Mining User Dwell Time for Personalized Web Search Re-Ranking

Songhua Xu‡ **Hao Jiang**§ **Francis C.M. Lau**§

‡: Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
USA

§: Department of Computer Science
The University of Hong Kong
Hong Kong S.A.R., P.R. China

Abstract

We propose a personalized re-ranking algorithm through mining user dwell times derived from a user's previously online reading or browsing activities. We acquire document level user dwell times via a customized web browser, from which we then infer concept word level user dwell times in order to understand a user's personal interest. According to the estimated concept word level user dwell times, our algorithm can estimate a user's potential dwell time over a new document, based on which personalized webpage re-ranking can be carried out. We compare the rankings produced by our algorithm with rankings generated by popular commercial search engines and a recently proposed personalized ranking algorithm. The results clearly show the superiority of our method.[1]

1 Introduction and Related Work

In web search, personalized ranking can significantly improve user experiences, which has thus attracted a great deal of interest in both academia and industry. A personalized ranking method or algorithm needs to estimate individual users' interests over search result documents according to some types of user feedback. However, acquiring explicit user feedback to determine the relevance of a search result document to a web search query is not practical if not infeasible. In comparison, implicit user feedbacks, such as browsing and click history, and user dwell times over a document are more practical, which are most suitable for developing personalized web search algorithms and services, due to their uninstrusiveness in the user search process and the abundance of data available using the approach. In this paper, we propose a personalized webpage re-ranking algorithm through exploring a user's dwell times in his/her previous readings over individual documents. Based on them we then infer concept word level user dwell times for personalized search result re-ranking for the user.

The closest related work to this paper is the personalized webpage ranking algorithm proposed in [Xu *et al.*, 2008b], which measures a user's reading interest based on user attention time, i.e., user dwell time. However, their method represents user dwell times only at the level of individual documents, which could result in significant computational overheads when the number of documents a user has previously read is large. The lack of representation of user dwell times at a finer granularity also affects their algorithm's accuracy. In this paper, we address both issues by inferring concept word level user dwell times from originally captured document level user dwell times. Liu et al. [2002] built a personalized web search system by mapping user queries to multiple categories. Teevan et al. [2005] constructed a user interest model for personalized search from various types of search-related information, e.g., user submitted queries, browsed webpages, and personal emails. Sugiyama et al. [2004] understood user interest according to each user's personal needs in his/her previously searches. Kelly and Belkin [2004] studied the relationship between user interests and user dwell time, focusing on understanding task effects of treating display time as implicit feedback. Speretta and Gauch [2005] located multiword phrases to enhance the bag-of-words representation of a document for predicting a user's personal interest from the user's previous webpage access patterns. Dou et al. [2007] utilized five different personalized search approaches to perform searches over 12-day MSN query logs, and concluded that personalized searches cannot always outperform common web searches. They further pointed out that click entropy provides a good criterion for whether a query should be personalized. Unlike their algorithm, which mines browsing history for personalized web searches, our approach analyzes a user's past reading and search activities in terms of user dwell times for personalized re-ranking.

2 Mining User Dwell Time for Personalized Re-Ranking

2.1 Acquiring User Dwell Time

We implemented the customized web browser suggested in [Xu *et al.*, 2008b] for acquiring user dwell time. In the up-

[1]This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

coming sections, we will not distinguish between documents and webpages as they are treated the same in our method.

2.2 Representing a Document as a Collection of Concept Words

In our method, we represent each document D_i previously read by a user u_k, as a collection of concept words. We use this representation because we mine user dwell times at the concept word level in our method. Compared with [Xu *et al.*, 2008b]'s approach, which represents user dwell times at the level of an entire document, our concept word level user dwell time representation makes the training samples more reusable. Given a document D_i, we first extract its main content text using an HTML filter. In this step, we carry out several preprocessing substeps to remove non-content information such as tags, advertisements, website navigation bars, and redirection links from the webpage text using the method proposed in [Ntoulas *et al.*, 2006]. For the remaining words, we check whether there is a corresponding definition page for the word in Wikipedia. If such a definition page exists, we recognize the word as a concept word. In this way, we can extract all the concept words contained in a document as well as each word's occurrence number. (The connection between concepts and Wikipedia articles has been intensively studied and well established in prior studies, e.g., [Gabrilovich and Markovitch, 2007].) Formally, we can represent D_i as a collection of concept words, i.e., $CV(D_i) \triangleq \{(C_j, n_j)|j = 1, 2, \cdots, z(D_i)\}$, where C_j is a concept word in D_i, which occurs n_j times, and $z(D_i)$ is the total number of concept words in D_i.

