Personalization of Search Results Using Interaction Behaviors in Search Sessions

Chang Liu
School of Communication and Information, Rutgers University
4 Huntington Street,
New Brunswick, NJ 08901, USA
changl@eden.rutgers.edu

Nicholas J. Belkin
School of Communication and Information, Rutgers University
4 Huntington Street,
New Brunswick, NJ 08901, USA
belkin@rutgers.edu

Michael J. Cole
School of Communication and Information, Rutgers University
4 Huntington Street,
New Brunswick, NJ 08901, USA
m.cole@rutgers.edu

ABSTRACT

Personalization of search results offers the potential for significant improvement in information retrieval performance. User interactions with the system and documents during information-seeking sessions provide a wealth of information about user preferences and their task goals. In this paper, we propose methods for analyzing and modeling user search behavior in search sessions to predict document usefulness and then using information to personalize search results. We generate prediction models of document usefulness from behavior data collected in a controlled lab experiment with 32 participants, each completing uncontrolled searching for 4 tasks in the Web. The generated models are then tested with another data set of user search sessions in radically different search tasks and constrains. The documents predicted useful and not useful by the models are used to modify the queries in each search session using a standard relevance feedback technique. The results show that application of the models led to consistently improved performance over a baseline that did not take account of user interaction information. These findings have implications for designing systems for personalized search and improving user search experience.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – relevance feedback.

General Terms

Measurement, Human Factor

Keywords

Implicit feedback, search behaviors, document usefulness, task type, personalization

1. INTRODUCTION

User search interactions are complex, but they are also coherent over segments of a search session, for example when evaluating results from a query. Overall, one expects the actions of a user to reflect their task goal and the process they carry out to achieve that goal. It is reasonable, then, to think that evidence of the user's search interests and task goal may be available in observations of their behaviors during the search session. In this study, we investigate the effectiveness of using the observation of people's behaviors during information-seeking sessions for improving and personalizing their interactions with information. We generated document usefulness prediction models from a laboratory experiment of the influence of different types of tasks on task search session behaviors, including a general model and several specific models tailored to different types of tasks. Then these prediction models were applied to TREC 2011 Session Track data to evaluate their retrieval performance.

Our study makes several contributions to the literature. First, we examined user behavioral measures as predictors of document usefulness and built prediction models. Second, we demonstrate the prediction models perform well on an entirely different dataset of user interactions. Third, we found that the prediction accuracy of document usefulness and some of users' search behaviors are positively related to the consistent increase in retrieval performance by the personalization model.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

2. RELATED WORK

2.1 Sources for Implicit Relevance Feedback

Users of information retrieval systems can provide explicit information about their interests and task intent via relevance feedback. However, users may not be willing to provide explicit relevance feedback, so research has concentrated on techniques to gain relevance feedback information via implicit processes. These implicit relevance feedback (IRF) techniques observe user search behaviors unobtrusively and infer the relevance/usefulness of documents and other information objects from the user interactions. Evidence for IRF comes from various measures of user search behaviors while interacting with individual content pages, search result pages, and other behaviors during a search session [11] [9].

Document usefulness for IRF may be learned from behaviors such as dwell time on each content page, the number of clicks and scrolling on each content page, number of visits to each content page, and further usage of content pages [9]. User behaviors on search result pages have been found to be good indicators of user document preferences [9]. These behaviors include the time on a search result page before the first click, total time on a search result page, click-through and click order of each content page on the search result page, and so on [1] [7]. There is also evidence that the nature of query reformulation during a search episode can be predictive of document usefulness, and perhaps of task type [13]. In the research to date, these sources for IRF have typically been investigated individually, with little work integrating multiple user behaviors over all types of Web pages to predict document usefulness.

