Personalized Document Re-ranking Based on Bayesian Probabilistic Matrix Factorization

Fei Cai[*†]
f.cai@uva.nl

Shangsong Liang[†]
s.liang@uva.nl

Maarten de Rijke[†]
derijke@uva.nl

[*†]Key lab of Information System Engineering, National University of Defense Technology, Hunan, China
[†]University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

A query considered in isolation provides limited information about the searcher's interest. Previous work has considered various types of user behavior, e.g., clicks and dwell time, to obtain a better understanding of the user's intent. We consider the searcher's search and page view history. Using search logs from a commercial search engine, we (i) investigate the impact of features derived from user behavior on reranking a generic ranked list; (ii) optimally integrate the contributions of user behavior and candidate documents by learning their relative importance per query based on similar users. We use dwell time on clicked URLs when estimating the relevance of documents for a query, and perform Bayesian Probabilistic Matrix Factorization as smoothing to predict the relevance. Considering user behavior achieves better rankings than non-personalized rankings. Aggregation of user behavior and query-document features with a user-dependent adaptive weight outperforms combinations with a fixed uniform value.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Personalization; user behavior; document re-ranking

1. INTRODUCTION

There is a growing interest in personalized search in order to better account for a searcher's individual information need. This kind of personalization improves retrieval performance by tailoring, or *re-ranking*, the ranked results provided by a generic ranker for an individual user based on the models of her previous or current interests. Teevan et al. [7] find that retrieval performance can be improved as more data becomes available about the searcher's interests. And White et al. [11] investigates the effectiveness of a task-based approach in predicting the searcher's interests. They explore the value of modeling current task behavior, finding a significant opportunity in leveraging the on-task behavior to identify web pages to promote in the current ranking. They also explore the use of the on-task behaviors of particular user groups who are experts in the topic currently being searched, rather than all other users, yielding a promising gain in retrieval performance. Other recent work focuses on the role of domain expertise [3] and the use of long-term behaviors for personalized web search by modeling search interests from previous queries [9].

Understanding searchers' information needs requires a thorough understanding of their interests expressed explicitly through search queries, implicitly through clicks on the search engine result page (SERP) or through post-SERP browsing behavior such as dwell time. When sufficient data of a given user is unavailable, the search behavior of other users may be beneficial. Teevan et al. [8] explores the similarity of queries and explicit relevance judgments across a small group. They find that some group members share documents that are relevant to a query because of a shared group focus, but that it is difficult to identify implicitly defined valuable groups.

We address the following personalized web document re-ranking task: re-rank the URLs of a SERP returned by the search engine according to the personal preferences of the users. That is, we aim to personalize search using the long-term (user history based) and short-term (session-based) user context. We use Bayesian Probabilistic Matrix Factorization (BPMF) to estimate the relevance of a URL for a query and to estimate the user's preference for a URL. We begin with a probabilistic graphical model to model the relationship between the searcher, the issued query and the document to be re-ranked. Then we estimate the probability through Bayesian networks using the aggregated dwell time and automatically assign a probability to each query-document and user-document pair using BPMF. We combine users' short- and long-term behaviors in a linear fashion, and adaptively aggregate user and document information depending on similar users.

We demonstrate the effectiveness of our approach to personalized document re-ranking based on a real world dataset that was made available as part of the Web Search Click Data workshop (at WSDM 2014).[1] We find that combining short- and long-term behaviors of users achieves higher scoring rankings than non-personalized rankings and that aggregating user behavior and document features with a user-dependent adaptive weight outperforms combinations with a uniform fixed value.

2. APPROACH

The relationship between *user*, *query* and *URLs* (or *documents*) to be re-ranked can be modeled by a graphical model as Fig. 1. The user submits a query, in response to which a list of 10 URLs are returned by a search engine. Our task is to re-rank the top 10 URLs.

SIGIR '14, July 06–11, 2014, Gold Coast, QLD, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2257-7/14/07 ... $15.00.
http://dx.doi.org/10.1145/2600428.2609453

[1] http://www.wsdm-conference.org/2014/accepted-workshops/.