2.3 Modeling Document Level User Dwell Time and Concept Word Level User Dwell Time

A Reduced Case

To quantitatively model the relationship between document level user dwell time and concept word level user dwell time, we first consider an extremely simplified case where the whole document consists of only one concept word, which may occur multiple times. Without loss of generality, we assume the concept word is C_i and the number of its occurrence in the document is n_i. According to the cognitive neuroscience phenomenon of semantic satiation [Satiation, 2011], human brain has a fatigue mechanism where the more times a stimulus is repeatedly received by our brain in a short span of time, the less aroused our brain becomes. After a large number of exposures to the same stimulus, our brain may eventually become irresponsive to the stimulation for a certain period of time, where the period length depends on the frequency and each duration of the repeatedly exerted stimulus. Based on this study result from the cognitive science community, we introduce the following modified logistic sigmoid function to approximately model this saturation effect of our brain due to brain fatigue:

$$\vartheta(C_i, u_k, n_i) \triangleq \frac{\alpha_2 \vartheta(C_i, u_k, 1)}{\alpha_2 - 1 + \exp\big(\alpha_1(1 - n_i)\big)}, \qquad (1)$$

where $\vartheta(C_i, u_k, 1)$ and $\vartheta(C_i, u_k, n_i)$ respectively represent the dwell times of the user u_k due to the exposure of the concept word C_i for the first time and for the first n_i times when reading a document. α_1, α_2 are modulating parameters that control the numerical characteristics of our modeled brain fatigue mechanism, whose determination will be discussed later.

The Full Case

Now we consider the general case where a document consists of multiple concept words of which some may occur multiple times. For a given document D, we assume it carries the concepts $C_1, C_2, \cdots, C_{z(D)}$. We also denote the number of occurrences of the concept word C_j in D as n_j. We order all the concept words according to the respective occurrence numbers of these words in a descending order, i.e., $n_1 \geqslant n_2 \geqslant \cdots \geqslant n_{z(D)}$. In our modeling of the relationship between document level user dwell time and concept word level user dwell time, we introduce some auxiliary versions of the original document D where occurrences of certain concept words are removed. Let $D^{\widehat{i}}$ be the auxiliary version of the original document D where only occurrences of the first i concept words $C_1, C_2, \cdots, C_i$ are retained. By this definition, $D^{\widehat{z(D)}}$ is the original document D itself since the full set of concept words is retained; $D^{\widehat{1}}$ is the auxiliary version of the original article D that only consists of occurrence of the first concept word C_1, which is the simplified case examined above.

Let $\phi(u_k, D^{\widehat{i}})$ be the dwell time of the user u_k over the auxiliary document version $D^{\widehat{i}}$. We assume the dwell time of the user u_k over the auxiliary document version $D^{\widehat{i}}$, denoted as $\phi(u_k, D^{\widehat{i}})$, is the dwell time of the user u_k over the document version $D^{\widehat{i-1}}$, $\phi(u_k, D^{\widehat{i-1}})$, plus the dwell time of the user u_k due to the exposure of the concept word C_i after seeing $D^{\widehat{i-1}}$, denoted as $\varphi(C_i, u_k, D^{\widehat{i-1}})$. This relationship can be formulated as follows:

$$\phi(u_k, D^{\widehat{i}}) \triangleq \begin{cases} \phi(u_k, D^{\widehat{i-1}}) + \varphi(C_i, u_k, D^{\widehat{i-1}}) & \text{if } i > 1; \\ \vartheta(C_i, u_k, n_i) & \text{if } i = 1. \end{cases} \qquad (2)$$

