PERSONALIZED SEARCH: POTENTIAL AND PITFALLS

Susan Dumais, Microsoft Research
Overview

- Importance of context in search
- Potential for personalization framework
- Examples
  - Personal navigation
  - Client-side personalization
  - Short- and long-term models
  - Temporal dynamics
- Challenges and new directions
Search and Context

User Context

Query Words

Task Context

Document Context

Query Words

Ranked List

NIPS 2013: Personalization Workshop
Context Improves Query Understanding

- Queries are difficult to interpret in isolation
- Easier if we 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 | Jul. 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 individual relevance for the same query

- Different ways to measure individual relevance:
  - Explicit judgments from different people
  - Implicit judgments (clicks, content analysis, etc.)

- Personalization can lead to large improvements:
  - Study with explicit judgments
  - 46% gain with single ranking
  - 72% gain with personalized ranking

NIPS 2013: Personalization Workshop
Potential For Personalization

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

- Learn when to personalize
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
  ▪ When used: Always, sometimes, context learned
  ▪ How used: Ranking, query support, presentation, etc.
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
  - When used: Always, sometimes, context learned
  - How used: Ranking, query support, presentation, etc.
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/sigir2013
    - SIGIR (for Bowen Jr.) -> www.sigir.mil

<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>
</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, was there a unique click on same result?
    - Behavior consistent over time
  - Coverage and accuracy
    - Many such queries: ~12% of queries
    - Prediction accuracy high: ~95% accuracy
    - High coverage, low risk personalization
- Can be used to re-rank, or augment presentation
- Online evaluation

NIPS 2013: Personalization Workshop
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 (lots 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

NIPS 2013: Personalization Workshop
PSearch Details

- **Ranking Model**
  - Score: Weighted combination of personal and global scores
    
    \[ \text{Score}(\text{result}_i) = \alpha \text{PersonalScore}(\text{result}_i) + (1 - \alpha) \text{WebScore}(\text{result}_i) \]
  - Personal score: Content and interaction history features
    - Content score - log odds of term in personal vs. web content
    - Interaction history score - visits to the specific URL, with backoff to domain

- **Evaluation**
  - Offline evaluation, using explicit judgments
  - Online evaluation, using PSearch prototype
    - Internal deployment; 225+ people for several months
    - Coverage: Results personalized for 64% of queries
    - Effectiveness:
      - CTR 28% higher, for personalized results
      - CTR 74% higher, when personal evidence is strong
      - Learned model for when to personalize

NIPS 2013: Personalization Workshop
Example 3: Short + Long

- **Short-term interests**
  - Behavior: Queries, clicks within current session
    - (Q= sigir | information retrieval vs. iraq reconstruction)
    - (Q= nips | icml vs.
    - (Q= acl | computational linguistics vs.
  - Content: Language models, topic models, etc.

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

- Developed unified model for both
- Sometimes short-term activity consistent with long-term interests, sometimes not
Short + Long Details

- User model (features)
  - Related queries, clicked 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 4: Temporal Dynamics

- Queries are not uniformly distributed over time
  - Often triggered by events in the world
- Relevance changes over time
  - E.g., *US Open* ... in 2013 vs. in 2012
  - E.g., *US Open 2013* ... in May (golf) vs. in Sept
  - E.g., *US Tennis Open 2013* ... before vs. during vs. after
    - Before event: Schedules and tickets, e.g., stubhub
    - During event: Real-time scores or broadcast, e.g., espn
    - After event: General sites, e.g., wikipedia, usta
Temporal Dynamics Details

- 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
  - Learn appropriate weighting of historical data
  - Useful for queries with local or global trends
Challenges in Personalization

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

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

- Profile on client (e.g., PSearch)
  - Profile is private
  - Query to server, many documents returned, local computations

- Profile in cloud
  - Transparency about what’s stored
  - Control over what’s stored … including nothing

- Other approaches
  - Light weight profiles (e.g., queries in a session)
  - Public or semi-public profiles (e.g., tweets, Facebook status)
  - Matching an individual to group
Serendipity

- 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 and Feedback

- **External judges, e.g., crowdworkers**
  - Lack diversity of intents and backgrounds

- **Actual searcher**
  - **Offline**
    - Allows safe exploration of many different alternatives
    - Labels can be explicit or implicit judgments (log analysis)
  - **Online**
    - Explicit judgments: Nice, but annoying and may change behavior
    - Implicit judgments: Scalable, but can be very noisy
    - Note ... limited experimental bandwidth; not directly repeatable; requires production-level code; mistakes costly

- **Diversity of methods important**
  - User studies, log analysis, and A/B testing
Summary

- Queries difficult to interpret in isolation
- Augmenting query with context can help
  - Who, what, where, when?
- Potential for improving search using context is large
- Examples
  - PNav, PSearch, Short/Long, Time
- Challenges and new directions

NIPS 2013: Personalization Workshop
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, Krysta Svore, Kira Radinski, Jon Elsas, Sarah Tyler, Alex Kotov, Anagha Kulkarni, David Sontag, Carsten Eickhoff
References

**Short-term models**
- Eickhoff et al., WSDM 2013. *Personalizing atypical search sessions.*

**Long-term models**
- 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.*

**Temporal models**