Selectively Personalizing Query Auto-Completion

Fei Cai[†‡]
f.cai@uva.nl

Maarten de Rijke[‡]
derijke@uva.nl

[†]Science and Technology on Information Systems Engineering Laboratory,
National University of Defense Technology, Hunan, China

[‡]Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Query auto-completion (QAC) is being used by many of today's search engines. It helps searchers formulate queries by providing a list of query completions after entering an initial prefix of a query. To cater for a user's specific information needs, personalized QAC strategies use a searcher's search history and their profile. Is personalization consistently effective in different search contexts?

We study the QAC problem by selectively personalizing the query completion list. Based on a lenient personalized QAC strategy that encodes the ranking signal as a trade-off between query popularity and search context, we propose a model for selectively personalizing query auto-completion (SP-QAC) to study this trade-off. We predict effective trade-offs based on a regression model, where the typed query prefix, clicked documents and preceding queries in the same session are used to weigh personalization in QAC. Experiments on the AOL query log show the SP-QAC model can significantly outperform a state-of-the-art personalized QAC approach.

Keywords

Personalization; Query auto completion; Web search

1. INTRODUCTION

Personalization techniques have been widely adopted by today's search engines [1, 3, 13]. Query auto-completion (QAC) is no exception [2]. In general, personalized lists of query completions produced by a generic QAC ranking approach are obtained by distinguishing individuals and their contexts, either based on their search history [1, 3, 8] or on their profiles [13]. Such personalization strategies have been proven to be effective as evidenced by a series of personalized query auto-completion approaches [1, 3, 8, 11, 13].

Generally, when ranking query completions in response to a prefix entered by a user, most existing personalized QAC approaches use an interpolation parameter λ that controls the trade-off between query popularity and the searcher's personal search context [1, 3]. This trade-off is uniformly applied to all typed prefixes [1, 3], so as to achieve a modest performance across all cases. It is unclear whether personalization will always boost the ranking performance over a generic QAC approach. As an aside, in web search and recommendation systems, it has been shown that not all queries should be personalized equally as personalization strategies occasionally harm the search accuracy [5, 14, 15]. Similarly, we hypothesize that in QAC personalization of query prefixes should not be handled in a uniform manner because:

(1) a user's initial information needs may be addressed by previous interactions;
(2) users may change their search intent during a search session.

Such clues can be explicitly expressed by the clicks or directly revealed by the query flow in a session.

We propose a model for selectively personalizing query auto-completion (SP-QAC) to re-rank the top N query completions produced by the MPC (Most Popular Completion) model [1]. In particular, personalization in the proposed SP-QAC model is individually weighed when combined with ranking signals from search popularity. We study the following factors for weighing personalization: the typed prefix for which we recommend query completions, the clicked documents for inferring a user's satisfaction, and the topic changes of preceding queries in the same session for detecting search intent shifts. We use the description of each URL from the ODP data[1] to represent documents and queries based on the word2vec model [10] when inferring a user's satisfaction as well as shifts in query intent in a session. Finally, we test the improvements of our proposal over state-of-the-art QAC baselines on a publicly available query log dataset. We find that SP-QAC outperforms a traditional non-personalized QAC approach and a uniformly personalized QAC approach with a fixed trade-off controlling the contribution of search popularity and search context. Our contributions in this paper are summarized as:

(1) We propose a model for selectively personalizing query auto-completion (SP-QAC) that flexibly outweighs or depresses the contribution of personalization in QAC.
(2) We study the role of typed prefixes, clicked documents, and preceding queries in the same search session when estimating the weight of personalization in QAC.

2. APPROACH

A straightforward approach to ranking query completions is based on the popularity of queries. Bar-Yossef and Kraus [1] refer to this type of ranking as the Most Popular Completion (MPC) model:

$$MPC(p) = \underset{q \in S(p)}{\arg\max}\, w(q),\ w(q) = \frac{f(q)}{\sum_{q_i \in L} f(q_i)}, \qquad (1)$$

where $f(q)$ denotes the number of occurrences of query q in search log L, and $S(p)$ is a set of query completions that start with prefix p.