When estimating the term $\varphi(C_i, u_k, D^{\widehat{i-1}})$, we also consider an *inhibition effect*—the more related C_i is to a user's previously encountered concept words $C_1, C_2, \cdots, C_{i-1}$, the less excited a user, or more precisely the user's brain, will be when encountering the concept word C_i. We therefore modify the single concept word user dwell time modeling equation (1) defined in the above to take into consideration such an inhibition effect. The form of our model is:

$$\varphi(C_i, u_k, D^{\widehat{i-1}}) \triangleq \frac{\alpha_2\ \vartheta(C_i, u_k, 1)}{\alpha_2 - 1 + \exp\Big(\alpha_1(1 - n_i - \sum_{C_j \in D^{\widehat{i-1}}}(s(C_i, C_j)n_j))\Big)}. \qquad (3)$$

In (3), $s(C_i, C_j) \in [0, 1]$ is the pairwise concept word semantic relatedness between the concept word pair C_i and C_j, which is calculated according to [Gabrilovich and Markovitch, 2007]. $s(C_i, C_j) = 0$ if C_i and C_j are completely irrelevant; $s(C_i, C_j) = 1$ if C_i and C_j are identical. n_j is the number of times C_j occurs in the document version $D^{\widehat{i-1}}$.

2.4 Inferring Concept Word Level User Dwell Time

Problem Formulation

Let $\mathcal{D}_k = \{D_1, D_2, \cdots\}$ be the full set of documents the user u_k has read previously, and the user's dwell time over each document has been captured. We denote user u_k's captured dwell time over the document D_i as $t(u_k, D_i)$ where $D_i \in \mathcal{D}_k$. Let $|\mathcal{D}_k|$ be the cardinality of the set $\mathcal{D}_k$. Let $\mathcal{C}_k = \{C_1, C_2, \cdots\}$ be the full set of concept words that occur in user u_k's previous read document set $\mathcal{D}_k$. According to our modeled relationship between document level user dwell time and concept word level user dwell time introduced in Sec. 2.3, we have the following relationships:

$$\phi(u_k, D_i) = t(u_k, D_i), \ (i = 1, 2, \cdots, |\mathcal{D}_k|). \quad (4)$$

The variables involved in (4) include user u_k's concept word level dwell time $\vartheta(C_i, u_k, 1)$ for all $C_i \in \mathcal{C}_k$ (see (1)) as well as the parameters α_1, α_2. We will look at how to configure α_1, α_2 at the end of Sec. 2.5. For the time being, we assume α_1, α_2 are pre-fixed.

Due to noises in our data capturing procedure, we understand that the user dwell time data we acquire will likely exhibit internal contradiction, making the strict satisfaction of (4) often impossible. Therefore, instead of attempting to find an ideal solution, which may not exist, we aim at finding a most satisfying solution for (4). For this purpose, we convert (4) into the following target function:

$$\Xi(u_k) = \sum_{D_i \in \mathcal{D}_k} \omega(D_i)\big(\phi(u_k, D_i) - t(u_k, D_i)\big)^2, \quad (5)$$

where $\omega(D_i)$ is an observation recency factor, which specifies how long ago the user dwell time sample $t(u_k, D_i)$ was collected. The more aged a sample is, the less reliable and weighted the sample shall be considered as it is well understood that users' interests change over time. In our current implementation, we empirically choose $\omega(D_i)$ to be $\omega(D_i) \triangleq \frac{1}{\exp(\text{\# of days passed since } t(u_k, D_i) \text{ was collected})}$. Overall, by finding a solution that minimizes $\Xi(u_k)$, we can find a most satisfying solution for (4).

It is also noted that each concept word C_i that occurs in a user's past reading activities will introduce a variable $\vartheta(C_i, u_k, 1)$ into (4). Depending on the number of concept words ever appearing in a user's past readings, the total number of unknowns involved in (4) could be quite large, which poses another challenge for solving (4). Lastly, (4) could be under-constrained as well. To address the above challenges in resolving (4), we introduce a constraint based approach.

