Personalization via Probabilistic Adaptation

David Sontag
dsontag@cs.nyu.edu

Kevyn Collins-Thompson
kevynct@umich.edu

Paul N. Bennett, Ryen W. White, Susan Dumais, Bodo Billerbeck
{paul.n.bennett, ryenw, sdumais, bodob}@microsoft.com

Abstract

We present a new approach for personalizing Web search results to a specific user. Ranking functions for Web search engines are typically trained by machine learning algorithms using either direct human relevance judgments or indirect judgments obtained from click-through data from millions of users. The rankings are thus optimized to this generic population of users, not to any specific user. We propose a generative model of relevance which can be used to infer the relevance of a document to a specific user for a search query, and show how to learn these profiles from a user's long-term search history. Our algorithm for computing the personalized ranking is simple and has little computational overhead. We evaluate our personalization approach using historical search data from thousands of users of a major Web search engine.[1]

1 Introduction

Our paper proposes an end-to-end system for learning personalization models and performing extremely fast adaptation of a general ranking into a personalized ranking. We formalize the problem using a probabilistic model for predicting document relevance for a specific user and query. The user representation corresponds to user-specific parameters for part of the model. Our formalization is general and assumes only that there are document-specific latent variables (i.e., document features), user-specific latent variables (i.e., information need for this query), and some way of combining them to determine whether a document's features satisfy the user's information need.

Our approach begins with the assumption that the Web search engine provides a generic estimate of the probability that a document is relevant to the query. Since relevance is subjective, different people will find different documents relevant for the same query and no single ranking can satisfy all users [3, 4]. We explicitly consider the distribution of users for which the global ranking function was trained, and identify how a specific user is different from the population as a whole. Using this, we deconvolve the relevance score into a probability that a page is relevant for any specific query intent. Then, we recompute the probability of relevance taking into consideration the user's profile.

Although our formalization is general, we specifically consider its application to the task of personalization using topic-based profiles. We have one discrete variable for each document whose states specify the topic of the document. The state space that we use corresponds to the top two levels of the human-generated ontology provided by the Open Directory Project. In a pre-processing step, we use a text-based classifier, trained with logistic regression, to obtain the distribution over topics for each document in the index. This allows the personalized ranking to be computed extremely quickly at query time. In addition to having one variable per document, we have one variable for the user whose states specify the topic of the documents being searched for using the query. Even before seeing the query, the user's history provides a prior distribution for this variable.

[1] This workshop paper is a summarization of the full paper previously presented at WSDM 2012 [2].

2 Probabilistic Models for Personalization

We have a single variable for the document, T_d, and a single variable for the user, T_u. These discrete-valued variables refer to the document's topic and the topic that the user is searching for, respectively. The conditional distribution $\Pr(T_d \mid d)$ specifies the topic of each document and is assumed to be given to us. A document about topic T_d is assumed *relevant* to a user looking for topic T_u if both: (1) topic T_d satisfies a user with information need T_u, and (2) given that the document's topic matches that of the search intent, the document is relevant to the query.

The first criterion is measured by the variable $\text{cov}_u(d, q) \in \{0, 1\}$, which represents the extent to which T_d "covers" the information need T_u. The distribution $\Pr(\text{cov}_u(d, q) \mid T_u, T_d)$ could simply be given by $1[T_u = T_d]$, the indicator function for whether T_u is the same as T_d. This choice would imply that a document is irrelevant for queries outside of its topic area.

The second criterion is measured by the variable $\psi(d, q) \in [0, 1]$, which we call the *non-topical* relevance score, corresponding to the user-independent probability that the document is relevant to this query. This score is assumed to be comprised of a large number of user-independent signals such as the match of the query to document text or anchor text, aggregate user behavior for this query, etc. We intentionally do not model how this score arises, taking a black-box view of it.

The variable $\text{rel}_u(d, q) \in \{0, 1\}$ combines the two criteria, taking the value 1 when the user finds the document relevant to the query, and 0 otherwise. Specifically, we have $\Pr(\text{rel}_u(d, q) = 1 \mid \text{cov}_u(d, q), \psi(d, q)) = 0$ if $\text{cov}_u(d, q)$ is 0, and $\psi(d, q)$ if $\text{cov}_u(d, q)$ is 1. The user's personalized ranking is then obtained by sorting all of the documents by this probability.

Model 1 (No Background Model). θ_u refers to user-specific parameters, also called the user profile, that are learned from the user's historical data in an offline step. The user profile together with the current query q are used to come up with a *distribution* over the user's search intent (i.e. a distribution over topics), $\Pr(T_u \mid \theta_u, q)$. Here, the variables θ_u, q, d, and $\psi(d, q)$ are observed. Marginalizing out the latent variables, we obtain the following formula to use during ranking:

$$\Pr(\text{rel}_u(d, q) = 1 \mid \theta_u, q, d, \psi(q, d)) = \psi(d, q) \sum_{T_d} \Pr(T_d \mid d)\alpha(T_d) \tag{1}$$

where $\alpha(T_d) = \sum_{T_u} \Pr(T_u \mid \theta_u, q)\Pr(\text{cov}_u(d, q) = 1 \mid T_u, T_d)$, and can be computed just once for each query, regardless of the number of documents to be ranked.

