Behavior-Based Personalization in Web Search*

Fei Cai
Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Hunan, China and Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands. E-mail: f.cai@uva.nl

Shuaiqiang Wang
Department of Computer Science and Information Systems, University of Jyväskylä, Jyväskylä, Finland. E-mail: shuaiqiang.wang@jyu.fi

Maarten de Rijke
Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands. E-mail: derijke@uva.nl

Personalized search approaches tailor search results to users' current interests, so as to help improve the likelihood of a user finding relevant documents for their query. Previous work on personalized search focuses on using the content of the user's query and of the documents clicked to model the user's preference. In this paper we focus on a different type of signal: We investigate the use of behavioral information for the purpose of search personalization. That is, we consider clicks and dwell time for reranking an initially retrieved list of documents. In particular, we (i) investigate the impact of distributions of users and queries on document reranking; (ii) estimate the relevance of a document for a query at 2 levels, at the query-level and at the word-level, to alleviate the problem of sparseness; and (iii) perform an experimental evaluation both for users seen during the training period and for users not seen during training. For the latter, we explore the use of information from similar users who have been seen during the training period. We use the dwell time on clicked documents to estimate a document's relevance to a query, and perform Bayesian probabilistic matrix factorization to generate a relevance distribution of a document over queries. Our experiments show that: (i) for personalized ranking, behavioral information helps to improve retrieval effectiveness; and (ii) given a query, merging information inferred from behavior of a particular user and from behaviors of other users with a user-dependent adaptive weight outperforms any combination with a fixed weight.

Introduction

Personalized web search aims to better account for an individual's information needs than generic web search (Goker & He, 2003; Liu & Belkin, 2015; Liu & Turtle, 2013; Shapira & Zabar, 2011). It is meant to boost retrieval performance by reranking the results ranked by a generic ranker for a particular user, based on a model of their previous and/or current interests. Personalized web search strategies take into account both static information, for example, the content of documents and queries, and dynamic information, such as user behavior. Static information can reflect the intrinsic similarity between documents and queries but may fail to capture a user's real-time interests, which are more directly reflected by the user's interactions with a search engine. Behavioral information can help search engines tune the ranking strategy to improve result rankings (Agichtein, Brill, & Dumais, 2006).

So far, a notable number of approaches to search personalization have been proposed. The dominant approach primarily focuses on content similarity between query and document (Bennett et al., 2012; Sontag et al., 2012; White, Bennett, & Dumais, 2010). In contrast, user behavior-based

*A preliminary version of this paper was published in the proceedings of SIGIR '14 (Cai, Liang, & de Rijke, 2014). In this extension, we (i) extend the behavioral personalization search model introduced there to deal with queries issued by new users for whom long-term search logs are unavailable; (ii) examine the impact of sparseness on the performance of our model by considering both word-level and query-level modeling, as we find that the word-document relevance matrix is less sparse than the query-document relevance matrix; (iii) investigate the effectiveness of our behavior-based reranking model with and without assuming a uniform distribution of users as users may behave differently; (iv) include more related work and provide a detailed discussion of the experimental results.

Received September 24, 2015; revised January 30, 2016; accepted February 8, 2016

 DOI: 10.1002/asi.23735

personalization has not been as well exploited for the purpose of improving rankings. However, it has been shown that incorporating user behavior information may boost the ranking performance (Agichtein et al., 2006). For instance, click models have been well studied for personalized search (Chuklin, Markov, & de Rijke, 2015), where clicks with a reasonable dwell time on a particular document may suggest that a user favors this result (Xu, Jiang, & Lau, 2011; Yi, Hong, Zhong, Liu, & Rajan, 2014), while it might be nonrelevant for other users. In this study we personalize rankings based on user behavior to address their real-time information needs.

In addition, previous work on document ranking (Kurland & Lee, 2004; Wang et al., 2013) often assumes that several key aspects of users and queries are uniformly distributed. However, users may behave differently, as some of them are very active, frequently producing clicks on documents, while others rarely interact with result pages. Similarly, queries may differ, as some are associated with many documents or are often issued by users, receiving relatively more attention from users than others. So we do not assume that the activity of users and the attractiveness of queries are uniformly distributed. Such factors could affect the reranking of documents; below, we investigate the impact of such assumptions.

Previous research on personalized document retrieval has found that implicitly gathered information such as long-term browser history, query history, and click history can be used to improve the ranking accuracy for a given user (Agichtein et al., 2006; Kim, Hassan, White, & Zitouni, 2014). Thus, in this paper we address the personalized web document reranking task by considering users' search behavior both in the current session as well as in their previous search history. In other words, we generate a personalized search result list based on users' long-term (history-based) and short-term (session-based) search contexts.

We apply Bayesian probabilistic matrix factorization (Salakhutdinov & Mnih, 2008a, BPMF) to estimate the relevance of a document to a query and to predict the user's preference for a document, based on their dwell time on clicked documents rather than the content of queries and documents. We begin with a probabilistic graphical model to build the relationships between a user, an issued query, and a document to be reranked, and then define a reranking criterion for documents conditional on a given query and a user. We combine users' short- and long-term behaviors in a linear fashion, and adaptively merge information inferred from the behavior of a specific user and information inferred from the behaviors of other users. We answer the following research questions:

RQ1: Does the combination of document and user information as expressed by their short-term and long-term behavior help improve the document ranking performance?
RQ2: What is the effect on document reranking of the uniform assumption of the distribution of users and queries in our proposed personalized reranking model?
RQ3: Does sparseness of the relevance matrix affect document reranking? That is, do we see a difference in performance between using the query-document matrix and using the word-document matrix?
RQ4: For the document reranking task, what is the impact on the performance of the trade-off between the contribution of the current searcher vs. that of other users?

We demonstrate the effectiveness of our approach to personalized document reranking by using a real-world data set that was made available as part of the Web Search Click Data workshop (at WSDM 2014).[1] We find that combining a user's short- and long-term behaviors achieves higher ranking scores than baselines for document reranking. In addition, merging information inferred from behavior of a particular user and information inferred from behaviors of other users with a user-dependent adaptive weight outperforms any combinations of these two parts with a fixed value. Our contributions can be summarized as follows:

1. We propose an adaptive personalized reranking model that considers a user's short- and long-term interaction behaviors in which the relevance of a document to a query and the preference of a user for certain documents are adaptively combined for document reranking.
2. We perform an investigation of the document ranking performance affected by the assumption that user activity and query attractiveness are uniformly distributed and find that it does have an effect on personalized document reranking.
3. We examine how ranking performance is impacted by sparseness when estimating document relevance to a given query, and find that our model works better by incorporating a word-level relevance matrix that is less sparse than a query-level relevance matrix.

Next, we describe the related work and the details of our proposed personalized reranking model, which is followed by our experimental setup and the experimental results. Finally, we conclude our work and suggest a number of future research directions.

Related Work

In this section we summarize related work on personalized web search, where information from a user's search context in the current session and from their long-term historical search behaviors is exploited for document ranking.

Short-Term Search Context-Based Personalization

Most modern search engines return their results by employing information not only about the query itself but also about the user's preferences as expressed in their current search context, which has the potential to significantly improve the ranking quality (Bennett et al., 2012).

The search context could be user behavior-related, for example, the clicks and dwell time, which conveys a strong

[1]https://www.kaggle.com/c/yandex-personalized-web-search-challenge.

DOI: 10.1002/asi

signal for modeling a user's recent interests. Liu, White, and Dumais (2010) model the dwell time using a Weibull distribution and bring a new approach to analyzing implicit feedback for personalization. Moreover, dwell time distributions can be predicted reasonably well based on low-level page features, which broadens the possible applications to personalization where log data are unavailable. In addition, dwell time may relate to judging the result relevance to a query. Collins-Thompson, Bennett, White, de la Chica, and Sontag (2011) verified that reading-related user behavior features can provide a valuable relevance signal for personalized web search. These publications motivate us to focus on user behavior for web search.

