Probabilistic Models for Personalizing Web Search (WSDM '12)

David Sontag, Kevyn Collins-Thompson,
Paul N. Bennett, Ryen W. White, Susan Dumais,
Bodo Billerbeck

NEW YORK UNIVERSITY

Microsoft® Research

Personalizing web search

Query "Michael Jordan"

Results	Pr(relevance)
en.wikipedia.org/wiki/ ***Michael_Jordan***	.9
www.nba.com/playerfile/ ***michael_jordan***	.7
www.nba.com/history/players/ ***jordan***_*summary.html*	.6
…	…
www.eecs.berkeley.edu/ Faculty/Homepages/ ***jordan***.*html*	.0001

Personalized results	Pr(relevance to me)
www.eecs.berkeley.edu/Faculty/ Homepages/ ***jordan***.*html*	.8
…	…
en.wikipedia.org/wiki/ ***Michael_Jordan***	.1
www.nba.com/playerfile/ ***michael_jordan***	.08
www.nba.com/history/players/ ***jordan***_*summary.html*	.07

Key problems to solve:

- Representation – how to compactly summarize user preferences?
- Learning – how to discover user profiles from historical data?
- Ranking – how to balance preferences with other relevance signals?

Previous approaches

- Re-Finding (Teevan '04)
  - Remember user's browsing history
  - Re-rank search results, boosting score of previously visited pages
- Term-based profiles (Teevan *et al.* '05, Tan *et al.* '06, Matthijs and Radlinski '11)
  - Construct personalized vocabulary from browsing history
  - Use to re-weight term-based scoring methods such as BM25 or TF-IDF
- Topic-based profiles (Gauch *et al.* '03, Liu *et al.* '04, Chirita *et al.* '05, Dou *et al.* '07)
  - Learn distribution over a priori query intents for user
  - Re-rank search results using linear combination of user topic-document topic match and other relevance scores

Our representation

- Fundamental goal is to predict **query intent**

(top 2 levels of human-generated ontology, dmoz.org)

- User preferences summarized as (user specific) parameters,

User *u* history → θ_u

of the conditional distribution:

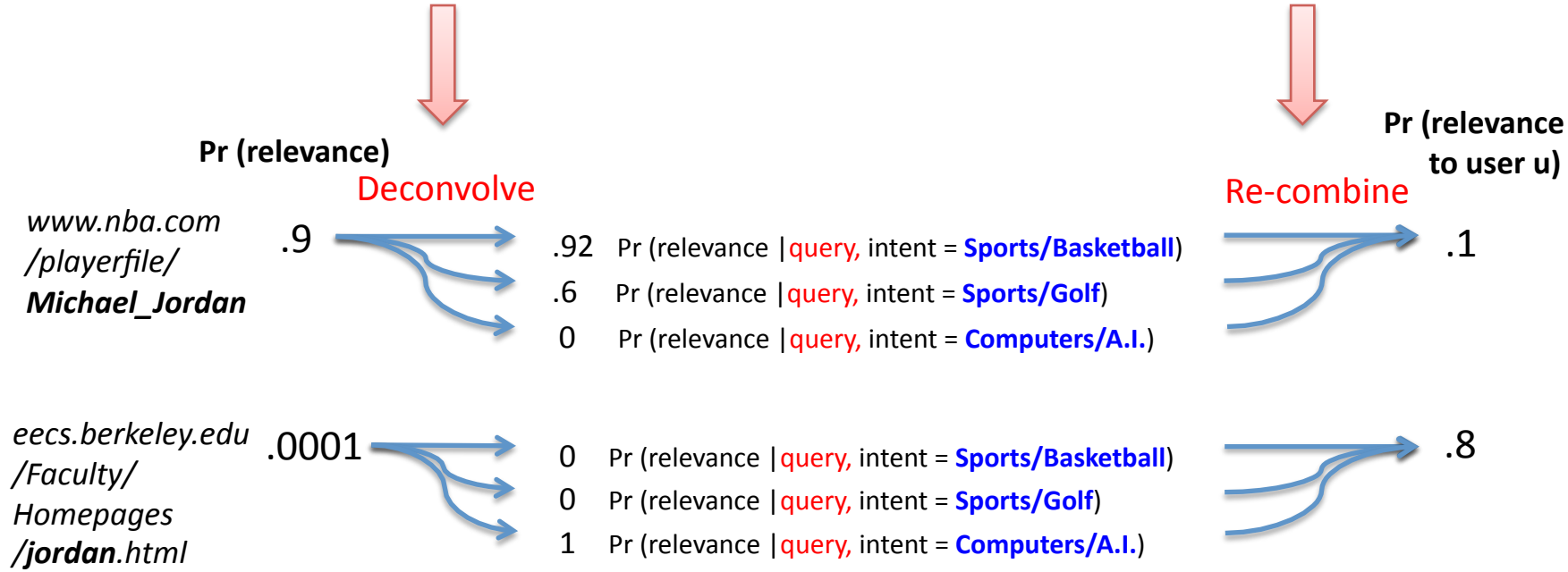
$$\Pr(\text{intent} \mid \text{query}; \theta_u)$$