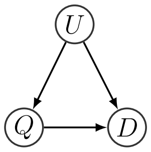


Figure 1: Probabilistic graphical model (U: user, Q: query and D: document)

We measure the relevance of a URL (d) to a query (q) submitted by a user (u) as a probability $P(d|q,u)$ and use this probability as the final ranking score to output a ranking list of the top 10 URLs.

From Fig. 1, we calculate the joint probability $p(u,q,d)$:

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

The relevance of d given q and u, $p(d|q,u)$, is

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

To estimate $p(q,u|d)$, we use a linear mixture governed by a free parameter λ: $p(q,u|d) = (1-\lambda) \cdot p(q|d) + \lambda \cdot p(u|d)$ [4].

In the simplest case, $p(u)$ and $p(d)$ are assumed to be uniform, and hence, do not affect the document ranking, so that $p(d|q,u)$ can be estimated by

$$p(d|q,u) \propto \frac{(1-\lambda) \cdot p(q|d) + \lambda \cdot p(u|d)}{p(q|u)}, \quad (3)$$

where $p(q|d) = \prod_{t_i \in q} p(t_i|d)^{N(t_i,q)}$, with $N(t_i,q)$ being the number of query terms t_i in q and $p(u|d) = \frac{p(d|u) \cdot p(u)}{p(d)} \propto p(d|u)$ because $p(u)$ and $p(d)$ are uniform.

As our task is to re-rank the top 10 documents, the contribution to each document from $p(q|u)$ remains the same and will not change the ranking. Hence, we have:

$$p(d|q,u) \propto (1-\lambda) \cdot \prod_{t_i \in q} p(t_i|d)^{N(t_i,q)} + \lambda \cdot p(d|u). \quad (4)$$

We estimate $p(d|u)$ from the short- and long-term behaviors of the user u. Again, we use a linear combination $p(d|u) = (1-\omega) \cdot p(d|u)_{short} + \omega \cdot p(d|u)_{long}$, as suggested by Bennett et al. [1] to achieve the final outcome. Therefore, our final re-ranking criteria is (5):

$$\begin{aligned} p(d|q,u) \propto (1-\lambda) \cdot \prod_{t_i \in q} p(t_i|d)^{N(t_i,q)} + \\ + \lambda \cdot [(1-\omega) \cdot p(d|u)_{short} + \omega \cdot p(d|u)_{long}], \quad (5) \end{aligned}$$

The way in (5) we estimate $p(d|u)_{short}$ and $p(d|u)_{long}$ is described in §2.2.

2.1 Smoothing with Bayesian Probabilistic Matrix Factorization (BPMF)

Many smoothing methods have been proposed in the setting of language models for IR. For our re-ranking task, it makes sense to use dwell time rather than term frequency for smoothing because the contents of query and document are unavailable.

We use Bayesian Probabilistic Matrix Factorization (BPMF) [5] to predict the relevance of a document for a query as well as the preference of a user for a document. Taking the former, for example, we first take the logarithm of the aggregated dwell time of known query-document pairs to dampen sharp peaks and then label the relevance of each pair as $\min(\lfloor \lg(t+10) \rfloor, 5)$, where t is the aggregated dwell time, and $\lfloor \cdot \rfloor$ is the floor function. BPMF is then applied to the query-document matrix to assign a non-zero value to each element in the matrix. This completes our smoothing method.

The original matrix query-document matrix is replaced by:

$$R = Q_{N \times k} \times D^T_{M \times k}, \quad (6)$$

where $Q_{N \times k}$ and $D_{M \times k}$ represent the query- and user-specific latent feature matrix, where N, M and k indicate the number of queries, documents and latent features, respectively.

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

$$\begin{aligned} p(R^*_{ij}|R,\Theta_0) = \int\int p(R^*_{ij}|Q_i,D_j) p(Q,D|R,\Theta_Q,\Theta_D) \\ p(\Theta_Q,\Theta_D|\Theta_0) d_Q d_D d_{\Theta_Q} d_{\Theta_D}, \quad (7) \end{aligned}$$

where $\Theta_Q = \{\mu_Q, \Sigma_Q\}$ and $\Theta_D = \{\mu_D, \Sigma_D\}$ are query and document hyperparameters, and the prior distributions over the query and document feature vectors 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. BPMF introduces priors for the hyperparameters, which allows the model complexity to be controlled based on the training data [6]. When the prior is Gaussian, the hyperparameters can be updated by performing a single EM step [2], which scales linearly with the number of observations without significantly affecting the time to train the model.

