PUTTING THE SEARCHERS BACK INTO SEARCH
Overview

- The changing IR landscape
- Search increasingly pervasive and important
  - Characterized by diversity of tasks, searchers and interactivity
- Methods for understanding searchers
  - Lab, panels, large-scale logs
  - Examples from Web and desktop search, and contextualized search
- New trends and opportunities
20 Years Ago ...

- **Web in 1994:**
  - **Size of the web**
    - # web sites: 2.7k (13.5% .com)
  - **Mosaic** 1 year old (pre Netscape, IE, Chrome)

- **Search in 1994:**
  - 17th SIGIR
  - TREC 2.5 years old
  - **Size of Lycos search engine**
    - # web pages in index: 54k
    - This was about to change rapidly
  - **Behavioral logs**
    - # queries/day: 1.5k
Today ... Search is Everywhere

- Trillions of pages discovered by search engines
- Billions of web searches and clicks per day
- Search a core fabric of people’s everyday lives
  - Diversity of tasks, searchers, and interactivity
  - Pervasive (desktop, enterprise, web, apps, etc.)
- We should be proud, but …
- Understanding and supporting searchers more important now than ever before
  - Requires both great results and experiences
Where are the Searchers in Search?
Search in Context

Searcher Context

Query

Task Context

Ranked List

Document Context
Evaluating Search Systems

- Cranfield/TREC-style test collections
  - Fixed: Queries, Documents, Relevance Judgments, Metrics
  - Goal: Compare systems, w/ respect to metric(s)

- What’s missing?
  - Characterization of queries/tasks
    - How selected? What can we generalize to?
  - Searcher-centered metrics
    - Implicit models in: AvgPr vs. Pr@10 vs. DCG or RBP vs. time
  - Rich models of searchers
    - Current context, history of previous interactions, preferences, expertise
  - Presentation/Interaction
    - Snippets, composition of the whole page, search support (spelling correction, query suggestions), speed of system, etc.

[Voorhees, HCIR 2009] A test collection is (purposely) a stark abstraction of real user search tasks that models only a few of the variables that affect search behavior and was explicitly designed to minimize individual searcher effects. … this ruthless abstraction of the user …
Filling the Gaps in Evaluation

- Methods for understanding and modeling searchers
  - Experimental lab studies
  - Observational log analysis
  - … and many more

- What can learn from each?

- How can we use these insights to improve search systems and evaluation paradigms?

- How can we bridge the gap between “offline” and “online” experiments?
Kinds of Behavioral Data

**Lab Studies**
In lab, controlled tasks, with detailed instrumentation and interaction

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

Dumais et al., 2014
Kinds of Behavioral Data

Lab Studies
In lab, controlled tasks, with detailed instrumentation and interaction

Panel Studies
In the wild, real-world tasks, ability to probe for detail

- 100-1000s of people (and tasks)
- In-the-wild
- Special client instrumentation
- Can probe about specific tasks, successes/failures
Kinds of Behavioral Data

**Lab Studies**
In lab, controlled tasks, with detailed instrumentation and interaction

**Panel Studies**
In the wild, real-world tasks, ability to probe for detail

**Log Studies**
In the wild, no explicit feedback but lots of implicit feedback

- Millions of people (& 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></th>
<th>Observational</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lab Studies</strong></td>
<td>Controlled tasks, in laboratory, with detailed instrumentation</td>
<td>In-lab behavior observations</td>
</tr>
<tr>
<td><strong>Panel Studies</strong></td>
<td>In the wild, real-world tasks, ability to probe for detail</td>
<td>Ethnography, case studies, panels (e.g., Nielsen)</td>
</tr>
<tr>
<td><strong>Log Studies</strong></td>
<td>In the wild, no explicit feedback but lots of implicit feedback</td>
<td>Logs from a single system</td>
</tr>
<tr>
<td></td>
<td></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
What Are Behavioral Logs?

