PERSONALIZED SEARCH: POTENTIAL AND PITFALLS
Overview

- Context in search
- “Potential for personalization” framework

Examples
- Personal navigation
- Client-side personalization
- Short- and long-term models
- Personal crowds

Challenges and new directions
20 Years Ago … In Web Search

- NCSA Mosaic graphical browser 3 years old, and
web search engines 2 years old
20 Years Ago … In Web Search

- NCSA Mosaic graphical browser 3 years old, and web search engines 2 years old
- Online presence ~1996
20 Years Ago … In Web Search

- NCSA Mosaic graphical browser 3 years old, and web search engines 2 years old
  - Online presence ~1996
- Size of the web
  - # web sites: 2.7k
- Size of Lycos search engine
  - # web pages in index: 54k
- Behavioral logs
  - # queries/day: 1.5k
  - Most search and logging client-side
Today ... Search is Everywhere

- A billion web sites
- Trillions of pages indexed by search engines
- Billions of web searches and clicks per day
- Search is a core fabric of everyday life
  - Diversity of tasks and searchers
  - Pervasive (web, desktop, enterprise, apps, etc.)
- Understanding and supporting searchers more important now than ever before
Search in Context

Searcher Context

Query

Task Context

Ranked List

Document Context
Context Improves Query Understanding

- Queries are difficult to interpret in isolation

- Easier if we can model: who is asking, what they have done in the past, where they are, when it is, etc.

  **Searcher:** *(SIGIR | Susan Dumais ... an information retrieval researcher)*

  vs. *(SIGIR | Stuart Bowen Jr. ... the Special Inspector General for Iraq Reconstruction)*
Context Improves Query Understanding

- Queries are difficult to interpret in isolation

- Easier if we can model: who is asking, what they have done in the past, where they are, when it is, etc.

Searcher: (SIGIR | Susan Dumais ... an information retrieval researcher) vs. (SIGIR | Stuart Bowen Jr.)

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

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

Time: (SIGIR | Jan. submission) vs. (SIGIR | Aug. conference)

- 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

- A single ranking for everyone limits search quality
- Quantify the variation in relevance for the same query across different individuals

CIKM Oct 26, 2016
Potential For Personalization

- A single ranking for everyone limits search quality
- Quantify the variation in relevance for the same query across different individuals
- Different ways to measure individual relevance
  - Explicit judgments from different people for the same query
  - Implicit judgments from click entropy or content analysis
- Personalization can lead to large improvements
  - Study with explicit judgments
  - 46% improvements for core ranking
  - 70% improvements with personalization

CIKM Oct 26, 2016
Potential For Personalization

- Not all queries have high potential for personalization
  - E.g., facebook vs. sigir
  - E.g., * maps

- Learn when to personalize
Potential for Personalization

- Query: CIKM
- What is the “potential for personalization”?
- How can you tell different intents apart?
  - Contextual metadata
    - E.g., Location, Time, Device, etc.
  - Past behavior
    - Current session actions, Longer-term actions and preferences
User Models

- Constructing user models
  - Sources of evidence
    - Content: Queries, content of web pages, desktop index, etc.
    - Behavior: Visited web pages, explicit feedback, implicit feedback
    - Context: Location, time (of day/week/year), device, etc.
  - Time frames: Short-term, long-term
  - Who: Individual, group

- Using user models
  - Where resides: Client, server
  - How used: Ranking, query support, presentation, etc.
  - When used: Always, sometimes, context learned
Example 1: Personal Navigation

- Re-finding is common in Web search
  - 33% of queries are repeat queries
  - 39% of clicks are repeat clicks
- Many of these are navigational queries
  - E.g., facebook -> www.facebook.com
  - Consistent intent across individuals
  - Identified via low click entropy, anchor text

- “Personal navigational” queries
  - Different intents across individuals ... but consistently the same intent for an individual
    - SIGIR (for Dumais) -> www.sigir.org/sigir2017
    - SIGIR (for Bowen Jr.) -> www.sigir.mil
Personal Navigation Details