A Constraint based Approach for Inferring Concept Word Level User Dwell Time

We introduce the following constraint to represent the internal relationships between a user's dwell time over closely related concept words. Our assumption is: *concept words that are more related should possess more similar concept word level user dwell times for the same user*. To mathematically represent the above assumption, we first measure the difference between a pair of concept words, C_i and C_j, in terms of their user dwell time and semantic relatedness respectively. For the former, we use the relative difference between the two concept words' user dwell times as the measurement, which is denoted as: $\delta(u_k, C_i, C_j) \triangleq \frac{|\vartheta(C_i,u_k,1) - \vartheta(C_j,u_k,1)|}{\max\{\vartheta(C_i,u_k,1), \vartheta(C_j,u_k,1)\}}$. For the latter, recall that we use $s(C_i, C_j)$ to define the pairwise relatedness between a pair of concept words C_i and C_j (see Sec. 2.3). Based on the two measurement terms, we can further measure the satisfaction of our above assumption through examining every concept word triple C_i, C_j, C_l. For the this purpose, we introduce the auxiliary function $\Psi(C_i, C_j, C_l, u_k)$ for measuring the satisfaction of our aforementioned assumption by the three concept words C_i, C_j, C_l, which is defined as follows:

$$\Psi(C_i, C_j, C_l, u_k) \triangleq \sum_{m=1}^{3} \Psi_m(C_i, C_j, C_l, u_k) \quad (6)$$

$$\Psi_1\big(C_i, C_j, C_l, u_k\big) \triangleq \big(\delta(u_k, C_i, C_j) - \delta(u_k, C_j, C_l)\big)\big(s(C_j, C_l) - s(C_i, C_j)\big) \quad (7)$$

$$\Psi_2\big(C_i, C_j, C_l, u_k\big) \triangleq \big(\delta(u_k, C_i, C_l) - \delta(u_k, C_j, C_l)\big)\big(s(C_j, C_l) - s(C_i, C_l)\big) \quad (8)$$

$$\Psi_2\big(C_i, C_j, C_l, u_k\big) \triangleq \big(\delta(u_k, C_j, C_i) - \delta(u_k, C_i, C_l)\big)\big(s(C_i, C_l) - s(C_j, C_i)\big) \quad (9)$$

To understand (7), let's examine the following case: the pair of concept words C_i and C_j are semantically more closely related than the pair of concept words C_j and C_l; while the difference in user dwell time of the concept word pair C_i and C_j is smaller than the dwell time difference of the concept word pair C_j and C_l. This is a case that aligns with our assumption; under this circumstance $\big(\delta(u_k, C_i, C_j) - \delta(u_k, C_j, C_l)\big)$ is negative and $\big(s(C_j, C_l) - s(C_i, C_j)\big)$ is negative. Therefore, $\Psi_1\big(C_i, C_j, C_l,\ u_k\big)$ is positive. Following the same reasoning, we can easily examine the other three cases and verify the general property that a positive $\Psi_1\big(C_i, C_j, C_l, u_k\big)$ value indicates a satisfying instance of our assumption while a negative $\Psi_1\big(C_i, C_j, C_l, u_k\big)$ value indicates a violating instance of our assumption. In addition, it is worth noticing that in all the situations, the larger the magnitudes of $\big(\delta(u_k, C_i, C_j) - \delta(u_k, C_j, C_l)\big)$ and $\big(s(C_i, C_j) - s(C_j, C_l)\big)$ are, meaning the more significant the qualitative relationships regarding semantic relatedness and user dwell time difference over the two concept word pairs are, the more our assumption would be satisfied. Such a property is quantitatively embodied in the multiplication operation in (7). All in all, we can see that $\Psi_1\big(C_i, C_j, C_l, u_k\big)$ calculates how well our assumption is satisfied by the pairwise relationship between the concept word pair C_i and C_j versus the concept word pair C_j and C_l. In the same way, we can verify that $\Psi_2\big(C_i, C_j, C_l, u_k\big)$ and $\Psi_3\big(C_i, C_j, C_l, u_k\big)$ respectively calculate how well our assumption is satisfied by the pairwise relationships demonstrated between the other two concept word pair comparison cases. Finally, we can see that $\Psi(C_i, C_j, C_l, u_k)$, defined in (6), measures the overall satisfaction of our assumption by the group of three concept words C_i, C_j, C_l.