- Traces of human behavior
  - ... seen through the lenses of whatever sensors we have
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
Benefits of Behavioral Logs

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

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

- Real-time
  - Feedback is immediate
Early log analysis …
- Excite logs 1997, 1999
- Silverstein et al. 1998, 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
- Queries are not independent, e.g., tasks
- Amazing diversity of information needs (long tail)
Queries Not Equally Likely

- **Excite 1999 data**
  - ~2.5 mil queries <time, user id, query>
  - Head: top 250 account for 10% of queries
  - Tail: ~950k occur exactly once

- **Zipf Distribution**

**Top 10 Q**
- sex
- yahoo
- chat
- horoscope
- pokemon
- hotmail
- games
- mp3
- weather
- ebay

Navigational queries, one-word queries

**Query Freq = 10**
- foosball AND Harvard
- sony playstation cheat codes
- breakfast or brunch menus
- australia gift baskets
- colleges with majors of web page design

Multi-word queries, specific URLs

**Query Freq = 1**
- acm98
- winsock 1.1 w2k compliant
- Coolangatta, Gold Coast newspaper
- email address for paul allen the seattle seahawks owner

Complex queries, rare info needs, misspellings, URLs
Queries Vary Over Time and Task

- **Time**
  - Periodicities
  - Trends
  - Events

- **Tasks/Individuals**
  - Sessions
  - Longer history

(Q = **SIGIR** | information retrieval vs. Iraq reconstruction)

(Q = **SIGIR** | Susan vs. Stuart)

Q = **tesla**

Q = **world cup**
What Observational Logs Can Tell Us

- **Summary measures**
  - Query frequency
  - Query length

- **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]

<table>
<thead>
<tr>
<th>retrieval function</th>
<th>bxx</th>
<th>tfc</th>
<th>hand-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. clickrank</td>
<td>6.26±1.14</td>
<td>6.18±1.33</td>
<td>6.04±0.92</td>
</tr>
</tbody>
</table>

[Joachims 2002]
From Observations to Experiments

- Observations provide insights about interaction with existing systems
- **Experiments** are the life blood of web systems
  - Controlled experiments to compare system variants
  - Used to study all aspects of search systems
    - Ranking algorithms
    - Snippet generation
    - Spelling and query suggestions
    - Fonts, layout
    - System latency
- Guide where to invest resources to improve search
Experiments At Web Scale

- Basic questions
  - What do you want to evaluate?
  - What metric(s) do you care about?
- Within- vs. between-subject designs
  - Within: Interleaving (for ranking changes); otherwise add temporal-split between experimental and control conditions
  - Between: More widely useful, but higher variance
- Some things easier to study than others
  - Algorithmic vs. Interface vs. Social Systems
- Counterfactuals, Power, and Ramping-Up important

Kohavi et al., DMKD 2009
Dumais et al., 2014
Uses of Behavioral Logs

- Provide (often surprising) insights about how people interact with search systems
  - Focus efforts on supporting actual (vs. presumed) activities
    - E.g., Diversity of tasks, searchers, contexts of use, etc.
  - Suggest experiments about important or unexpected behaviors
  - Provide input for predictive models and simulations
- Improve system performance
  - Caching, Ranking features, etc.
- Support new search experiences
- Changes how systems are evaluated and improved
How do you go from 2.4 words to great results?

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

- **Link structure**
  - Non-uniform priors on pages

- **Author/searcher behavior**
  - Anchor text
  - Query-click data
  - Query reformulations

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

*Powered by ... behavioral insights*
What Logs (Alone) Cannot Tell Us

- 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

- Lots about “what” people are doing, less about “why”

- Complement with other techniques to provide a more complete picture (e.g., lab, panel studies, modeling)
Understanding Searchers

- Using complementary methods to better understand and model searchers

- Examples from …
  - New domains
    - Web search vs. Library search
    - Desktop search vs. Web search
  - Contextual search
    - Personalization
    - Tasks/sessions
    - Temporal dynamics
Web Search != Library Search

- Traditional notions of “information needs” did not adequately describe web searcher behavior
- Alta Vista studies
  - Analysis of AV logs
    - yahoo
    - ebay
    - Hotmail
    - Yahoo.com
    - aol
    - maps
    - weather
    - Gold Coast
    - Pearl Jam lyrics
  - download free wallpaper
  - quicktime download
  - buy CD online
  - How can Jeeves help me shop for books?