Regarding click-related user behavior, Bilenko and White (2008) propose heuristic algorithms for identifying relevant websites using the combined history of queries and clicks of many web users. Jiang, Leung, and Ng (2011) optimize the search results towards each user's preferences by using search contexts to facilitate concept-based search personalization. They capture a user's preference in the form of concepts obtained by mining clicked web search results. Shen, Tan, and Zhai (2005a) propose several context-sensitive retrieval algorithms based on statistical language models that combine the preceding queries and clicked documents in a session with the current query to specify the actual information need of a searcher. Following Shen et al., Xiang et al. (2010) adopt a learning-to-rank framework and devise a short-term personalization ranking model by encoding context information as features of the model. In addition, user clicks on the results page in a session can be embedded into a framework for automatically weighting the relevance labels (i.e., confidence levels) of query-document pairs (Ustinovskiy, Gusev, & Serdyukov, 2015); these weights are further utilized in personalized web search. Such approaches incorporate click information with content-based search context for personalization, which differs from our approach, which solely explores user behavior information for personalized web search.

Another type of short-term personalization refers to content-based search context. Shen et al. (2005b) infer a user's interest from their search context for personalized search by introducing an intelligent client-side agent that uses a related preceding query and its corresponding search results to select appropriate terms to expand the current query. Ustinovskiy and Serdyukov (2013) restrict their attention to the set of initial queries of search sessions. They employ short-term browsing contents to enrich the current query, which has the largest potential for improvement on single-word queries. Moreover, Mihalkova and Mooney (2009) exploit relations of the current search session to previous similarly short sessions of other users in order to disambiguate the current search query. White et al. (2010) recover a user's short-term interests using both browsing and search context, where an optimal weight can be assigned to the context in order to combine it with the current query. Other content relates to location information. Bennett, Radlinski, White, and Yilmaz (2011) compute location information by using implicit user behavior and characterizing the most location-centric pages for personalization. They find that a substantial fraction of queries can be significantly improved by incorporating location-based features.

So far, a user's long-term behavior has not been taken into account to infer their preference. However, a user's interests expressed by their long-term historical behavior, that is, queries and clicks, can be exploited for differentiating between particular users' requests, which is what we will pursue in this paper.

Long-Term Search History-Based Personalization

Ever since commercial search engines started to record user activities in browsers, long-term browsing logs have been used extensively to create a precise picture of the information needs of users. Such data have become an important resource for search personalization (Dou, Song, & Wen, 2007; Sontag et al., 2012).

The first type of long-term personalization relates to a user's profile. Chirita, Nejdl, Paiu, and Kohlschütter (2005) personalize search by introducing an additional criterion for webpage ranking, namely, the distance between a user profile defined using ODP topics[2] and the set of ODP topics covered by each URL returned in regular web search. Bilenko and Richardson (2011) describe a learning-driven client-side personalization approach for advertising platforms, which relies on storing user-specific information entirely within the user's control as user's specific profile.

Another type of long-term personalization refers to a user's interactions. Tan, Shen, and Zhai (2006) exploit a user's long-term search history, for example, past queries, returned documents, and clickthrough rates, and conclude that recent history is more important than distant history, especially for recurring queries. Matthijs and Radlinski (2011) present a personalization approach that uses a set of features extracted from a user's long-term browsing history and then use this model to rerank web search results. Sontag et al. (2012) use a user's long-term search history to tune parameters of a user-specific ranking model. White, Bailey, and Chen (2009) present a systematic study of the effectiveness of different sources of contextual information for user interest modeling, for example, social, historic, and user interaction. The interest models are required to predict short-, medium-, and long-term user interests. In these models, a user's historical search context is used for modeling their particular preference. However, a user's in-session interests, that is, the short-term context, are not exploited.

Since both a user's short- and long-term search logs can be recorded, it is natural to combine them for personalization purposes. However, little is known about how these behaviors interact and when we should be leveraging them separately or in combination. Bennett et al. (2012) demonstrate that allowing the ranker to learn weights for short-term features, long-term features, and their combination can model a

[2]http://www.dmoz.org

DOI: 10.1002/asi

searcher's interests more effectively than leaving out one of these ingredients is able to achieve. In addition, the preference of a group of similar users has been studied by mining their long-term history (Pan & Chen, 2013), where the interests of a current user and of other close users are considered. Yan, Chu, and White (2014) propose to use so-called cohort modeling to enhance search personalization based on three predefined cohorts, namely, topic, location, and top-level domain preference by mining search logs. Teevan, Dumais, and Liebling (2008) show that there is a lot of variation across queries in the benefits that can be achieved through personalization. Hence, they characterize queries using a variety of features, of the query, of its returned results, and of people's interaction history with the query.

Generally, most previous work on personalization treats all queries and users equally, which means that queries are assumed to be issued uniformly regardless of their popularity and that users behave similarly in spite of their activity levels. However, some queries are more popular than others and some may be associated with many documents. Also, some users may behave actively when interacting with a search engine, producing a relative number of clicks compared to others. These factors are not well studied and we hypothesize that they may affect the document ranking. Hence, our models capture such aspects, allowing us to examine their impact on personalization.

Approach

In this section, we formally describe the problem of document reranking studied in this paper, which is followed by a section discussing our personalized document reranking model and by sections related to some practical issues.

Problem Formulation

First of all, we describe the task of *Document Reranking* (DRR), which we study in this paper. Assume that the following are given: (i) a search session with T queries $\{q_1, q_2, \ldots, q_T\}$ of a user u, where each query consists of a sequence of words, for example, $q_i=(w_{i1}, w_{i2}, \ldots, w_{im})$, $i=1, 2, \ldots, T$; and (ii) the list of top N documents to be reranked, that is, $(d_{T1}, d_{T2}, \ldots, d_{TN})$, which are returned by a search engine in response to the last query q_T in the session. Then the purpose of DRR is to return a reranked list of these N documents to the user u, where their previous short- and long-term search behavior, for example, clicks and dwell time, may be available. Obviously, for test sessions consisting of only one query, no short-term search behavior is available.

The main notations used in the paper are listed in Table 1. In the DRR task, the relationships between the *user*, the *query*, and the *URLs*[3] to be reranked can be simply modeled by a graphical model, as shown in Figure 1. The user u submits a query q, in response to which the top N URLs are returned by a search engine. Hence, our purpose is to rerank the top N URLs in response to this query. Because the variables u and q are known before document reranking, we can estimate the relevance of a document d to a query q issued by a user u using a conditional probability $P(d|q, u)$, based on which a reranked list of the top N URLs is generated as output.

[3]We use *URL* as if it were interchangeable with *document*, represented by d.

TABLE 1. Main notations used in this paper.

Notation	Description
λ	a trade-off parameter controlling the contributions of the document and of behavioral information of the user for reranking
ω	a trade-off parameter controlling the contributions of a user's short- and long-term behaviors for individual preference
T	the number of queries in a session
w	query term (word)
q	a query consisting of a sequence of m words, i.e., $\{w_1, w_2, \ldots, w_m\}$
u	a user
N	the number of documents to be reranked given a query
D	the set of documents to be reranked given a query, i.e., $\{d_1, d_2, \ldots, d_N\}$
d	a document in D
k_f	number of latent features in BPMF

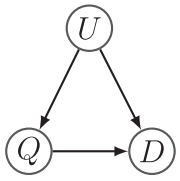

FIG. 1. Probabilistic graphical model indicating the relationships between user u, query q, and document d.

Personalized Document Reranking Model

From the graphical model shown in Figure 1, we can interpolate the joint probability $p(u, q, d)$ as follows:

$$p(u, q, d)=p(u) \cdot p(q|u) \cdot p(d|q, u). \quad (1)$$

Then, the relevance of document d to query q given user u, that is, $p(d|q, u)$, can be computed as:

$$p(d|q, u)=\frac{p(q, u, d)}{p(u) \cdot p(q|u)}=\frac{p(q, u|d) \cdot p(d)}{p(u) \cdot p(q|u)}, \quad (2)$$

based on Bayes's rule.