Model 2 (Probabilistic Adaptation using Background Model). We now present a more realistic model that does not assume knowledge of $\psi(d, q)$, instead treating it as a latent variable. We suppose that we know only $\text{obs\_rel}(d, q)$, the *expected* relevance with respect to the distribution of users who typically search for query q, which we assume is summarized by $\Pr_r(T \mid q)$, the distribution of their query intents (we call this the *background distribution*). Marginalizing over $\psi(d, q)$, we obtain the following for the posterior marginal (to use for ranking):

$$\text{obs\_rel}(d, q) \sum_{T_d} \Pr(T_d \mid d) \overbrace{\frac{\sum_{T_u} \Pr(T_u \mid \theta_u, q)\ \Pr(\text{cov}(d, q) = 1 \mid T, T_d)}{\sum_T \Pr_r(T \mid q)\ \Pr(\text{cov}(d, q) = 1 \mid T, T_d)}}^{\text{Re-weighting factor}}. \tag{2}$$

Importantly, when $\Pr(T_u \mid \theta_u, q)$ is the same as $\Pr_r(T \mid q)$, the ranking is unchanged. That is, if we cannot distinguish the user's query intent from that of the general population of users that search for this query, then the re-weighting factor has value 1 and the personalized probability of relevance is simply given by $\text{obs\_rel}(d, q)$. We believe that this invariance property is essential to our approach's success. To our knowledge, our approach is the first personalization algorithm that explicitly uses a background distribution and satisfies such an invariance property.

Example. We illustrate our personalization approach using a demo that we implemented to re-rank the top 200 Bing search results. The user profile used was learned from two months of search logs from one of the authors, a computer scientist. Fig. 1 shows the results of our algorithms for the ambiguous query `[kevin murphy]`. Fig. 1(a) shows the top five ODP categories from the background model, $\Pr_r(T \mid q)$, and also the top five ODP categories that our algorithms predict as the query intent for the computer scientist, given by $\Pr(T_u \mid \theta_u, q)$. There are marked differences. The distribution over query intents for the "generic user" for the `[kevin murphy]` query is centered

Pr(topic | query) for generic user
Business: 0.213
Society: 0.107
Shopping/Health: 0.096
Business/Consumer Goods+Services: 0.077
Arts: 0.062

Pr(topic | query) for CS researcher
Computers/Artificial Intelligence: 0.663
Arts/People: 0.098
Science: 0.044
Computers: 0.042
Arts/Performing Arts: 0.036

(a)

Web search engine results	Categories
1. http://www.kevinmurphy.com.au	Business, Shopping
2. http://en.wikipedia.org/wiki/Kevin_Murphy_(actor)	Arts
3. http://www.kevinmurphystore.com	Health, Shopping

Personalized re-ranking results (using Model 1)		Categories
1. http://en.wikipedia.org/wiki/Kevin_Murphy_(actor)	(2)	Arts
2. http://www.kevinmurphy.com.au	(1)	Business, Shopping
3. http://www.cs.ubc.ca/~murphyk	(13)	Reference, Computers

Personalized re-ranking results (using Model 2)		Categories
1. http://www.cs.ubc.ca/~murphyk	(13)	Reference, Computers
2. http://en.wikipedia.org/wiki/Kevin_Murphy_(actor)	(2)	Arts
3. http://www.kevinmurphystore.com	(3)	Health, Shopping

(b)

Figure 1: (a) Top categories based on Pr(topic | query) for the query `[kevin murphy]`. (b) The original top three results from Bing for the same query, and re-ranked results using Models 1 and 2. Also shown to the right of each result is the original rank in parentheses and the predicted top-level ODP categories.

around business, society, and health, whereas for the computer scientist, the predicted query intent involved artificial intelligence, people, and science.

Fig. 1(b) presents the search results returned for this query, issued by the computer scientist, to: (i) Bing, (ii) the same Web search engine with results re-ranked using Eq. 1, and (iii) the same engine with results re-ranked using Eq. 2. When a computer scientist issues this query, it is likely that the intent is to reach the website of the University of British Columbia (UBC) professor, and not the actor or hair stylist. As can be seen from the example, both Eq. 1 and Eq. 2 promote the UBC professor's page from outside the top 10 results.

We observed across a large number of ambiguous queries that Eq. 2 (Model 2) performs significantly better than Eq. 1 (Model 1), typically promoting the desired result directly to the top position. The algorithms appear to work particularly well for name queries (e.g., on the query `[Michael Jordan]`, promoting the website of the statistician to position 1 from position 198 when queried by the computer scientist) and acronyms (e.g., `[sigir]`).