Our ranking method

Query "Michael Jordan"

Background model,
Pr (intent | "Michael Jordan"; generic user)

User model,
Pr (intent | "Michael Jordan" ; θ_u)



Example

Query "Rockefeller" issued by Biologist

Pr (intent | "Rockefeller"; generic user)
Business: 0.213
Society: 0.107
Shopping/Health: 0.096
Business/Consumer Goods+Services: 0.077
Arts: 0.062

Web search engine results	**Categories**
1. http://en.wikipedia.org/wiki/John_D._Rockefeller	Society
2. http://en.wikipedia.org/wiki/Rockefeller_family	Science, Society
3. http://www.rockefeller.edu	Reference, Science

Original

Pr (intent | "Rockefeller" ; θ_u)
Science/Biology: 0.402
Science: 0.228
Society: 0.052
Reference: 0.040
Health: 0.031

(original rank)

Personalized re-ranking results		**Categories**
1. http://www.rockefeller.edu	(3)	Reference, Science
2. http://en.wikipedia.org/wiki/Rockefeller_family	(2)	Science, Society
3. http://bridges.rockefeller.edu/?page=news	(12)	Science, Health

Personalized

Document distributions

Query "Rockefeller" issued by Biologist

Web search engine results	Categories
1. http://en.wikipedia.org/wiki/John_D._Rockefeller	Society
2. http://en.wikipedia.org/wiki/Rockefeller_family	Science, Society
3. http://www.rockefeller.edu	Reference, Science

Original

- For every document we have a distribution,

  Pr (doc *d* about **topic T** | *d*'s text)

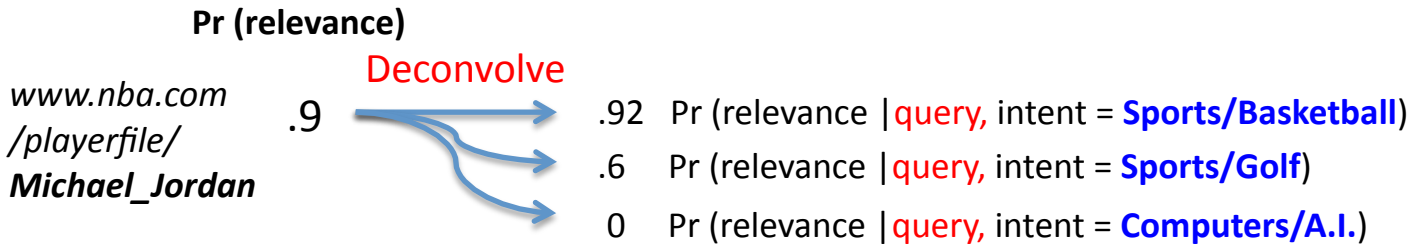
  over the topic that the document is about

- Learned using logistic regression; training data is from the Open Directory Project
- Stored in the index and accessible quickly

(Bennett, Svore, Dumais, WWW '10)

Our ranking method: intuition

Step 1.