- Pop up survey on AV, Jun-Nov 2001

2. Which of the following describes best what you are trying to do?
   - I want to get to a specific website that I already have in mind
   - I want a good site on this topic, but I don’t have a specific site in mind

3. Which of the following best describes why you conducted this search?
   - I am shopping for something to buy on the Internet
   - I am shopping for something to buy elsewhere than on the Internet
   - I want to download a file (e.g., music, images, programs, etc.)
   - None of these reasons

4. Which of the following describes best what you are looking for?
   - A site which is a collection of links to other sites regarding this topic
   - The best site regarding this topic
Web Search != Library Search

- Traditional notions of “information needs” did not adequately describe web searcher behavior
- Alta Vista studies
  - Analysis of AV logs
  - Pop up survey on AV, Jun-Nov 2001
- Three general types of search intents
  - Informational (find information about a topic)
  - Navigational (find a single known web page)
  - Transactional (find a site where web-mediated activities can be performed, e.g., download game, find map, shop)
Desktop Search != Web Search

- Desktop search, circa 2000
  - Easier to find things on the web than on your desktop

- Fast, flexible search over “Stuff I’ve Seen”
  - Heterogeneous info: files, email, calendar, web, IM
  - Index: full-content plus metadata
  - Interface: highly interactive rich list-view
    - Sorting, filtering, scrolling
    - Rich actions on results (open folder, drag)
    - Support re-finding vs. finding
**Stuff I’ve Seen: Example searches**

**Looking for:** recent email from Fedor that contained a link to his new demo  
**Initiated from:** Start menu  
**Query:** from:Fedor

**Looking for:** the pdf of a SIGIR paper on context and ranking (not sure it used those words) that someone (don’t remember who) sent me a month ago  
**Initiated from:** Outlook  
**Query:** SIGIR

**Looking for:** meeting invite for the last intern handoff  
**Initiated from:** Start menu  
**Query:** intern handoff kind:appointment

**Looking for:** C# program I wrote a long time ago  
**Initiated from:** Explorer pane  
**Query:** QCluster*.*
Stuff I’ve Seen: Evaluation

- Surveys and structured interviews
- Developed and deployed the system, and iterated
  - Log data [queries, interactions, time]
  - Questionnaire and interviews [pre- and post-]
  - Experiment [6 alternative systems]

Sort By Date vs. Rank
Top vs. Side
Preview vs. Not
Stuff I’ve Seen: Results

- **Queries**
  - Very short (1.6 words); People important (25%)

- **Opened items**
  - Type: Email (76%), Web pages (14%), Files (10%)
  - Age: Today (5%), Last week (21%), Last month (47%)

- **Interface expts: large effect of Date vs. Rank**
  - **Date** by far the most common sort order
  - Few searches for “best” matching object
  - Many other criteria – e.g., time, people

- **Abstractions important**
  - E.g., “image”, “people”, “useful date”
People remember many attributes in re-finding
- Seldom: only general overall topic
- Often: time, people, file type, etc.
- Different attributes for different tasks

Rich client-side interface
- Support fast iteration and refinement
- Fast filter-sort-scroll vs. next-next-next
- “Fluidity of interactions”

Desktop search \(!=\) Web search
Context: 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, when it is, etc.

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

*SIGIR*
Context: One Size Does Not Fit All

- Queries are difficult to interpret in isolation

![Bing](https://via.placeholder.com/150)

- Easier if we can model: who is asking, where they are, what they have done in the past, when it is, 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 | July 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
Potential for Personalization

- Framework to quantify the variation relevance for the same query across individuals
  - Measured individual relevance with explicit & implicit

- Personalized search study with explicit judgments
  - 46% potential increase in search quality with core ranking
  - 70% potential increase with personalization
Potential for Personalization (cont’d)

- Framework to quantify the variation relevance for the same query across individuals
  - Measured individual relevance w/ explicit & implicit
  - Personalized search study with explicit judgments
    - 46% potential increase in search quality with core ranking
    - 70% potential increase with personalization