Based on the notation of $\Psi(C_i, C_j, C_l, u_k)$, we can now define the constraint for our concept word level user dwell time inference task as follows:

$$\mathbf{\Psi}(u_k) \triangleq \sum_{C_i \in \mathcal{C}_k, C_j \in \mathcal{C}_k, C_l \in \mathcal{C}_k} \Psi(C_i, C_j, C_l, u_k). \quad (10)$$

The goal is to maximize $\mathbf{\Psi}(u_k)$ through finding an optimal assignment over the concept word level user dwell time $\vartheta(C_i, u_k, 1)$ estimate where $C_i \in \mathcal{C}_k$.

Finally, based on the constraint $\mathbf{\Psi}(u_k)$ formulated above, we can redefine the objective function of our optimization problem as follows:

$$\Upsilon(u_k) \triangleq \Xi(u_k) - \mathbf{\Psi}(u_k), \quad (11)$$

where the goal is to minimize $\Upsilon(u_k)$ through finding an optimal assignment over the concept word level user dwell time estimate $\vartheta(C_i, u_k, 1)$ for all $C_i \in \mathcal{C}_k$.

2.5 Optimizing Target Function

Initialization

For each document $D_i \in \mathcal{D}_k$ that the user u_k has read in the past whose document level user dwell time is $t(u_k, D_i)$, we extract all the concept words that appear in D_i: $C_1, C_2, \cdots, C_{z(D_i)}$ whose corresponding occurrence numbers in D_i are $n_1, n_2, \cdots, n_{z(D_i)}$ respectively (see Sec. 2.2). We then distribute the user dwell time $t(u_k, D_i)$ spent over the entire document D_i onto the concept words $C_1, C_2, \cdots, C_{z(D_i)}$ proportionally according to the respective concept word density in D_i, i.e.:

$$\Delta_{u_k}(D_i, C_i) \triangleq t(u_k, D_i) \frac{n_i}{\sum_{j=1}^{z(D_i)} n_j}. \quad (12)$$

By accumulating user dwell times assigned to a concept word from processing all the document level dwell times of the user in this way, we can derive an initial estimation over concept word level user dwell time, i.e.:

$$\vartheta(C_i, u_k, 1) \triangleq \sum_{D_j \in \mathcal{D}_k} \Delta_{u_k}(D_j, C_i). \quad (13)$$

Searching for Optimal Solutions

In the following step, we search for better solutions that can optimize the target function $\Upsilon(u_k)$ through a multidimensional gradient descent based procedure. Recall that earlier in Sec. 2.3, we mentioned in our model of concept word level user dwell time (1), there are free parameters α_1, α_2. In our experiments, we search for optimal value assignment over the parameters based on all the user data we collected. That is, we congregate the user dwell time data captured from all the users participated in our experiments and perform a top level search over our optimization procedure introduced above to minimize the target function $\Upsilon(u_k)$. The identified values for the parameters are $\alpha_1 = 0.33, \alpha_2 = 1.16$. We use the values as the default value assignment over α_1, α_2 for any user of our system. It is also possible to perform a personalized optimization process that only uses a specific user u_k's captured user dwell time data to search for an optimal value assignment over α_1, α_2 for the user. We envision this top level optimization procedure will take place when the user's PC has a long period of idling cycles, e.g., during the wee hours.

2.6 Predicting User Dwell Time for New Documents for Personalized Re-Ranking

Given our estimated concept word level user dwell time, for each document D_x new to the user u_k, we can predict the user's potential dwell time on the document. In our prediction, we first represent the document D_x as a collection of concept words via the document representation procedure discussed earlier in Sec. 2.2. And then we can simply apply (2) to derive the predicted user dwell time $\phi(u_k, D_x)$ on the new document D_x.

