Personalization of Web-search Using Short-term Browsing Context

Yury Ustinovskiy
Yandex
Leo Tolstoy st. 16, Moscow, Russia
yuraust@yandex-team.ru

Pavel Serdyukov
Yandex
Leo Tolstoy st. 16, Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

Search and browsing activity is known to be a valuable source of information about user's search intent. It is extensively utilized by most of modern search engines to improve ranking by constructing certain ranking features as well as by personalizing search. Personalization aims at two major goals: extraction of stable preferences of a user and specification and disambiguation of the current query. The common way to approach these problems is to extract information from user's search and browsing long-term history and to utilize short-term history to determine the context of a given query. Personalization of the web search for the first queries in new search sessions of new users is more difficult due to the lack of both long- and short-term data.

In this paper we study the problem of short-term personalization. To be more precise, we restrict our attention to the set of initial queries of search sessions. These, with the lack of contextual information, are known to be the most challenging for short-term personalization and are not covered by previous studies on the subject. To approach this problem in the absence of the search context, we employ short-term browsing context. We apply a widespread framework for personalization of search results based on the re-ranking approach and evaluate our methods on the large scale data. The proposed methods are shown to significantly improve non-personalized ranking of one of the major commercial search engines. To the best of our knowledge this is the first study addressing the problem of short-term personalization based on recent browsing history. We find that performance of this re-ranking approach can be reasonably predicted given a query. When we restrict the use of our method to the queries with largest expected gain, the resulting benefit of personalization increases significantly.

Categories and Subject Descriptors

D.4.6 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505679.

Keywords

Personalization, re-ranking, browsing sessions, search context, machine learning.

1. INTRODUCTION

The classical approach to *web retrieval* considers a user's request aside from her personality, context of the request, nature of the search intent. In the past decade the situation has drastically changed. Now modern commercial search engines tend to employ not only information about the query itself, but also all knowledge about searcher's personality expressed in her long-term interests, context of the current query and her engagement with certain web pages. Recent studies [15, 20, 8, 2] show that properly processed click-through and browsing logs can significantly improve quality of the classical ranking. Browsing and search sessions constitute the major part of user activity studied by researchers and provide one of the most powerful sources of data for personalization of search results and context-sensitive ranking.

Search session is a series of intent-related user's requests to a search engine. Similarly, a series of Web pages visited by a user with similar intent are referred to as a *browsing session.* Modern studies [20, 8] prove that usage of search click-through logs helps to disambiguate the current information need of a searcher as well as to construct her interests profile. At the same time, there is not much known about applicability of the browsing sessions to the same tasks, though several works [19, 1] provide comprehensive studies of the nature of the post-search browsing sessions and suggest various ways to incorporate them into the design of a search engine.

In the current study we address the problem of personalization of the Web search for *new users.* Since long-term search history is sparse or completely absent for them, we perform only short-term personalization. Moreover, we restrict our attention to the most arduous queries: first queries in the search session with lack of even short-term search context. The research questions motivated our study are:

1. What portion of queries cannot be covered by personalization methods based on the short-term search context?
2. To what extent could short-term browsing context improve ranking on those queries?
3. How does the length of the utilized browsing session affect the performance of short-term personalization?
4. How to select exactly those queries that have potential to benefit from the given personalization approach?

By analyzing the search and browsing logs of a major commercial search engine we demonstrate that a considerable volume of queries has very little short-term contextual information from the search session itself. These queries are challenging for the state-of-the-art short-term personalization methods based on *search context*, so we adopt a personalization framework based on re-ranking using short-term *browsing context*. We also extract various statistics from browsing sessions of other users to improve ranking for the current user. Each re-ranking approach has its applicability limits and inappropriate employment could even harm the quality of a search engine. To cope with this issue we present a method of automatic filtering of queries suitable for context-aware personalization.

Our study is completely *data-driven*, i.e. both sets for training and evaluation are extracted from user logs.

To sum up, the contributions of the present study are:

- Utilization of short browsing context in a personalization framework based on re-ranking. Most importantly, it allows to improve ranking for queries with the lack of search-context and for users without long-term interest profiles.
- Development of a framework for automatic learning of a filter that distinguishes queries to be personalized. Each re-ranking algorithm has its own scope, and the filtering directly depends on a personalization algorithm.
- Utilization of knowledge about recent *browsing experience* of other users to improve *search experience* of a new user.
- Comprehensive analysis of importance of various sources of data and features extracted from them for short-term personalization based on browsing context.

The rest of the paper is organized as follows. In Section 2 we describe related work relevant to our research. We briefly describe the data employed for further evaluations in Section 3. In Section 4 we give examples of short- and long-term-based personalization cases and observe some statistics about typical search sessions to motivate our study. In Section 5 we describe in detail our method based on a common framework for personalization. We discuss features employed in our study in Section 6 and experiments themselves in Section 7. Finally, we conclude our research in Section 8.

2. RELATED WORK

Personalization of web search gains more and more interest in the last years. It has been realized that employment of contextual and personal information from browsing and search logs has the potential to significantly improve ranking quality.

The first studies on the subject mostly used short-term search context and long-term search history. In [15] Shen *et al.* incorporated previous queries and their click-throughs to specify the actual information need of a searcher. Authors analyzed several context-sensitive language models based on queries submitted in the same session and snippets clicked on by a user. All methods were evaluated on the TREC collection augmented with click-through data. Xiang *et al.* [20] developed Shen's initial approach further by integrating learning to rank ideas into short-term personalization. They learned a model on click-based and text-based features extracted from search session and evaluated the model on both human labelled data and user clicks. Bennett *et al.* in [2] proposed a unified framework that combines long-term and short-term behaviour of a user in a search engine. It comprises queries issued by a user during a long period of activity and documents clicked on as well as during short-term search session interaction. In [16] Sontag *et al.* utilize user's long-term search history to tune parameters of user-specific ranking model. The model extracts the personal topic distribution on the ODP (Open Directory Project) categories and re-weights the results according to this. The method shows maximal performance on ambiguous acronyms. Let us remind that we tackle the problem of "zero" context in terms of the context studied in these papers, as we assume that we have neither short-term search session context, nor long-term user preferences.

