney spears	brinty spears	8 brottany spears	4 britinie spears	3 brytiny spears	2 britndy spears
269 bretney spear	brittaney spears	7 baritney spears	4 britinney spears	3 btney spears	2 britnch spears
269 britneys spea	brittnay spears	7 biritey spears	4 britnney spears	3 drittney spears	2 britneney spears
244 britne spears	britey spears	7 biteney spears	4 britneaz spears	3 pzrtney spears	2 britneyf spears
244 brytney spear	brittiny spears	7 biziny spears	4 britnel spears	3 zbritney spears	2 britneye spears
220 breatney spea	brtney spears	7 breateny spears	4 britneuy spears	2 baritany spears	2 britneyh spears
220 britiany spea	bretney spears	7 brianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears
199 britnney spea	britneys spears	7 brintye spears	4 britnney spears	2 bbitney spears	2 britneyny spears
169 britnry spear	britne spears	7 britianny spears	4 brittaby spears	2 bbritny spears	2 britnkey spears
147 breatny spear		7 britly spears	4 brittery spears	2 bbrittany spears	2 britnjey spears
147 brittnay spe		7 britnej spears	4 brittkey spears	2 britany spears	2 britnne spears
147 britty spears		7 britneyn spears	4 brittnasy spears	2 britny spears	2 britnu spears
147 brotney spear		7 britniey spears	4 brittnat spears	2 brtney spears	2 britoney spears
147 brutney spear		7 britnnay spears	4 brittneny spears	2 brtny spears	2 britrany spears
133 brittenay spe		7 brittian spears	4 brittnye spears	2 betney spears	2 britrany spears
133 briyney spear		7 briyny spears	4 brittteny spears	2 betny spears	2 britry spears
121 bittany spear		7 brzittany spears	4 briutney spears	2 bkriney spears	2 britsany spears
121 bridney spears	17 brittanie spears	7 brttiney spears	4 briyeny spears	2 biney spears	2 brittanay spears
121 britainy spears	15 brinney spears	7 briteny spears	4 brnity spears	2 bintey spears	2 brittany spears
121 britney spears	15 briten spears	7 bzrittany spears	4 britteny spears	2 biretny spears	2 brittans spears
109 brietney spears	15 briteney spears	6 beritny spears	4 brttiany spears	2 biritany spears	2 brittanyh spears
109 brithny spears	15 britheny spears	6 bhritney spears	4 bryney spears	2 biritany spears	2 brittanyn spears

Query-flow graph



- Correct
- Specialize
- Generalize
- Parallel Move

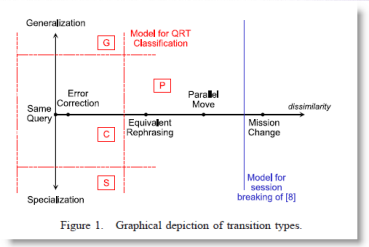


Figure 1. Graphical depiction of transition types.

[Boldi, Bonchi, Castillo and Vigna Web Intelligence 2009] *"From Dango to Japanese Cakes: Query Reformulation Models and Patterns*

Notification: 18/Sep/2009

15

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

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

  - 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

16

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, Web Intelligence 2004] *"Ranking Boosting based in Query Clustering"*
- 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"*

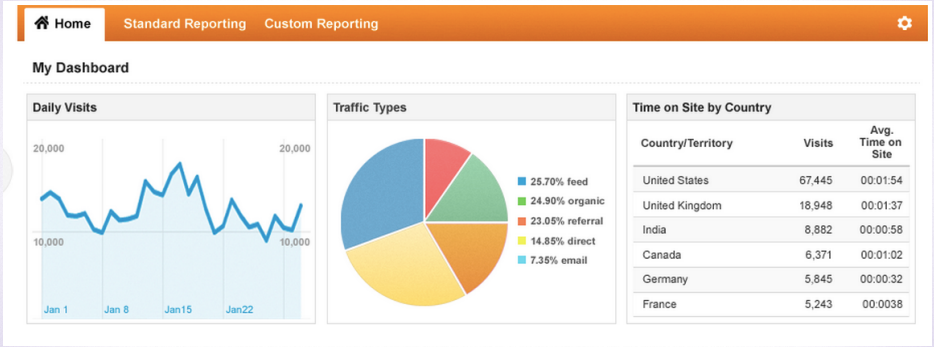
17

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.

18

Dashboards

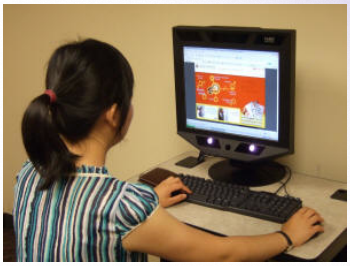


Example of dashboard from Google Analytics

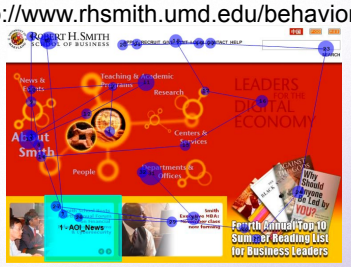
19

More Signals: Non-intrusive Eye-Tracking devices

- Camera by the monitor, facing the user



University of Maryland TOBII system
http://www.rhsmith.umd.edu/behaviorlab/

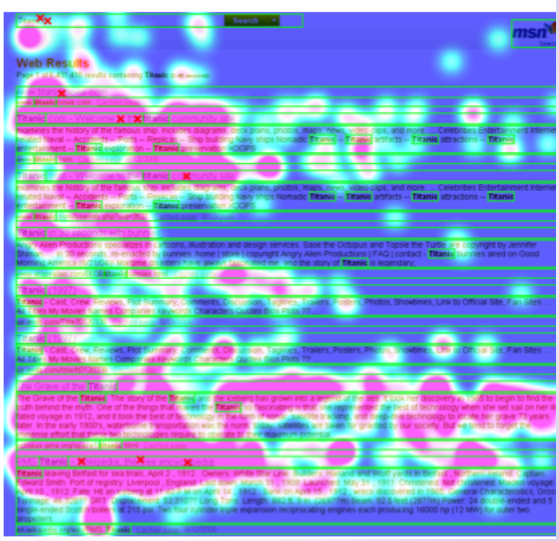


- Data Representation
  - Static “saccade path”
  - Animated “saccade path”
  - Blind zones maps
  - Heat maps

20

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.



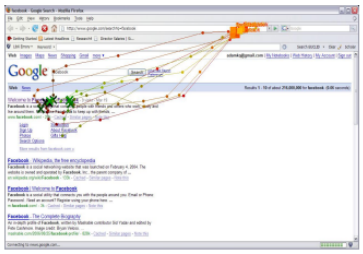
21

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"*



"facebook"

"black thunder"

22

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"

23

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.

24

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

25

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

26

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

27

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"*

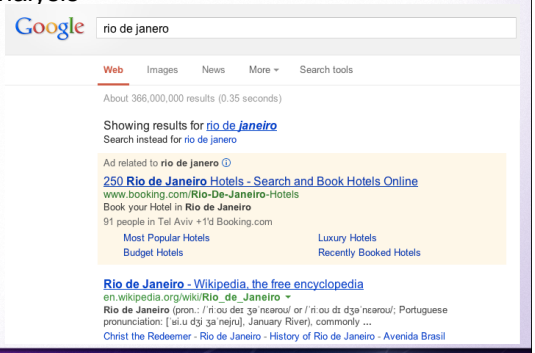
28

Benefits of Usage Data

YAHOO!