Obviously, for the document reranking task studied in this paper, given a user u and a query q, the probabilities $p(u)$ and $p(q|u)$ in Equation (2) do not affect the reranking of documents, which results in:

DOI: 10.1002/asi

$$p(d|q,u) \propto p(q,u|d) \cdot p(d). \quad (3)$$

Moreover, to estimate the probability $p(q,u|d)$, following Kurland and Lee (2004), we use a linear mixture governed by a free trade-off parameter λ as follows:

$$p(q,u|d) = (1-\lambda) \cdot p(q|d) + \lambda \cdot p(u|d), \quad (4)$$

where $p(q|d)$ denotes the relevance of document d to query q, and $p(u|d)$ reflects user u's preference for document d. Hence, based on Equations (3) and (4), we have:

$$p(d|q,u) \propto p(d) \cdot ((1-\lambda) \cdot p(q|d) + \lambda \cdot p(u|d)) \quad (5)$$

$$= (1-\lambda) \cdot p(d,q) + \lambda \cdot p(d,u). \quad (6)$$

For the DRR task, because the query and the user are known before document reranking, it makes sense to represent the ranking score $p(d|q,u)$ under the condition of a given q or u. Thus, based on Equation (6), we have:

$$p(d|q,u) \propto (1-\lambda) \cdot p(d|q) \cdot p(q) + \lambda \cdot p(d|u) \cdot p(u). \quad (7)$$

For the simplest case, if we further assume $p(q)$ and $p(u)$ to be uniform, based on Equation (7), $p(d|q,u)$ can be directly estimated by:

$$p(d|q,u) \propto (1-\lambda) \cdot p(d|q) + \lambda \cdot p(d|u). \quad (8)$$

where the probabilities $p(d|q)$, $p(d|u)$, $p(q)$ and $p(u)$ are to be estimated from the training data using Bayesian probabilistic matrix factorization (BPMF).

In practice, for calculating $p(d|q)$ we aggregate the behavioral information from other users, and then implement BPMF at two levels, that is, at the query-level and at the word-level, to alleviate the sparseness problem. We argue that the word-document relevance matrix (which is at the word-level) should be denser than the query-document relevance matrix (which is at the query-level). To verify this assumption we define the *sparseness* of an $m \times n$ matrix $X_{m \times n}$ as:

$$sparseness(X_{m \times n}) = \left(1 - \frac{\sum_i \sum_j \phi(x_{ij})}{m \times n}\right) \times 100\%, \quad (9)$$

where

$$\phi(x_{ij}) = \begin{cases} 1, \text{ if } x_{ij} > 1 \\ 0, \text{ otherwise,} \end{cases} \quad (10)$$

and x_{ij} is one of the entries of the input matrix X with m rows and n columns. We find that in our experimental setup, the sparseness of the query-document relevance matrix is $\sim$71%, while the sparseness of the word-document relevance matrix is $\sim$43%.

At the query-level, $p(d|q)$ can be directly computed by implementing BPMF on a query-document relevance matrix. In contrast, at the word-level, based on the assumption that the words in a query are independent of each other (Zhai & Lafferty, 2004), $p(d|q)$ can be obtained by:

$$p(d|q) = p(d|w_1, w_2, \ldots, w_m) = \prod_{w_i \in q} p(d|w_i)^{N(w_i,q)}, \quad (11)$$

where $N(w_i, q)$ is the number of words w_i in query q, and $p(d|w_i)$ Tab is similarly computed by implementing BPMF on a word-document relevance matrix.

Regarding $p(d|u)$, we first implement BPMF on a user-document preference matrix directly to obtain the distribution of user preference over documents. However, to generate the final value of $p(d|u)$ in Equation (8) for document reranking, the short- and long-term behaviors of a specific user u are considered as a linear combination as follows:

$$p(d|u) = (1-\omega) \cdot p(d|u)_s + \omega \cdot p(d|u)_l, \quad (12)$$

as suggested by Bennett et al. (2012), where $p(d|u)_s$ indicates user u's current preference for document d expressed by his previous behaviors in the session and $p(d|u)_l$ denotes user u's overall interests in document d expressed by his long-term behaviors in the training period.

In sum, with a query-level BPMF, the final reranking criterion is:

$$p(d|q,u) \propto (1-\lambda) \cdot p(d|q) + \lambda \cdot ((1-\omega) \cdot p(d|u)_s + \omega \cdot p(d|u)_l) \quad (13)$$

if users and queries are assumed to be uniformly distributed. Otherwise, meaning that we do not assume users and queries to be uniformly distributed, we have:

$$p(d|q,u) \propto \propto (1-\lambda) \cdot p(d|q) \cdot p(q) + \lambda \cdot ((1-\omega) \cdot p(d|u)_s + \omega \cdot p(d|u)_l) \cdot p(u). \quad (14)$$

Let us turn to word-level BPMF, we would conclude:

$$p(d|q,u) \propto \propto (1-\lambda) \cdot \prod_{w_i \in q} p(d|w_i)^{N(w_i,q)} + \lambda \cdot (1-\omega) \cdot p(d|u)_s + \omega \cdot p(d|u)_l \quad (15)$$

in case we assume users and queries to be uniformly distributed, and have:

$$p(d|q,u) \propto (1-\lambda) \cdot \prod_{w_i \in q} p(d|w_i)^{N(w_i,q)} \prod_{w_i \in q} p(w_i) + \ldots \ldots + \lambda \cdot ((1-\omega) \cdot p(d|u)_s + \omega \cdot p(d|u)_l) \cdot p(u), \quad (16)$$

in case we do not assume users and queries to be uniformly distributed. The way in which we estimate $p(d|u)_s$ and $p(d|u)_l$ is described later.

DOI: 10.1002/asi

Additionally, in practice, we have to face a *user cold-start problem*, that is, how to calculate the probability $p(d|u)$ if the test query is issued by a new user in the test period for whom no long-term search history is available from the training period. We tackle the user cold-start problem by finding the most similar users who submitted the same query in the training period. That is, given a test query q issued by a new user u, we first generate a set U_c of users who submitted the same query q in the training period by a function $\psi_u(q)$, that is, $U_c \leftarrow \psi_u(q)$, then select the optimal user candidate by:

$$u^* \leftarrow \arg\max_{u_i \in U_c} \phi(u_i, q), \tag{17}$$

where $\phi(u_i, q)$ returns the frequency of query q issued by user u_i. Notice that if more than one user submitted the same query q with the highest frequency when generating the optimal u^*, we select the user who has the most clicks. We do not use a clustering algorithm to find similar users for a new user because in our setting limited information of new users is available. We resort to their short-term search context in the current session to find similar users. Alternatively, we directly find similar users through the submitted queries that have been issued before by others, that is, via a query-user bipartite graph. This approach has previously been used to infer user's personal search interest (Dou et al., 2007; Volkovs, 2014). We address the user cold-start problem by finding the most similar users who submitted the same query in the training period.

Smoothing by Bayesian Probabilistic Matrix Factorization

In this study we use Bayesian probabilistic matrix factorization (Salakhutdinov & Mnih, 2008a, BPMF) to predict the relevance of a document to a query as well as the preference of a user for a document. Taking the former, for instance, we first take the logarithm of the aggregated dwell time of known query-document pairs to dampen sharp peaks and then label the relevance for this pair by:

$$\min\left(\lfloor \lg(t+10) \rfloor, 5\right),$$

where t is the aggregated dwell time, and $\lfloor \cdot \rfloor$ is the floor function. In this manner, a query-document relevance matrix R_{QD} can be built with each entry indicating the relevance of the corresponding document to a query. BPMF is then applied to this query-document relevance matrix R_{QD} to assign a nonzero value to each entry in the original matrix. This completes our smoothing method.

Thus, by applying BPMF we replace the original query-document relevance matrix R_{QD} by an approximation R^*_{QD} as:

$$R^*_{QD} = Q^*_{N_q \times k_f} \times {D^*_{M_d \times k_f}}^{\top}, \tag{18}$$

where $Q^*_{N_q \times k_f}$ and ${D^*_{M_d \times k_q}}^{\top}$ represent the query-specific and document-specific latent feature matrix, respectively, and N_q, M_d, and k_f indicate the number of queries, documents, and latent features, respectively.