Type\Source of data	Search session	Browsing session
Long-term	[15] [2] [16]	[11]
Short-term	[20] [2]	Current study

Table 1: Types of personalization employed in previous studies.

Recently commercial search engines started to distribute special toolbars that record user activity in browsers. These browsing logs give more specific picture of user's actions on the Web and become one of the major sources of personalization data nowadays. The nature of *search trails* — parts of a browsing log initiated by a query — is the subject of study by White *et al.* in [19]. Authors analyze the evolution of various characteristics (e.g. relevance, novelty) in the course of the trail. The observations in this work demonstrate the possible value of search trails for search engines. White *et al.* in [18] considered a problem of recovering short-term interest using both browsing and search context. They predicted an optimal weight to assign to context in order to combine it with query and predict the ODP label of a clicked document. Note that authors leave the problem of employment of derived context information in web search for a future work and also use the combination of browsing and search context, so do not consider the cases, where there is no short-term search context available. Matthijs and Radlinski in [11] study the usefulness of a set of features extracted from long-term browsing history by tuning a few parameters (without using any actual machine learning algorithm and only on 72 queries). Results of both offline experiment on this set of queries and online evaluation are reported. Unlike Matthijs and Radlinski we utilize *short-term browsing context* for personalization and conduct offline evaluation on the large-scale dataset of 500K queries.

Various types and sources of data for personalization utilized in the previous studies are summarized in Table 1.

3. DATA

As we mentioned above, search and browsing history constitutes the major source of data for short- and long-term personalization. While the entire query log is valuable for long-term personalization, only some recent records are important for short-term personalization. To be more precise, only the pages visited with the same or similar information need could possibly give proper context for the cur-

rent query. The idea of a series of pages (queries) with the same intent is usually formalised in the notion of a *browsing (search) logical session*.

DEFINITION. *Browsing (resp. search) logical session* (or *logical session*) is a subset of the user's browsing (resp. search) log, consisting of intent-related pages, i.e. pages visited with the same or similar search goal.

The problem of accurate identification of logical sessions is rather hard and is a subject of independent studies [3, 9]. To avoid unnecessary complications most studies apply common convention: a set of successive browsed pages (submitted queries) is attributed to one logical session, unless it is followed by 30 minutes of inactivity. This simple heuristic allows efficiently and effectively divide the whole history log into intent-related parts. It is an interesting research question (which is addressed in [3, 9, 18] for search sessions) how quality of the partition affects quality of the data extracted from logical sessions. However this question is out of the scope of our study and, following previous works, we divide the browsing log into sessions on the basis of 30 minutes inactivity timeout. Further these sessions are simply referred to as *browsing sessions*.

Similarly, the web pages visited by a user prior to formulating a query q in the same session are called *browsing context* and queries issued before q respectively are called *search context*.

Now we describe the data organization and information that is available in our study. Our personalization algorithm was trained and evaluated on the fully anonymized real browsing sessions collected from one of the major search engines via its special browser toolbar for 8 days. The toolbar store URLs of all non-personal pages visited by a user and links followed during the browsing. It does not store texts of the visited web pages. In total we have collected 200M browsing sessions. 175M sessions from first 7 days were used for session statistics extraction, see Section 6. The last 8th day was used for the training/test dataset construction: we keep all browsing sessions with at least one query and with at least one page in the browsing session preceding the first query. To purify our data for the 8th day we filtered out all browsing sessions with deliberately useless context, namely we deleted a query if all preceding browsed pages are private or from the same host as the search engine. Since our personalized ranking step and the final evaluation metrics rely solely on click-through data, we deleted all the queries with no clicks on their SERPs. After filtering and all purifications we sampled 500K browsing sessions for personalization training and evaluation.

4. MOTIVATION

Let us start this section with an example of cases covered by each of the long- and short-term personalization methods. First, assume that we have a researcher with specialization in Information Retrieval submitted query 'MRR' into a search engine. Then he is likely to be interested in an article on 'Mean Reciprocal Rank'. At the same time, if an identical query is submitted by an amateur sportsmen in UK, the search engine should return at the first position an article on 'Manchester Road Race'. This is a simple example of an ambiguous query, which is covered by long-term personalization. Second, let the same query be submitted by a searcher for which a search engine has no complete information about her personality, after reading a page with

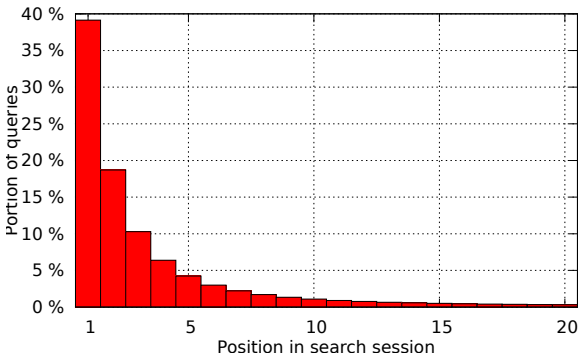

Figure 1: Histogram of positions of queries in their search sessions

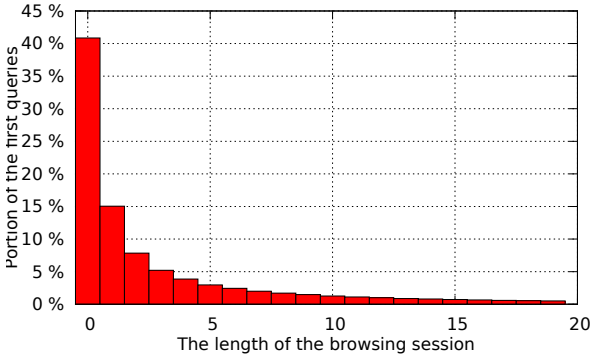

Figure 2: Histogram of lengths of browsing sessions prior to formulating the first query.

title 'Internet marketing'. Then probably the searcher is interested in an article on 'Master Resell Right'. Here is the case, where short-term personalization based on the browsing context comes out. In either case we improve the quality of the original non-personalized ranking by adding personal and contextual bias.