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

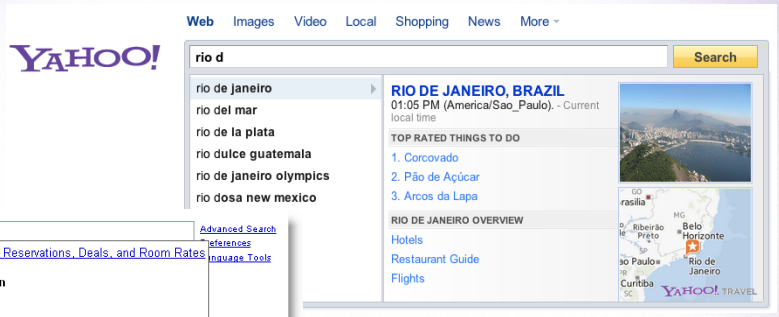


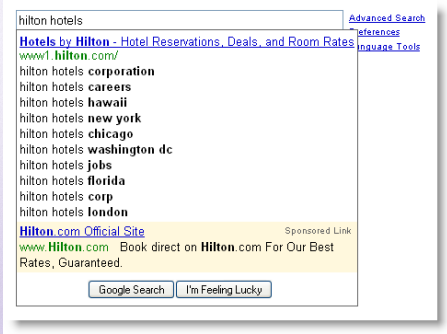
30

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



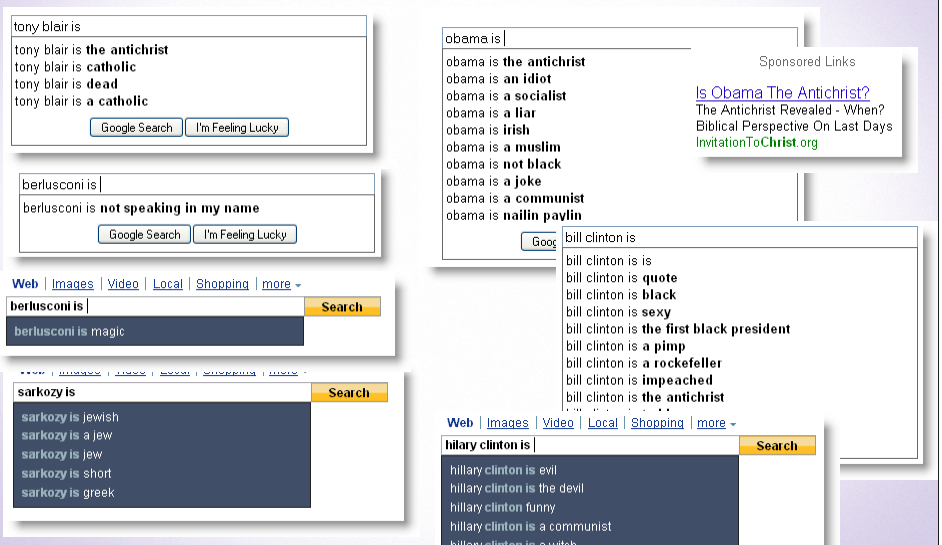


Dynamic Query Suggestions

- 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.

33

The "Voice of the Search Engine"???



34

Query Assistance

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

35

Help Users Reformulate their Queries

Related Queries

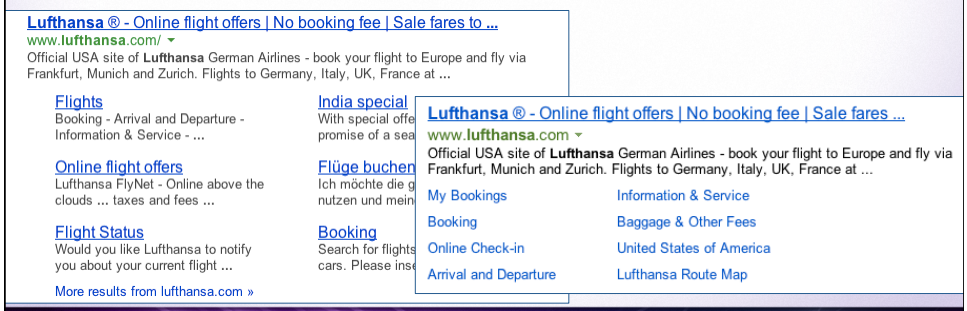
1. Content-aware approaches: Use SERP or target pages to measure query similarity
2. Content-ignorant approaches: Use Clicks
3. Query-flow approaches: Monitor the users' sequential search behavior to better understand query intent

36

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"



37

Identifying "research" sessions on the fly

- A research session R is a maximal order sequence

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

Where, for given thresholds s_θ, t_θ and c_θ, all (u_i)'s refer to the same user and

$$t_{i_1} \le \cdots \le t_{i_k} \le \tau$$

$$\forall i, j \in \{i_1, \ldots, i_k\} : s\left(f(q_l), f(q_j)\right) \ge s_\vartheta$$

$$|R| = k \ge k_\vartheta$$

$$\sum_{j=1}^{k} \left| C_{i_{j\beta}} \right| \ge c_\vartheta$$

No specific time limitation

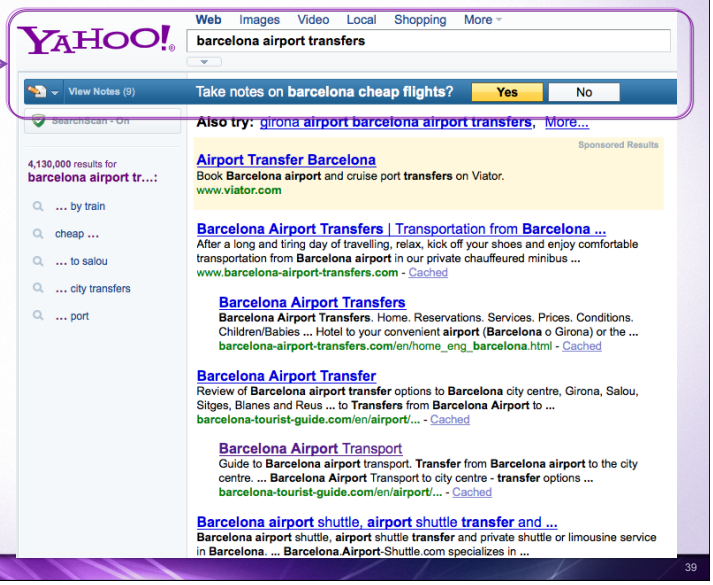
Topical coherence

Complexity

Engagement

38

Exploiting query research sessions: Search Pad*



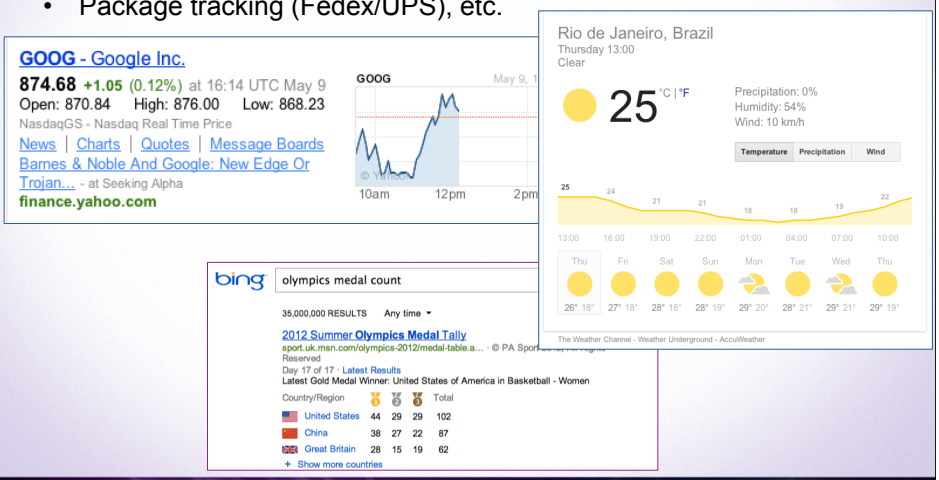
Appears when research session is identified

Discontinued!

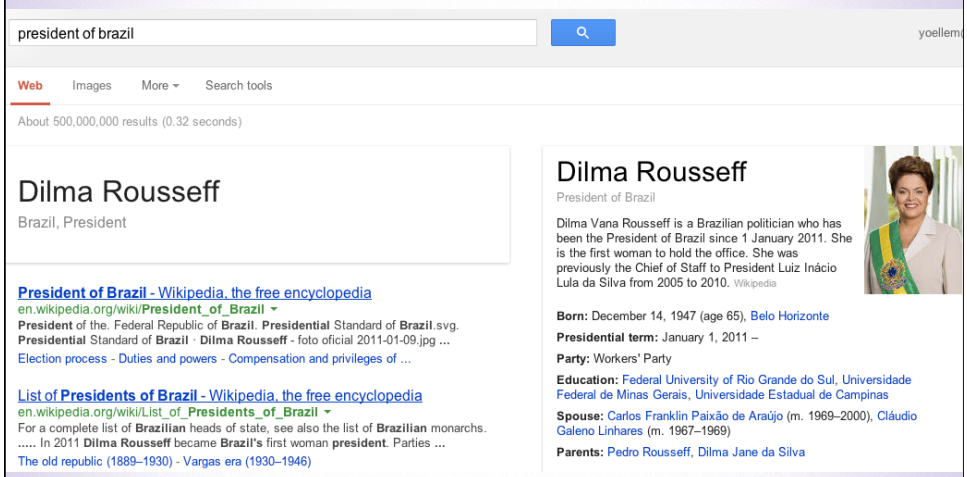
How to Measure when there is No Click?

40

- 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?

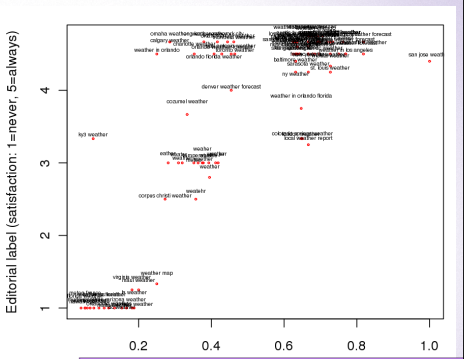


41

Analyze Change in User's Behavior

42

- 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 = (XQQ+XQC)/(XQQ +XQC+XQX)
- Generalize: Analyze behavior of "pitbulls", "poodles", "serial shoppers", kids etc.?