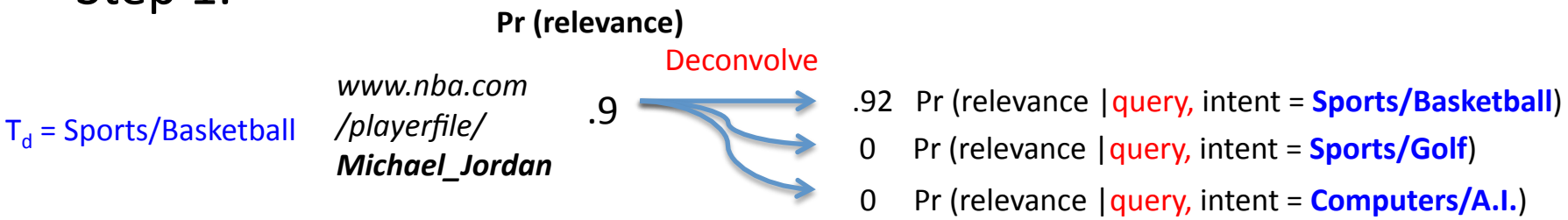
We assume that Pr (relevance) is the *expected* relevance across all users:

$$\Pr(d \text{ relevant}) = \sum_{T} \Pr(\text{intent} = T \mid \text{“Michael Jordan”; generic user}) * \Pr(d \text{ relevant} \mid \text{intent} = T)$$

Suppose that doc *d* is about topic T_d and $\Pr(d \text{ relevant} \mid \text{intent} \neq T_d) = 0$

Our ranking method: intuition

Step 1.



We assume that Pr (relevance) is the *expected* relevance across all users:

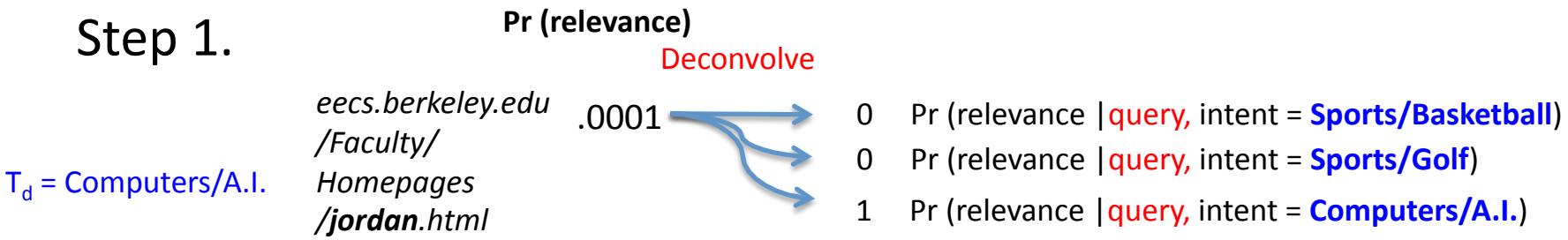
$$\Pr(d \text{ relevant}) = \sum_{T} \Pr(\text{intent} = T \mid \text{“Michael Jordan”; generic user}) * \Pr(d \text{ relevant} \mid \text{intent} = T)$$

Suppose that doc *d* is about topic T_d and $\Pr(d \text{ relevant} \mid \text{intent} \neq T_d) = 0$

$$\Pr(d \text{ relevant} \mid \text{query, intent} = \textbf{Sports/Basketball}) = \frac{\Pr(d \text{ relevant})}{\Pr(\text{intent} = \textbf{Sports/Basketball} \mid \text{“Michael Jordan”; generic user})}$$

Our ranking method: intuition

Step 1.



We assume that Pr (relevance) is the *expected* relevance across all users:

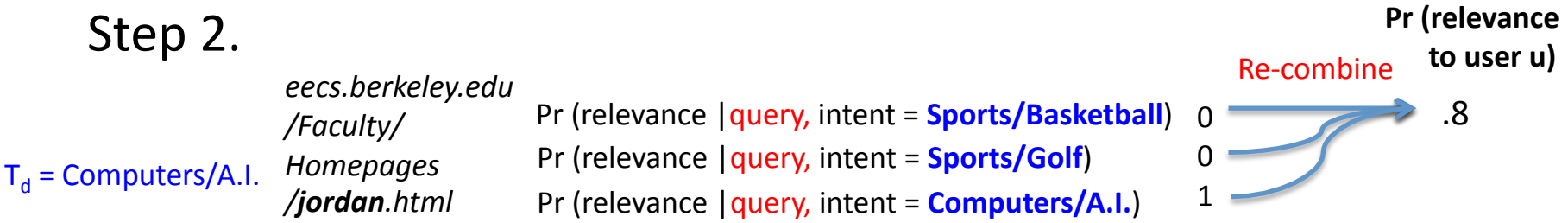
$$\Pr(d \text{ relevant}) = \sum_{T} \Pr(\text{intent} = T \mid \text{“Michael Jordan”; generic user}) * \Pr(d \text{ relevant} \mid \text{intent} = T)$$

Suppose that doc *d* is about topic T_d and $\Pr(d \text{ relevant} \mid \text{intent} \neq T_d) = 0$

$$\Pr(d \text{ relevant} \mid \text{query, intent} = \textbf{Computers/A.I.}) = \frac{\Pr(d \text{ relevant})}{\Pr(\text{intent} = \textbf{Computers/A.I.} \mid \text{“Michael Jordan”; generic user})}$$

Our ranking method: intuition

Step 2.



