Cohort Modeling for Enhanced Personalized Search

Jinyun Yan  
Rutgers University

Wei Chu  
Microsoft Bing

Ryen White  
Microsoft Research
Personalized Search

• Many queries have multiple intents
  – e.g., [H2O] can be a beauty product, wireless, water, movie, band, etc.

• Personalized search
  – Combines relevance and the searcher’s intent
  – Relevant to the user’s interpretation of query
Challenge

• Existing personalized search
  – Relies on the access to personal history
    • Queries, clicked URLs, locations, etc.

• Re-finding common, but not common enough
  – Approx. 1/3 of queries are repeats from same user [Teevan et al 2007, Dou et al 2007]
  – Similar statistics for <user, q, doc> [Shen et al 2012]

2/3 queries new in 2 mo. - ‘cold start’ problem
Motivation for Cohorts

• When encountering new query by a user
  – Turn to other people who submitted the query
  – e.g., Utilize global clicks

• Drawback
  – No personalization

• Cohorts
  – A group of users similar along 1+ dimensions, likely to share search interests or intent
  – Provide useful cohort search history
Situating Cohorts

Global

Cohort

Individual

Not personalized

Conjoint Analysis
Learning across Users
Collaborative
Grouping/Clustering
Cohorts ...

Hard to Handle New Queries
Hard to Handle New Documents
Sparseness (Low Coverage)
Related Work

• Explicit groups/cohorts
  – Company employees [Smyth 2007]
  – Collaborative search tools [Morris & Horvitz 2007]

• Implicit cohorts
  – Behavior based, $k$-nearest neighbors [Dou et al. 2007]
  – Task-based / trait-based groups [Teevan et al. 2009]

• Drawbacks
  – Costly to collect or small $n$
  – Uses information unavailable to search engines
  – Some offer little relevance gain
Problem

• Given search logs with <user, query, clicks>, can we design a cohort model that can improve the relevance of personalized search results?
Concepts

• **Cohort**: A cohort is a group of users with shared characteristics
  – E.g., a sports fan

• **Cohort cohesion**: A cohort has cohesive search and click preferences
  – E.g., search [fifa] → click fifa.com

• **Cohort membership**: A user may belong to multiple cohorts
  – Both a sports fan and a video game fan
Our Solution

- **Cohort Generation**: Identify particular cohorts of interest
- **Cohort Membership**: Find people who are part of this cohort
- **Cohort Behavior**: Mine cohort search behavior (clicks for queries)
- **Cohort Preference**: Identify cohort click preferences
- **Cohort Model**: Build models of cohort click preferences
- **User Preference**: Apply that cohort model to build richer representation of searchers’ individual preferences
Cohort Generation

• Proxies
  – **Location** (U.S. state)
  – **Topical interests**
    (Top-level categories in Open Directory Project)
  – **Domain preference**
    (Top-level domain, e.g., .edu, .com, .gov)
  – Inferred from search engine logs
    • Reverse IP address to estimate location
    • Queries and clicked URLs to estimate search topic interest and domain preference for each user
Cohort Membership

- Multinomial distribution
  - Smoothed

\[
p(C_j|u) = w(u, C_j) = \frac{SATClips(u, C_j) + 1}{\sum_j SATClips(u, C_j) + K}
\]

- Example:

\[
C = [\text{Arts, Business, Computers, Games}]
\]
\[
SATClips = [0, 1, 2, 5] \text{ (clicks w/ dwell \geq 30s)}
\]
\[
w(u, C) = [0.083, 0.167, 0.25, 0.5]
\]
Cohort Preference

• Cohort click preference
  – Cohort CTR:
  \[
  CTR(d, q, C_j) = \frac{\sum_u SATClicks(d, q, u) \cdot w(u, C_j)}{\sum_u Impressions(d, q, u) \cdot w(u, C_j)}
  \]
  – Global CTR:
  \[
  CTR(d, q) = \frac{\sum_u SATClicks(d, q)}{\sum_u Impressions(d, q)}
  \]
  – Simplified example:
    • Global preference:
      \[
      [CTR(d1, q), CTR(d2, q)] = \left[ \frac{4}{100}, \frac{3}{100} \right]
      \]
    • Cohort preference
      – Cohort 1: \([CTR_C(c1, d1, q), CTR_C(c1, d2, q)] = \left[ \frac{4}{100}, 0 \right]\)
      – Cohort 2: \([CTR_C(c2, d1, q), CTR_C(c2, d2, q)] = \left[ 0, \frac{3}{100} \right]\)
Cohort Model