The distribution of the values $R^*_{QD}(i,j)$ for query i and document j is computed by marginalizing over the model parameters and the hyperparameters:

$$\begin{aligned} &p(R^*_{QD}(i,j)|R_{QD}, \Theta_0) = \\ &= \int\int p(R^*_{QD}(i,j)|Q_i, D_j) \cdot p(Q, D|R_{QD}, \Theta_Q, \Theta_D) \cdot \\ &\quad \cdot p(\Theta_Q, \Theta_D|\Theta_0) \cdot d\{Q, D\} \cdot d\{\Theta_Q, \Theta_D\}, \end{aligned} \tag{19}$$

where $\Theta_Q = \{\mu_Q, \Sigma_Q\}$ and $\Theta_D = \{\mu_D, \Sigma_D\}$ are query and document hyperparameters; the prior distribution vectors over the queries and documents are assumed to be Gaussian; and $\Theta_0 = \{\mu_0, \Sigma_0, W_0\}$ is a Wishart distribution hyperparameter with $\Sigma_0 \times \Sigma_0$ scale matrix W_0 (Salakhutdinov & Mnih, 2008a). The intuition beyond this approximation is that the relevance of a query to a document is determined by a small number of unobserved hyperparameters. This means that taking a Bayesian approach to the prediction problem involves integrating the model hyperparameters. In addition, the use of Markov chain Monte Carlo (MCMC) methods (Neal, 1993) for approximating relevance comes from finding only point estimates of model hyperparameters instead of inferring the full posterior distribution over them, which results in a significant increase in predictive accuracy (Salakhutdinov & Mnih, 2008a). BPMF introduces priors for the hyperparameters, which allows the model complexity to be controlled automatically based on the training data (Salakhutdinov & Mnih, 2008b). In addition, as the priors are assumed to be Gaussian, the hyperparameters can be updated by performing a single step of EM, which finally results in a complexity of $O(N_q + M_d)$ if the Gibbs sampling count is small (Salakhutdinov & Mnih, 2008b).

Here we present a more detailed description of a single-step EM algorithm. Suppose we have the representation of query-document samples by a relevance matrix R_{QD} consisting of N_q queries and M_d documents. We fit the parameters of a model $p(R_{QD}, z)$ to the data, where the z's are latent random variables. The likelihood is given by:

$$l(\theta) = \sum_{i=1}^{N_q} \log p(R_{QD}(i); \theta) = \sum_{i=1}^{N_q} \log \sum_{z_i} p(R_{QD}(i), z_i; \theta),$$

where $\theta = \{\Theta_Q, \Theta_D\}$ are the parameters. As explicitly finding the maximum likelihood estimates of the parameters θ may be hard (Neal & Hinton, 1999), the EM algorithm gives an efficient method for maximum likelihood estimation by repeatedly constructing a lower-bound on $l(\theta)$ (E-step) and then optimizing that lower-bound (M-step). For each i, given a preassumed Gaussian distribution $\phi(z_i)$, we have:

$$\begin{aligned} \sum_i \log p(R_{QD}(i); \theta) &= \sum_i \log \sum_{z_i} \phi(z_i) \frac{p(R_{QD}(i), z_i; \theta)}{\phi(z_i)} \\ &\geq \sum_i \sum_{z_i} \phi(z_i) \log \frac{p(R_{QD}(i), z_i; \theta)}{\phi(z_i)}, \end{aligned}$$

DOI: 10.1002/asi

based on Jensen's inequality (Farenick & Zhou, 2007). The E-step is:

$$\phi_i(z_i) := p(z_i|R_{QD}(i) : \theta)$$

and the M-step is:

$$\theta := \arg\max_{\theta} \sum_i \sum_{z_i} \phi_i(z_i) \log \frac{p(R_{QD}(i), z_i; \theta)}{\phi_i(z_i)}$$

Together, these give us the maximum likelihood.

By a similar strategy, we can generate the approximations R^*_{WD} and R^*_{UD} corresponding to the original matrix R_{WD} (word-document relevance matrix) and R_{UD} (user-document preference matrix) by marginalizing over the respective model parameters and hyperparameters. In essence, BPMF is used to address the problem of sparseness when inferring the relevance distribution of documents to queries. Other methods, such as a probabilistic matrix factorization (PMF) approach in Salakhutdinov and Mnih (2008b), could also deal with this sparseness issue. However, we incorporate BPMF in our approach because BPMF can be successfully applied to large data sets and achieves significantly higher predictive accuracy than PMF models (Salakhutdinov & Mnih, 2008a).

Modeling Behavior

In the DRR task, short-term behavior, more specifically, the clicks on documents returned in response to previous queries in the current session, may provide a strong signal of a user's current interest. We aggregate the contributions of all clicked URLs in a current session to compute the probability $p(d|u)_s$ mentioned in Equation (13) and Equation (15) as follows:

$$p(d|u)_s = \sum_{d_i \in D_s} \omega_i \cdot p(d_i|u), \tag{20}$$

where D_s is the set of clicked URLs in current search session and

$$\omega_i = \frac{1}{Z_\omega} \cdot \frac{\sum_{d_j \in D_s \setminus \{d_i\}} Dis(d_j, d)}{\sum_{d_k \in D_s} Dis(d_k, d)}$$

depends on the similarity between the clicked document d_i and the document d to be reranked, where:

$$Z_\omega = \sum_{d_i \in D_s} \frac{\sum_{d_j \in D_s \setminus \{d_i\}} Dis(d_j, d)}{\sum_{d_k \in D_s} Dis(d_k, d)}$$

is a normalization factor. Furthermore, $D_s \setminus \{d_i\}$ denotes the subset of D_s excluding d_i and $Dis(d_j, d)$ returns the Euclidean distance between d_i and d; documents are represented by vectors returned by the BPMF process on a user-document preference matrix.

Similarly, for the long-term behavior of user u, we estimate the probability $p(d|u)_l$ as follows:

$$p(d|u)_l = \sum_{d_j \in D} p(d_j|u)^{c(d_j,u)}, \tag{21}$$

where $c(d_j, u)$ indicates the number of clicks of user u on document d_j. The probability $p(d|u)$ can be returned by the BPMF process on a user-document preference matrix.

Adaptive Weighting

Previous work (White et al., 2010) uses a fixed weight λ in Equation (8), that is, the same trade-off for all test users when combining the contributions from the user and from a specific document. This choice shows good reranking results in the setting discussed by White et al. (2010). However, we treat the weight differently, as different users behave differently. We propose an adaptive weight solution to assign a specific weight λ in Equation (8) to each user u in the test period, which depends on the users in the training period who are similar to u.

First, each user u in the training period should be assigned an optimal λ^*_u, which is obtained by:

$$\lambda^*_u = \arg\max_{\lambda \in [0,1]} MAP(u, \lambda), \tag{22}$$

where the function $MAP(u, \lambda)$ returns the mean average precision value of all training queries from user u when changing λ from 0 to 1 with step-size 0.1. Then, in the test phase, for a seen user, we directly use this value λ^*_u in Equation (22) to calculate the final ranking score based on Equation (8). However, for an unseen user u, we first find their closest user u^* by Equation (17) and then select a group G of N_u users who are most similar to u^*. Finally, we assign an adaptive weight:

$$\lambda'_u = \sum_{u_i \in G} \alpha_i \cdot \lambda^*_{u_i} \tag{23}$$

to this unseen user u, where:

$$\alpha_i = \frac{1}{Z_\alpha} \cdot \frac{\sum_{u_j \in G \setminus \{u_i\}} Dis(u_j, u^*)}{\sum_{u_k \in G} Dis(u_k, u^*)}$$

depends on the similarity between u_i in group G and u^*, while:

$$Z_\alpha = \sum_{u_i \in G} \frac{\sum_{u_j \in G \setminus \{u_i\}} Dis(u_j, u^*)}{\sum_{u_k \in G} Dis(u_k, u^*)}$$

is a normalization factor, and $\lambda^*_{u_i}$ is the optimal weight of u_i learnt from the training period. Again, users are represented by the latent vectors returned by applying BPMF on a user-document preference matrix.

