UNDERSTANDING AND
IMPROVING WEB SEARCH USING
LARGE-SCALE BEHAVIORAL LOGS

Susan Dumais, Microsoft Research
Overview

- The big data revolution
  - ... examples from Web search
- Large-scale behavioral logs
  - Observations: Understand behavior
  - Experiments: Improve a system or service
- Limitations of logs
- Challenges
20 Years Ago … (Not Such Big) Data

- In popular media …
  - Mt St Helen’s eruption, *Friends* debut, OJ trial

- In web and search …
  - Mosaic one year old (pre Netscape, IE,
  - Size of the web
    - # web sites:
  - Size of Lycos search engine
    - # web pages in index:
  - Behavioral logs
    - # queries/day:
    - Most logging client-side
Today ... Big Data

- One trillion web sites
- Trillions of pages indexed by search engines
- Billions of posts and likes per day
- Billions of web searches and clicks per day
- Behavioral logs increasingly prevalent and changing our “ways of knowing”
What Are Behavioral Logs?

- Traces of human behavior
  - ... seen through the lenses of whatever sensors we have
  - Web search: queries, results, clicks, dwell time, etc.

- Actual, real-world (in situ) behavior
  - Not ...
    - Recalled behavior
    - Subjective impressions of behavior
    - Controlled experimental task
Kinds of Behavioral Data

- **Lab Studies**
  - 10-100s of people (and tasks)
  - Known tasks, carefully controlled
  - Detailed information: video, gaze, think-aloud
  - Can evaluate experimental systems

- **Field Studies**
  - 100-1000s of people (and tasks)
  - In-the-wild
  - Special instrumentation
  - Can probe about specific tasks, successes/failures

- **Log Studies**
  - Millions of people (and tasks)
  - In-the-wild
  - Diversity and dynamics
  - Abundance of data, but it’s noisy and unlabeled (what vs. why)
# Kinds of Behavioral Data

<table>
<thead>
<tr>
<th>Lab Studies</th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled tasks, in laboratory, with detailed instrumentation</td>
<td>In-lab behavior observations</td>
<td>In-lab controlled tasks, comparisons of systems</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field Studies</th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the wild, real-world tasks, ability to probe for detail</td>
<td>Ethnography, case studies, panels (e.g., Nielsen)</td>
<td>Clinical trials and field tests</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Log Studies</th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the wild, no explicit feedback but lots of implicit feedback</td>
<td>Logs from a single system</td>
<td>A/B testing of alternative systems or algorithms</td>
</tr>
</tbody>
</table>

**Goal:** Build an abstract picture of behavior  
**Goal:** Decide if one approach is better than another
Benefits of Behavioral Logs

- Real-world
  - Portrait of real behavior, warts and all

- Large-scale
  - Millions of people and tasks
  - Rare behaviors are common
  - Small differences can be measured
  - Tremendous diversity of behaviors and information needs (the “long tail”)

- Real-time
  - Feedback is immediate
How do you go from 2.4 words to anything sensible?

- **Content**
  - Match (query, page content)

- **Link structure**
  - Used to set non-uniform priors on pages

- **User behavior**
  - Anchor text
  - Query-click data

- **Contextual metadata**
  - Who, what, where, when, …

Understanding what people want to do and whether they are successful

- Behavioral logs (and more)

Driven by ... behavioral log data
Surprises In (Early) Search Logs

- Early log analysis ...
  - Silverstein et al. 1999, Broder 2002
- Web search != library search
  - Queries are very short, 2.4 words
  - Lots of people search for sex
  - “Navigating” is common, 30-40%
    - Getting to web sites vs. finding out about things
  - “Re-finding” is common, 30-40%
  - Amazing diversity of information needs
Queries Not Equally Likely

- **Excite 1999 data**
  - ~2.5mil queries
  - Head: top 250 accounts for 10% of queries
  - Tail: ~950k occur exactly once

- **Zipf Distribution**

<table>
<thead>
<tr>
<th>Q Rank</th>
<th>Q Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Top 10 Q</th>
<th>Query Freq = 10</th>
<th>Query Freq = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>bahia AND brazil</td>
<td>‘coren, s’</td>
</tr>
<tr>
<td>yahoo</td>
<td>games</td>
<td>UNC neuroscience</td>
</tr>
<tr>
<td>chat</td>
<td>mp3</td>
<td>hormones in memory loss</td>
</tr>
<tr>
<td>horoscope</td>
<td>weather</td>
<td>electronic roladex memory</td>
</tr>
<tr>
<td>pokemon</td>
<td>ebay</td>
<td>email address for paul allen</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the seattle seahawks owner</td>
</tr>
</tbody>
</table>

Navigational queries, one-word queries
Multi-word queries, specific URLs
Complex queries, rare info needs, misspellings, URLs
Queries Vary Over Time (and Location)