- Large-scale log analysis (offline)
  - Identifying personal navigation queries
    - Use consistency of queries & clicks within an individual
    - Specifically, the last two times a person issued the query, did they have a unique click on same result?
  - Coverage and prediction
    - Many such queries: ~12% of queries
    - Prediction accuracy high: ~95% accuracy
    - High coverage, low risk personalization

- A/B in situ evaluation (online)
  - Confirmed benefits
Example 2: PSearch

- Rich client-side model of a person’s interests
  - Model: Content from desktop search index & Interaction history
    - Rich and constantly evolving user model
  - Client-side re-ranking of web search results using model
  - Good privacy (only the query is sent to server)
    - But, limited portability, and use of community
PSearch Details

- Personalized ranking model
  - Score: Global web score + personal score
  - Personal score: Content match + interaction history features

- Evaluation
  - Offline evaluation, using explicit judgments
  - Online (in situ) A/B evaluation, using PSearch prototype
    - Internal deployment, 225+ people several months
    - 28% higher clicks, for personalized results
    - 74% higher, when personal evidence is strong
    - Learned model for when to personalize
Example 3: Short + Long

- Long-term preferences and interests
  - Behavior: Specific queries/URLs
  - Content: Language models, topic models, etc.

- Short-term context or task
  - 60% of search session have multiple queries
  - Actions within current session (Q, click, topic)
    - (Q=sigir | information retrieval vs. iraq reconstruction)
    - (Q=cikm | knowledge mgt vs. twin star aircraft vs. discount tickets)
    - (Q=ego | id vs. eldorado gold corporation vs. dangerously in love)

- Personalized ranking model combines both
Short + Long Details

- User model (temporal extent)
  - Session, Historical, Combinations
  - Temporal weighting
- Large-scale log analysis
- Which sources are important?
  - Session (short-term): +25%
  - Historic (long-term): +45%
  - Combinations: +65-75%
- What happens within a session?
  - 1st query, can only use historical
  - By 3rd query, short-term features more important than long-term
Atypical Sessions

- Example user model
  - 55% Football ("nfl", "philadelphia eagles", "mark sanchez")
  - 14% Boxing ("espn boxing", "mickey garcia", "hbo boxing")
  - 9% Television ("modern familiy", "dexter 8", "tv guide")
  - 6% Travel ("rome hotels", "tripadvisor seattle", "rome pasta")
  - 5% Hockey ("elmira pioneers", "umass lax", "necbl")

- New Session 1:
  - Boxing ("soto vs ortiz hbo")
  - Boxing ("humberto soto")

- New Session 2:
  - Dentistry ("root canal")
  - Dentistry ("dental implant")
  - Healthcare ("dental implant recovery")

- ~6% of sessions are atypical
  - Common topics: Medical (49%), Computers (24%)
  - Tend to be more complex, and have poorer quality results
  - What you “need” to do vs. what you “choose” to do
Atypical Sessions Details

- Learn model to identify atypical sessions
  - Logistic regressions classifier
- Apply different personalization models for them
  - If typical, use long-term user model
  - If atypical, use short-term session user model
- Change in precision by typicality of session
Example 4: A Crowd of Your Own

- Personalized judgments from crowd workers
  - Taste “grokking”
    - Ask crowd workers to understand (“grok”) your interests
  - Taste “matching”
    - Find workers who are similar to you (like collaborative filtering)

- Useful for: personal collections, dynamic collections, or collections with many unique items

- Studied several subjective tasks
  - Item recommendation (purchasing, food)
  - Text summarization, Handwriting recognition

Organisciak et al., HCOMP 2015, IJCAI 2015
A Crowd of Your Own

- “Personalized” judgments from crowd workers

Requester

Workers

...
A Crowd of Your Own Details

- **Grokking**
  - Requires fewer workers
  - Fun for workers
  - Hard to capture complex preferences