Change in the tenacity of "pitbulls" on "weather" queries

[C. Castillo, A. Gionis, R. Lempel and Y. Maarek, Industrial Track, SIGIR'2010] "*When no clicks are good news*"

42

Explicit Usage Data
Relevance feedback Revisited

Google Searchwiki

- User's feedback
  - Comment
  - Promote
  - Remove

Discontinued!

Web Images News Translate Scholar YouTube Gmail more ▾

Google haifa Search

Search: the web pages from Israel

Web Show options...

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 -

Type your comment here

Make a public comment Cancel

Haifa Wehbe - Wikipedia, the free encyclopedia
Haifa Wehbe was born in Mahrouna, a small Shiite farming town in Southern ... **Haifa** Wehb has also performed with artists from outside the Arab world. ...
en.wikipedia.org/wiki/**Haifa**_Wehbe - Cached - Similar -

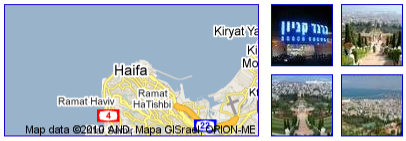
Gooooooooooogle
1 2 3 4 5 6 7 8 9 10 Next

Add a result - See all my SearchWiki notes - See all notes for this SearchWiki - Learn more

43

To "Stars in Search"

Haifa Israel maps.google.com



Haifa - Wikipedia, the free encyclopedia
Haifa (Hebrew: חֵיפָה About this sound Heifa (help·info); Arabic: حيفا About this sound Ḥayfā (help·info)) is the largest city in northern Israel, ...
Etymology - Early history - Modern history - Demographics
en.wikipedia.org/wiki/**Haifa** - Cached - Similar

Haifa Wehbe - Wikipedia, the free encyclopedia
Haifa Wehbe (Arabic: هيفاء وهبي) is a Lebanese model, actress, and singer who rose to fame in the Arab world as runner up for Miss Lebanon and later ...
en.wikipedia.org/wiki/**Haifa**_Wehbe - 9 hours ago - Cached - Similar

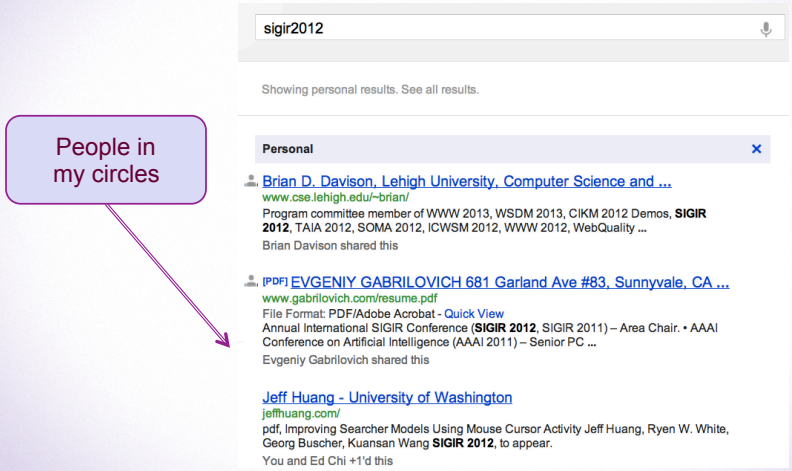
Discontinued!

[Dupont, The official Google Blog, 2010] *"Stars in search replace SearchWiki."*

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

44

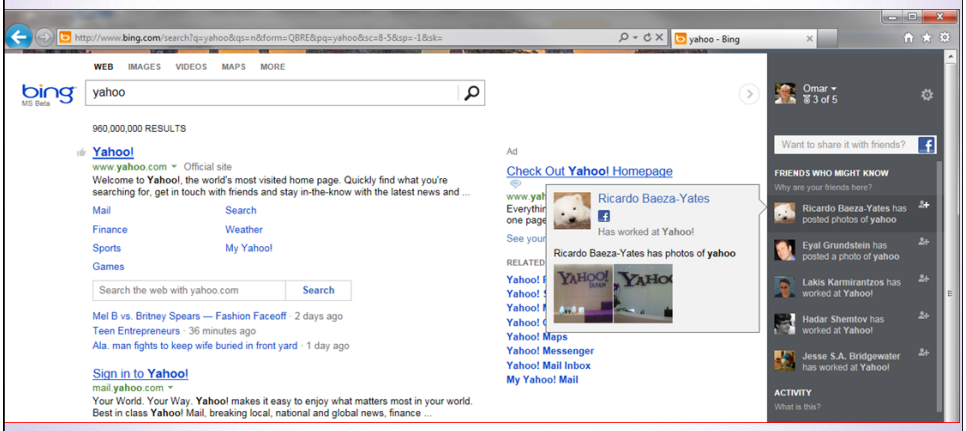
To +1 in Social Search



45

Bing Social Search –Facebook integration

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

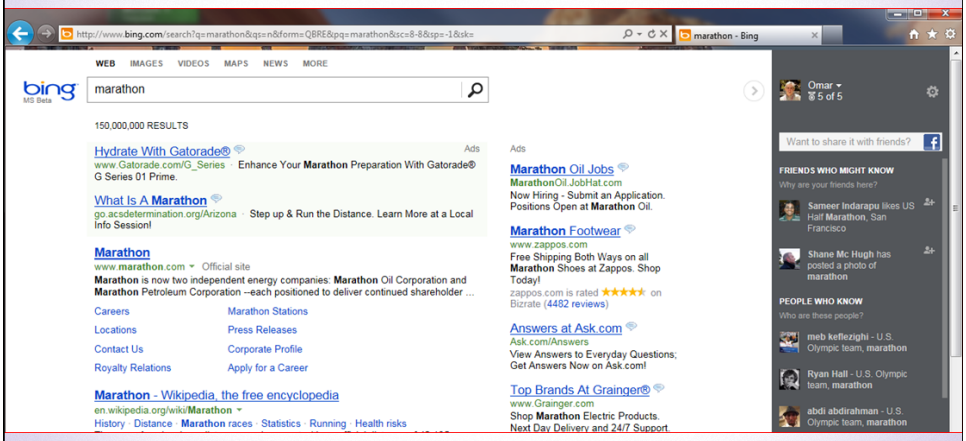


http://www.bing.com/community/site_blogs/b/search/archive/2012/05/15/start-doing-more-now-try-the-new-bing-today.aspx

Screendump courtesy of Omar Alonso

46

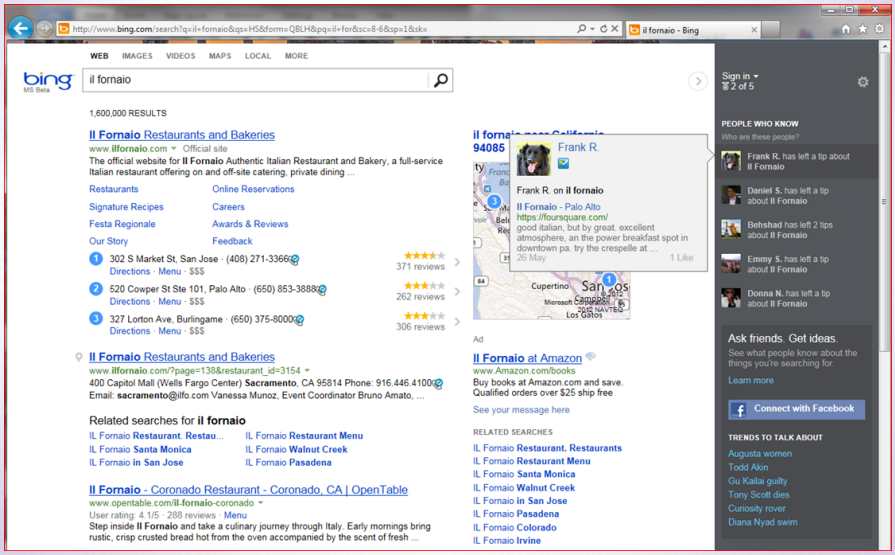
Bing Social Search – Friends and experts (twitter)



http://www.bing.com/community/site_blogs/b/search/archive/2012/07/02/bing-search-quality-insights-friends-and-experts.aspx

Screendump courtesy of Omar Alonso 47

Bing Social Search – foursquare integration



Screendump courtesy of Omar Alonso 48

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

$$\textbf{query share} = \left(\frac{\textit{number of queries for q}}{\textit{number of queries}} \right)_{(t_i, geo_j)}$$