- Periodicities
  - Daily
  - Weekly
  - Longer

- Trends

- Predicted events

- Surprising events

Q = flu

Q = IRS taxes

Q = pizza
<table>
<thead>
<tr>
<th>Query</th>
<th>Time</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>aps 2014</td>
<td>10:41 am 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>social science</td>
<td>10:44 am 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>computational social science</td>
<td>10:56 am 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>aps 2014</td>
<td>11:21 am 5/15/14</td>
<td>659327</td>
</tr>
<tr>
<td>hilton san francisco</td>
<td>11:59 am 5/15/14</td>
<td>659327</td>
</tr>
<tr>
<td>restaurants seattle</td>
<td>12:01 pm 5/15/14</td>
<td>318222</td>
</tr>
<tr>
<td>pikes market restaurants</td>
<td>12:17 pm 5/15/14</td>
<td>318222</td>
</tr>
<tr>
<td>stuart shulman</td>
<td>12:18 pm 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>daytrips in seattle, wa</td>
<td>1:30 pm 5/15/14</td>
<td>554320</td>
</tr>
<tr>
<td>aps 2014</td>
<td>1:30 pm 5/15/14</td>
<td>659327</td>
</tr>
<tr>
<td>aps 2014 program</td>
<td>2:32 pm 5/15/14</td>
<td>435451</td>
</tr>
<tr>
<td>aps 2014.org</td>
<td>2:42 pm 5/15/14</td>
<td>435451</td>
</tr>
<tr>
<td>computational social science</td>
<td>4:56 pm 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>aps 2014</td>
<td>5:02 pm 5/15/14</td>
<td>312055</td>
</tr>
<tr>
<td>xxx clubs in seattle</td>
<td>10:14 pm 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>sex videos</td>
<td>1:49 am 5/16/14</td>
<td>142039</td>
</tr>
<tr>
<td>Query</td>
<td>Time</td>
<td>User</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>aps 2014</strong></td>
<td>10:41 am</td>
<td>142039</td>
</tr>
<tr>
<td>social science</td>
<td>10:44 am</td>
<td>142039</td>
</tr>
<tr>
<td>computational social science</td>
<td>10:56 am</td>
<td>142039</td>
</tr>
<tr>
<td><strong>aps 2014</strong></td>
<td>11:21 am</td>
<td>659327</td>
</tr>
<tr>
<td><strong>hilton san francisco</strong></td>
<td>11:59 am</td>
<td>659327</td>
</tr>
<tr>
<td>restaurants seattle</td>
<td>12:01 pm</td>
<td>318222</td>
</tr>
<tr>
<td>pikes market restaurants</td>
<td>12:17 pm</td>
<td>318222</td>
</tr>
<tr>
<td>stuart shulman</td>
<td>12:18 pm</td>
<td>142039</td>
</tr>
<tr>
<td>daytrips in seattle, wa</td>
<td>1:30 pm</td>
<td>554320</td>
</tr>
<tr>
<td><strong>aps 2014</strong></td>
<td>1:30 pm</td>
<td>659327</td>
</tr>
<tr>
<td>aps program</td>
<td>2:32 pm</td>
<td>435451</td>
</tr>
<tr>
<td><strong>aps 2014.org</strong></td>
<td>2:42 pm</td>
<td>435451</td>
</tr>
<tr>
<td>computational social science</td>
<td>4:56 pm</td>
<td>142039</td>
</tr>
<tr>
<td><strong>aps 2014</strong></td>
<td>5:02 pm</td>
<td>312055</td>
</tr>
<tr>
<td>xxx clubs in seattle</td>
<td>10:14 pm</td>
<td>142039</td>
</tr>
<tr>
<td>sex videos</td>
<td>1:49 am</td>
<td>142039</td>
</tr>
</tbody>
</table>