• Estimate individual click preference by cohort preference

\[ z(d, q, u, C_j) = p(d, q, C_j) \cdot p(C_j | u) = CTR(d, q, C_j) \cdot w(u, C_j) \]
Experiments

• Setup
  – Randomly sampled 3% of users
  – 2-month search history for cohort profiling: cohort membership, cohort CTR
  – 1 week for evaluation:
    3 days training, 2 days validation, 2 days testing
  – 5,352,460 query impressions in testing

• Baseline
  – Personalized ranker used in production on Bing
  – With global CTR, and personal model
Experiments

• Evaluation metric:
  – Mean Reciprocal Rank of first SAT click (MRR)*
  \[ \Delta \text{MRR} = \text{MRR(cohort model)} - \text{MRR(baseline)} \]

• Labels: Implicit, users’ satisfied clicks
  – Clicks w/ dwell ≥ 30 secs or last click in session
  – 1 if SAT click, 0 otherwise

* \( \Delta \text{MAP} \) was also tried. Similar patterns to MRR.
Results

• Cohort-enhanced model beats baseline

<table>
<thead>
<tr>
<th>Group Type</th>
<th>ΔMRR ±SEM</th>
<th>Re-Ranked@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODP (Topic interest)</td>
<td>0.0187 ±0.00143</td>
<td>0.91%</td>
</tr>
<tr>
<td>TLD (Top level domain)</td>
<td>0.0229 ±0.00145</td>
<td>0.96%</td>
</tr>
<tr>
<td>Location (State)</td>
<td>0.0113 ±0.00142</td>
<td>0.90%</td>
</tr>
<tr>
<td>ALL (ODP + TLD + Location)</td>
<td>0.0211 ±0.00146</td>
<td>0.98%</td>
</tr>
</tbody>
</table>

– Positive MRR gain over personalized baseline
  • Average over many queries, with many ΔMRR = 0
  • Gains are highly significant (p < 0.001)

– ALL has lower performance, could be noisier:
  • Re-ranks more often, Combining different signals
Performance on Query Sets

• **New queries**
  – Unseen queries in training/validation
  \[\uparrow 2 \times \text{MRR gain vs. all queries}\]

• **Queries with high click-entropy**
  \[\text{ClickEntropy}(q) = - \sum_{d} CTR(d, q) \cdot \log(CTR(d, q))\]
  \[\uparrow 5 \times \text{MRR gain vs. all queries}\]

• **Ambiguous queries**
  – 10k acronym queries, all w/ multiple meanings
  \[\uparrow 10 \times \text{MRR gain vs. all queries}\]
Cohort Generation: *Learned* Cohorts

- **Thus far:** Pre-defined cohorts
  - Manual control of cohort granularity
- **Next:** Automatically learn cohorts
  - User profile
    <location, search interests, domain preference>
  - Cluster users into cohorts: *K*-means
  - Cohort membership:
    - Soft cluster membership
      \[ w(u, C_j) = p(C_j | u) = \frac{\exp\left(-\frac{d(x_u, \mu_j)^2}{\sigma^2}\right)}{\sum_{i=1}^{K} \exp\left(-\frac{d(x_u, \mu_i)^2}{\sigma^2}\right)} \]
    - Simplified version of Gaussian mixture model w/ identity covariance
      \[ w(u, C_j) = \frac{1}{\sqrt{(2\pi)^k \sigma^k}} \exp\left(-\frac{1}{2} \sum_{i=1}^{K} (x_u - \mu_i)^2 \right) \]
Finding Best $K$

- **Baseline**: Predefined cohorts (from earlier)
- Focus on different query sets
  - e.g., those with higher click entropy
- Probed $K = 5, 10, 30, 50, 70$
- **Learned** (for one set)
  - Top gain at $K=10$, sig
- Future work:
  - Need more exploration of results at $5 < K < 30$
Summary

• Cohort model enhanced personalized search
  – Enrich models of individual intent using cohorts
  – Automatically learn cohorts from user behavior

• Future work:
  – More experiments, e.g., parameter sweeps
  – More cohorts: Age, gender, domain expertise, political affiliation, etc.
  – More queries: Long-tail queries, task-based and fuzzy matching rather than exact match
Thanks

• Questions?