DOI: 10.1002/asi

In the next section we examine the effectiveness of our behavior-based document reranking method on a real-world query log data set and compare it to previously proposed methods.

Experimental Setup

In this section, we describe the data set used in our experiments, the metrics and baselines used for comparisons, as well as the experimental settings.

Data Set

The data set for this study consists of anonymous logs of users provided by the Personalized Web Search Challenge.[4] This challenge is a part of the Web Search Click Data workshop (WSCD 2014),[5] in which participants are required to rerank documents of each SERP originally returned by the search engine according to the personal preferences of each user. The evaluation relies on clicks and dwell time, which has been widely used in state-of the-art research on search personalization (Xu et al., 2011; Yi et al., 2014). The logs, collected for 1 month, contain a unique user identifier, a search session identifier, a query identifier, the top-10 URLs returned by the search engine for that query, and the dwell time on clicked results. The training period covers the first 27 days in the data set and the last 3 days constitute the test period. Users with more than 20 queries during the training span are kept in order to have a rich long-term search history of users. In addition, only sessions with multiple queries are kept for testing in order to have access to the short-term behavior of users in our test set. All test queries are required to have been issued in the training period to obtain the ground truth (relevance of documents to queries). Table 2 lists the statistics of the processed data set.

The ground truth we use is obtained as follows. To determine the relevance of a document to a query, we aggregate all dwell times on the document given this query. Figure 2 shows that the majority of the logarithmic aggregated dwell time on documents given a query is smaller than 6. Thus, we generate relevance labels by:

$$relevance \leftarrow \min\left(\lfloor \lg\left(t+10\right)\rfloor, 5\right),$$

where t is the aggregated dwell time, and $\lfloor\cdot\rfloor$ is the floor function. By doing so, we assign a positive judgment 5 (highly relevant) grade to clicked documents given a query whose logarithmic dwell time units are not shorter than 5. We also assign a grade 4, 3, and 2 (relevant, normally relevant, and slightly relevant, respectively) corresponding to documents with clicks whose logarithmic dwell time units are between 4, 3, 2, respectively, and a grade equal to 1 (irrelevant) to documents with no clicks or clicks whose logarithmic dwell time units are strictly less than 1.

[4]https://www.kaggle.com/c/yandex-personalized-web-search-challenge.
[5]http://research.microsoft.com/en-us/um/people/nickcr/wscd2014/

TABLE 2. Statistics of the processed data set.

Variable	Training	Test
# log records	6,113,430	1,773,167
# queries	449,079	119,328
# unique queries	69,597	69,597
# query terms	318,253	318,253
# unique query terms	43,162	43,162
# documents	5,215,272	1,354,184
# unique documents	231,671	126,756
# unique users	168,863	89,328

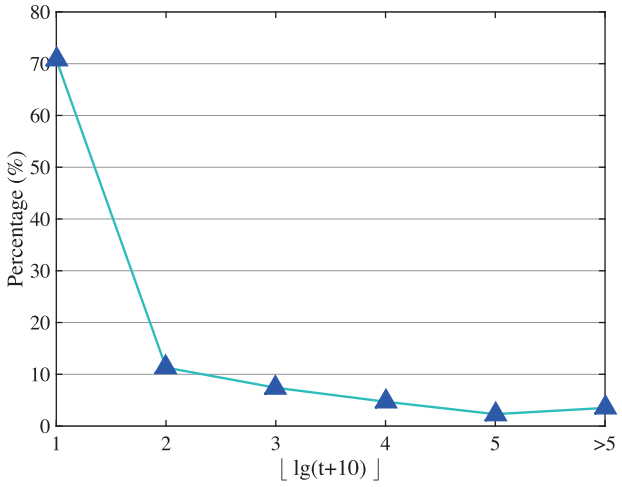

FIG. 2. Ratio of logarithmic aggregated dwell time on documents. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Using dwell time on clicked documents to infer their relevance to a given query is widely used in web search (e.g., Cai, Liang, & de Rijke, 2014; Kim et al., 2014; Xu et al., 2011; Yi et al., 2014), where typically the relevance label is assigned based on a fixed time span, for example, 30 seconds. However, in our approach we use an aggregated dwell time on clicked documents to indicate the relevance to a given query because this scheme is based on long-term observation and could be insensitive to noise. In other words, our scenario takes the dwell time from a large population of users and could reflect the intrinsic relevance of a document to a given query. We first take the logarithm of the aggregated dwell time of a known query-document pair to dampen sharp peaks and then label the relevance for this pair; this approach has also been used previously (Cai et al., 2014).

Metrics

For evaluation, we report our performance in terms of MAP, P@5, NDCG@5, and NDCG@10; we use the trec_eval script for ranking evaluation as provided by TREC.[6]

Statistical significance of observed differences between the performance of two approaches is tested using a two-tail

[6]http://trec.nist.gov/trec_eval/.

DOI: 10.1002/asi

paired t-test and is denoted using ▲/▼ for significant differences at level $\alpha = .01$, or △/▽ at level $\alpha = .05$.

Models Used for Comparison

We write ComP to refer to our proposed reranking model with a fixed value of λ (see Equation [8]) and use aComP to denote our proposal with an adaptive value of λ, with the adaption governed by Equation (22) and Equation (23), respectively.

We compare the results generated by our proposed models to: (i) those originally returned by the search engine, denoted SE; (ii) those reranked by the current user's personal preference (Rendle, Freudenthaler, Gantner, & Schmidt-Thieme, 2009), denoted BPR; (iii) those reranked by the preference of a group of similar users and of the current user, denoted as GBPR, where following (Pan & Chen, 2013), we set the tradeoff $\rho = 0.6$, which means 60% of the contribution to the user's preference for a document is from the user group and the rest is from the current user. In addition, the size of the group is set to 5 because under these settings GBPR works relatively better than others. Table 3 provides an overview of the reranking models used in the paper.

Reports from the three winners in the Kaggle Challenge are available. The winners all focus on feature engineering rather than ranking algorithms. In addition, because of limited space, the participants did not provide the details of the features used nor how they are computed (Masurel, Lefevre-Hasegawa, Bourguignat, & Scordia, 2014; Song, 2014). Some features descriptions are totally missing, such as the session features in Masurel et al. (2014). In Volkovs (2014), although the formulas of most used features are provided, how to generate the final ranking score is missing: The approach is based on an aggregated model, which incorporates several learning models with model-specific weights. (The participants pointed out that their paper should not be considered as proper research [Masurel et al.] but as an engineering report). For the reasons just given, it is hard to reproduce the top-performing methods. Instead, we select other alternatively related methods dealing with similar tasks as our baselines to compare (e.g., Pan & Chen, 2013; Rendle et al., 2009). These two baseline approaches rank documents based on user behaviors rather than on the exact content of documents and queries, which is similar to our scheme used in this paper, that is, user behavior-based personalization. In addition, these two approaches both consider personalization for document ranking and achieve good performance. Hence, they can be taken as state-of-the-art approaches that rank documents only by considering user behavior.

Settings and Parameters

The search engine initially returns a list of the top-10 URLs, that is, $N = 10$; we use a fixed $\lambda = 0.5$ in our ComP model to answer RQ1–RQ3 and then study its impact on the document ranking for RQ4. Following Cai et al. (2014) and Salakhutdinov and Mnih (2008b), the number of latent features in BPMF is set to 10, that is, $k_f = 10$. In order to derive a user's long-term preference, that is, $p(d|u)_l$ in Equation (21), we use at most 10 unique documents clicked by user u. For an adaptive λ'_u in Equation (23), following Salakhutdinov and Mnih (2008a), we select a group of similar users with $N_u = 5$, that is, $|G|=5$.

TABLE 3. An overview of reranking models compared in this paper.

Model	Description	Source
SE	The original ranking.	Search engine
BPR	User preference Bayesian personalized ranking.	(Rendle et al., 2009)
GBPR	Group preference based Bayesian personalized ranking.	(Pan & Chen, 2013)
ComP	Personalization method reranks results of SE with a fixed $\lambda = 0.5$ in Equation (8).	This paper.
aComP	Personalization method reranks results of SE with an adaptive λ in Equation (8).	This paper.