*Query typology*  
E.g., “navigational queries”
<table>
<thead>
<tr>
<th>Query</th>
<th>Time</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>aps 2014</td>
<td>10:41 am</td>
<td>142039</td>
</tr>
<tr>
<td>social science</td>
<td>10:44 am</td>
<td>142039</td>
</tr>
<tr>
<td>computational social science</td>
<td>10:56 am</td>
<td>142039</td>
</tr>
<tr>
<td>aps 2014</td>
<td>11:21 am</td>
<td>659327</td>
</tr>
<tr>
<td>hilton san francisco</td>
<td>11:59 am</td>
<td>659327</td>
</tr>
<tr>
<td>restaurants seattle</td>
<td>12:01 pm</td>
<td>318222</td>
</tr>
<tr>
<td>pikes market restaurants</td>
<td>12:17 pm</td>
<td>318222</td>
</tr>
<tr>
<td>stuart shulman</td>
<td>12:18 pm</td>
<td>142039</td>
</tr>
<tr>
<td>daytrips in seattle, wa</td>
<td>1:30 pm</td>
<td>554320</td>
</tr>
<tr>
<td>aps 2014</td>
<td>1:30 pm</td>
<td>659327</td>
</tr>
<tr>
<td>aps program</td>
<td>2:32 pm</td>
<td>435451</td>
</tr>
<tr>
<td>aps 2014.org</td>
<td>2:42 pm</td>
<td>435451</td>
</tr>
<tr>
<td>computational social science</td>
<td>4:56 pm</td>
<td>142039</td>
</tr>
<tr>
<td>aps 2014</td>
<td>5:02 pm</td>
<td>312055</td>
</tr>
<tr>
<td>xxx clubs in seattle</td>
<td>10:14 pm</td>
<td>142039</td>
</tr>
<tr>
<td>sex videos</td>
<td>1:49 am</td>
<td>142039</td>
</tr>
</tbody>
</table>

Query typology
E.g., “navigational queries”

Query behavior
E.g. “repeat Q”
<table>
<thead>
<tr>
<th>Query</th>
<th>Time</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>aps 2011</td>
<td>10:41 am 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>social science</td>
<td>10:44 am 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>computational social science</td>
<td>10:56 am 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>aps 2011</td>
<td>11:21 am 5/15/14</td>
<td>659327</td>
</tr>
<tr>
<td>hilton san francisco</td>
<td>11:59 am 5/15/14</td>
<td>659327</td>
</tr>
<tr>
<td>restaurants seattle</td>
<td>12:01 pm 5/15/14</td>
<td>318222</td>
</tr>
<tr>
<td>pikes market restaurants</td>
<td>12:17 pm 5/15/14</td>
<td>318222</td>
</tr>
<tr>
<td>stuart shulman</td>
<td>12:18 pm 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>daytrips in seattle, wa</td>
<td>1:30 pm 5/15/14</td>
<td>554320</td>
</tr>
<tr>
<td>aps 2011</td>
<td>1:30 pm 5/15/14</td>
<td>659327</td>
</tr>
<tr>
<td>aps program</td>
<td>2:32 pm 5/15/14</td>
<td>435451</td>
</tr>
<tr>
<td>aps 2011.org</td>
<td>2:42 pm 5/15/14</td>
<td>435451</td>
</tr>
<tr>
<td>computational social science</td>
<td>4:56 pm 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>jitp 2011</td>
<td>5:02 pm 5/15/14</td>
<td>312055</td>
</tr>
<tr>
<td>xxx clubs in seattle</td>
<td>10:14 pm 5/15/14</td>
<td>142039</td>
</tr>
<tr>
<td>sex videos</td>
<td>1:49 am 5/16/14</td>
<td>142039</td>
</tr>
</tbody>
</table>

**Query typology**
E.g., “navigational queries”

**Query behavior**
E.g. “common Q”

**Long-term trends**
E.g. “repeat Q or topic”
What Observational Logs Can Tell Us

- Summary measures
  - Query frequency
  - Query length
- Analysis of query intent
  - Query types and topics
- Temporal patterns
  - Session length
  - Common re-formulations
- Click behavior
  - Relevant results for query
  - Queries that lead to clicks

- Queries appear 3.97 times
  [Silverstein et al. 1999]
- Queries 2.35 terms
  [Jansen et al. 1998]
- Informational, Navigational, Transactional
  [Broder 2002]
- Sessions 2.20 queries long
  [Silverstein et al. 1999]
- [Lau and Horvitz, 1999]

- [Joachims 2002]
Uses of Observational Logs

- Provide insights about how people interact with existing systems and services
- Make it possible to design systems to support actual (rather than presumed) activities
- Enable design of more detailed experiments to focus on things that matter
- Support new user experiences
Observations provide insights about behavior with existing systems

**Experiments** are the life blood of web services

- Controlled experiments to compare system variants
- Used to study all aspects of search systems
  - System latency
  - Fonts, layout
  - Snippet generation techniques
  - Ranking algorithms
- Data-driven design
Experiments At Web Scale

- Basic questions
  - What do you want to evaluate?
  - What metrics do you care about?
- Within- vs. between-”subject” design
  - Between: More widely used, conditions can run concurrently
  - Within: Temporal-split vs. Interleaving
- Controls, Counterfactuals, Power are important
- Some things easier to study than others
  - Algorithmic changes easy
  - Interface changes harder
  - Social systems even harder

Kohavi et al., 1999
Dumais et al., 2014
Examples from Contextual Search

- Personal navigation
  - Simple repeat behavior
- Adaptive ranking
  - Rich user model with varied features and temporal extent
- Temporal dynamics
One Size Does Not Fit All

- Queries are difficult to interpret in isolation
  - Easier if we can model: who is asking, where they are, what they have done in the past, etc.