2.2 Implicit Relevance Feedback

Correlation of individual measures of user behavior with users' document preferences have been found to vary for different task types [10] [16] and in different search stages [14]. White and Kelly [16] analyzed retrieval performance after implementing an IRF system based on dwell time on content pages when task type information was considered and when the user information was considered. They found that dwell time combined with knowledge of task information improved retrieval performance. White, Ruthven and Jose [18] examined the effect of search stages on the utility of implicit relevance feedback (IRF) and explicit relevance feedback (ERF) in two separate systems. Their results showed that in the IRF system, IRF is used more in the middle of the search than at the beginning or end, whereas in ERF system, ERF is used more towards the end. Liu and Belkin [14] examined the interaction between decision time on content pages, search stage, and document usefulness, and found knowledge of search stage could help improve the interpretation of decision time as IRF, as could knowledge of task and topic. This work suggests that a high performance search personalization model should predict document usefulness taking account of the users' search context, including task type, user task and topic knowledge, search stage, and so on. Generally, the performance of a personalization algorithm should be improved if it incorporates the ability to predict aspects of the user's task and knowledge.

3. USER EXPERIMENT

3.1 Experimental Design

For our initial study, we recruited 32 participants from a domain relevant to our work/search tasks. They were informed in advance that they would receive $20 for participation. To ensure they treated their assigned tasks seriously, they were told that the top 25% who saved the best set of pages for all four tasks, as judged by an external expert, would receive an additional $20.

The participants were undergraduates at Rutgers university majoring in journalism and were between 18 and 27 years old (26 female, and 6 male). Most were native English speakers (78%) with the balance indicating a high proficiency in English. Participants had an average search experience of 8.85 years using a range of browsers (IE, Firefox, Safari, Chrome, and others). Generally, participants rated their search experience high with more Web search experience as compared to online library catalog search. They were generally positive about their average success when conducting online searches.

Each participant was given a tutorial as a warm-up task and then performed four Web search tasks (described in section 3.2). A pre-search questionnaire elicited their prior familiarity with each task, and their knowledge of the task topic. Participants were asked to search using IE 6.0 on our lab computer, were free to go anywhere on the Web to search for information, and were asked to continue the search until they had gathered enough information to accomplish the task. There was a time limit of 20 minutes for each task.

For each task, participants were asked to **save** content pages that were useful for accomplishing the assignment, and could delete saved pages that were later found to be not useful. When participants decided they had found and saved enough information objects for purposes of the task, they were then asked to evaluate the usefulness of the information objects they saved, or saved and then deleted, through replaying the search using a screen capture program. An online questionnaire was then administered to ask about their searching experience, including their subjective evaluation of their performance, and reasons for that evaluation. The order of the four tasks was systematically rotated for each participant following a Latin Square design. After completing four different tasks, an exit questionnaire was administered asking about their overall search experience.

3.2 Search Tasks

The four work tasks and associated search tasks that we identified are presented below. These tasks follow the normal scenario practice as proposed by Borlund [4], and are couched in journalism terms; that is, journalists are typically given an assignment, and an associated task to complete. Each task was constructed using the faceted task classification scheme proposed by Li & Belkin [12], (modified slightly by us) in order to vary tasks systematically by facet values. In the task descriptions, below we identify the facet values for each (indicated in *italics*). These are also specified in Table 1.

Background Information Collection (BIC)

Your assignment: You are a journalist for the New York Times, working with several others on a story about "whether and how changes in US visa laws after 9/11 have reduced enrollment of international students at universities in the US". You are supposed to gather background information on the topic, specifically, to find what has already been written on this topic. **Your task**: Please find and save all the stories and related materials that have already been published in the last two years in the Times on this topic, and also in five other important newspapers.

The BIC task is a *Mixed Product* because identifying "important" newspapers is intellectual, but finding topical documents is factual. It is *Document Level* because whole stories are judged. It has the *Specific Goal* of finding documents on a well-defined topic, but *Unnamed* because the search targets are not specifically identified (compare with CPE, below).

Interview Preparation (INT)

Your assignment: Your assignment editor asks you to write a news story about "whether state budget cuts in New Jersey are affecting financial aid for college and university students. **Your Task:** Please find the names of two people with appropriate expertise that you are going to interview for this story and save just the pages or sources that describe their expertise and how to contact them.

INT is a *Mixed Product*, because defining expertise is intellectual, and contact information is a fact. It is at the *Document Level*, because expertise is determined by a whole page. The *Goal Quality is Mixed*, because determining expertise is amorphous but contact information is specific. It is *Unnamed* because the search targets are not specifically identified in the task.