2.2 Modeling behavior

For our task of re-reranking the top 10 URLs returned by the search engine, short-term behaviors, more specifically the clicks, may provide a strong signal of the user's interest. We aggregate the contributions of all clicked URLs to compute the term $p(d|u)_{short}$ mentioned before as:

$$p(d|u)_{short} = \sum_{d_i \in D} \omega_i \cdot p(d_i|u), \quad (8)$$

where D is the set of clicked URLs inside the current search session, and

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

depends on the similarity between the clicked document d_i and the document d to be re-ranked, while

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

is a normalization factor. Further, $D \setminus \{d_i\}$ denotes the subset of D except d_i and $Dis(d_j,d)$ returns the Euclidean distance between d_i and d. Both documents are represented by the latent feature vectors returned by the BPMF process.

For the long-term behaviors of the user, we estimate the probability $p(d|u)_{long}$ by accumulating all his dwell time on d from the historical logs.

2.3 Adaptive weights

Previous work [10] uses a fixed weight λ in (5), i.e., the same for all users, when integrating the contributions from the user and a specific document. This choice shows good re-ranking results. However, we treat the weight differently as different users behave differently. We propose an adaptive weight solution to assign a specific weight λ in (5) per user u, which depends on users that are similar to u.

We first cluster the users in the training set using the k-Nearest Neighbors algorithm (KNN), and the users have been assigned the optimal λ using a sweep from 0 to 1 with step-size 0.1 while maximizing MAP. We set the number of clusters to $k = 10$ as our result performs the best with this setting. In the test phase, an unseen $user_u$ is assigned to the nearest cluster C and allocated a weight λ as:

$$\lambda = \textstyle\sum_{user_i \in C} \alpha_i \cdot \lambda_i, \tag{9}$$

where

$$\alpha_i = \frac{1}{Z_\alpha} \times \frac{\sum_{user_j \in C \setminus \{user_i\}} Dis(user_j, user)}{\sum_{user_k \in C} Dis(user_k, user)}$$

depends on the similarity between the $user_i$ in cluster C and the test $user_u$ while

$$Z_\alpha = \textstyle\sum_{user_i \in C} \frac{\sum_{user_j \in C \setminus \{user_i\}} Dis(user_j, user)}{\sum_{user_k \in C} Dis(user_k, user)}$$

is a normalization factor, and λ_i is the weight for $user_i$. Again, users are represented by latent feature vectors returned by the BPMF.

3. EXPERIMENTS

We begin by describing the research questions that we aim to answer. We then report the results of experiments aimed at answering the questions and present the findings of an analysis of the results.

3.1 Research questions

We are particularly interested in the contributions of two types of information obtained from user logs for personalized re-ranking: short- and long-term behaviors. We train our model and evaluate it by re-ranking the top 10 results from a web search engine. This original ranking is used as one of our baselines for comparisons. We address the following two research questions:

1. Does the combination of short- and long-term behaviors help improve the quality of re-ranking results?
2. What are the optimal relative contributions for document re-ranking of user and document information?

Answers to these questions provide valuable insights about the relative utility of the historical logs and can help inform decisions about when and how to use the historical logs for search personalization.

3.2 Experimental setup

The primary source of data for this study consists of anonymized logs of users provided by the Personalized Web Search Challenge.[2] The logs, collected for four weeks, contain a unique user identifier, a search session identifier, a query identifier, the specific query term identifier, the top 10 URLs returned by the search engine for that query, and the dwell time on clicked results. The information of users with more than 6 participations during the specific span is kept. Besides, we remove user records without short-term behavior. Finally, BPMF is performed on the whole dataset and, we randomly split the dataset into five partitions, such that 80% is used as training data and the remaining 20% is used as test data for the latter experiments. Table 1 shows some statistics of the training dataset. Each query is different from the others between sessions; the numbers of search sessions and unique queries are the same.

Relevance labels are obtained automatically based on the aggregated dwell time units, with a graded relevance scale. We assign a grade 5 (highly relevant) to documents with clicks whose logarithmic dwell time is at least 5 or with the last result click in the session. We assign a grade 4, 3 and 2 (relevant, normal relevant and slightly relevant, respectively) to documents with clicks with a logarithmic dwell time between 5 and 2. We assign grade 1 (irrelevant) to documents with no clicks or clicks whose logarithmic dwell time is less than 1.