To re-combine, we marginalize over the user's intent,

$$\Pr(d \text{ relevant to } u) = \sum_T \Pr(u\text{'s intent} = T \mid \text{"Michael Jordan"}; \theta_u) * \Pr(d \text{ relevant} \mid \text{intent} = T)$$

Putting steps 1 and 2 together, we obtain:

$$\Pr(d \text{ relevant to } u) = \Pr(d \text{ relevant}) \frac{\Pr(u\text{'s intent} = \text{Computers/A.I.} \mid \text{"Michael Jordan"}; \theta_u)}{\Pr(\text{intent} = \text{Computers/A.I.} \mid \text{"Michael Jordan"}; \text{generic user})}$$

Our ranking method

- Recall our simplifying assumptions:

~~Suppose that doc *d* is about topic T_d~~ and ~~Pr(d relevant | intent ≠ T_d) = 0~~

↓ Treat T_d as a random variable and marginalize over it

↓ Use the distribution Pr (d relevant | *u*'s intent = t_u, doc about topic t_d)

↑ Call this f (t_u, t_d)

- Ranking formula:

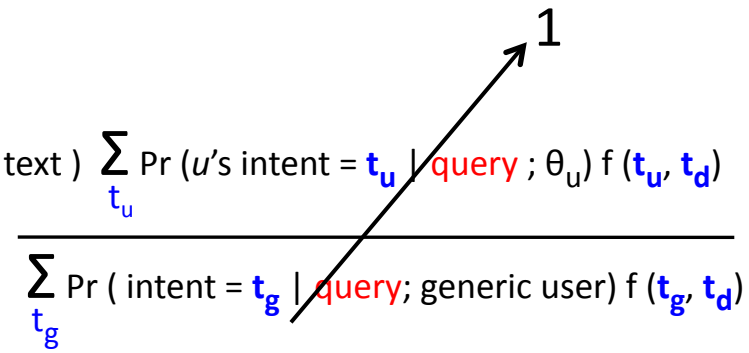
$$\Pr(d \text{ relevant to } u) = \Pr(d \text{ relevant}) \sum_{t_d} \Pr(\text{doc } d \text{ about topic } \mathbf{t_d} \mid d\text{'s text}) \frac{\sum_{t_u} \Pr(u\text{'s intent} = \mathbf{t_u} \mid \text{query}; \theta_u) f(\mathbf{t_u}, \mathbf{t_d})}{\sum_{t_g} \Pr(\text{intent} = \mathbf{t_g} \mid \text{query; generic user}) f(\mathbf{t_g}, \mathbf{t_d})}$$

- Obtained by probabilistic inference – see paper for the formal probabilistic model

Algorithm properties

- Ranking formula:

$$\Pr(d \text{ relevant to } u) = \Pr(d \text{ relevant}) \sum_{t_d} \Pr(\text{doc } d \text{ about topic } \mathbf{t_d} \mid d\text{'s text}) \frac{\sum_{t_u} \Pr(u\text{'s intent} = \mathbf{t_u} \mid \text{query}; \theta_u) f(\mathbf{t_u}, \mathbf{t_d})}{\sum_{t_g} \Pr(\text{intent} = \mathbf{t_g} \mid \text{query; generic user}) f(\mathbf{t_g}, \mathbf{t_d})}$$



- Ranking **unchanged** if user intent = generic user's for query
- Ranking can exhibit **big effects** for less-common intents vs. generic user
- Very **fast** to compute:
  - Ranking N docs takes time O($Nk + T^2$)
  - k = # topics per doc (e.g. 3), and T is total number of topics (e.g. 300)

What is left to specify?