Previous studies on personalization either omit the first queries in search sessions or process them only with the long-term personalization. For instance, in [2] authors consider these queries and report zero performance of the personalization based on the search history and respectively modest improvement of personalization based on long history.

Obviously, the distribution of browsing sessions lengths and of number of queries per sessions heavily depend on various specific characteristics of a search engine. However, we report these distributions on our data to demonstrate the possible benefit from short-term personalization based on the browsing sessions in our setting.

Figure 1 shows the distribution of search positions of queries in their sessions. It follows that 39% of queries have no short-term search context for personalization (the same numbers are reported in [18]) and 58% of queries have ≤ 1 queries for contextualization. It proves that the large portion of queries suffers from the lack of search context for short-term personalization. To overcome these limits of short-term personalization based on search context one needs to find some new source of data.

One obvious solution is to incorporate long-term history. It covers all queries of users with sufficient search history and in that way allows to cope with first requests in search sessions, though it does not help new users. Therefore we decided to analyze the utility of the preceding browsing session for short-term personalization. Figure 2 shows the distribution of lengths of browsing sessions prior to formulating the first query. From this plot we conclude that the majority of first queries in search sessions has very rich browsing context: all except 41% of queries have at least one preceding web page in the same session. Although most queries have some browsing context, we are interested in the range of applicability of this data and restrict ourselves considering $K = 1, \dots, 10$ pages visited prior to the current query.

5. METHOD

In this section we describe our method of personalization. We start with presenting the general personalization framework based on a re-ranking approach [20, 8, 2]. This framework is rather popular due to the simplicity of its implementation, the transparency of the ranking process and possibility of additional tuning. This framework is the core part of our approach and we also discuss its specifications. Essentially, the method runs in two steps: click prediction and re-ranking. Afterwards, we describe the additional query filtering step, which identifies requests that can be potentially enhanced by our personalization method.

5.1 Click prediction problem

Let $\mathcal{L} = \{(u, q, \mathcal{S}(q), \mathcal{C}(q))\}$ be a set of records from a click-through log: each record contains user's u query q, search engine result page $\mathcal{S}(q) = \{s_1, \dots, s_{10}\}$ and her clicks $\mathcal{C}(q) = \{c_1, \dots, c_{10}\}$ on the SERP, $c_i \in \{0, 1\}$. At the *click prediction* step we aim to recover c_i for each sample of the form $\{u, q, s_i\}$. Namely, a sample $\{u, q, s_i\}$ receives a positive judgement if document s_i was clicked by the user for the query q and gets negative judgement otherwise. Our goal is to learn labels c_i from the set of features. That is a standard way to get personalized relevance judgements [2, 16, 11].

The features employed in this step are determined by the type of personalization performed and by the available data. For instance, in the case of long-term personalization we can leverage historical information about clicks, queries and pages visited by the user. On the contrary, for short-term personalization we use user's actions in the search engine and browser during the current session right before formulating the query. Note that the effective combination of these methods is an interesting and challenging research question [18, 2], although it is not the subject of the current study. As we have mentioned, we restrict our attention to the first queries of search sessions, so only browsing context could by employed for short-term personalization. For these queries with lack of search context we consider K last web pages from the browsing session as a context:

$$\underbrace{\mathcal{B} = \{b_K, b_{K-1}, \dots, b_1\}}_{\text{Browsing session}} \to q \to \mathcal{S}(q) \to \underbrace{\mathcal{C}(q)}_{\text{Ground truth}} \quad (1)$$

In general click prediction problem is very hard and admits accurate solution in very restrictive cases (see e.g. [12]). However, we are interested in ranking quality of our predictor rather than in its recall, precision or perplexity.

5.2 Re-ranking problem

From now on let's assume that the click prediction problem is solved, i.e. for each sample $\{u, q, s_i\}$ we have value p_i that estimates c_i. The predicted values are utilized to produce new ranking of the top k documents retrieved for q by original ranking. There is a number of methods in previous studies that suggest various ways to perform the re-ranking step. Proper approach depends on many factors including the quality of estimators p_i and features used at the prediction step. Also some freedom in re-ranking step gives the opportunity for additional tuning of personalized ranking.

The simplest way to personalize SERP is just to rank all documents according to the value of p_i, assuming that the higher p_i the more likely it will be clicked. The main disadvantage of the approach is the lack of flexibility: often personalization harms the initial ranking's quality, so we would like get rid of undesirable re-rankings. To approach that problem one can control aggressiveness of personalization by introducing smoothing parameter $\mu \in [0; 1]$ and ranking using Borda's ranking fusion (see e.g. [20]):

$$\text{score}(s_i) = (1 - \mu)\frac{1}{r(s_i)} + \mu\frac{1}{i}. \quad (2)$$

where i is the original ranking position, $r(s_i)$ — the new position. Large smoothing parameter μ results into "almost" baseline non-personalized ranking, i.e. we swap documents only if we are strongly convinced in superiority of one over the other. On the contrary, small values of μ lead to very aggressive personalization which affects large volume of queries. Further in our experiments (see Section 7.4.2) we evaluate the effect of smoothing (2) for various values of μ.

5.3 Query filtering

Any re-ranking method is potentially beneficial only for a certain subset of queries. Some users, sessions and queries in principle could not benefit from certain re-ranking methods: often the latter only harms the baseline ranking quality. To avoid undesirable losses we construct a classifier that predicts the possible gain from personalization for a given query. The core difference from similar classifiers described in previous studies (see e.g. [14, 17]) is its direct dependence on the constructed re-ranking algorithm.