Table 1: Dataset statistics.

#log records	#unique users	#search sessions
175406	522	3328
#unique URLs	#unique query terms	#unique queries
81006	11344	3328

For evaluation purposes, with a graded relevance scale, we report our performance with NDCG@5 plus NDCG@10 as well as p@5 and MAP. The metrics for all queries are averaged to obtain a final measure across the top 10 results retrieved before re-ranking (for one baseline) and after re-ranking. Statistical significance of observed differences between the performance of two runs is tested using a two-tailed paired t-test and is denoted using $^\blacktriangle/^\blacktriangledown$ for significant differences for $\alpha = .01$, or $^\vartriangle/^\triangledown$ for $\alpha = .05$.

3.3 Results and analysis

We begin by investigating the influence of short- and long-term behavior using a fixed value $\lambda = 0.5$. It is clear from Table 2 that the performance shows very minor differences when the weight ω in (5) changes; it reaches a peak for $\omega = 0.3$. Consequently, we choose $\omega = 0.3$ in later experiments.

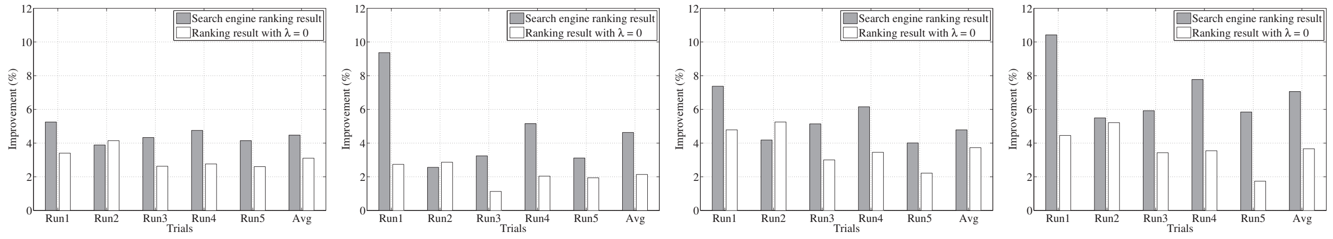
To verify the effectiveness of personalization, we test our model with 5 runs, representing five different splits of the available data in training and test set, and choose $\lambda = 0.5$ for the experiment (personalization is active whenever $\lambda > 0$). For $\lambda = 0.5$, document and user-based scoring receive equal weights. We plot the improvements over two baselines in Fig. 2. One baseline ranking is simply the ranked list produced by the search engine and another is produced when $\lambda = 0$ (non-personalization). We report the improvements of each run and the average as well. Our proposed method for re-ranking outperforms the two baselines on all metrics and the differences are statistically significant at significance level $\alpha = .01$. We can see that user information can effectively be used to boost the ranking performance when historical behaviors are available. Another interesting observation from Fig. 2 is that the relative improvements over the $\lambda = 0$ setting are smaller than those over simply choosing the results from the input ranker.

Finally, we take a closer look at the effect of the free parameter λ in (5) that governs the relative contribution of searcher information and document information to the overall performance of our re-ranker. We report our results by averaging the outcomes of 5 separate runs of adaptive λ and each fixed weight in Table 3, respec-

Table 2: Evaluation with a fixed parameter $\lambda = 0.5$.

ω	MAP	p@5	NDCG@5	NDCG@10
0.0	.4642	**.3082**	.4183	.5460
0.1	.4646	.3081	.4182	.5464
0.2	.4648	.3081	.4181	.5462
0.3	**.4649**	**.3082**	**.4187**	**.5467**
0.4	.4648	**.3082**	.4183	**.5467**
0.5	.4645	.3080	.4181	.5464
0.6	.4647	.3081	.4185	.5466
0.7	.4648	**.3082**	**.4187**	.5464
0.8	.4646	.3081	.4183	.5465
0.9	.4644	.3080	.4181	.5463
1.0	.4645	.3080	.4182	.5462

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



(a) MAP **(b) NDCG@5** **(c) P@5** **(d) NDCG@10**

Figure 2: Relative improvements over two baselines, using four metrics. The grey bar indicates improvements over the search engine ranking result, and the white bar shows improvements over ranking results with $\lambda = 0$. (Settings used: $\lambda = 0.5$ and $\omega = 0.3$.)