- Construct individual models considering different
  - Sources of evidence: Content, behavior
  - Time frames: Short-term, long-term
  - Who: Individual, group

Personalized Nav
Adaptive Ranking
Re-finding common in web search
- 33% of queries are repeat queries
- 39% of clicks are repeat clicks

Teevan et al., SIGIR 2007
Tyler & Teevan, WSDM 2010
**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., sigir 2014 -> sigir.org/sigir2014

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

- **Online A/B experiments**

---

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

- Queries do not occur in isolation
  - 60% of sessions contain multiple queries
  - 50% of search time spent in sessions of 30+ mins
  - 15% of tasks continue across sessions or devices

- Unified model to represent

- Short-term session context
  - Previous actions (queries, clicks) within current session
    - (Q = SIGIR | information retrieval vs. Iraq reconstruction)
    - (Q = ACL | computational linguistics vs. knee injury vs. country music)

- Long-term preferences and interests
  - Behavior: Specific queries, URLs, sites
  - Content: Language models, topic models, etc.
Adaptive Ranking (cont’d)

- **Searcher model (content)**
  - Specific queries, URLs
  - Topic distributions, using ODP

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

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

- **Searcher model (time)**
  - Session, Historical, Combinations
  - Temporal weighting

---

Diagram: Adaptive Ranking

- Historic
- Session
- Aggregate

Bar charts showing MAP gain over query position in session: Session, Historic, Aggregate, Union.
Building Predictive Models

- Collect searcher behavior
  - From lab, panel, or log studies
- Identify variables of interest
  - E.g., doc relevance, session success, task continuation
- Collect some labeled data
  - From searcher (ideal), or annotator
- Learn models to predict variables of interest
  - Curious Browser [doc relevance, session success]
  - Cross-session/device continuation [task continuation]
- Evaluate, validate and generalize
Summary of Examples

- Complementary methods (from lab studies, to panels, to large-scale behavioral logs) can be used to understand and model searchers.

- Especially important in new search domains, and in accommodating the variability that we see across individuals and tasks.
Looking Forward: What’s Next?

- Importance of spatio-temporal contexts
- Richer representations and dialogs
  - E.g., knowledge graphs, Siri, Cortana
- More proactive search, especially in mobile
- Tighter coupling of digital and physical worlds
- Computational platforms that seamlessly couple human and algorithmic components
  - E.g., IM-an-Expert, Tail Answers, VizWiz
- Richer task support
Summary

- Search is an increasingly important part of people’s everyday lives
  - Traditional test collections are very limited, especially with respect to modeling searchers
  - Need to extend evaluation methods to handle the diversity of searchers, tasks, and interactivity that characterize search

- To understand and support searchers requires varied behavioral insights, and a broad inter-disciplinary perspective

- If search doesn’t work for people, it doesn’t work. Let’s make sure that it does !!!
Thank you!

More info at:

http://research.microsoft.com/~sdumais
Voorhees, I come not to bury Cranfield, but to praise it. *HCIR 2009*

Dumais et al., Understanding user behavior through log and data analysis. *Ways of Knowing 2014*

Kohavi et al., Controlled experiments on the Web: Survey and practical guide *DMKD 2009*

Broder, A taxonomy of Web search. *SIGIR Forum 2002*

Rose & Levinson, Understanding user goals in Web search. *WWW 2004*

Dumais et al., Stuff I’ve Seen: A system for personal information retrieval and re-use. *SIGIR 2003*

Teevan et al., Potential for personalization. *ToCHI 2010*

Teevan et al., Information re-retrieval: Repeat queries in Yahoo’s logs. *SIGIR 2007*

Tyler & Teevan, Large scale query log analysis of re-finding. *WSDM 2010*

Bennett et al., Modeling the impact of short- and long-term behavior on search personalization. *SIGIR 2012*

Elsas & Dumais, Leveraging temporal dynamics of document content in relevance ranking, *WSDM 2010*

Radinski et al., Behavioral dynamics on the Web: Learning modeling and predicting. *TOIS 2013*

Fox et al., Evaluating implicit measures to improve the search experience. *TOIS 2005*