Let R_0 and R_1 be two rankings: the baseline and the new personalized rankings respectively. Assume that for each ranking and any query q submitted by a user u we have some quality measure $\mathcal{M}(R_0, q)$. Then we assign to every query q a difference $\mathcal{M}(R_1, q) - \mathcal{M}(R_0, q)$. This is our ground truth, which distinguishes the queries that benefit from reranking from the others. Our aim is to predict whether the difference is greater or smaller than zero. We extract a set of features for each pair (u, q) and apply a machine learning approach to estimate the value $\mathcal{M}(R_0, q) - \mathcal{M}(R_1, q)$. Any feature $F(u, q, s_i)$ (including s_i-independent features) employed at the click prediction step results into a new feature averaged over all documents from $\mathcal{S}(q)$:

$$AvF(u, q) = \frac{1}{10}\sum_{s_i \in \mathcal{S}(q)} F(u, q, s_i).$$

In our setting we again take user's clicks on SERP $\mathcal{S}(q)$ as implicit relevance judgements and use *reciprocal rank* of the last user's click as a performance measure $\mathcal{M}$ (see equation (3) below). It is important to note that feature and

training set generation for query filtering step can be done automatically given click prediction features and its validation set. So, learning of filter does not require any additional data and is specific for our re-ranking method.

Learned predictor $f(u, q, S)$ allows subtle tuning of a personalization method: re-ranking is applied exactly for the queries q with $f(u, q, S)$ greater than some fixed threshold θ. Setting θ is equivalent to the choice of portion of queries affected by personalization.

6. FEATURES

In this section we fix some notations and give a detailed account of features derived from the available data. Let u be a user who started her search session with a query q. Assume that pages $\mathcal{B} = \{b_K, \dots, b_1\}$ were visited by u in the same logical session prior to formulating query q (see equation (1)). As above, let $\mathcal{S}(q) = \{s_1, \dots, s_{10}\}$ be a result page for q and $\mathcal{C}(q) = \{c_1, \dots, c_{10}\}$ user's clicks on the result page. We pick out up to 10 pages b_i visited before q as the source of contextual information. The tuple $\{u, \mathcal{B}, q, \mathcal{S}(q), \mathcal{C}(q)\}$ is referred to as *configuration*. Our primary goal is to find the best ranking for every configuration. Each configuration plus a document from the result page s_i forms a sample for the click prediction problem with ground truth c_i.

Now we list all features used by our personalized ranker and give a brief motivation for their employment. In general, for accurate short-term personalization it is necessary to solve two problems: 1) understand whether available contextual information is useful for the ranking of the current configuration, and 2) if so, find out which documents of $S(q)$ are relevant to the contextualized information need. We extract features for both of these problems and use them twice: during the click prediction step (Section 5.1) and during the query filtering step (Section 5.3). We report coverage of each feature, as it is important to understand the quality of sparse features such us click-through-based statistics, see Table 2. Some of personalization algorithms utilize textual information from long- and short-term data.However, a browser toolbar records visited URLs only and does not send the content of pages to the search engine in real-time (due to privacy and network latency related reasons).

We summarise employed classes of features as follows:

- **Characteristics of the query q. Features 1-4.** The better non-personalized ranking for a query the less likely we are able to improve its quality, so we use $NumDocs$ as a measure of quality of a query. *Navigational queries* are requests with intention of finding a particular web page. Thus, if a search engine has recognised navigational intent of a query, it is very undesirable to affect first position during re-ranking. To avoid it, we use $Nav(q)$ — predicted probability of query q being navigational, produced with the proprietary classifier inside our search engine. Ambiguous queries are known to benefit from various types of personalization, so we added the entropy of the distribution over topics of the proprietary directory of web sites (similar to ODP) of 10 SERP pages retrieved for a query as a measure of ambiguity.

- **Usefulness of browsing session $\mathcal{B}$ as the context of q. Features 5-12.** Temporal and session-based proximity between documents b_j's and query q is an important signal for potential improvement of initial ranking of $\mathcal{S}(q)$ on the basis of the context derived from b_j. We extract $BrowseTime(b_j)$ as the indicator of possibility of intent shift and the time spent on b_j prior to opening a new tab — $Time(b_j)$.

  Navigational behaviour of the user in the browser prior to formulating a query gives strong evidence of her commitment to the current topic. At any moment of time user has several opportunities: she can follow a link from the current page, go to a previously opened tab, click 'back' button or start a new tab (e.g. by manually typing address, following a bookmark or an external link). We store activities of a user in a *browsing graph*, consisting of visited URL's and followed links (so we do not know which tab user is actually viewing). That is: out degrees of vertices — $Deg(b_j)$, number of leaves — $Leaves$, edge count — $Links$, number of vertices $BrLen$. The larger $BrLen$, $Leaves$, $Time(b_j)$, $Deg(b_j)$ and $Links$ the higher user's interest in the current topic and therefore the closer its relationship with a subsequent query.

- **Click-through-based proximity between s_i and b_j. Features 13-22.** User behaviour during browsing session constitutes a rich source of knowledge about latent semantic connections among web pages. We processed the large-scale browsing log of 175M sessions (see Section 3) and aggregated session statistics about co-occurrence of pages and hosts in one browsing session. The known issue in employment of click-through- and browsing-based features is their sparsity. To overcome this problem, besides basic *page-to-page* statistics we also extracted their *host-to-host* analogues (see "coverage" column in Table 2). Collected raw counts are: $Num_s(p_1, p_2)$ — the number of cooccurrences of pages p_1 and p_2 in one browsing session in the same order, $Num_h(h_1, h_2)$ — the number of cooccurrences of hosts h_1 and h_2 in one browsing session in the same order.

  Our aim is to use recent *browsing experience* of other users to better understand the current *search experience* of a new user. For each pair of pages b_j, s_i from $\mathcal{B}$ and $\mathcal{C}(q)$ we compute various statistics on the basis of these counts: *Conditional probability* of browsing from b_j to s_i:

$$P_O(s_i|b_j) = \frac{Num_s(b_j, s_i)}{\sum Num_s(b_j, \cdot)};$$

  Pointwise mutual information of occurrences of b_j and s_i in one session in the same order:

$$Pmi_O(s_i, b_j) = \log \frac{P(b_j, s_i)}{P(s_i)P(b_j)}.$$

  Also we compute their analogues for hosts and for unordered cooccurrence ($P_U(\cdot|\cdot)$ and $Pmi_U(\cdot, \cdot)$).