To cater for a user's particular information need, personalization

SIGIR '16 Pisa, Italy

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4069-4/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2911451.2914686

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

has been incorporated into the MPC model, as described in [1, 3]. A generic personalized QAC approach employs a fixed parameter λ to control the contribution of personal information to generate the final ranking of query completions. For instance, Bar-Yossef and Kraus [1] compute a hybrid score for each query candidate q_c, which is a convex combination of two scores, i.e., a query popularity score $MPCsco(q_c)$ and a personalization score $Psco(q_c)$:

$$hybsco(q_c) = \lambda \cdot MPCsco(q_c) + (1 - \lambda) \cdot Psco(q_c), \qquad (2)$$

where $MPCsco(q_c)$ is estimated by candidate q_c's frequency in the query log and $Psco(q_c)$ is measured by q_c's similarity to the search context in session. This approach handles each prefix uniformly. However, users may modify their search intent during a session and personalization may harm the quality of the ranking of query completions if we continue to use the previous search context. Hence, we propose a model for selectively personalizing query auto-completion (SP-QAC) that varies the importance of personalization when generating the final hybrid score of a query completion:

$$hybsco(q_c) = \phi(\cdot) \cdot MPCsco(q_c) + (1 - \phi(\cdot)) \cdot Psco(q_c), \qquad (3)$$

where $\phi(\cdot)$ outputs a trade-off in $[0, 1]$ and is parameterized by three arguments: the typed prefix, clicked documents and preceding queries in the same session; see below.

2.1 Signal from typed prefix

Most previous work on query auto-completion [1, 3, 8, 13] only considers the typed prefix for generating a list of query completions, ignoring the potential signal hidden in the typed prefix for personalization. However, the typed prefix normally reveals a strong clue for inferring a user's personal query activity, such as query expansion and query repetition, etc. We introduce a factor f_p to model the signal from the typed prefix p on weighing the personalization for query auto-completion as follows:

$$f_p = \frac{|\mathcal{W}(p)|}{|\mathcal{S}|} + c, \qquad (4)$$

where $|\mathcal{W}(p)|$ and $|\mathcal{S}|$ indicate the number of words that start with p and that appear in the current session, respectively; c is a small constant used for smoothing. A larger value of f_p could imply a higher importance of personalization in the final ranking score in (3).

2.2 Inferring search satisfaction from clicked documents

User's clicks on documents retrieved in response to a query are a widely used behavioral signal for measuring search satisfaction [9]. Search satisfaction can be further approximated by the closeness between a submitted query and its clicked documents: the closer they are, the more satisfied the user could be [6]. The cosine similarity can be applied to model a factor f_d for measuring closeness as follows:

$$f_d = \begin{cases} c, & \text{no clicks} \\ \frac{1}{|\mathcal{Q}|}\sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{D}_q|}\sum_{d_c \in \mathcal{D}_q} \cos(q, d_c), & \text{otherwise,} \end{cases} \qquad (5)$$

where $\mathcal{D}_q$ is a set of clicked documents corresponding to a submitted query q in a set $\mathcal{Q}$ of previous queries in the session. To each clicked document $d_c \in \mathcal{D}_q$ we assocaite a short description T extracted from the ODP. We vectorize this document description consisting of a sequence of words using the word2vec model [10], where each word is represented by a vector v_w. In doing so, a clicked document can be vectorized by averaging the words in T as $d_c = \frac{1}{|T|}\sum_{v_w \in T} v_w$. After that, a query q is then similarly represented by averaging its clicked documents $\mathcal{D}$ in the training log, i.e., $q = \frac{1}{|\mathcal{D}|}\sum_{d_c \in \mathcal{D}} d_c$. In practice, for a query q that has no clicked documents, the same representation of its most semantically similar query q_o, which is identified by the word2vec model [10] and has been vectorized in the training period, is assigned to it by

$$\begin{aligned} q_o &\leftarrow \operatorname*{arg\,max}_{q_l \in Q_L} \cos(q, q_l) \qquad (6) \\ &= \operatorname*{arg\,max}_{q_l \in Q_L} \frac{1}{|q|} \times \frac{1}{|q_l|} \sum_{w_k \in q} \sum_{w_j \in q_l} \cos(w_k, w_j), \end{aligned}$$

where Q_L is a set of queries that have clicked documents in the training period.

Intuitively, a large score of f_d in (5) indicates a high probability that the user is satisfied with the results, thus resulting in a low weight of personalization in QAC. We assign a small constant c to f_d when no clicks are available, where personalization could make sense because the user's request has not been addressed and they may continue to submit similar queries.