3 Evaluation

To evaluate these methods, we used 25 days of search logs for the Bing Web search engine for the English-speaking United States locale. We used the first 20 days to construct profiles for any user having at least 100 clicks and use the final 5 days as a test set. The methods were evaluated by comparing a reranking of the original top 10 results with Bing's default (*i.e.* personalization disabled) ranking with the goal of moving the last long dwell-time clicked item (satisfied or SAT click) by the user higher in the ranking. Specifically, we measure our performance using the inverse of the rank of the relevant (last SAT click) document, otherwise known as the mean reciprocal rank (MRR). We labeled each of top-10 results with ODP categories using a text-based classifier, described in [1]. To optimize parameters (*e.g.* β), we use a non-overlapping set of five weeks of search log data. To focus on underspecified queries which [4] have found especially amenable to personalization, we filtered the test queries to only include one word queries that were non-navigational (estimated by classifier) with high topic ambiguity (using entropy of ODP distribution of top 10) and that we saw sufficiently (50) many times in the training set to reliably estimate the language model. After these filters, our test set consisted of 54581 users with at least one query, and 102417 queries in total.

Table 1 shows the change in MRR of the queries in the test set for each of three methods of predicting the user's query intent, $\Pr(T_u \mid \theta_u, q)$, and for both Models 1 and 2. The generative method refers to

using a language modeling approach to predict the user's query intent. The discriminative method refers to a user-specific re-weighting of $\Pr_r(T \mid q)$ which attempts to maximize the probability of the user's actual intent conditioned on the query over the user's previous search history, and the interpolation method refers to using the convex combination of the distributions predicted by the generative and discriminative methods.

Table 1: Performance on ambiguous, one word non-navigational queries. Bold face indicates significant improvement ($p = 0.05$, Bonferroni correction) improvement over the baseline according to a sign test.

Model	Last SAT Moved	Moved MRR Δ	MRR Δ
Generative, Model 1	8.93%	**0.0753**	**0.0067**
Generative, Model 2	18.41%	0.0187	0.0034
Discriminative, Model 1	4.22%	**0.0732**	**0.0031**
Discriminative, Model 2	7.96%	**0.1808**	**0.0144**
Interpolated, Model 1	5.23%	**0.0957**	**0.0050**
Interpolated, Model 2	11.18%	**0.1686**	**0.0189**

The baseline for these experiments is the original ranking provided by the Bing Web search engine. The results are shown relative to the MRR of the baseline.[2] All of the methods improve on the baseline, with the best results achieved by using the background model (Model 2) together with the interpolation method. In general, Model 2 appears to be more aggressive than Model 1, re-ranking more often (as can be seen in the first column). This is because while Model 1 may change scores, it often does not change scores enough to change the ranking. Because Model 2 normalizes by the background model, a document whose topic is substantially more likely to be the intent of the user than of the generic user has its score dramatically amplified, even if the absolute probability of this topic being the user intent is small. It is thus essential that we correctly predict the user's query intent. For the generative method, this aggressiveness results in lower performance (as seen in the third column), while the discriminative method is much more reliable and actually gains in performance. The interpolation method provides the best overall estimate of the user's query intent: when applied together with Model 2 it achieves good performance with high coverage, yielding the highest total gain of 0.0189, much higher than the next best method. Out of the total 102417 ambiguous, non-navigational queries in the test set, for the *Interpolated, Model 2* method, 11448 queries (11%) had a change in position of the relevant item, with 7881 (69%) of these queries helped – significantly higher than the null hypothesis of 50%.

4 Discussion

Although in this paper we specifically applied the framework to the problem of single topic-based personalization, the same ranking formula can be directly applied to a number of different types of personalization criteria such as multiple topics, geographic location, and reading proficiency. The ranking approach can also be used for personalizing based on short-term user profiles, by simply plugging in a different distribution for $\Pr(T_u \mid q, \theta_u)$.

The objective functions that we optimize to learn the user profiles are convex, making it straightforward to design online learning algorithms for the user profiles. This would give a simple update to use for θ_u after observing each new search query, and would guarantee low regret relative to the best possible θ_u chosen in hindsight.

References

[1] P. Bennett, K. Svore, and S. Dumais. Classification-enhanced ranking. In *WWW '10*, pages 111–120, 2010.

[2] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck. Probabilistic models for personalizing web search. In *Proceedings of WSDM*. ACM, 2012.

[3] J. Teevan, S. Dumais, and E. Horvitz. Potential for personalization. *ACM TOCHI*, 17(1), 2010.

[4] J. Teevan, S. Dumais, and D. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *SIGIR '08*, pages 163–170, 2008.

[2] To help interpret MRR Δ, if the last satisfied clicked document was always returned in the fourth position by the baseline and the personalized ranking always returned it in the third position, then this would be an MRR Δ of 0.0833. Moving from third to second would yield a Δ of 0.1667.