- Ranking formula:

Next slides

$$\Pr(d \text{ relevant to } u) = \Pr(d \text{ relevant}) \sum_{t_d} \Pr(\text{doc } d \text{ about topic } \mathbf{t_d} \mid d\text{'s text}) \frac{\sum_{t_u} \Pr(u\text{'s intent} = \mathbf{t_u} \mid \text{query}; \theta_u) f(\mathbf{t_u}, \mathbf{t_d})}{\sum_{t_g} \Pr(\text{intent} = \mathbf{t_g} \mid \text{query}; \text{generic user}) f(\mathbf{t_g}, \mathbf{t_d})}$$

If not available, can use 1/rank

We learn this (details in paper)

These statistics can be pre-computed for common queries

Alternatively, use weighted average of document topics for top docs in **original** ranking,

$$\sum_{d} \Pr(d \text{ relevant}) * \Pr(\text{doc } d \text{ about topic } \mathbf{t} \mid d\text{'s text})$$

(White, Bennett, Dumais, CIKM '10)

Predicting user intent

- Long term personalization, using historical user click-through data:

  User *u*'s history = **{** (query$_i$, intent$_i$), i=1, …, c **}**

- Pr (*u*'s intent = **t** | query ; θ_u) estimated using two approaches,
  1. Generative model
  2. Discriminative model

- We use an ensemble method, interpolating between the predictions of both methods

Predicting user intent: Generative model

- From single user *u*'s history, **{** ($query_i$, $intent_i$), i=1, …, c **}**, estimate **a priori query intents**:

$$Pr_u (intent)$$

- Using data from all users, estimate **language model**,

$$Pr (query \mid intent),$$

  i.e. statistics on frequency of queries seen for each intent

- Use Bayes' rule to "invert":

$$Pr (u\text{'s } intent = t \mid query) = \frac{Pr_u (intent = t)^* Pr (query \mid intent = t)}{\sum_{t'} Pr_u (intent = t')^* Pr (query \mid intent = t')}$$

Predicting user intent: Discriminative model

- Choose user-specific parameters θ_u which correctly predict intent on historical data:

Complexity penalty to avoid over-fitting

$$\max_{\theta_u} \sum_{i=1}^{c} \log \Pr(\text{intent}_i \mid \text{query}_i ; \theta_u) - C|\theta_u|^2$$

- We assume log-linear model for the distribution
- T+1 dimensional feature vector, where T = # topics
- Parameters specify how to re-weight Pr (intent | query; generic user)

Large-scale Evaluation

- Data set: September 2010 search logs (Bing)
  - 20 days training, 6 days test
  - ~600K queries, ~200K users

	average	stdev	median
num days	16.21	3.72	17
num queries	229.60	112.28	204
num SAT clicks	143.82	52.80	128

User statistics for training data

- Re-rank top 10 results
  - Assign positive judgment to URL in top 10 if it is the last satisfied result click in the session
  - Negative judgment to other 9 URLs
- We report the **mean reciprocal rank (MRR)**:

3rd → 2nd position:
ΔMRR = 0.1667

$$MRR = (1/|Q|) \sum_{q \text{ in } Q} \frac{1}{\text{rank of last satisfied click URL}}$$

- Compare original **Bing** ranking with personalized ranking

Large-scale Evaluation

Filter conditions	Set Size	Change in Filter Set MRR	Change in Overall MRR
One Word	100.00%	0.1213	0.1213

MRR on subset of queries where last satisfied result click moves position

Large-scale Evaluation

Filter conditions	Set Size	Change in Filter Set MRR	Change in Overall MRR
One Word	100.00%	0.1213	0.1213
One Word. Ambig.	68.21%	0.1361	0.0928
One Word, non-Nav	73.28%	0.1442	0.1064
One Word, Ambig., non-Nav	54.81%	0.1686	0.0924