- **SERP-aggregated features. Features 33-42.** Besides query- and session-based features we also add their analogues averaged over all documents on SERP (see Section 5.3). These features are the same for all documents s_i corresponding to a given configuration $\{u, \mathcal{B}, q, \mathcal{S}(q), \mathcal{C}(q)\}$.

  Their employment helps personalization for two reasons. First, the averaged features give the evidence on

Table 2: List of features used in predicting SERP clicks. Coverage is the fraction of records with available data for the feature.

	Feature name	Feature description	Coverage
Characteristics of a query			
1	$NumDocs(q)$	Number of documents found for a given query	100%
2	$Nav(q)$	Probability that query q is navigational	100%
3	$NumWords(q)$	Number of words in q	100%
4	$CatEntropy(q)$	Entropy of categories of SERP pages	100%
Characteristics of the current browsing session and its individual pages			
5	$Time(b_j)$	Time spent on page b_i	100%
6	$BrowseTime(b_j)$	Time spent in the browser after viewing b_i prior to formulating a query	100%
7	$P_q(b_j)$	Probability of formulating any query after visiting b_i	100%
8	$BrLen$	Length of the browsing session	100%
9	$Leaves$	Number of *leaves* in the browse graph i.e. pages without outbound links	100%
10	$Links$	Number of links followed during the browsing session	100%
11	$Deg(b_j)$	Number of outbound links followed by the user on page b_j	100%
12	$Deg(h_{b_j})$	Number of outbound links followed by the user on host h_{b_j}	100%
Click-through-based proximity between s_i and b_j, $j \in 1, \dots K$			
13	$\log P_O(s_i\|b_j)$	Log of probability of browsing during the session to s_i after being on b_j	7%
14	$\log P_U(s_i\|b_j)$	Log of probability of browsing during the session to or from s_i being on b_j	8%
15	$\log P_O(h_{s_i}\|h_{b_j})$	Log of probability of browsing during the session to host h_{s_i} after being on h_{b_j}	55%
16	$\log P_U(h_{s_i}\|h_{b_j})$	Log of probability of browsing during the session to or from host h_{s_i} being on h_{b_j}	60%
17	$Pmi_O(b_j, s_i)$	PMI of coocurrence of b_j and s_i in the same order	7%
18	$Pmi_U(b_j, s_i)$	PMI of coocurrence of b_j and s_i in any order	8%
19	$Pmi_O(h_{b_j}, h_{s_i})$	PMI of coocurrence of hosts of b_j and s_i in the same order	55%
20	$Pmi_U(h_{b_j}, h_{s_i})$	PMI of coocurrence of hosts of b_j and s_i in any order	60%
21	$I(s_i = b_j)$	Sites s_i and b_j are the same	100%
22	$I(h_{s_i} = h_{b_j})$	Hosts of s_i and b_j are the same	100%
Click-through-based proximity between s_i and the whole $\mathcal{B}$			
23	$\max_{j=1}^{K} \log P_O(s_i\|b_j)$	The maximal value of $\log P_O(s_i\|b_j)$ over b_j	100%
24	$\max_{j=1}^{K} \log P_U(s_i\|b_j)$	The maximal value of $\log P_U(s_i\|b_j)$ over b_j	100%
25	$\max_{j=1}^{K} \log P_O(h_{s_i}\|h_{b_j})$	The maximal value of $\log P_O(h_{s_i}\|h_{b_j})$ over b_j	100%
26	$\max_{j=1}^{K} \log P_U(h_{s_i}\|h_{b_j})$	The maximal value of $\log P_U(h_{s_i}\|h_{b_j})$ over b_j	100%
27	$\max_{j=1}^{K} Pmi_O(b_j, s_i)$	The maximal value of $Pmi_O(b_j, s_i)$ over b_j	100%
28	$\max_{j=1}^{K} Pmi_U(b_j, s_i)$	The maximal value of $Pmi_U(b_j, s_i)$ over b_j	100%
29	$\max_{j=1}^{K} Pmi_O(h_{b_j}, h_{s_i})$	The maximal value of $Pmi_O(h_{b_j}, h_{s_i})$ over b_j	100%
30	$\max_{j=1}^{K} Pmi_U(h_{b_j}, h_{s_i})$	The maximal value of $Pmi_U(h_{b_j}, h_{s_i})$ over b_j	100%
External features			
31	$Pos(s_i)$	Position of s_i on SERP, namely i	100%
32	$Rel(s_i)$	Relevance of s_i predicted by the search engine	100%
Feature averages over $\mathcal{S}(q)$			
33-42	AvF	Averages of features 13-22 over documents $s_i \in \mathcal{S}(q)$	

how good s_i is in comparison with other documents: the fact some feature $F(s_i)$ is "better" than on average, i.e $F(s_i) > AvF(s_i)$ indicates that s_i should be promoted in the ranking list. Second, averaged features measure the relation between browsing session $\mathcal{B}$ and query q. For instance, large value of $AvP_O(s_i|b_j)$ implies large number of browsing sessions containing page b_j and some page from $\mathcal{S}(q)$, therefore b_j is likely to be related to q.

Features 5, 6, 7, 11-22 were computed for each $j \in 1, \dots K$. Afterwards they were averaged over the most resent k browsed pages b_j for each $k \leq K$. For example, feature $\log P_O(s_i|b_j)$ results into K features:

$$\log P_O(s_i|b_1), \quad \frac{1}{2}(\log P_O(s_i|b_1) + \log P_O(s_i|b_2)), \quad \dots,$$
$$\frac{1}{K}(\log P_O(s_i|b_1) + \dots + \log P_O(s_i|b_K)).$$

Thus, each feature 14-23 and its analogue among features 33-42 results into K features. If there are l browsed pages and $l < 10$ for a given configuration, then we put averages over all l browsed documents instead of averages over $k > l$ documents into the final feature vector. This results into $17 + 25 \cdot K$ features, depending on the number of browsed pages under consideration. To measure the value of the browsing context, we collect the set of features for each $K = 1, \dots, 10$ and train 10 separate personalization algorithms.