TABLE 4. A sanity check to determine an optimal value of ω for later experiments by varying ω from 0 to 1 manually.

ω	MAP	p@5	NDCG@5	NDCG@10
0.0	.4018	.2814	.3756	.5013
0.1	.4153	.2956	.3812	.5108
0.2	.4186	.3014	.3867	.5167
0.3	**.4207**	**.3082**	**.3921**	**.5208**
0.4	.4201	.3069	.3894	.5183
0.5	.4204	.3075	.3903	.5192
0.6	.4183	.3011	.3855	.5154
0.7	.4172	.2981	.3824	.5138
0.8	.4144	.2926	.3794	.5086
0.9	.4137	.2918	.3767	.5042
1.0	.4003	.2789	.3723	.4980

Note. The best performer in each column is boldfaced. (Settings: query-level BPMF and uniform assumption of user.)

In addition, from previous work (Bennett et al., 2012; Cai et al., 2014), we know that the parameter controlling the contributions of a user's short-term and long-term behavior when estimating their preference for a document does affect the document ranking performance. Hence, as a sanity check, we report the results of our ComP model in Table 4 when ω varies from 0 to 1 with steps of 0.1 to select an optimal value of ω for later experiments. It is clear from Table 4 that the performance shows very minor differences when the weight ω in Equation (13) changes; it reaches a peak at $\omega = 0.3$. In addition, a small value of ω, for example, $\omega = -0.2$, often results in better performance than a large value of ω, for example, $\omega = -0.8$, which is consistent with previous findings (Bennett et al., 2012), where in web search better performance can be achieved when more attention is paid to short-term behavior rather than long-term behavior of a searcher. Therefore, we use $\omega = 0.3$ in our model in later experiments.

Results and Discussion

In this section we compare the results of our models with those of the baselines, and then examine the impact on

DOI: 10.1002/asi

ranking performance of the distribution of users and of the level where BPMF works (queries vs. words), respectively. Finally, we zoom in on the relative importance of user and document for personalized ranking.

Performance of Ranking Models

In this section we compare the performance of various models in Table 3. Our models, ComP and aComP, work based on the uniform assumption of user and on the query-level BPMF. We report the results in Table 5. Notice that the NDCG scores here are somewhat lower than those obtained at the personalized web search challenge[7] after submitting the rankings. This could be due to differences in (i) generating the ground truth or in (ii) processing the data for our experiments.

As we can see from Table 5, among the three baselines the group preference-based personalized ranking approach, that is, GBPR, performs best, indicating that close users may behave similarly in web search. Compared to BPR, where only the preference of the current user rather than a group of similar users is considered for personalization, GBPR reports higher scores in terms of all four metrics, MAP, P@5, NDCG@5, and NDCG@10. Generally, user preference does indeed help to boost the ranking performance. For instance, GBPR and BPR present notably better results than the original ranking of the search engine, that is, SE, showing near 7.2% and 5.4% improvements in terms of MAP over SE, respectively. Below we only use GBPR as a baseline for comparisons with our model.

However, by considering a user's short- and long-term behavior, our ComP model can further improve the performance over GBPR. For instance, ComP achieves a near 3% improvement in terms of MAP over GBPR. The improvements in terms of all four metrics are significant at the $\alpha = .05$ level. We believe that these improvements are due to the following: (i) ComP considers the relevance of a document for a query (see Equation [8]) as well as user preference for document ranking, while GBPR only incorporates the latter; (ii) ComP exploits both a short-term and long-term user behavior, whereas user preference in GBPR is based only on a user's long-term history. Regarding aComP, which optimizes the trade-off λ in Equation (13) for each user, it achieves slight improvements over ComP. In particular, it reports less than 1% improvement over ComP but more than 4% improvement over GBPR in terms of NDCG@5. Similar findings are obtained for the other metrics. It is worth noting that aComP presents a significant improvement over GBPR at level $\alpha = .01$ in terms of NDCG@5, but at level $\alpha = .05$ in terms of the other metrics. In other words, compared to other models, aComP can return the most relevant results early in the ranked list, for example, in the top five.

TABLE 5. Performance comparison.

Method	MAP	p@5	NDCG@5	NDCG@10
SE	.3817	.2829	.3611	.4621
BPR	.4023	.2926	.3731	.4973
GBPR	.4092	.2981	.3778	.5024
ComP	.4207$^{\triangle}$	.3082$^{\triangle}$	.3921$^{\triangle}$	.5208$^{\triangle}$
aComP	**.4243**$^{\triangle}$	**.3107**$^{\triangle}$	**.3955**$^{\blacktriangle}$	**.5226**$^{\triangle}$

Note. The best performer in each column is in boldface. Statistically significant differences between the results of our models and those of the best baselines, which are underlined, are indicated. (Settings used: $\omega = 0.3$.)

In absolute terms, the NDCG scores generated by our models are lower than those produced by the winners of the Kaggle Challenge, from which we obtained the data for our experimental evaluation. For instance, according to the challenge leaderboard,[8] the default ranking baseline achieves an NDCG@10 of 0.7913, while the team winning the third prize in the competition achieves an NDCG@10 of 0.8047. Both NDCG scores are higher than ours (e.g., an NDCG@10 of 0.5226 reported in Table 5). However, the scores are not comparable, for two reasons: (i) In our setup we use 5-point scales to estimate the relevance of a query to a document while the Kaggle Challenge uses 3-point scales. Our setup allows us to make more fine-grained distinctions; (ii) The scenarios for generating the relevance labels are different, as we use the aggregated dwell time for a particular pair of a query and a document from the entire history while the competition uses the interaction in a single session.

Effect of Uniform Distributions

In this section we examine the performance of aComP with and without making the uniform distribution assumption about users and queries, which works by incorporating the query-level BPMF. We report the results in terms of four metrics at various query positions in a session as well as the average scores, in Figure 3.

Generally, as shown in Figure 3, the assumption of uniform distributions of users and queries does affect the performance of aComP, as it performs better without making the assumption, although the differences are not statistically significant. In particular, without the assumption, aComP shows more than 1% improvement in terms of P@5 over aComP with the assumption but less than 1% improvement in terms of other metrics.

It is worth noting that the performance of aComP shows little fluctuation as the query position changes, as shown in Figure 3. However, it achieves its peak performance when the query position is 4. The performance of aComP increases monotonously as the position of test queries increases from 2 to 4. But later, the performance drops slowly. This observation can be explained by the fact that in long search sessions users may change their search intent from their

[7]https://www.kaggle.com/c/yandex-personalized-web-search-challenge/details/evaluation.