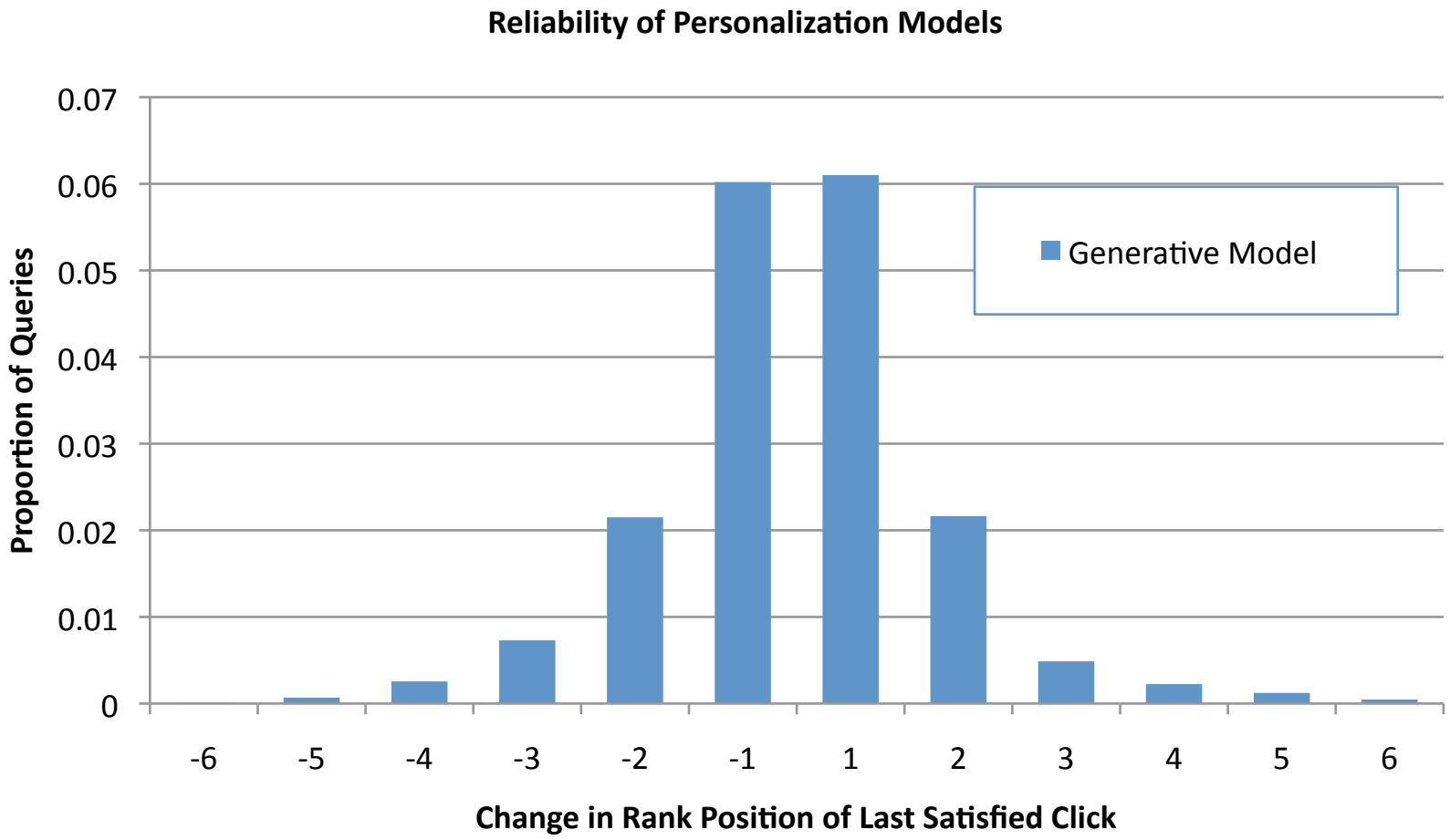
MRR on subset of queries where last satisfied result click moves position

Large-scale Evaluation

Filter conditions	Set Size	Change in Filter Set MRR	Change in Overall MRR
One Word	100.00%	0.1213	0.1213
One Word. Ambig.	68.21%	0.1361	0.0928
One Word, non-Nav	73.28%	0.1442	0.1064
One Word, Ambig., non-Nav	54.81%	0.1686	0.0924
Acronym	31.73%	0.1745	0.0554
Acronym, Ambig, non-Nav	21.08%	0.2269	0.0478