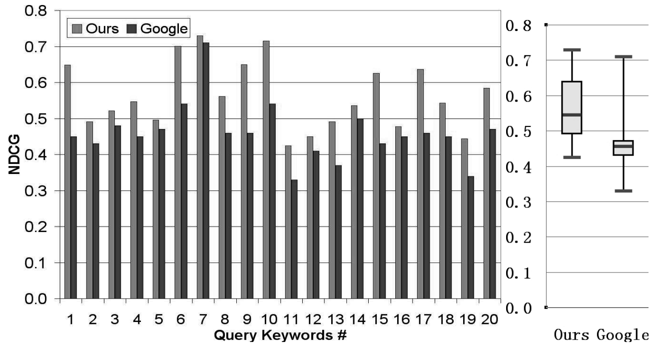
Given the respective predicted dwell times for user u_k over all the search result documents in a search session, we can now re-rank the search result set. We notice however that re-ranking search results only based on a user's reading interest is not sufficient as other factors should also be accounted for, such as the relatedness of a webpage to the query, the reputation of the webpage's source site, and the link structure around the webpage. Fortunately, these issues have been considered by mainstream commercial search engines. In this work, we utilize rankings produced by the Google web search engine to observe these factors indirectly. Also, like any learning based methods, our new algorithm suffers from the cold-start problem, i.e., if there are not enough webpages browsed by a user in the past, our method would not be able to produce reliable prediction on the user's dwell time over new documents due to lack of training data. Thanks to the hybrid re-ranking mechanism that leverages a commercial search engine's ranking results, we can also overcome the cold-start problem.

3 Experiments

In our experiment, to apply our algorithm to re-rank web search results, each time when a web search query is submitted by a user, we first use the Google search engine to acquire the top 300 search result webpages. And then we predict the potential user interest over these 300 webpages using the hybrid version of our method which has incorporated the original Google ranking result. These webpages are then recommended to the user in descending order of the overall predicted user interest. Inspired by [Xu *et al.*, 2008a], we compute an overall predicted user interest $\Gamma(u_k, D_i)$ for user u_k over webpage D_i as follows:

$$\Gamma(u_k, D_i) \triangleq (1-\lambda) \cdot \phi(u_k, D_i) + \lambda \cdot \frac{2\exp(-\gamma \cdot TR_i)}{1+\exp(-\gamma \cdot TR_i)}, \quad (14)$$

where D_i is one of the top 300 webpages fetched by the Google search engine. Recall $\phi(u_k, D_i)$ denotes user u_k's predicted dwell time over the document D_i. The parameter λ ranging from 0 to 1 is a user adjustable value that tunes the balance between our learning based results and the commercial ranking produced by Google. In our experiments, we set λ as $\lambda = exp(-\frac{n_k}{100})$ where n_k is the number of the articles the user u_k has read in the past whose dwell times are known to our algorithm. TR_i denotes the rank of the document D_i produced by Google among these 300 webpages of search results. The parameter γ specifies how sharply this drop-off is, whose value in our experiment is empirically set to 0.2.



(a) (b)

Figure 1: (a) NDCG scores for twenty example queries' search result rankings by our algorithm (with one week's user dwell time as the training data) and by Google respectively. On average, our algorithm achieves a 26% gain compared with Google's ranking results. These twenty example query words respectively are: 1) apple, 2) car, 3) barcelona, 4) da vinci, 5) ETS, 6) gnome linux, 7) greenhouse effect, 8) happy new year, 9) NBA, 10) olympics, 11) wow, 12) great wall, 13) hurricane, 14) iron man, 15) moon, 16) national treasure, 17) porsche, 18) forbidden kingdom, 19) tiger, 20) west lake. (b) A box-plot diagram illustrating NDCG score distribution on rankings of the twenty example search sessions produced by Google and our algorithm respectively.

#	Keywords	NDCG	NDCG/Gain(%)			
		Google	Yahoo	Bing	AT08	Ours
1	blizzard	.58	.53/-9%	.60/3%	.68/17%	.72/24%
2	diving	.39	.48/23%	.38/-3%	.57/46%	.60/53%
3	dweep	.60	.56/-7%	.55/-8%	.66/11%	.82/37%
4	earthquake	.43	.46/7%	.46/7%	.44/3%	.62/43%
5	everest	.41	.40/-2%	.42/2%	.46/13%	.60/47%
6	eyes-on-me	.76	.60/-21%	.78/3%	.82/7%	.89/17%
7	gnome-linux	.54	.44/-19%	.50/-7%	.58/8%	.72/33%
8	grand-canyon	.46	.43/-7%	.46/0%	.47/3%	.58/26%
9	phoenix	.35	.33/-6%	.34/-3%	.42/21%	.50/42%
10	prison-break	.61	.52/-15%	.56/-8%	.68/11%	.75/23%
11	RISC	.45	.48/7%	.47/4%	.45/1%	.63/39%
12	the-beach	.56	.55/-2%	.49/-13%	.64/14%	.77/37%
13	tomb-raider	.64	.56/-13%	.70/9%	.70/9%	.75/17%
14	transformers	.56	.56/0%	.59/5%	.65/17%	.83/48%
15	world-cup	.66	.58/-12%	.56/-15%	.64/-3%	.80/21%
Average		**0.53**	**.50/-5%**	**.52/-1%**	**.59/12%**	**.70/34%**