- **Matching**
  - Requires many workers to find a good match
  - Easy for workers
  - Data reusable

- Crowdsourcing promising in domains where lack of prior data limits established personalization methods

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Grok</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salt shakers</td>
<td>1.64</td>
<td>1.07 (34%)</td>
<td>1.43 (13%)</td>
</tr>
<tr>
<td>Food (Boston)</td>
<td>1.51</td>
<td>1.38 (9%)</td>
<td>1.19 (22%)</td>
</tr>
<tr>
<td>Food (Seattle)</td>
<td>1.58</td>
<td>1.28 (19%)</td>
<td>1.26 (20%)</td>
</tr>
</tbody>
</table>
Challenges in Personalization

- **User-centered**
  - Privacy
  - Serendipity and novelty
  - Transparency and control

- **Systems-centered**
  - Evaluation
    - Measurement, experimentation
  - System optimization
    - Storage, run-time, caching, etc.
Privacy

- Profile and content need to be in the same place
- Local profile (e.g., PSearch)
  - Private, only query sent to server
  - Device specific, inefficient, no community learning
- Cloud profile (e.g., Web search)
  - Need transparency and control over what’s stored
- Other approaches
  - Public or semi-public profiles (e.g., tweets, public FB status, blog posts, papers)
  - Light weight profiles (e.g., queries in a session)
  - Matching to a group cohort vs. an individual
Serendipity and Novelty

- Does personalization mean the end of serendipity?
  - ... Actually, it can improve it!

- Experiment on Relevance vs. Interestingness
  - Personalization finds more relevant results
  - Personalization also finds more interesting results
    - Even when interesting results were not relevant

- Need to be ready for serendipity
  - ... Like the Princes of Serendip
Evaluation

- External judges, e.g., assessors
  - Lack diversity of intents and realistic context
  - Crowdsourcing can help some

- Actual searchers are the “judges”
  - Offline
    - Labels from explicit judgments or implicit behavior (log analysis)
    - Allows safe exploration of many different alternatives
  - Online (A/B experiments)
    - Explicit judgments: Nice, but annoying and may change behavior
    - Implicit judgments: Scalable and natural, but can be very noisy

- Linking implicit actions and explicit judgments
Queries difficult to interpret in isolation
  - Augmenting query with context helps

Potential for improving search via personalization is large

Examples
  - PNav, PSearch, Short/Long, Crowd

Challenges
  - Privacy, transparency, serendipity
  - Evaluation, system optimization

Personalization/contextualization prevalent today, and increasingly so in mobile and proactive scenarios
Thanks!

Questions?

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

Collaborators:

- Eric Horvitz, Jaime Teevan, Paul Bennett, Ryen White, Kevyn Collins-Thompson, Peter Bailey, Eugene Agichtein, Sarah Tyler, Alex Kotov, Paul André, Carsten Eickhoff, Peter Organisciak
References

- **Short-term models**
  - Eickhoff et al., WSDM 2013. *Personalizing atypical search sessions.* *
  - André et al., CHI 2009. *From x-rays to silly putty via Uranus: Serendipity and its role in Web search.* *
  - Fox et al., TOIS 2005. *Evaluating implicit measures to improve web search.* *

- **Long-term models**
  - Teevan et al., SIGIR 2005. *Personalizing search via automated analysis of interests and activities.* *
  - Teevan et al., SIGIR 2008. *To personalize or not: Modeling queries with variations in user intent.* *
  - Teevan et al., TOCHI 2010. *Potential for personalization.* *
  - Teevan et al., WSDM 2011. *Understanding and predicting personal navigation.* *
  - Bennett et al., SIGIR 2012. *Modeling the impact of short- & long-term behavior on search personalization.* *

- **Personal crowds**
  - Eickhoff et al., ECIR 2013. *Designing human-readable user profiles for search evaluation.* *
  - Organisciak et al., HCOMP 2015. *A crowd of your own: Crowdsourcing for on-demand personalization.* *