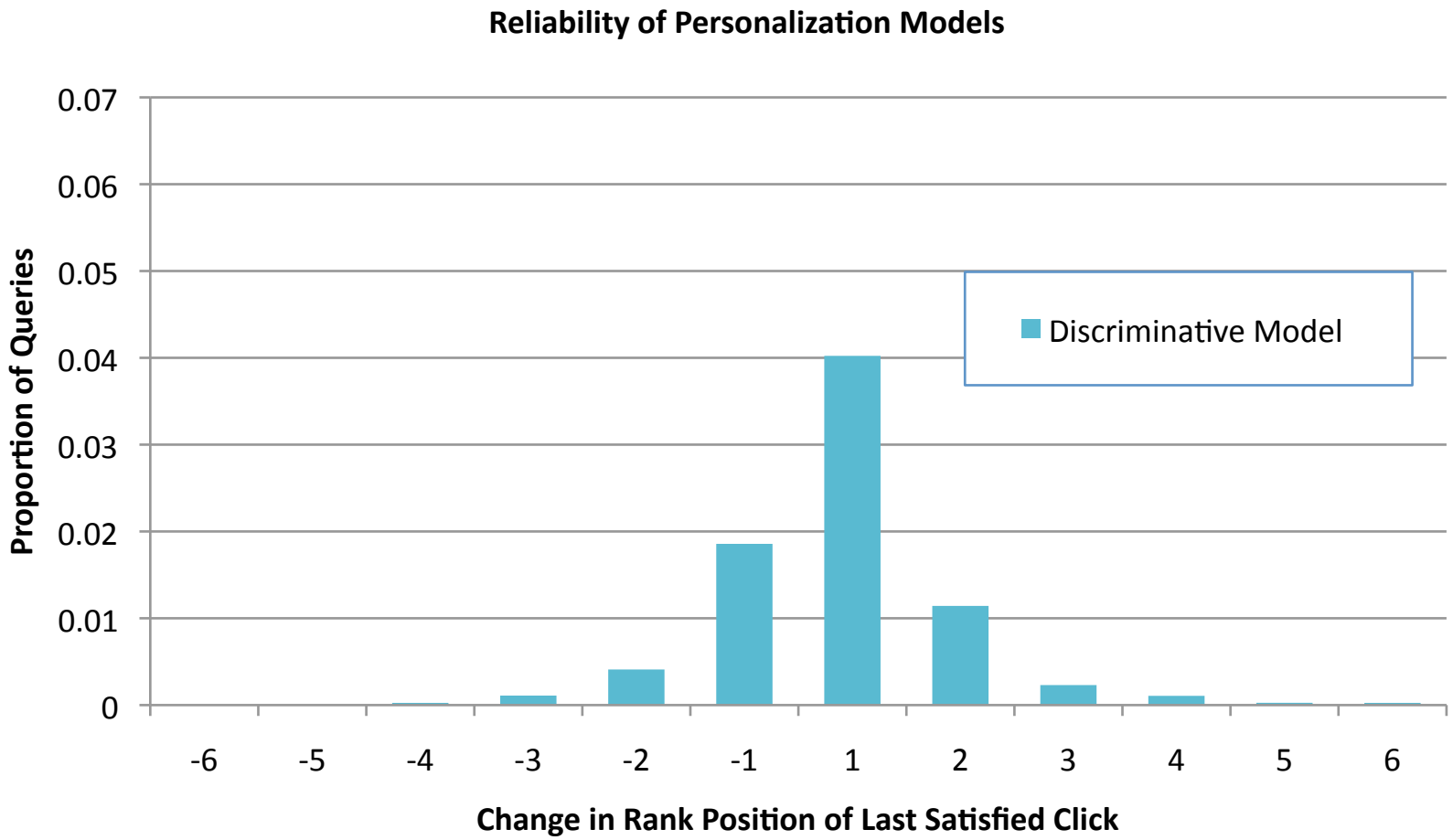
MRR on subset of queries where last satisfied result click moves position