7. EXPERIMENTS

7.1 Evaluation

During the evaluation step we compare our re-ranking algorithm with the baseline — non-personalized ranking of a commercial search engine. We stress that commercial engine is highly tuned, thus any (even relatively small) significant improvement of its results is notable achievement. There are several common methods of re-ranking evaluation. One way is to run some online comparison test i.e. present to users both rankings (e.g. interleaved, side by side or via AB-testing [13, 11, 10]) and measure their preferences. However, since presenting possibly inferior results to users is highly undesirable, it should be supported by strong evidence originating from a certain offline evaluation step.

In our study we have conducted offline experiments based on the common *click-through-based evaluation*, see e.g. [5, 11, 2]. In this approach we assign positive judgements exactly to the results which were clicked on by a user and compute some standard click metric (e.g. average first click position, mean reciprocal rank) before and after re-ranking — to compute the metric on a re-ranked list we assume that the user clicks the same set of documents in spite of re-ranking. The relative change of a click metric measures the quality of re-ranking. The method allows effective tuning of parameters and does not require any additional judgements or users' effort besides implicit relevance feedback. However, click-through-based evaluation is known to have certain weak points. Since users tend to click on the documents ranked higher in the result page independently on their relevance and to skip possibly relevant result in the bottom of the ranking list, this evaluation method just gives the lower bound on the algorithm performance, as was also pointed out in [16].

Rk.	Feature	Sc.
1	$Pos(s_i)$	46
2	$Nav(q)$	20
3	$NumWords(q)$	4.7
4	$NumDocs(q)$	1.7
5	$CatEntropy(q)$	0.9
6	$\max Pmi_U(h_{b_j}, h_{s_i})$	0.63
7	$\max \log P_U(h_{s_i}\|h_{b_j})$	0.57
8	$AvPmi_U(h_{b_2}, h_{s_i})$	0.5
9	$BrowseTime(b_5)$	0.5
10	$Links$	0.47
11	$\max Pmi_U(b_j, s_i)$	0.44
12	$Deg(h_{b_{10}})$	0.44
13	$Deg(h_{b_9})$	0.42
14	$BrowseTime(b_6)$	0.42
15	$Av \log P_U(s_i\|b_3)$	0.40

Table 3: Top 15 features according to the contribution to the click prediction model.

Further we use *Min Reciprocal Rank* ($MinRR$) as the basic click metric and tune all algorithms relying on it.

$$MinRR = \frac{1}{Q}\sum_q \frac{1}{rk(s_i(q))}, \tag{3}$$

where $rk(s_i(q))$ is the new rank of the last clicked document for a query q. This metric is very similar to the one used in previous works on personalization (see e.g. [16, 4]) and gives more weight to relevant documents at the top of the list, thus improvements at the top are more important. We also report improvements in terms of several popular click metrics: *Mean Reciprocal Rank* ($MeanRR$) and *First Click Position* (FCP), considering all clicked documents as relevant.

7.2 Click prediction

On the collected dataset we run a proprietary implementation of Friedman's gradient boosted decision tree-based machine learning algorithm [6]. The processed dataset, collected during the 8th day, contains 500K configurations and 5M samples for learning. We divided it into 5 equal parts on the basis of unique user ids and performed 5-fold cross validation.

In Table 3 we report top 15 features according to the weighted contribution into the final click prediction model (see [7, Section 10.13] for the description of those weights). It measures weighted improvement of the loss function over all employments of a feature during the learning process. Similar method of evaluating feature strengths was used in [4]. Apparently, Pos is the strongest feature for click prediction, since it represents the relevance score of the search engine and thus is highly correlated with the fact of a click. Similarly, Nav is correlated with clicks on the first result. Interestingly, features $Deg(h_{b_{10}})$ and $Deg(h_{b_{10}})$ are relatively high, representing the value of the whole browsing session for personalization

7.3 Re-ranking

In this subsection for a given set of features and the click prediction algorithm we just reorder documents in $\mathcal{S}(q)$ according to the predicted probability of a click, i.e. we do not

	All queries	Meas. different
+/=/-	3.5K/94K/2.5K	3.5K/0/2.5K
MinRR*	+0.05%	+3.7%
MeanRR*	+0.013%	+0.5%
FCP*	-0.18%	-1.7%

Table 4: Improvements of click metrics ($K = 10$) on the whole stream and on measurably different queries. *significant with $p < 0.01$ (t-test).

apply any smoothing or query filtering and report all metric changes on the test log. Further we use the similar way to analyze the benefit of personalization as [2, 16]. A query q is said to be *measurably different* if the original ranking and the personalized one have different $MinRR$ scores. Obviously, the major part of queries is not affected by personalization, since the large portion of queries either has already the optimal value of a click metric (like the only click at top 1), or it is supported by a weakly relevant browsing context. However, even in this case we detect significant improvements of click metrics on the full query stream for all models and affect up to 6% of all queries. Interestingly, in [2] authors report 5,42% of queries being affected by search session-based short-term personalization.

We re-rank documents according to the learned models and evaluate its quality. In Table 4 we report relative change of click metrics among all queries: min reciprocal rank, mean reciprocal rank and first click position. Due to proprietary nature of our data we do not report absolute values of the metrics. The numbers in the second column are improvements on the whole dataset and the numbers in the third column are improvements only over measurably different queries. Further we show how to increase the degree of improvement by focusing on specific queries.