- Google trends
  - "Google Trends provides an index of the volume of Google queries by geographic location and category"
  - [Varian and Choi 2009, in Google Search Blog]
    *"Pr*edicting 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"

49

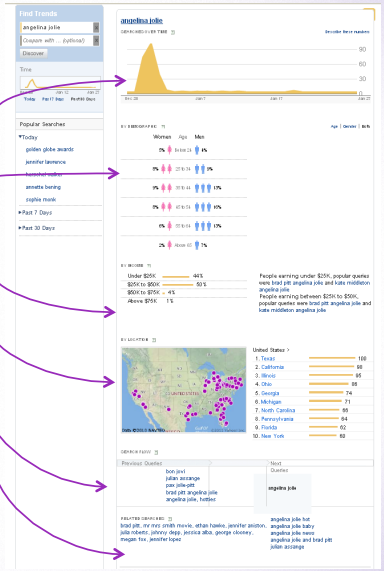
Query Trends Exposed to Users

50

Yahoo! Clues clues.yahoo.com

Consider other facets

- Searches over time
- **By demographic**
- **By Income**
- By Location
- **Search Flow**
- Related Search



51

By Demographic and By Income

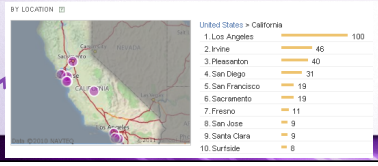
Combine Yahoo! search query log with:

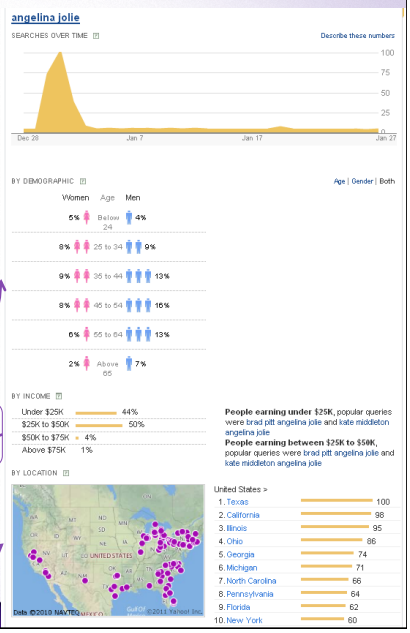
- profile information provided by Yahoo! (birth year, gender, ZIP code)
- US-census information aggregated by ZIP code

Annotate each query with the average per-capita income in the ZIP code area

[Weber & Castillo SIGIR'2010] *"The Demographics of Web Search"* [Weber & Jaimes, WSDM'2011] *"Who Uses Web Search for What? And How?"*

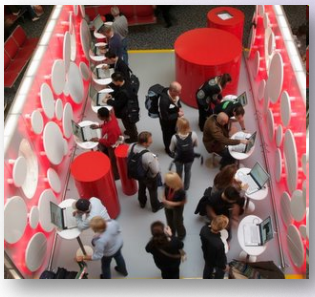
[Serdyukov, Murdock & van Zwol SIGIR'2009] *"Placing Flickr Photos on a Map"*





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"*

53

Wisdom of Crowds and the Long Tail

The Wisdom of Crowds

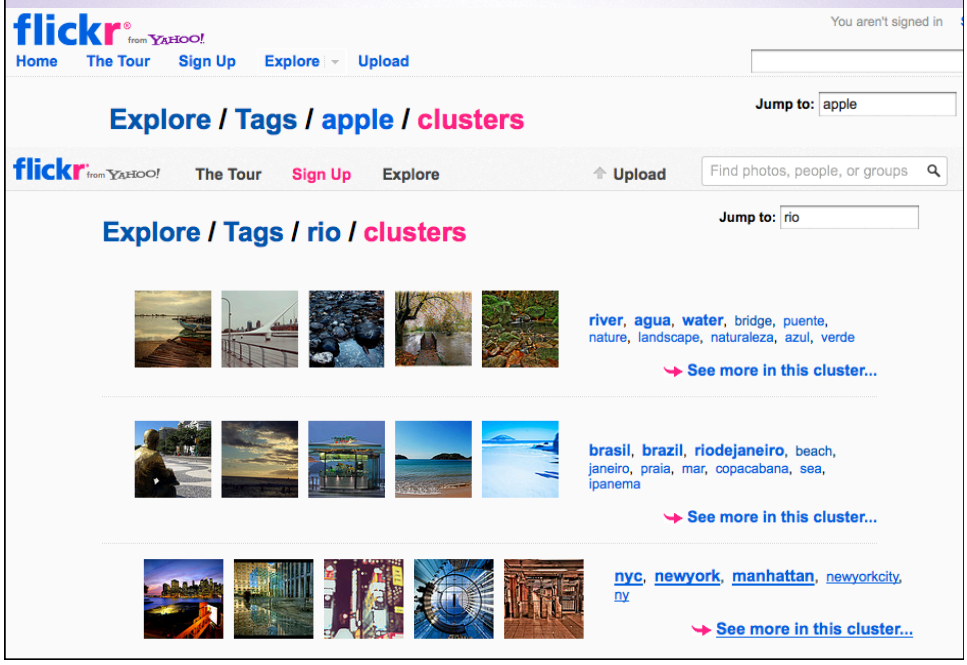
- James Surowiecki, a ***New Yorker*** columnist, published this book in 2004
  - **"Under the right circumstances, groups are remarkably intelligent"**
- Importance of diversity, independence and decentralization

Aggregating data

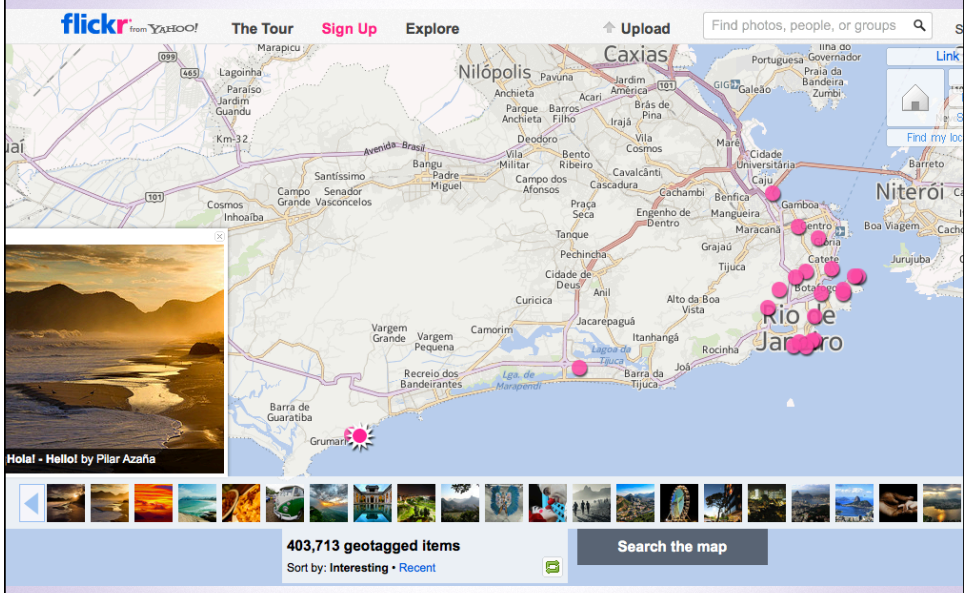
"large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future".

55

Flickr: Clustering Pictures



Flickr: Geo-tagged pictures



57

Wisdom in Web Search

Generation	Technology	Wisdom
First: 1994-98	Classical IR	Writers
Second: 1997-2003	+Link Analysis +Anchor text +Click-through voting	+Webmasters +Readers
Third: 2003-2010	+Usage data mining	Everyone
Fourth: 2008-??	+Query intent detection +Learning to rank	Everyone