2.3 Detecting topic shifts from preceding queries

A long session may contain queries on multiple topics [7]. We use this observation to infer signals for weighing personalization in QAC. A strong topical shift in the preceding queries implies a low-weight personalization in QAC because the user may have shifted topics. Hence, we model a factor f_q based on topic shifts in preceding queries in the session:

$$f_q = \begin{cases} c, & r = 1 \text{ or } 2 \\ \cos(q_1, q_2), & r = 3 \\ \cos(q_{r\text{-}1} - q_{r\text{-}2}, q_{r\text{-}2} - q_{r\text{-}3}), & r > 3, \end{cases} \qquad (7)$$

where r is the query position in session and each query is vectorized by the scheme described in §2.2. For queries at the beginning of a session, i.e., $r = 1$ or 2, the topic shift from queries is unavailable, making no impact on personalization, and thus we assign a small constant c to f_q. Similarly, at query position $r = 3$, the relative topical shift of queries is still unavailable. Instead, we use the absolute query similarity between $q_{r\text{-}1}$ and $q_{r\text{-}2}$ as an indicator of query topical shift. For queries at position $r > 3$, we study the topic shift from their preceding queries, i.e., $q_{r\text{-}1}$, $q_{r\text{-}2}$ and $q_{r\text{-}3}$.

A large value of f_q is produced if the topical similarity of preceding queries is high, which, in turn, implies with high probability that the user's search intent has remained unchanged from previous queries in the current session. In this case, personalization should be emphasized in (3).

2.4 Weighing personalization

Taking these discussed factors into account, i.e., the typed prefix, clicked documents and preceding queries in the same session, we adopt logistic regression to model how likely each factor should affect the weight of personalization in a QAC task. In the training period, for each typed prefix, we manually change the value of λ in (3) from 0 to 1 (with steps of size 0.1) to guarantee that the final submitted query is ranked at the top position. By doing so, we obtain an optimal weight for personalization, which is used as a label in the regression model.

Regarding the inputs to the regression model, the factors discussed above, i.e., f_p, f_d and f_q, are involved. For the cases that no word appearing in the current session starts with the typed prefix, we perform the regression model based on the factors f_d and f_q; otherwise, we perform the regression model based on all three factors. Hence, the selective weight $\phi(\cdot)$ in (3) is determined as follows:

$$\phi(\cdot) = \begin{cases} Reg(f_d, f_q), & \text{if } f_p = c \\ Reg(f_p, f_d, f_q), & \text{otherwise.} \end{cases} \tag{8}$$

We use the personalization scenario proposed in [3] to compute the $Psco(q_c)$ score in (3) when generating the optimal personalization weights, i.e.,

$$Psco(q_c) = p(q_c \mid \mathcal{Q}) = \sum_{q_s \in \mathcal{Q}} Z_n \cdot p(q_c \mid q_s),$$

where $\mathcal{Q}$ is a set of preceding queries in the current session and Z_n is used for normalization; $p(q_c \mid q_s)$ is measured using the common strings of query terms in q_c and q_s.

3. EXPERIMENTS

We write SP-QAC for our proposed selectively personalized query auto completion model; it personalizes query completions based on a regression model considering the factors described in §2.1, §2.2 and §2.3, respectively. Our research questions are:

(**RQ1**) Does selective personalization help improve the accuracy of ranking query completions?

(**RQ2**) What is the performance of the proposed SP-QAC model under various inputs to the regression model for weighing the importance of personalization?

3.1 Experimental setup

We use the publicly available AOL query log dataset [12] in our experiments, which is split into three parts: a training set, a validation set and a test set consisting of the first 60%, the following 20% and the last 20% of the query log, respectively. A large volume of navigational queries containing URL substrings (.com, .net, .org, http, etc.) are removed. One-query and no-click sessions are both excluded in our experiments as not enough search context is available. In addition, we follow previous QAC work and adapt a commonly used evaluation methodology in QAC [3, 8, 13] by only keeping cases where the final submitted query is included in the top N query completions returned by the MPC approach,

For comparison, the following baselines are selected: (1) the most popular completion (MPC) method, which ranks query candidates by their frequency [1]; (2) a personalized QAC approach based on session context with a fixed tradeoff $\lambda = 0.5$ in (2), denoted as P-QAC [3].

We use Mean Reciprocal Rank (MRR) for evaluating the performance of QAC models. For a prefix p associated with a list of query completions $\mathcal{L}(p)$ and the user's finally submitted query q', the Reciprocal Rank (RR) is computed as:

$$RR = \begin{cases} \frac{1}{\text{rank of } q' \text{ in } \mathcal{L}(p)}, & \text{if } q' \in \mathcal{L}(p) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Then, MRR is computed as the mean of RR for all prefixes.

In addition, we set $N = 10$ in our experiments, which means the top ten query completions returned by the MPC approach are to be re-ranked. We randomly assign a small value 0.01 to the constant c in our experiments.