[8]https://www.kaggle.com/c/yandex-personalized-web-search-challenge/leaderboard

DOI: 10.1002/asi

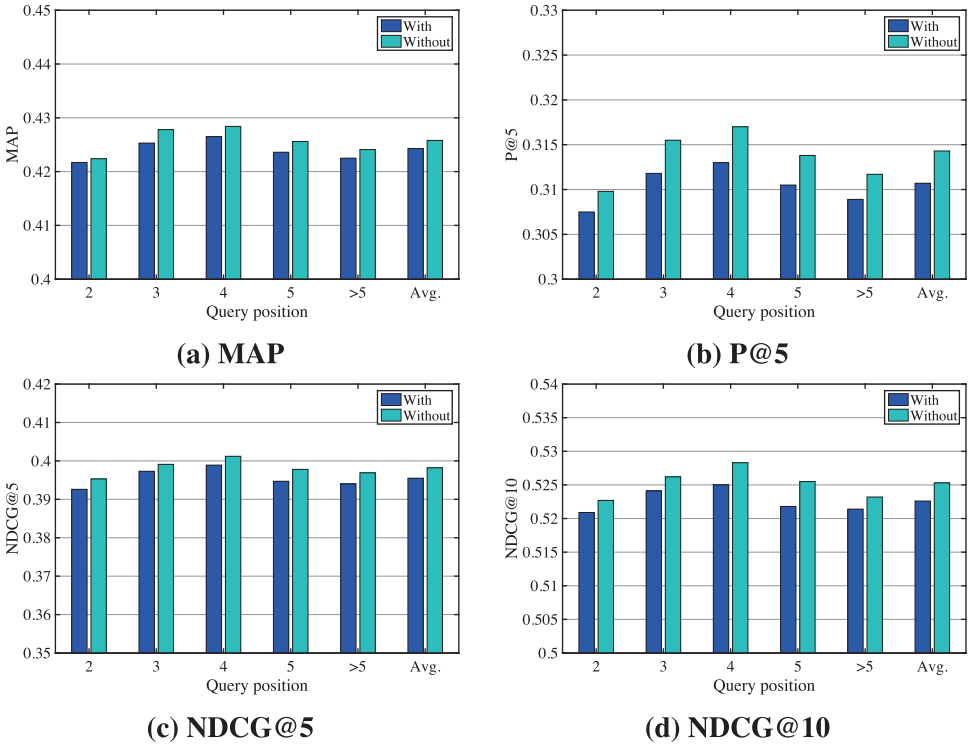

FIG. 3. Performance of the aComP model with and without making the uniform assumption of users and queries, tested at different query positions. (Settings used: $\omega = 0.3$.) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

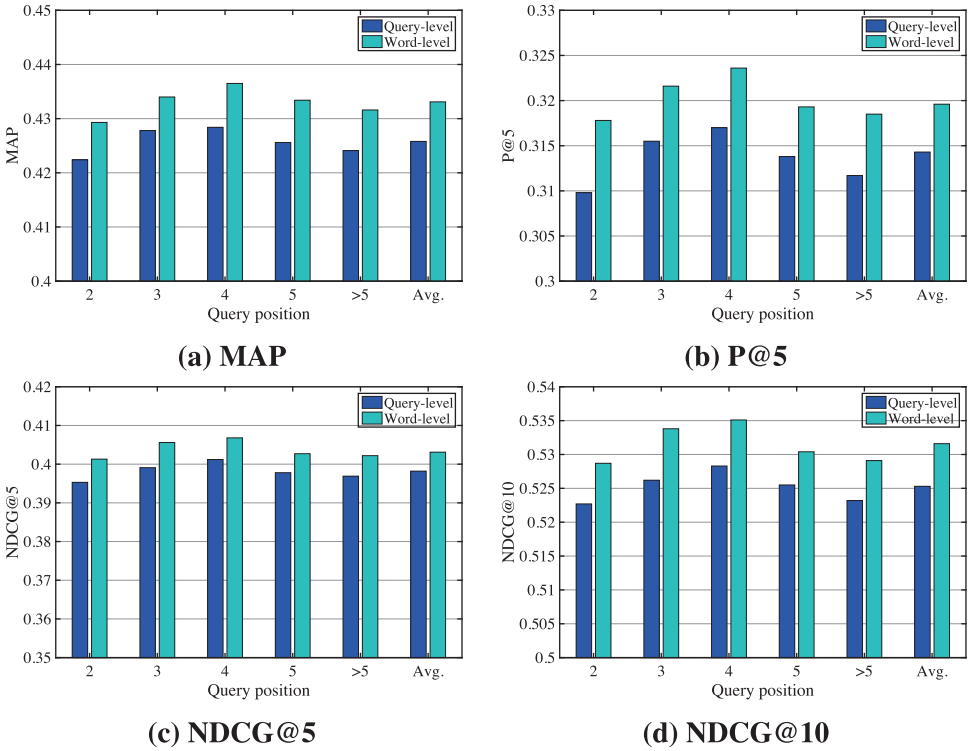

FIG. 4. Performance of the aComP model, with BPMF at the query-level and word-level, tested at different query positions. (Settings used: $\omega = 0.3$.) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

TABLE 6. Performance of the ComP model under different settings of a fixed $\lambda \in (0, 1)$ with step-size 0.1.

Trade-off λ	MAP	P@5	NDCG@5	NDCG@10
$\lambda = 0.1$	.4104	.2951	.3792	.5094
$\lambda = 0.2$	.4130	.2986	.3814	.5117
$\lambda = 0.3$	.4163	.3027	.3851	.5140
$\lambda = 0.4$	.4185	.3053	.3897	.5182
$\lambda = 0.5$	.4207	.3082	.3921	.5208
$\lambda = 0.6$	.4218	.3087	.3923	.5212
$\lambda = 0.7$	**.4232**	**.3095**	**.3946**	**.5217**
$\lambda = 0.8$	.4219	.3091	.3935	.5214
$\lambda = 0.9$	.4211	.3088	.3924	.5186

Note. The result of the best performer is boldfaced. (Settings used: $\omega = 0.3$.)

original ones as their needs have been (partially) addressed by previous queries and clicked documents. However, we still find that aComP works better in long sessions than sessions with only a single previous query, as aComP achieves higher scores at query position 5 (or >5) than at position 2: more rich context in the current search session can help detect a user's intents and generate reasonable rankings.

Zooming in on BPMF at Different Levels

Following the earlier discussion, we run aComP work without the uniform assumption and then implement BPMF at different levels to examine the impact of sparseness on the ranking performance. We report the scores at different query positions (see Figure 4).

Clearly, the problem of sparseness in BPMF affects the performance of aComP. In particular, aComp favors incorporating word-level BPMF, as it performs better than using query-level BPMF. For instance, word-level BPMF based aComP reports around 1.6% improvements over query-level BPMF based aComP in terms of MAP and P@5 scores. We believe that this can be attributed to the difference between the sparseness of word- or query-document relevance matrix as query-document relevance matrix is more sparse than the word-document relevance matrix in our data set, which has been reported. Hence, with more valuable information available on judging relevance, for example, of query-document pairs, the aComP model performs better.

Impact of Contribution Weight λ

Finally, we take a closer look at the impact of the free parameter λ in Equation (8) that governs the relative contribution of the current searcher and of other users to the overall performance of our reranker. For simplicity, we report the results of the ComP model for various values of λ in Table 6, integrating query-level BPMF and making the uniform assumption of users and queries. Notice that a larger value of λ indicates that behavioral information from current user makes a bigger contribution to the overall performance.

As shown in Table 6, generally, the ComP model with a big value of λ (>0.5) shows better performance than with a small value of λ (<0.5). It achieves its peak performance for $\lambda = 0.7$. Interestingly, the component of current user contributes more than that of other users under our personalized web search settings. Therefore, in web search, paying more attention to individual preference could further enhance a user's search experience by providing a personalized ranking given a query. However, compared to the result of aComP that was presented in Table 5, ComP with a fixed tradeoff λ still loses the comparison. The aComP model boosts the ranking performance by slightly improving 0.3%, 0.4%, 0.2%, and 0.2% for MAP, P@5, NDCG@5, and NDCG@10, respectively, over the best ComP with a fixed $\lambda = 0.7$. It seems that aComp is relatively close to ComP with $\lambda = 0.7$. Hence, we conclude that merging behavioral information of current user and of all users together is helpful to document reranking, especially with an optimal weight controlling the contribution of each part, and paying more attention to behavioral information of current user can generate a better personalized ranking given a query.

Conclusion

In this paper we studied user behavior for personalized web search. In particular, user's short- and long-term search behaviors are integrated for reranking documents initially returned by a search engine. Bayesian probabilistic matrix factorization (BPMF) is applied at various levels to derive the relevance of documents to queries. In addition, we adaptively merge information inferred from behaviors of a specific user and information inferred from behaviors of other users with a user-dependent weight for reranking documents.

Our experimental results show that: (i) reranking performance is indeed affected by assuming that users and queries are uniformly distributed; (ii) BPMF works better at the word-level than at the query-level when estimating document relevance to a query; and (iii) for a document reranking task, behavioral information of a specific user contributes more than information derived from the behavior of other users. Together, these findings make an important step towards unifying prior work on personalization and could be incorporated with content-based personalized approach.