58

The Wisdom of Crowds

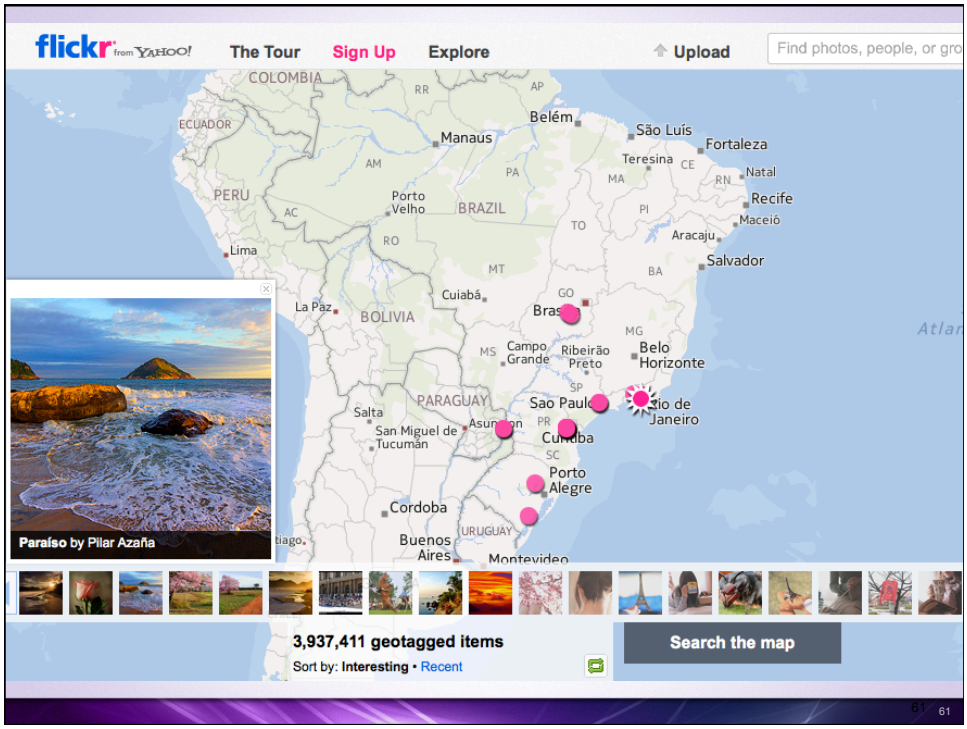
- Popularity
- Diversity
- Quality
- Coverage

59

Popularity

animals architecture art asia australia autumn baby band barcelona beach berlin bike bird birds birthday black blackandwhite blue bw california canada canon car cat chicago china christmas church city clouds color concert dance day de dog england europe fall family fashion festival film florida flower flowers food football france friends fun garden geotagged germany girl graffiti green halloween hawaii holiday house india instagramapp iphone iphoneography island italia italy japan kids la lake landscape light live london love macro me mexico model museum music nature new newyork newyorkcity night nikon nyc ocean old paris park party people photo photography photos portrait raw red river rock san sanfrancisco scotland sea seattle show sky snow spain spring square squareformat street summer sun sunset taiwan texas thailand tokyo travel tree trees trip uk unitedstates urban usa vacation vintage washington water wedding white winter woman yellow zoo

60

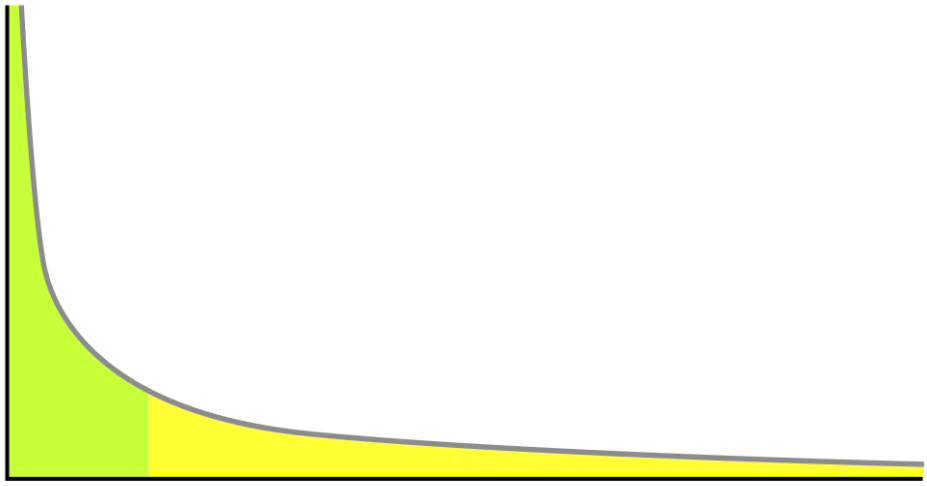


The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

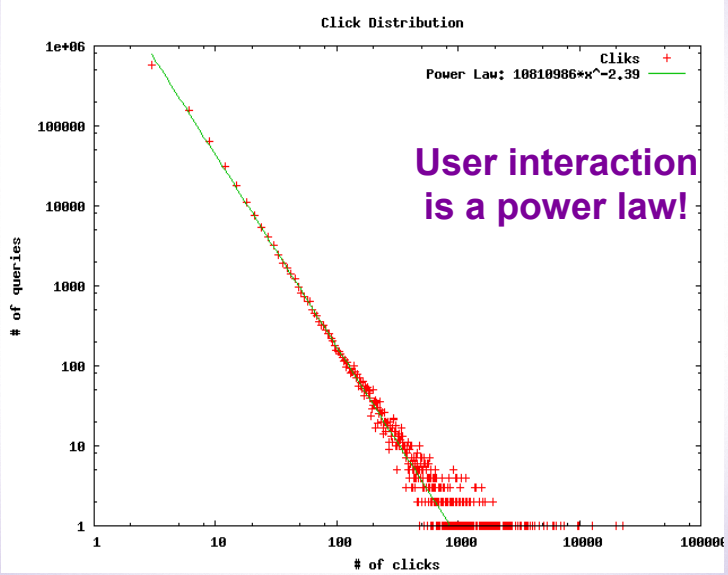
Long tail
Heavy tail

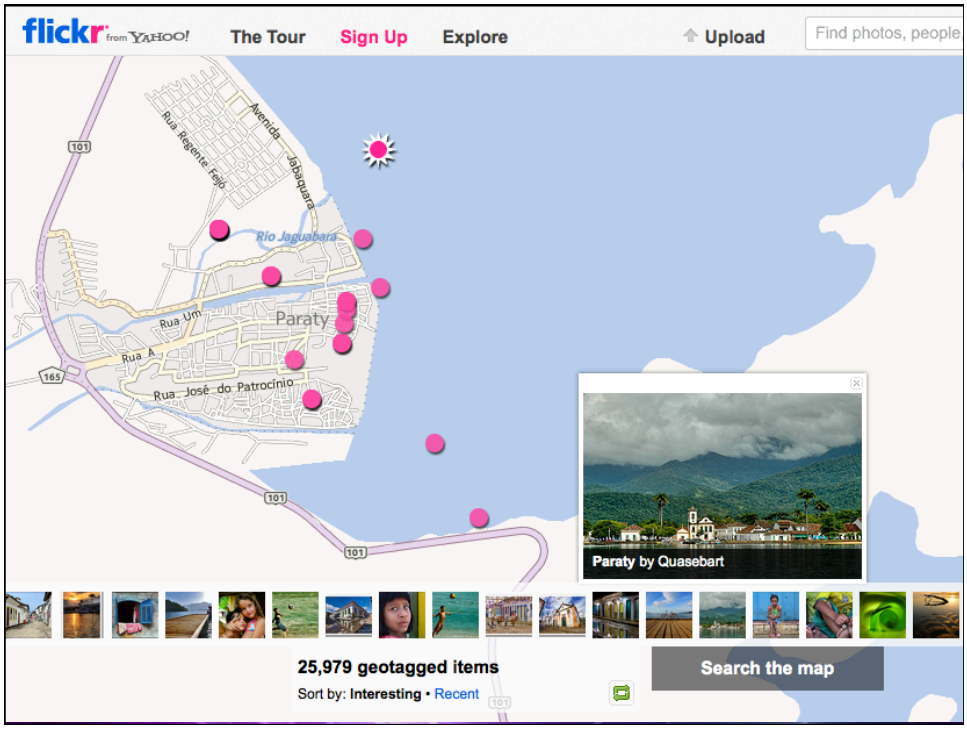
The Long Tail



Most measures in the Web follow a power law

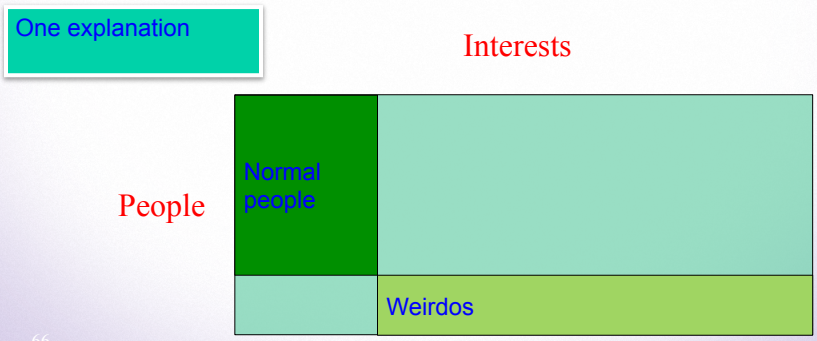
Example: Click Distribution





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, …



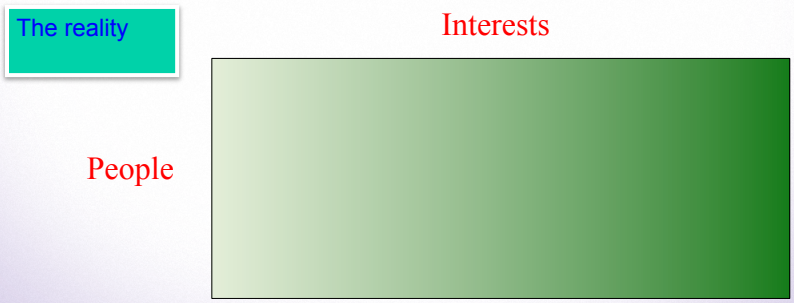
66

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



Broder, Gabrilovich, Goel, Pang; WSDM 2009

67

Who is in the Long Tail?