7.3.1 *Value of browsing context.*

Now we report the impact of K — the number of pages extracted from the browsing session for each query on the performance of personalization. Intuitively the more pages we consider, the better the quality of the personalized ranking. At the same time we expect some *saturation* with the growth of K, since distant browsed pages give less and less new and useful information for personalization. To support these expectations for each $K = 1, \dots 10$ we extract the set of features from Section 6 and learn separate personalization algorithm. Improvements of $MinRR$ on measurably different queries and fractions of affected queries as functions of K are shown on figures 3 and 4 respectively. Both plots confirm monotonic behaviour of the functions. Thus, with the growth of browsing context both coverage of personalization and its quality increase. Further we consider only $K = 10$ as the most beneficial case.

7.3.2 *Impact of the query characteristics*

Usually personalization is most beneficial on underspecified ambiguous queries (like 'MRR'), and it is widely accepted to measure its value for a search engine by analyzing not the entire query flow, but those hard cases [2, 16]. We evaluated our personalization independently on navigational/informational queries (according to our proprietary classifier) and on one word / two words / verbose (≥ 3 words) queries. Relative improvements of click metrics are

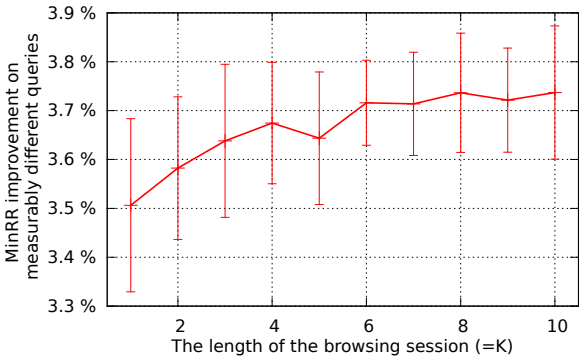

Figure 3: Performance of personalization learned on $K = 1, \dots 10$ browsed pages with 95% confidence intervals.

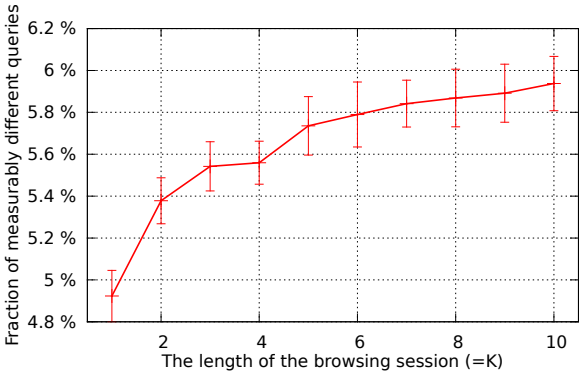

Figure 4: Fraction of measurably different queries for personalization learned on $K = 1, \dots 10$ browsed pages with 95% confidence intervals.

Words	Meas. diff.%	MinRR	MeanRR	FCP
1	4%	+5.2%	+1.6%	-3.7%
2	5.5%	+4.4%	+0.9%	-2.3%
≥ 3	6.8%	+3.1%	+0.16%	-0.96%

IsNav	Meas. diff.%	MinRR	MeanRR	FCP
+	2%	+5.3%	+1.5%	-4%
-	7%	+3.6%	+0.4%	-1.4%

Table 5: Performance of personalization on various sets of queries. Relative improvements are computed on the measurably different configurations. All improvements over baseline are significant with $p = 0.01$.

Rank	Feature	Contribution
1	$NumWords(q)$	7.5
2	$NumDocs(q)$	6.0
3	$CatEntropy(q)$	4.2
4	$Nav(q)$	3.1
5	$BrowseTime(b_1)$	2.7
6	$P_q(b_1)$	2.0
7	$Time(b_3)$	1.8
8	$AvPmi_U(h_{b_9}, h_{s_i})$	1.8
9	$AvPmi_U(h_{b_1}, h_{s_i})$	1.7
10	$Av \log P_U(s_i \vert b_1)$	1.6

Table 6: Top 10 features according to the contribution to the query filtering model

reported in Table 5. All improvements are reported on *measurably different* queries.

As we expected, the one word queries receive significantly more gain than any other class of queries, because of their ambiguity. Expectedly, portion of navigational queries affected by personalization is relatively small. Though, surprisingly, its improvement (on measurably different queries) even more than for non-navigational queries. Reported metric improvements are comparable with other works on personalization, see e.g. [16, 4].

7.4 Query filtering

We use every click prediction model learned on one of the 10 training folds to measure the change of MRR on the configurations from the corresponding validation fold. On this data we learn and tune query filtering using gradient boosting decision tree learning approach. The resulting function $f(u, d, q, S)$ estimates expected gain from the personalization for the corresponding configuration.

7.4.1 Feature contribution into the query filtering

In Table 6 we report contributions of top 10 features to the performance of the query filtering model $f(\cdot)$. Characteristics of a query are still the most important for the query filtering model. Expectedly, features of the most recent browsed page b_1 are more valuable then others.

7.4.2 Impact of smoothing and query filtering

Personalization of search often harms the quality of ranking, thus one needs some tools to control its aggressiveness. We have implemented two methods: smoothing from Equation (2) and query filtering from Section 5.3.

The smoothing just levels out the impact of re-ranking on all configurations and gives more weight to the original ranking. This approach reduces detriment of re-ranking, though it is not much known about its influence on the set of queries still affected by personalization. The extent of smoothing is controlled by parameter μ. On the contrary, query filtering does not change re-ranking, but bounds the volume of configurations affected by personalization. To tune and control it, one chooses threshold θ and applies re-ranking to configuration $\{u, d, q, S\}$, if and only if $f(u, d, q, S) > \theta$.

Dependence of re-ranking performance on smoothing parameter is not as interesting, thus instead we report its dependence on a portion of measurably different configurations. On Figure 5 we demonstrate two plots: performance of the smoothing and the query filtering as functions of a

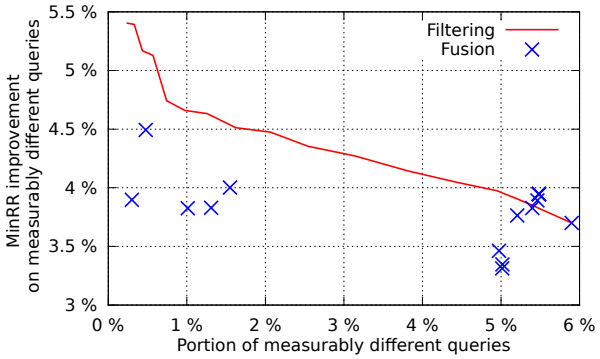

Figure 5: Impact of the volume of queries affected by re-ranking on the MRR performance.