As to limitations of this work, we implemented our ComP and aComP models for the document reranking, both with and without making the (joint) assumption that users and queries are uniformly distributed. However, we can split these two assumptions. In addition, here we only focus on the last query for document reranking, based on the data available to us; if sufficient data are available, we can handle all queries in a session.

As to future work, we plan to run our model on other query log data sets, where the ground truth, that is, the relevance of documents to a query, is provided. In addition, we plan to incorporate our work into a learning to rank framework by exploring useful features to improve search personalization. We could also undertake an investigation on the assumption by using a learning to rank approach.

DOI: 10.1002/asi

Acknowledgments

We thank our reviewers for thoughtful suggestions.

This research was partially supported by the Innovation Foundation of NUDT for Postgraduate under No. B130503, the Academy of Finland (268078), the Natural Science Foundation of China (71402083), Amsterdam Data Science, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the ESF Research Network Program ELIAS, the Royal Dutch Academy of Sciences (KNAW) under the Elite Network Shifts project, the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nos. 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, 652.002.001, 612.001.551, the Yahoo! Faculty Research and Engagement Program, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Agichtein, E., Brill, E., & Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 19–26). New York: ACM.

Bennett, P.N., Radlinski, F., White, R.W., & Yilmaz, E. (2011). Inferring and using location metadata to personalize web search. In Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 135–144). New York: ACM.

Bennett, P.N., White, R.W., Chu, W., Dumais, S.T., Bailey, P., Borisyuk, F., & Cui, X. (2012). Modeling the impact of short- and long-term behavior on search personalization. In Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 185–194). New York: ACM.

Bilenko, M., & Richardson, M. (2011). Predictive client-side profiles for personalized advertising. In Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 413–421). New York: ACM.

Bilenko, M., & White, R.W. (2008). Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In Proceedings of the 17th International World Wide Web Conference (pp. 51–60). New York: ACM.

Cai, F., Liang, S., & de Rijke, M. (2014). Personalized document reranking based on bayesian probabilistic matrix factorization. In Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 835–838). New York: ACM.

Chirita, P.A., Nejdl, W., Paiu, R., & Kohlschütter, C. (2005). Using odp metadata to personalize search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 178–185). New York: ACM.

Chuklin, A., Markov, I., & de Rijke, M. (2015). Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers.]

Collins-Thompson, K., Bennett, P.N., White, R.W., de la Chica, S., & Sontag, D. (2011). Personalizing web search results by reading level. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (pp. 403–412). New York: ACM.

Dou, Z., Song, R., & Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In Proceedings of the 16th International World Wide Web Conference (pp. 581–590). New York: ACM.

Farenick, D.R., & Zhou, F. (2007). Jensen's inequality relative to matrix-valued measures. Journal of Mathematical Analysis and Applications, 327(2), 919–929.

Goker, A., & He, D. (2003). Personalization via collaboration in web retrieval systems: A context based approach. Journal of American Society for Information Science and Technology, 40(1), 357–365.

Jiang, D., Leung, K.W.-T., & Ng, W. (2011). Context-aware search personalization with concept preference. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (pp. 563–572). New York: ACM.

Kim, Y., Hassan, A., White, R.W., & Zitouni, I. (2014). Modeling dwell time to predict click-level satisfaction. In Proceedings of the Seventh ACM International Conference on Web Search and Data Mining (pp. 193–202). New York: ACM.

Kurland, O., & Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 194–201). New York: ACM.

Liu, C., White, R.W., & Dumais, S. (2010). Understanding web browsing behaviors through weibull analysis of dwell time. In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 379–386). New York: ACM.

Liu, J., & Belkin, N.J. (2015). Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. Journal of American Society for Information Science and Technology, 66(1), 58–81.

Liu, X., & Turtle, H. (2013). Real-time user interest modeling for real-time ranking. Journal of American Society for Information Science and Technology, 64(8), 1557–1576.

Masurel, P., Lefevre-Hasegawa, K., Bourguignat, C., & Scordia, M. (2014). Dataiku's solution to yandex's personalized web search challenge. Technical Report, Dataiku, New York, USA.

Matthijs, N., & Radlinski, F. (2011). Personalizing web search using long term browsing history. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (pp. 25–34). New York: ACM.

Mihalkova, L., & Mooney, R. (2009). Learning to disambiguate search queries from short sessions. In Proceedings of the 18th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (pp. 111–127). New York: ACM.

Neal, R.M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neal, R.M., & Hinton, G.E. (1999). A view of the em algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan (Ed.), Learning in graphical models (pp. 355–368). Cambridge, MA: MIT Press.

Pan, W., & Chen, L. (2013). Gbpr: Group preference based bayesian personalized ranking for one-class collaborative filtering. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (pp. 2691–2697). Palo Alto, CA: AAAI Press.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In Proceedings of the 25th Conference Annual Conference on Uncertainty in Artificial Intelligence (pp. 452–461). New York: ACM.

Salakhutdinov, R., & Mnih, A. (2008a). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In Proceedings of the 25th Annual International Conference on Machine Learning (pp. 880–887). New York: ACM.

Salakhutdinov, R., & Mnih, A. (2008b). Probabilistic matrix factorization. In Advances in Neural Information Processing Systems (Vol. 20, pp. 1–8). Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Shapira, B., & Zabar, B. (2011). Personalized search: Integrating collaboration and social networks. Journal of American Society for Information Science and Technology, 62(1), 146–160.

Shen, X., Tan, B., & Zhai, C. (2005a). Context-sensitive information retrieval using implicit feedback. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 43–50). New York: ACM.

Shen, X., Tan, B., & Zhai, C. (2005b). Implicit user modeling for personalized search. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (pp. 824–831). New York: ACM.

Song, G. (2014). Point-wise approach for yandex personalized web search challenge. Technical Report, Playground Global, San Francisco, CA.

Sontag, D., Collins-Thompson, K., Bennett, P.N., White, R.W., Dumais, S. & Billerbeck, B. (2012). Probabilistic models for personalizing web search. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (pp. 433–442). New York: ACM.

Tan, B., Shen, X., & Zhai, C. (2006). Mining long-term search history to improve search accuracy. In Proceedings of the 12th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 718–723). New York: ACM.

Teevan, J., Dumais, S.T., & Liebling, D.J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 163–170). New York: ACM.

Ustinovskiy, Y., Gusev, G., & Serdyukov, P. (2015). An optimization framework for weighting implicit relevance labels for personalized web search. In Proceedings of the 24th International World Wide Web Conference (pp. 1144–1154). New York: ACM.

Ustinovskiy, Y., & Serdyukov, P. (2013). Personalization of web-search using short-term browsing context. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (pp. 1979–1988). New York: ACM.

Volkovs, M.N. (2014). Context models for web search personalization. Technical Report, Department of Computer Science, University of Toronto, Toronto, Canada.

Wang, H., He, X., Chang, M.-W., Song, Y., White, R.W., & Chu, W. (2013). Personalized ranking model adaptation for web search. In Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 323–332). New York: ACM.

White, R.W., Bailey, P., & Chen, L. (2009). Predicting user interests from contextual information. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 363–370). New York: ACM.

White, R.W., Bennett, P.N., & Dumais, S.T. (2010). Predicting short-term interests using activity-based search context. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (pp. 1009–1018). New York: ACM.

Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., & Li, H. (2010). Context-aware ranking in web search. In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 451–458). New York: ACM.

Xu, S., Jiang, H., & Lau, F.C.M. (2011). Mining user dwell time for personalized web search reranking. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (pp. 2367–2372). Palo Alto, CA: AAAI Press.

Yan, J., Chu, W., & White, R.W. (2014). Cohort modeling for enhanced personalized search. In Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 505–514). New York: ACM.

Yi, X., Hong, L., Zhong, E., Liu, N.N., & Rajan, S. (2014). Beyond clicks: Dwell time for personalization. In Proceedings of the Eighth ACM Conference on Recommender Systems (pp. 113–120). New York: ACM.

Zhai, C., & Lafferty, J. (2004). 'A study of smoothing methods for language models applied to information retrieval.' ACM Transactions on Information Systems, 22(2), 179–214.