- Popular hypothesis:
  - Majority of consumers consistently follow the crowd, only eccentrics issue tail queries
- Wrong!
  - See *"Anatomy of the Long Tail: Ordinary People with Extraordinary Tastes"* [Goel et al. WSDM'2009]
  - 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.

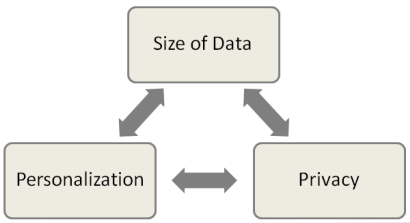
68

Risks and Limitations

Conflicting Factors

70

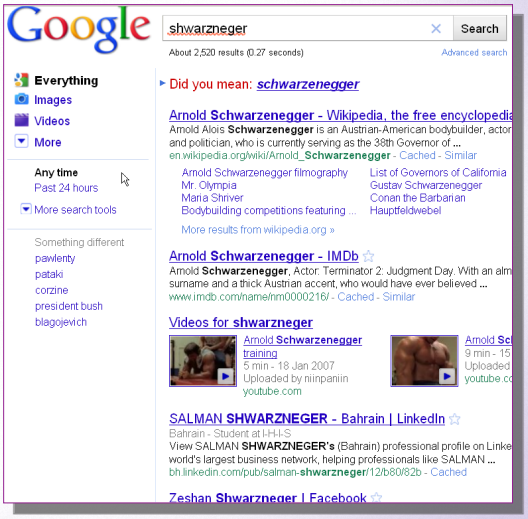
- Size
  - Web Search requires usage data over larger and larger populations over longer and longer periods of time
- Personalization
  - Size is not there on tail (sparse data), yet tail is critical
  - Personalization exposes private facets
- Privacy
  - Long-term logs endanger privacy



Risks of size: When the crowd dominates

71

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



71

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.

72

Personalization "Facets"

73

Successful "facets":

- Language (not always)
- Geo

liver	
liverpool fc	5,180,000 results
liverpool echo	5,660,000 results
liverpool airport	584,000 results
liverpool echo arena	210,000 results
liverpool one	11,200,000 results
liverpool empire	379,000 results
liverpool university	4,070,000 results
liverpool victoria	645,000 results
liverpool city council	798,000 results
liverpool community college	357,000 results
	close

kan	
kanye west	25,300,000 results
kangaroo island	1,130,000 results
kangaroo valley	461,000 results
kangan batman tafe	20,600 results
kangaroo tent city	125,000 results
kangaroo	14,400,000 results
kanye west lyrics	794,000 results
kangaroo photos	2,800,000 results
kangaroo island ferry	103,000 results
kangaroo island accommodation	1,090,000 results
	close

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

73

Social Personalization: personal results (discontinued?)

74

barcelona

barcelona

Showing personal results. See all results.

10 personal results. 1,1

Personal

Renaissance **Barcelona** | |
www.marriott.com/Spain
Luxury **Barcelona** Hotel Near Air

Official FC **Barcelona** Web
www.fc**barcelona**.com/
Official FC **Barcelona** website. A
supporters club services and info
Football - Calendar - Tickets -
6,688 people +1'd this

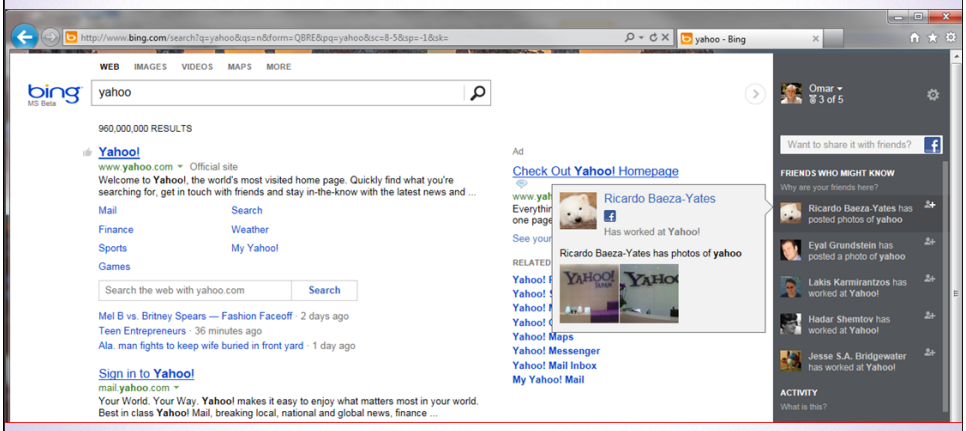
La Granja - No hace falta la direction
https://plus.google.com
Jackie Kettner · Jul 9, 2011 · Limited

2 must see movies: Slumdog Millionnaire and Vicky Cristina **Barcelona**
ouriel.typepad.com/myblog/2009/01/2-must-see-movi.html
Jan 18, 2009
I have seen a lot of movies but 2 are really striking and ...
Ouriel Ohayon shared this · Public

I didn't see this coming. Humbled to have Linus posting kernel ...
https://plus.google.com
Vic Gundotra · Nov 7, 2011 · Public
I didn't see this coming. Humbled to have Linus posting kernel updates to his new Google+ page.

Social Personalization

75

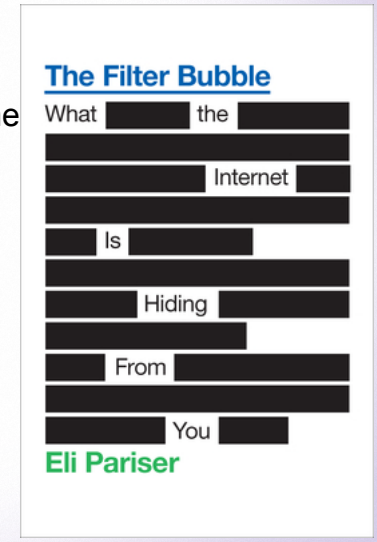


75

Risks of Personalization

- The Filter "Bubble", Eli Pariser
- Avoid the Poor get Poorer Syndrome

- Solutions:
- Diversity
- Serendipity
- Explore & Exploit



76

Risks of Privacy: AOL Query Logs Release Incident

A Face Is Exposed for AOL Searcher No. 4417749,
By MICHAEL BARBARO and TOM ZELLER Jr,
The New York Times, Aug 9 2006

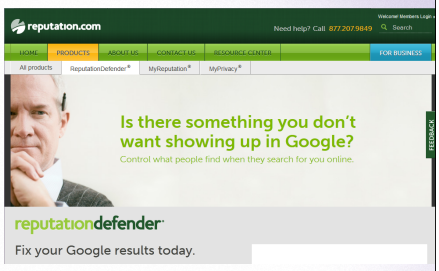
- No. 4417749 conducted hundreds of searches over a
- three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”
- Other queries: “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett 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.

77

Risks of Privacy

78

- (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



78

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

79

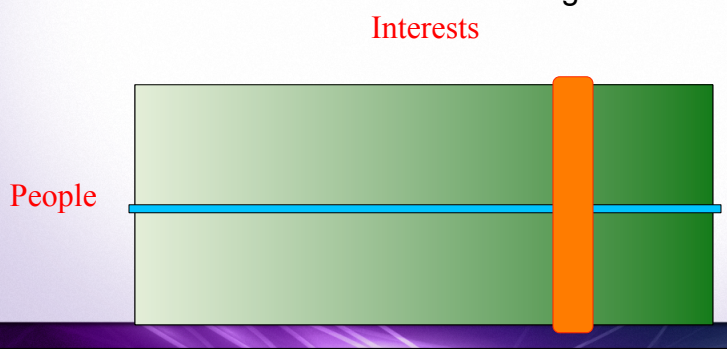
Wisdom of Ad-hoc Crowds

YAHOO!