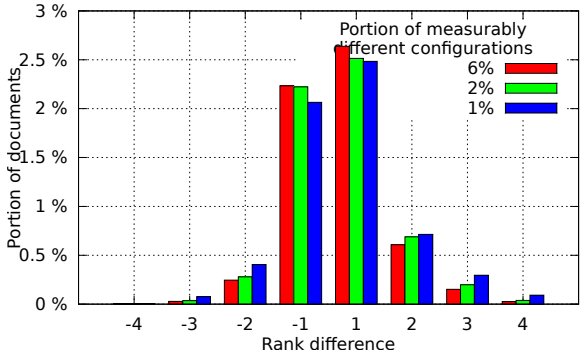

Figure 6: Histogram of rank differences.

portion of affected queries. Again we compute performance only on measurably different configurations. Maximal portion is achieved at $\mu = 0$ (respectively $\theta = -\infty$) and it is equal to 6%. By increasing μ (resp. θ) we reduce the volume of measurably different configurations and move from the right to the left on the plot. While smoothing (blue points) is rather noisy and performs almost equally on all portions of queries, the filtering (red line) method significantly increases its performance on the affected configurations. Therefore query filtering effectively identifies the queries that benefit from re-ranking and significantly advances the quality of the standard Borda fusion.

7.5 Re-ranking variance

Besides average improvement of a click metric (e.g. MRR) it is crucial to control its variance. As in the previous researches [16, 11] we plot a histogram of rank differences, see Figure 6. It demonstrates the distribution of gains and losses in positions of clicked documents. To compare these distributions with previous computations, we perform also query filtering and reduce portion of measurably different configurations from maximum 6% to 2% and 1%.

On the x axis we have difference of the original rank of a document and the rank after personalization, on y axis we have the portion of documents with given rank change among all documents with non-zero rank change. The histogram shows that personalization is more aggressive on the configurations that pass filtering. It is expected, since

these are exactly configurations with maximal possible benefit from personalization.

8. CONCLUSIONS

In the current paper we developed a framework for personalization based on the short-term browsing context. Main target of this personalization are queries with the lack of short-term search session context, particularly, the first queries in the search sessions. We show, that, in the absence of the previous searches, the previously browsed pages provide a rich contextual data for the current query. Our personalization algorithm comprises several orthogonal types of features and finally affects up to 6% of queries. The algorithm demonstrates significant improvements of the basic click metrics over the competitive baseline on the real browsing log. To avoid undesirable harm of personalization on certain queries, we explore the problem of preliminary query filtering and construct a framework for automatic model-specific query filtering. This filtering advances a common smoothing method and significantly increases metrics on the impacted queries. As well as the most of the context-aware re-ranking algorithms our method has the largest potential for improvement on one word queries and informational queries.

Our study can be developed in several directions. First, it will be interesting to estimate the range of applicability of filtering methods and evaluate them for other re-ranking algorithms. Second, the present work touches on the problem of query formulation rationales. That is, prediction of the appearance of an information need during the browsing session and detection of the dissatisfaction with the current browsing session. In this study we have employed just several simple statistics to measure the probability of switching from a browsing session to a search session, so in the future we are planning to develop a more complicated approach to predict the emergence of an information need.

9. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. SIGIR '11, pages 345–354, New York, NY, USA, 2011. ACM.

[2] 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. SIGIR '12, pages 185–194, New York, NY, USA, 2012. ACM.

[3] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. CIKM '08, pages 609–618, New York, NY, USA, 2008. ACM.

[4] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. CIKM '11, pages 403–412, New York, NY, USA, 2011. ACM.

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

[6] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, Feb. 2002.

[7] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001.

[8] D. Jiang, K. W.-T. Leung, and W. Ng. Context-aware search personalization with concept preference. CIKM '11, pages 563–572, New York, NY, USA, 2011. ACM.

[9] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. CIKM '08, pages 699–708, New York, NY, USA, 2008. ACM.

[10] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: five puzzling outcomes explained. KDD '12, pages 786–794, New York, NY, USA, 2012. ACM.

[11] N. Matthijs and F. Radlinski. Personalizing web search using long term browsing history. WSDM '11, pages 25–34, New York, NY, USA, 2011. ACM.

[12] B. Piwowarski and H. Zaragoza. Predictive user click models based on click-through history. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 175–182, New York, NY, USA, 2007. ACM.

[13] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? CIKM '08, pages 43–52, New York, NY, USA, 2008. ACM.

[14] M. Rahurkar and S. Cucerzan. Predicting when browsing context is relevant to search. SIGIR '08, pages 841–842, New York, NY, USA, 2008. ACM.

[15] X. Shen, B. Tan, and C. X. Zhai. Context-sensitive information retrieval using implicit feedback. SIGIR '05, pages 43–50, New York, NY, USA, 2005. ACM.

[16] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck. Probabilistic models for personalizing web search. WSDM '12, pages 433–442, New York, NY, USA, 2012. ACM.

[17] J. Teevan, S. T. Dumais, and D. J. Liebling. 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*, SIGIR '08, pages 163–170, New York, NY, USA, 2008. ACM.

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

[19] R. W. White and J. Huang. Assessing the scenic route: measuring the value of search trails in web logs. SIGIR '10, pages 587–594, New York, NY, USA, 2010. ACM.

[20] B. Xiang, D. Jiang, J. Pei, X. Sun, E. Chen, and H. Li. Context-aware ranking in web search. SIGIR '10, pages 451–458, New York, NY, USA, 2010. ACM.