Table 1: Fifteen example searches and their corresponding NDCG scores by Google, Yahoo, Bing, the "AT08" algorithm, and our algorithm. The NDCG gain measures improvement over Google's NDCG score. In this experiment, "AT08" and our algorithm have access to the first two weeks of user dwell time data as the training. The statistical distributions of these NDCG scores are also graphically reported in Figure 2.

To explore the effectiveness of our algorithm for personalized webpage ranking, we conducted a series of web search experiments using different query words. For each web search session in our experiment, we applied our algorithm to generate a personalized webpage re-ranking over the search results returned by Google. We adopted the normalized discounted cumulative gain (NDCG) to measure ranking quality. To conduct this measurement, we asked each user to assign an integer label between 0 and 4 to every webpage among the top N_{top} search result pages, where N_{top} denotes the number of search result items in the first page of a search result list, which is fixed to be 20 in all our experiments. These labels indicate the user's personal judgement on the corresponding search result's relevance to the query, where a label value of 4 indicates a strongly relevant document; 0 means an entirely irrelevant result. According to these user provided groundtruth labels on search result relatedness to a query, we can compute the NDCG score to measure the ranking quality. We also compared the performance of our personalized webpage re-ranking algorithm with that of the original ranking produced by Google, as well as the ranking produced by a recent user-oriented webpage ranking algorithm [Xu *et al.*, 2008b].

In our experiments, we invited sixteen participants from the graduate school of our university to use our personalized webpage ranking system in their daily life and work for three weeks. These participants are all engineering majored students with rich Internet search experiences. During these three weeks, they were encouraged to search information from the Internet as much as possible, and our system would collect the dwell times of individual users on each search result document. In the third week, each user was asked to perform web search using some given query words. After finishing a web search session using an assigned query word, a user would then be asked to label the top N_{top} search results' respective relevance to the query according to his/her personal judgement and preferences, as discussed above. Such data collected in the third week were used as groundtruth data for individual users in our experiments. For the training data, we prepared two settings for the experiments. In the first setting, we only used a participant's dwell time collected in the first week of our experiment as the training data; in the second setting, we used the participant's dwell time collected in the first two weeks of our experiment period as the training data. No dwell time data collected during the third week was used for training purpose in either setting. We call the first experiment setting "running our algorithm with limited user dwell time data" and the second experiment setting "running our algorithm with increased user dwell time data." In the following, we report experiment results from both environments respectively.

Running our algorithm with limited user dwell time data. Figure 1 reports example query words and NDCG scores of their respective search result rankings produced by our algorithm when running our algorithm with limited user dwell time data. For comparison purpose, we also report NDCG scores of rankings over the same sets of search results generated by *personalized Google search*, i.e., running the Google search function with web search history option turned on.

Running our algorithm with increased user dwell time data.