Table 3: Re-ranking performance for fixed values of λ and an adaptive setting of λ. Boldface marks the best result per column; a statistically significant difference between the Adaptive λ setting and the rankings produced with a fixed value of λ or the baseline search engine (SE) ranking is marked. (Settings used: $\omega = 0.3$.)

	MAP	P@5	NDCG@5	NDCG@10
SE	▼.4451	▼.2916	▼.4017	▼.5128
λ=0.1	▼.4640	▼.3081	▼.4181	▼.5462
λ=0.2	▼.4648	▼.3082	▼.4184	▼.5467
λ=0.3	▼.4649	▼.3082	▼.4186	▼.5467
λ=0.4	▼.4649	▼.3082	▼.4189	▼.5469
λ=0.5	▼.4650	▼.3082	▼.4185	▼.5469
λ=0.6	▼.4658	▽.3087	▼.4203	▼.5490
λ=0.7	▼.4655	▼.3084	▼.4187	▼.5475
λ=0.8	▼.4656	▽.3084	▼.4189	▼.5469
λ=0.9	▼.4648	▼.3082	▼.4185	▼.5465
Adaptive λ	**.4976**	**.3160**	**.4463**	**.5698**

tively. We also report the performance of the search engine (SE) as a baseline. A high value of λ indicates that user information makes a big contribution to the overall performance. As shown in Table 3, the experiments with a big λ (> 0.5) achieve better performance than those with a small λ (< 0.5) except for $\lambda = 0.9$. The user components contributes more than the document itself under our personalized web search settings. With λ adaptive, our model effectively boosts ranking performance by improving 6.83%, 2.36%, 6.19%, and 3.79% for MAP, P@5, NDCG@5, and NDCG@10 respectively, over the best fixed value ($\lambda = 0.6$). Additionally, it increases the MAP, P@5, NDCG@5, and NDCG@10 scores by 11.79%, 8.36%, 11.10%, and 11.12%, respectively, over the search engine ranking result. We conclude that adding user information and optimizing the weight λ that controls the contribution of user information helps to improve the effectiveness of re-ranking.

4. CONCLUSION

Previous work on search personalization has exploited user behaviors to model searcher interests. Relatively little was known about the relative contribution of document or user features for optimal personalized re-ranking. In this paper we have investigated how the contributions of document and user can be combined. We have demonstrated that historic behavior yields benefits for personalized re-ranking and that user information contributes more than the document itself for personalized re-ranking. This work makes an important step toward unifying prior work on personalization. Future work will further explore different features on how to best improve search performance through personalization.

Acknowledgments. This research was supported by the Innovation Foundation of NUDT for Postgraduate under No. B130503, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nrs 288024 and 312827, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, the Center for Creation, Content and Technology (CCCT), the QuaMerdes project funded by the CLARIN-nl program, the TROVe project funded by the CLARIAH program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), the Netherlands eScience Center under project number 027.012.-105, the Yahoo! Faculty Research and Engagement Program, the Microsoft Research PhD program, and the HPC Fund.

REFERENCES

[1] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *SIGIR '12*, pages 185–194, 2012.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Stat. Soc., Series B*, 39(1):1–38, 1977.

[3] B. Huurnink, L. Hollink, W. van den Heuvel, and M. de Rijke. Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *J. Amer. Soc. Inf. Sci. & Techn.*, 61(6):1180–1197, 2010.

[4] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *SIGIR '04*, pages 194–201, 2004.

[5] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *ICML '08*, pages 880–887, 2008.

[6] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS 20*, pages 1–8, 2008.

[7] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR '05*, pages 449–456, 2005.

[8] J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In *WSDM '09*, pages 15–24, 2009.

[9] H. Wang, X. He, M.-W. Chang, Y. Song, R. W. White, and W. Chu. Personalized ranking model adaptation for web search. In *SIGIR '13*, pages 323–332, 2013.

[10] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context. In *CIKM '10*, pages 1009–1018, 2010.

[11] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *WWW '13*, pages 1411–1420, 2013.