3.2 Results and discussions

To answer research question **RQ1**, we compare the results of our proposed model, i.e., SP-QAC, with those of the baselines. We report the results in Table 1. Clearly, as shown in Table 1, a generic personalization scheme helps to improve the QAC performance in terms of MRR as the MRR scores of the P-QAC model are higher than that of the MPC approach at every prefix length. In addition, when a selective personalization strategy is embedded into the P-QAC model, the QAC performance is further boosted as the MRR scores of the SP-QAC model are higher than those of MPC and P-QAC. Compared to the MPC approach, significant MRR improvements of the SP-QAC approach are observed at level $\alpha = .05$ for the short prefixes, i.e., $\#p = 1, 2, 3$, and at level $\alpha = .01$ for the long prefixes, i.e., $\#p = 4, 5$. Compared to short prefixes, long prefixes are able to reveal a stronger signal for search personalization, like query repetition. However, compared to the results of the P-QAC model, significant MRR improvements of the SP-QAC model are only observed at short prefixes, i.e., $\#p = 1, 2$. This is due to the fact that, compared to short prefixes, long prefixes often return the correct query early in the list of query completions by both models.

Table 1: Performance of QAC models in terms of MRR at a prefix length $\#p$ ranging from 1 to 5 characters. The best performance per row is highlighted. Statistical significance of pairwise differences of SP-QAC vs. MPC and SP-QAC vs. P-QAC are detected using a two-tailed t-test ($^{\blacktriangle}$/$^{\blacktriangledown}$ for α = .01, or $^{\vartriangle}$/$^{\triangledown}$ for α = .05) and marked in the upper left and upper right hand corners of SP-QAC scores, respectively.

$\#p$	MPC	P-QAC	SP-QAC
1	0.5368	0.5422	$^{\vartriangle}$**0.5535**$^{\vartriangle}$
2	0.5556	0.5628	$^{\vartriangle}$**0.5744**$^{\vartriangle}$
3	0.5944	0.6046	$^{\vartriangle}$**0.6165**
4	0.6294	0.6427	$^{\blacktriangle}$**0.6547**
5	0.6589	0.6646	$^{\blacktriangle}$**0.6762**

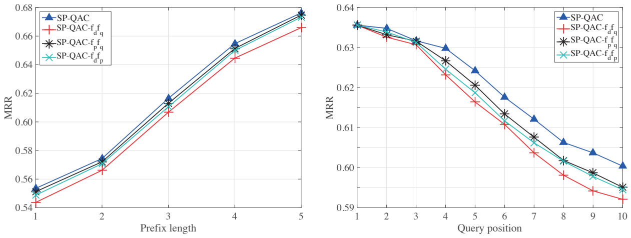
Table 2: Performance of QAC models in terms of MRR at various query positions. The best performer per row is highlighted. Statistical significance of pairwise differences of SP-QAC vs. MPC and SP-QAC vs. P-QAC are detected using a two-tailed t-test ($^{\blacktriangle}$/$^{\blacktriangledown}$ for α = .01, or $^{\vartriangle}$/$^{\triangledown}$ for α = .05) and marked in the upper left and upper right hand corners of SP-QAC scores, respectively.

Query position	MPC	P-QAC	SP-QAC
{1, 2, 3}	0.6283	0.6327	**0.6340**
{4, 5, 6}	0.6005	0.6113	$^{\blacktriangle}$**0.6257**$^{\vartriangle}$
{7, 8, 9, ⋯ }	0.5737	0.5863	$^{\blacktriangle}$**0.6058**$^{\vartriangle}$

To further examine the effectiveness of selective personalization for QAC, we examine the performance of QAC models at different query positions, i.e., at the beginning (1, 2 or 3), in the middle (4, 5 or 6) and in later ($\geq$ 7) parts of a session. See Table 2. At the start of a session, these three models return competitive MRR scores as limited information from search context is available for the P-QAC and SP-QAC models. However, as the search context becomes richer, significant improvements in terms of MRR are achieved by both personalization models.

Next, we turn to research question **RQ2** and examine the performance of the SP-QAC model under different inputs to the regression model for generating the personalization weights for QAC. We manually remove one factor and keep the other two for the regression model, resulting in the SP-QAC-$f_d f_q$, SP-QAC-$f_p f_q$ and SP-QAC-$f_d f_p$ models, which corresponds to the SP-QAC model without considering the factors f_p, f_d and f_q for selectively personalizing QAC, respectively. We plot the results in Figure 1, including the model incorporating all three factors for selective personalization (i.e., SP-QAC).

