PERSONALIZED SEARCH: POTENTIAL AND PITFALLS
Overview

- Context in information retrieval
- “Potential for personalization” framework
- Examples
  - Personal navigation
  - Client-side personalization
  - Short- and long-term models
- Challenges and new directions
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)
  
  **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

Potential for Personalization

Teevan et al., SIGIR 2008, ToCHI 2010
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 (search result clicks entropy, content analysis)
- Personalization can lead to large improvements
  - Study with explicit judgments
  - 46% improvements for core ranking
  - 70% improvements with personalization

HCII 2015
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: HCII
- What is the “potential for “personalization”?
- How can you tell different intents apart?
  - Past behavior
    - Current session actions, Longer-term actions and preferences
  - Contextual metadata
    - E.g., Location, Time, Device, etc.
20 Years Ago ... In Web and Search

- NCSA Mosaic graphical browser 2 years old, web search engines <1 year old
- CHI 1995 online presence

[Website Screenshot]

CHI '95 was the 1995 Conference on Human Factors in Computing Systems and was sponsored by ACM SIGCHI. CHI '95 was held in Denver, Colorado, USA (May 7 - 11, 1995). CHI '95 will take place in Vancouver, Canada (April 14-18, 1996).

Do you have complaints, comments, suggestions or praise for something about CHI '95? We want to hear from you. Send email to chi95-feedback@acm.org.

What's New?
- 95-12-06 The proceedings are back, re-designed but still not completely done.

Contents

- Electronic Proceedings
- Advance Program

The conference proceedings are being re-designed. The content of the papers and such is finished, but not all of the "surrounding" aspects are finished yet.

Quick jumps into interesting parts of the Advance Program:
- Topics of Conference
- Conference-at-a-Glance (requires graphical browser)
- Registration Information
- Important Conference Information
20 Years Ago … In Web and Search

- NCSA Mosaic graphical browser 2 years old, web search engines <1 year old
- CHI 1995 online presence
- 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
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

- “Personal navigational” queries
  - Different intents across individuals ... but consistently the same intent for an individual
    - SIGIR (for Dumais) -> www.sigir.org/sigir2015
    - SIGIR (for Bowen Jr.) -> www.sigir.mil

<table>
<thead>
<tr>
<th>Repeat</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td>33%</td>
</tr>
<tr>
<td>New Query</td>
<td>67%</td>
</tr>
</tbody>
</table>
Personal Navigation Details

- Large-scale log analysis
- Identifying personal navigation queries
  - Use consistency of 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
    - Consistent over time
  - High coverage, low risk personalization
- A/B experiments confirm benefits
Example 2: PSearch

- Rich client-side model of a user’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

User profile:
- Content
- Interaction history
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) 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

HCII 2015
Example 3: Short + Long

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

- Short-term context
  - Actions within current session (Q, click, topic)
    - (Q=sigir | information retrieval vs. iraq reconstruction)
    - (Q=ego | id)
    - (Q=hcii | chi conference)
  - 60% of session have multiple queries

- Personalized ranking model combines both
Short + Long Details

- User model (temporal extent)
  - Session, Historical, Combinations
  - Temporal weighting

- 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 family","dexter 8","tv guide")
6% Travel ("rome hotels","tripadvisor seattle","rome pasta")
5% Hockey ("elmira pioneers","umass lax","necll")

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

Typical

New Session 2:
Dentistry ("oral sores")
Dentistry ("aphthous sore")
Healthcare ("aphthous ulcer treatment")

Atypical

~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
A Crowd of Your Own

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

- Useful for: personal collections, dynamic collections, or unique information needs

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

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

- “Personalized” judgments from crowd workers

<table>
<thead>
<tr>
<th>Requester</th>
<th>Taste-Matching:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Worker I</td>
</tr>
<tr>
<td></td>
<td>Worker II</td>
</tr>
<tr>
<td></td>
<td>Worker III</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taste-Matching:</th>
<th>Taste-Grokking:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worker I</td>
<td>Sees ratings from requester</td>
</tr>
</tbody>
</table>

HCII 2015
A Crowd of Your Own Details

- **Images: Taste Matching**

<table>
<thead>
<tr>
<th></th>
<th>Products</th>
<th>Food Set #1</th>
<th>Food Set #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1.64</td>
<td>1.51</td>
<td>1.58</td>
</tr>
<tr>
<td>Best matched of five-workers</td>
<td><strong>1.43 (13%)</strong></td>
<td><strong>1.19 (22%)</strong></td>
<td><strong>1.26 (20%)</strong></td>
</tr>
<tr>
<td>Best matched of ten workers</td>
<td><strong>1.35 (18%)</strong></td>
<td><strong>1.08 (29%)</strong></td>
<td><strong>1.08 (31%)</strong></td>
</tr>
</tbody>
</table>

- **Images: Taste Grokking**

<table>
<thead>
<tr>
<th></th>
<th>Products</th>
<th>Food Set #1</th>
<th>Food Set #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1.64</td>
<td>1.51</td>
<td>1.58</td>
</tr>
<tr>
<td>Any grokker</td>
<td>1.29 <strong>(21%)</strong></td>
<td>1.53 (1%)</td>
<td>1.57 (0.5%)</td>
</tr>
<tr>
<td>5 rand grokkers</td>
<td>1.07 <strong>(34%)</strong></td>
<td>1.38 (9%)</td>
<td>1.28 (19%)</td>
</tr>
<tr>
<td>5 good grokkers</td>
<td>1.02 <strong>(34%)</strong></td>
<td>1.22 (19%)</td>
<td>1.13 (28%)</td>
</tr>
</tbody>
</table>

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

- User profile and content need to be in the same place
- Local profile (e.g., PSearch)
  - Local profile, local computation
  - Only query sent to server
- Cloud profile (e.g., Web search)
  - Cloud profile, cloud computation
  - Transparency and control over what’s stored
- Other approaches
  - Public or semi-public profiles (e.g., tweets, Facebook status)
  - Light weight profiles (e.g., queries in a session)
  - Matching to a group 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

- Actual searcher “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, but can be very noisy

- Diversity of evaluation methods important

Linking Implicit and Explicit

- “Curious Browser” plug-in to learn relationship between implicit and explicit behavior
  - Capture many implicit actions (e.g., click, click position, dwell time, scroll)
  - Probe for explicit judgments of relevance of a page to the query
- Deployed client software to ~4k people in US and Japan
“Curious Browser” plug-in to learn relationship between implicit and explicit behavior
- Capture many implicit actions (e.g., click, click position, dwell time, scroll)
- Probe for explicit judgments of relevance of a page to the query

Deployed client software to ~4k people in US and Japan

Learned models to predict explicit judgments (page relevance, session success) from implicit indicators
- Relevance of web page to query
  - 45% accuracy with just click
  - 75% accuracy with click + dwell time + session end
- Success of the session
  - Highly correlated w/ SAT clicks
  - ... but misses effort, expectations, delight, etc.
Summary

- Queries difficult to interpret in isolation
  - Augmenting query with context helps

- Potential for improving search using context is large

- Examples
  - PNav, PSearch, Short/Long

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

- Personalization 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
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.* *

**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.* *