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



81

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

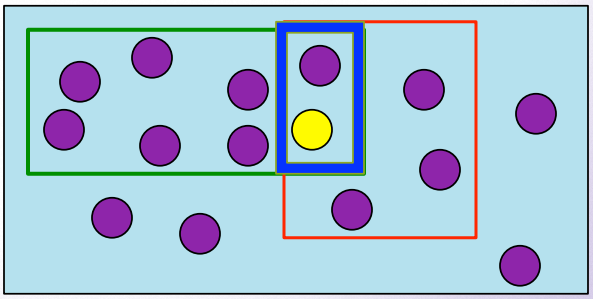
- What is the right interface?

82

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"

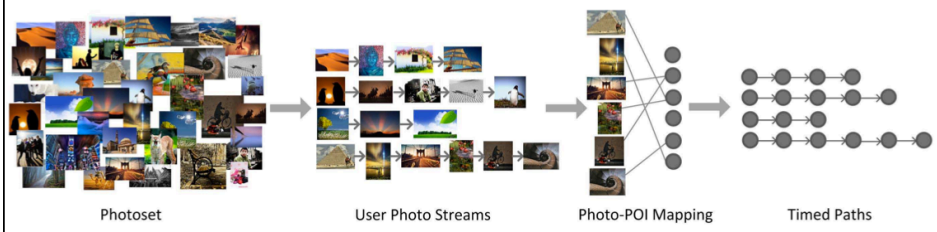
- Middle age men
- Fans of Messi



83

Example: Mining Geo/time Data

- Optimal Touristic Paths from Flickr

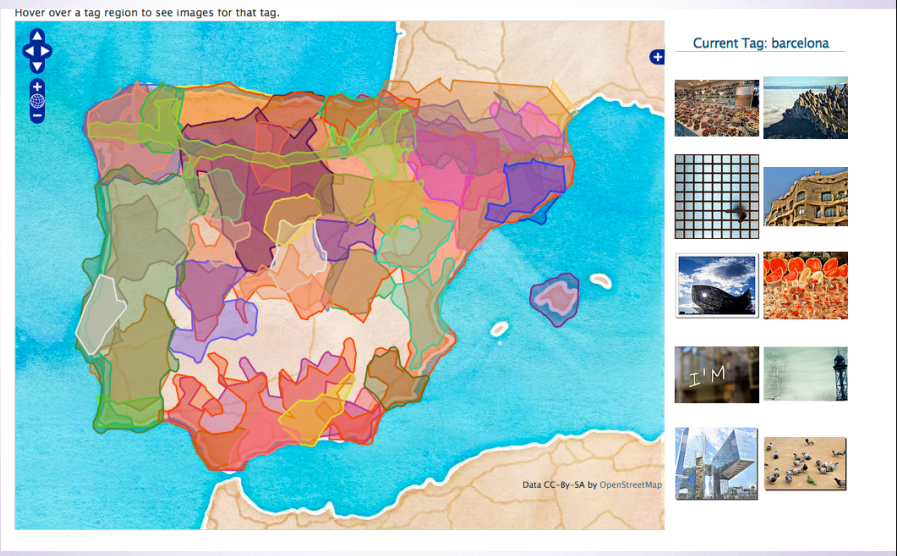


- Good for tourists and locals

De Choudhury et al, HT 2010

84

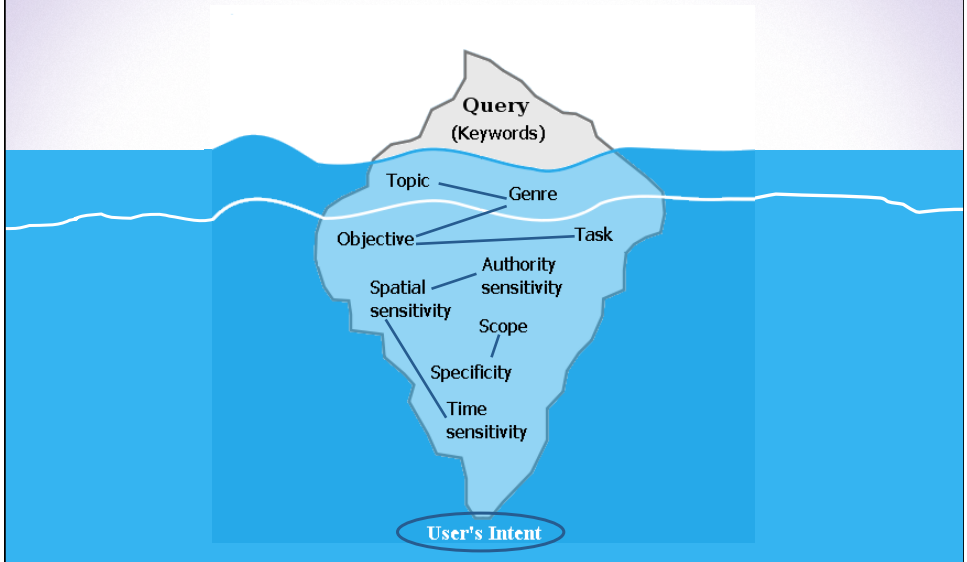
Another example: Regions from Pictures



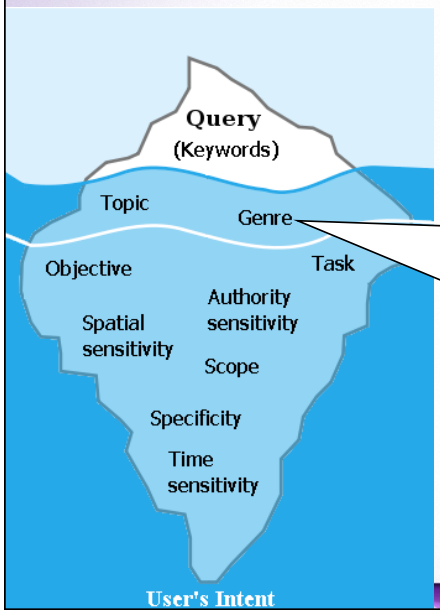
Search Example: Query Intent Prediction

- Users have complex interests that go beyond to 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

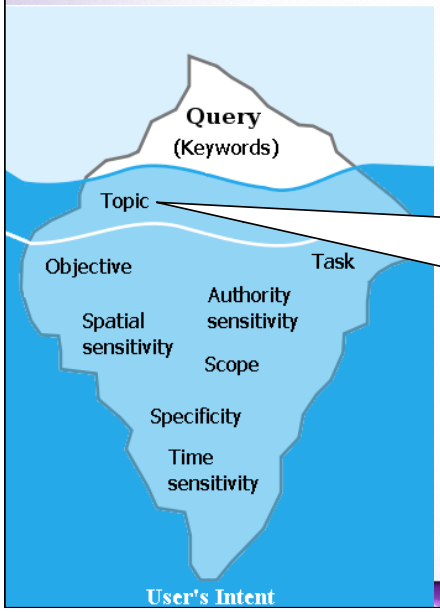


Meta-Dimension

- **News**. *University hunger's strike*
- **Business**. *Sale of furniture*
- **Reference**. *Contraceptive method*
- **Community**. *Your favorite singer*

88

Dimensions of the User's Intent

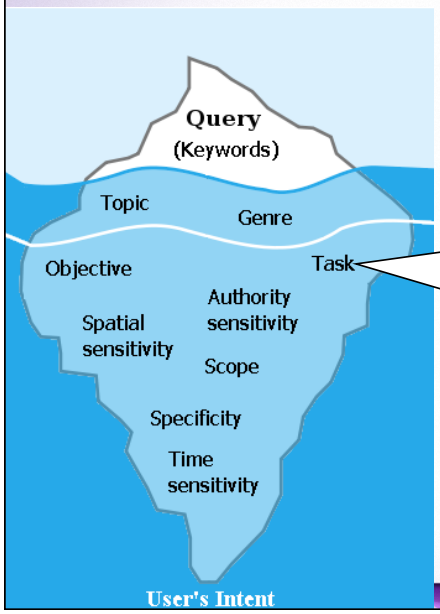


A list of 18 topics built from the first level of **categories** offered by ODP, Yahoo!, and Wikipedia.