Generally, the SP-QAC model achieves the best performance in terms of MRR at any prefix length and any query position. As



(a) QAC performance at varying prefix lengths.

(b) QAC performance at varying query positions.

Figure 1: Performance in terms of MRR of the SP-QAC models under different schemes for weighing personalization, tested at varying prefix lengths (left) and varying query positions (right).

shown in Figure 1a, as the prefix length increases, the MRR scores monotonously go up because a long prefix can sharply cut down the space of possible completions matching the input prefix, resulting in increasing MRR scores. In contrast, as shown in Figure 1b, the MRR scores decrease when users continue to query in a session. In the latter part of a search session, users are inclined to submit uncommon queries, making it difficult for MPC to return a correct completion early, which our proposal depends on.

Next, we zoom in on the three factors considered in selective personalization for QAC. As shown in Figure 1a, the MRR scores of the SP-QAC-$f_p f_q$ and SP-QAC-$f_d f_p$ models approximate those of the SP-QAC model as the prefix length increases. The MRR scores of SP-QAC-$f_d f_q$ lag behind those of the SP-QAC model: the signal for personalization from the typed prefix becomes stronger as the prefix length increases, which benefits SP-QAC-$f_p f_q$ and SP-QAC-$f_d f_p$ (both consider the factor f_p). Regarding the QAC performance at varying query positions, Figure 1b shows that the models perform similarly at early positions of a session, as insufficient search context is available to tell them apart. In addition, SP-QAC-$f_p f_q$ outperforms SP-QAC-$f_d f_q$ and SP-QAC-$f_d f_p$ at most query positions. For later queries in a session, where a richer search context is available, information from the typed prefix and previous queries helps for QAC.

In essence, from the results in Figure 1, the SP-QAC model that does not consider the factor f_p works worst, from which we infer that f_p is the most important factor for selectively personalizing QAC. Similarly, we infer that f_q is more important than f_d.

4. CONCLUSION

We have proposed a selectively personalized approach for query auto-completion (QAC). Our model predicts whether a specific prefix prefers a personalized approach when ranking the query completions. We have explored several factors that influence the weight of personalization in a generic personalized QAC model, such as the typed prefix, the clicked documents and the preceding queries in the same session. The typed prefix yields the most benefits for weighing personalization in query completion re-ranking, and that the preceding queries contributes more than click information.

This work makes an important step towards unifying prior work on personalized QAC by studying when and how to incorporate personalization in a QAC task. As to future work, other sources can be explored for investigating how to best personalize query auto-completion, e.g., a user's dwell time or their long-term search history. In addition, it is interesting to zoom in on individual users to determine whether they are likely to benefit from personalization in QAC and whether they stand to gain from diversifying query completions [4].

Acknowledgments. This research was supported by Ahold, Amsterdam Data Science, the Bloomberg Research Grant program, 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 nrs 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

[1] Z. Bar-Yossef and N. Kraus. Context-sensitive query auto-completion. In *WWW '11*, pages 107–116, 2011.

[2] F. Cai and M. de Rijke. Query auto completion in information retrieval. *Found. Trends in Inform. Retr.*, 2016. To appear.

[3] F. Cai, S. Liang, and M. de Rijke. Time-sensitive personalized query auto-completion. In *CIKM '14*, pages 1599–1608, 2014.

[4] F. Cai, R. Reinanda, and M. de Rijke. Diversifying query auto-completion. *ACM Trans. on Inform. Syst.*, 2016. To appear.

[5] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *WWW '07*, pages 581–590, 2007.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, 2005.

[7] J. Jiang, D. He, and J. Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *SIGIR '14*, pages 607–616, 2014.

[8] J.-Y. Jiang, Y.-Y. Ke, P.-Y. Chien, and P.-J. Cheng. Learning user reformulation behavior for query auto-completion. In *SIGIR '14*, pages 445–454, 2014.

[9] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM '14*, pages 193–202, 2014.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 26*, pages 3111–3119, 2013.

[11] B. Mitra. Exploring session context using distributed representations of queries and reformulations. In *SIGIR '15*, pages 3–12, 2015.

[12] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06*, pages 1–7, 2006.

[13] M. Shokouhi. Learning to personalize query auto-completion. In *SIGIR '13*, pages 103–112, 2013.

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

[15] W. Zhang, J. Wang, B. Chen, and X. Zhao. To personalize or not: A risk management perspective. In *RecSys '13*, pages 229–236, 2013.