We repeated our experiments described above, but using the user dwell time data collected from the first two weeks' user web search experiences, to explore the impact of increased amount of user dwell time data on the performance of our

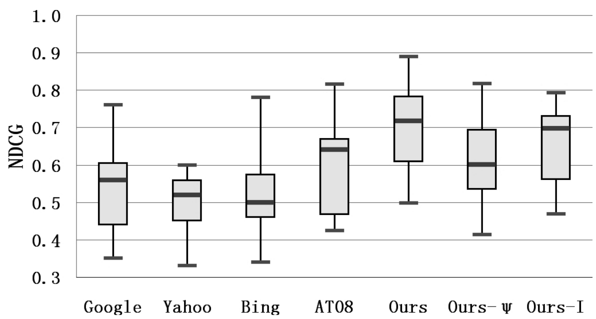


Figure 2: Boxplots on NDCG scores of webpage rankings by three popular commercial web search tools, the "AT08" algorithm, and our algorithm respectively for the fifteen example queries reported in Table 1. We also report NDCG scores of our algorithm without considering the constraint term $\Psi(u_k)$ when inferring concept word level user dwell time, which is denoted as "Ours$-\Psi$", as well as NDCG scores of our algorithm when the iterative optimal solution searching procedure is disabled (see Sec. 2.5), which is denoted as "Ours$-I$".

algorithm for personalized re-ranking. Table 1 reports our experiment results. This table also reports NDCG scores of search result rankings produced by three popular commercial web search engines—Google, Yahoo and Bing. For comparison purpose, we also report NDCG scores of rankings produced by a recent personalized webpage ranking algorithm [Xu *et al.*, 2008b], which also utilized user dwell time information for personalized webpage ranking. In their paper, they call the user dwell time on a document the attention time of the user over the document. Hence we refer to their algorithm by the abbreviation "AT08" in the following. Both our algorithm and the "AT08" algorithm have access to the user dwell time data collected in the first two weeks of the user's web search activities as the algorithm's training data. Figure 2 visually compares statistic distribution of NDCG scores of these methods using boxplots. The comparison results clearly show the advantage of our algorithm for generating personalized webpage rankings that best reflect a user's ideally expected webpage rankings. We also ran two simplified versions of our algorithm, one without imposing the constraint term $\Psi(u_k)$ in (11) when inferring concept word level user dwell time, and the other one without performing the iterative optimal solution search procedure when optimizing our target function in Sec. 2.5. The same set of fifteen example queries were used in these two sets of comparison experiments. We report the comparison results in the same figure, according to which we can clearly see the effectiveness of the two design features of our algorithm.

4 Conclusion

In this paper, we propose a new personalized webpage ranking algorithm through mining dwell times of a user. We introduce a quantitative model to derive concept word level user dwell times from the observed document level user dwell times. Once we have inferred a user's interest over the set of concept words the user has encountered in previous readings, we can then predict the user's potential dwell time over a new document. Such predicted user dwell time allows us to carry out personalized webpage re-ranking. To explore the effectiveness of our algorithm, we measured the performance of our algorithm under two conditions—one with a relatively limited amount of user dwell time data and the other with a doubled amount. Both evaluation cases put our algorithm for generating personalized webpage rankings to satisfy a user's personal preference ahead of those by Google, Yahoo!, and Bing, as well as a recent personalized webpage ranking algorithm.

Acknowledgement

Songhua Xu performed this research as a Eugene P. Wigner Fellow and staff member at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

References

[Dou *et al.*, 2007] 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.

[Gabrilovich and Markovitch, 2007] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07*, pages 1606–1611, 2007.

[Kelly and Belkin, 2004] D. Kelly and N.J. Belkin. Display time as implicit feedback: understanding task effects. In *SIGIR '04*, pages 377–384, 2004.

[Liu *et al.*, 2002] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *CIKM'02*, pages 558–565, 2002.

[Ntoulas *et al.*, 2006] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *WWW '06*, pages 83–92, 2006.

[Satiation, 2011] Semantic Satiation. Semantic satiation. http://en.wikipedia.org/wiki/Semantic_satiation, last visited: March 2011.

[Speretta and Gauch, 2005] M. Speretta and S. Gauch. Personalized search based on user search histories. In *Web Intelligence'05*, pages 622–628, 2005.

[Sugiyama *et al.*, 2004] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *WWW '04*, pages 675–684, 2004.

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

[Xu *et al.*, 2008a] S. Xu, H. Jiang, and F.C.M. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *RecSys '08*, pages 83–90, 2008.

[Xu *et al.*, 2008b] S. Xu, Y. Zhu, H. Jiang, and F.C.M. Lau. A user-oriented webpage ranking algorithm based on user attention time. In *AAAI '08*, pages 1255–1260, 2008.