Re-ranking win/loss distribution



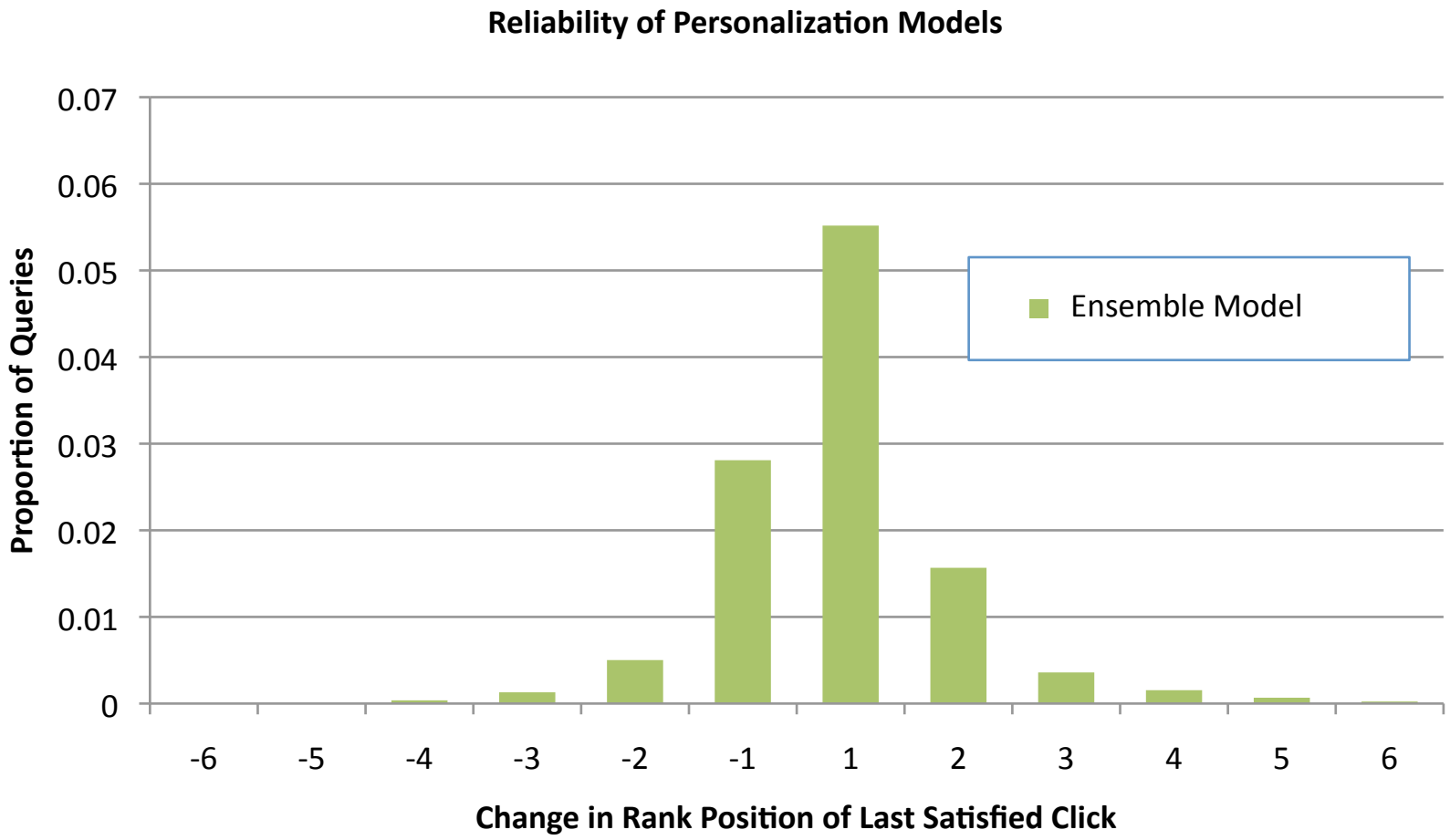
Filter = ambiguous, one word non-navigational queries

Re-ranking win/loss distribution



Filter = ambiguous, one word non-navigational queries

Re-ranking win/loss distribution



Filter = ambiguous, one word non-navigational queries

Framework is broadly applicable

- Short-term personalization (within session)
- Different personalization criteria
  - Geographic location
  - Reading proficiency
  - Multiple topics per document or user intent

Summary of contributions

- Probabilistic framework for personalization
- Learning user profiles formalized as intent prediction (conditioned on query)
- Use of a background model (generic user's intent) to interpret ranker's relevance scores
- Large-scale evaluation of long-term personalization using query logs
- Substantial gains over competitive baseline on ambiguous queries such as acronyms and names

Many directions to explore!

- Predicting intent given query and user history:
  - Expand the set of features used in conditional model
  - Transfer learning across users
  - Learn from non-search data, e.g. browsing, mobile, social
  - Online learning of user profiles
- Understanding when and how to personalize
  - Consider both potential for personalization and confidence in user's query intent
- Representation
  - (Un)supervised learning of topics rather than using ODP
  - Use relational classification to improve accuracy of web page classification
  - Cross-product of many variables, e.g. topic and reading proficiency