  **Searcher:** (SIGIR | Susan Dumais ... an information retrieval researcher) vs. (SIGIR | Stuart Bowen Jr. ... the Special Inspector General for Iraq Reconstruction)

  **Previous actions:** (SIGIR | information retrieval) vs. (SIGIR | U.S. coalitional provisional authority)

  **Location:** (SIGIR | at SIGIR conference) vs. (SIGIR | in Washington DC)

  **Time:** (SIGIR | Aug conference) vs. (SIGIR | Iraq news)

- Using a **single ranking for everyone, in every context, at every point in time** limits how well a search engine can do
Example 1: Personal Navigation

- Re-finding common in web search
  - 33% of queries are repeat queries
  - 39% of clicks are repeat clicks

- Many are navigational queries
  - E.g., nytimes-> www.nytimes.com

- “Personal” navigational queries
  - Different intents across individuals, but consistently same intent for an individual
    - E.g., SIGIR (for Dumais) -> www.sigir.org
    - E.g., SIGIR (for Bowen Jr.) -> www.sigir.mil
  - Very high prediction accuracy (~95%)
  - High coverage (~15% of queries)

<table>
<thead>
<tr>
<th></th>
<th>Repeat Click</th>
<th>New Click</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeat Query</td>
<td>33%</td>
<td>29%</td>
</tr>
<tr>
<td>New Query</td>
<td>67%</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>39%</td>
<td>61%</td>
</tr>
</tbody>
</table>
Example 2: Adaptive Ranking

- **Short-term context**
  - Previous actions (queries, clicks) within current session
    - (Q = Rich Shiffrin | psychology vs. lawyer)
    - (Q = APS | psychology vs. physics vs. public utility vs. public schools)
    - (Q = ACL | computational linguistics vs. knee injury vs. country music)

- **Long-term preferences and interests**
  - Behavior: Specific queries/URLs
    - (Q=weather) -> weather.com vs. weather.gov vs. intellicast.com
  - Content: Language models, topic models, etc.

- Unified model for both
Adaptive Ranking (cont’d)

- User model (content)
  - Specific queries/URLs
  - Topic distributions, using ODP

- Log-based evaluation, MAP

- Which sources are important?
  - Session (short-term): +25%
  - Historic (long-term): +45%
  - Combinations: +65-75%

- What happens within a session?
  - 60% of sessions involve multiple queries
    - By 3rd query in session, short-term features more important than long-term
    - First queries in session are different – shorter, higher click entropy

- User model (temporal extent)
  - Session, Historical, Combinations
  - Temporal weighting
Example 3: Temporal Dynamics

- Queries are not uniformly distributed over time
  - Often triggered by events in the world
- What’s relevant changes over time
  - E.g., US Open … in 2014 vs. in 2013
  - E.g., US Open 2014 … in June (golf) vs. in Sept (tennis)
  - E.g., US Golf Open 2014 …
    - Before event: Schedules and tickets, e.g., stubhub
    - During event: Real-time scores or broadcast, e.g., espn, cbssports
    - After event: General sites, e.g., wikipedia, usta
Temporal Dynamics (cont’d)

- Develop time-aware retrieval models

- Leverage **content** change on a page
  - Pages have different *rates of change* (influences document priors, $P(D)$)
  - Terms have different *longevity* on a page (influences term weights, $P(Q|D)$)
  - 15% improvement vs. LM baseline

- Leverage time-series modeling of **user interactions**
  - Model Query and URL clicks as time-series
  - Enables appropriate weighting of historical interactions
  - Useful for queries with local or global trends
Uses of Behavioral Logs

- Characterize information seeking behavior
- Enable practical improvements of search engines
  - Offline observations
    - E.g., Re-finding is common, Long tail of info needs
  - Behavioral features used in algorithms or interface
    - E.g., Previously clicked results boosted, query suggestion
  - Online experiments
    - E.g., Compare two algorithms or interfaces
- Change how systems are evaluated and improved
What Logs (Alone) Cannot Tell Us

- Lots about “what” people are doing, less about “why”
- Limited annotations
  - People’s intent
  - People’s success
  - People’s experience
  - People’s attention
- Behavior can mean many things
- Limited to existing systems and interactions
- Complement with other techniques to provide a more complete picture (e.g., lab, field studies)
Summary

- Large-scale behavioral logs
  - Provide traces of human behavior in situ at a scale and fidelity previously unimaginable
  - Observations and experiments enable us to characterize behavior and improve web search
  - Revolutionized how web-based systems are designed and evaluated

- Complementary methods important to develop more complete understanding
Thank you!

More info at:
http://research.microsoft.com/~sdumais