News - *Reference*

89

Dimensions of the User's Intent

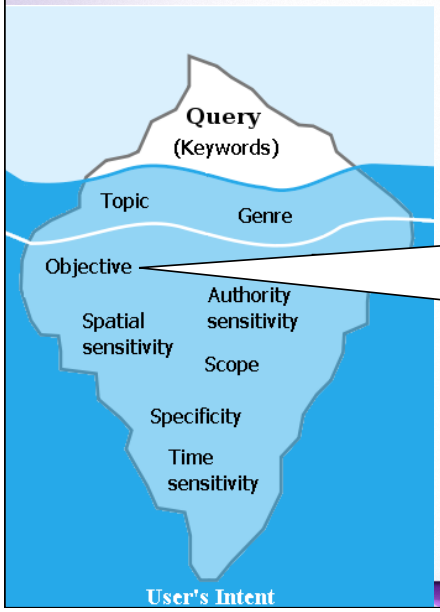


The primary need of the user:

- **Informational**. *What is thermodynamics?*
- **Not-Informational**. *Notebook sales*
- **Both**. *Airbrush*

90

Dimensions of the User's Intent

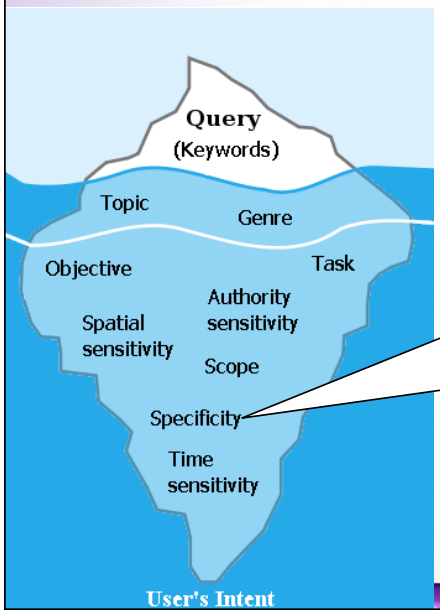


Is the query aimed to:

- Obtain an informational **resource**? *street map of Barcelona*
- Perform an **action**? *Renting an apartment*

91

Dimensions of the User's Intent

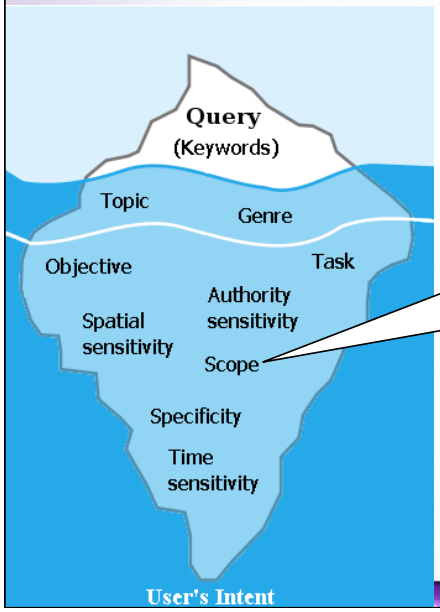


How specialized is a query:

- **Specific**. *Autonomous Univ. of Bucaramanga.*
- **Medium.** *Private universities*
- **Broad**. *Universities*

92

Dimensions of the User's Intent

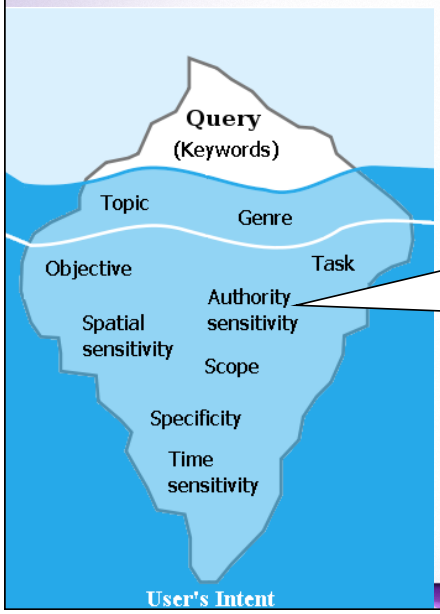


Does the query contains **polysemic** words?

wood, dove

93

Dimensions of the User's Intent

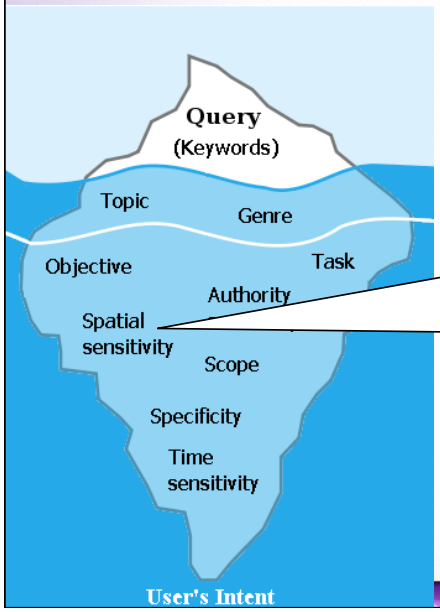


Is the query designed to retrieve **authoritative** and trusted answers?

Hospitals in Goteborg

94

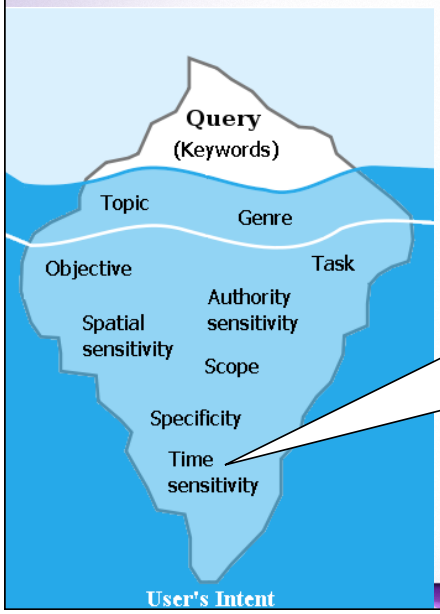
Dimensions of the User's Intent



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

95

Dimensions of the User's Intent



Could the results be influenced by **time**?

U.S. President vs. U.S. President 2010

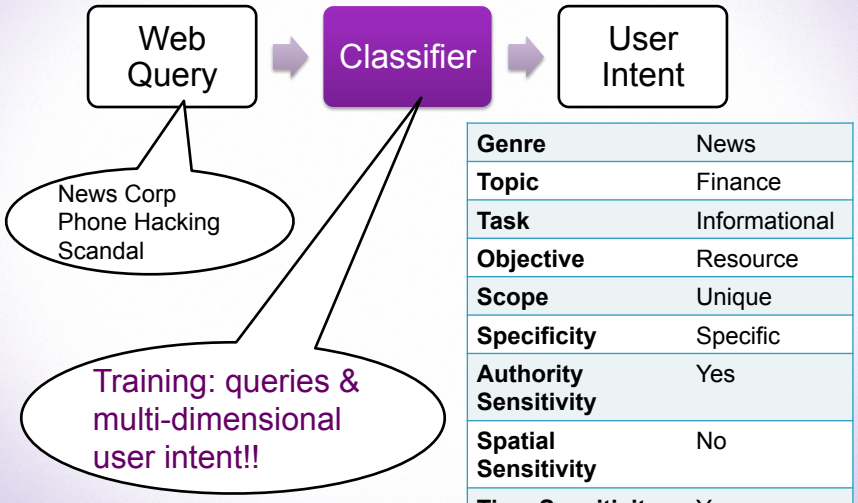
96

Dimensions of the User's Intent

Facet	Values
Genre: provides a generic context to the user's query intent. and can be thought as a meta-facet.	*News. Business. Reference. Community*
Topic: a list of topics built from the first level of categories offered by ODP . Yahoo! and Wikipedia.	*Adult. Arts & Culture. Beauty Style. Cars & Transportation. Charity. Computers & Internet. Education. Entertainment. Music & Games. Finance. Food & Drink. Health. Home & Garden. Industrial Goods & Services. Politics & Government. Religion & belief systems. Science & Mathematics. Social Science. Sports. Technology & Electronic. Travel. Undefined. Work*
Task: this facet is related with the type of resource associated with the query.	*Informational. Not Informational. Ambiguous*
Objective: represents if the query is aimed to do some action or to obtain a resource.	*Resource. Action*
Specificity: this facet describes how specialized is a query.	*Specific. Medium. Broad*
Scope: the scope aims at capturing whether the query contains polysemic words or not.	*Yes. No*
Authority Sensitivity: through this facet it is possible to determining whether the query is designed to retrieve authoritative and trusted answers	*Yes. No*
Spatial Sensitivity: this facet reflects the interest of the user to get a resource related to an explicit spatial location.	*Yes. No*
Time Sensitivity: this facet captures the fact that some queries require different results when posed at different times	*Yes. No*

97

The Problem



Genre	News
Topic	Finance
Task	Informational
Objective	Resource
Scope	Unique
Specificity	Specific
Authority Sensitivity	Yes
Spatial Sensitivity	No
Time Sensitivity	Yes

98

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

99

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.

Method	Size	Accuracy
Topic/homepage [17]‡	200	91%
FastQ (*Task*)	5249	76%
Rule-based hierarchy [15]	1500000	74%
Query-log based [19]	50	54%
Manual classification [2]	6042	50%

60%

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)

101

Conclusions

YAHOO!

Plenty of Open Problems

- Algorithm scalability
- Dependency on training data
- Privacy Issues

- More on query intent detection
  - Quality
  - Online

[Baeza-Yates & Maarek, SSDBM 2012]

103