Copy Editing (CPE)

Your assignment: You are a copy editor at a newspaper and you have only 20 minutes to check the accuracy of the three underlined statements in the excerpt of a piece of news story below.

Topic: New South Korean President Lee Myung-bak takes office

Body: Lee Myung-bak is the 10th man to serve as South Korea's president and the first to come from a business background. He won a landslide victory in last December's election. He pledged to make the economy his top priority during the campaign. Lee promised to achieve 7% annual economic growth, double the country's per capita income to US$4,000 over a decade and lift

the country to one of the topic seven economies in the world. Lee, 66, also called for a stronger alliance with top ally Washington and implored North Korea to forgo its nuclear ambitions and open up to the outside world, promising a better future for the impoverished nation. Lee said he would launch massive investment and aid projects in the North to increase its per capita income to US$3,000 within a decade "once North Korea abandons its nuclear program and chooses the path to openness."

Your Task: Please find and save an authoritative page that either confirms or disconfirms each statement.

CPE is a *Factual Product*, because facts have to be identified. It is at the *Segment Level*, because items within a document need to be found. It has the *Specific Goal* of confirming facts, and it is *Named* because the search targets are specified.

Advance Obituary (OBI)

Your assignment: Many newspapers commonly write obituaries of important people years in advance, before they die, and in this assignment, you are asked to write an advance obituary for a famous person. **Your task**: Please collect and save all the information you will need to write an advance obituary of the artist Trevor Malcolm Weeks.

OBI is a *Factual Product*, because facts about the person are needed. It is at the *Document Level* because entire documents need to be examined. It is *Unnamed* because the search targets are not specifically identified in the task. The *Goal Quality is Amorphous* because "all the information" is undefined.

Table 1 Variable facet values for the search tasks

Task	Naming	Product	Level	Goal (quality)
BIC	Unnamed	Mixed	Document	Specific
CPE	Named	Factual	Segment	Specific
INT	Unnamed	Mixed	Document	Mixed
OBI	Unnamed	Factual	Document	Amorphous

3.3 Data Collection and Behavioral Measures

During the search, all of the participants' interactions with the computer system were logged during the searches on the client side: eye-tracking; mouse movement capture; clicks and URL requests; scrolling; keystrokes; queries and other behaviors. This logging required multiple systems, including Tobii eyetracking software, UsaProxy, the Morae screen-capture program [3], as well as our own software. From these logs, we extracted a variety of measures of the users' search behaviors on search result pages (SERPs) and content pages throughout the session, and during query intervals. A query interval is defined as the interval between two successive queries issued in one search session; and the last query interval is the period after issuing the last query till the completion of the task. The behavioral variables used in the work reported here are listed in Table 2. We classified these behaviors into two groups: behavioral measures on the clicked documents and behavioral measures during the query intervals.

Table 2 Behavioral measures considered for learning predictive models

Behavioral measures	Definition
Behavioral measures on clicked documents	
dwell time	the dwell time on the document, from the point the page is opened till the point the page is closed
number.of. mouseclick	the number of mouse clicks, including Left button down, Right button down, and scrolling, while the page is opened
number.of. keystrokes	the number of keystrokes while the page is opened
visit_id	the number of times a content page has been visited during one search session
Behavioral measures during query intervals	
time_to_first_click	the time before first click on pages after issuing a query
content_mean	the average dwell time on content pages during the query interval
content_sum	the total dwell time on content pages during the query interval
content_count	the total number of content pages visited during the query interval
serp_mean	the average dwell time on SERPs during the query interval
serp_sum	the total dwell time on SERPs during the query interval
serp_count	the total number of SERPs examined during the query interval
prop_content	the proportion of time on content pages of the total dwell time during the interval
interval_time	the total time in each query interval
diff_content	the difference between the dwell time on a content page and the average dwell time on all content pages during the query interval

4. GENERATION OF PREDICTION MODELS

4.1 Building the Models

We used binary recursive partitioning analysis [5] to identify the most important predictors of useful pages. Recursive partitioning is a stochastic learning technique for non-parametric classification problems. It grows decision trees by examining all independent variables and splits a node using the variable that best distinguishes the remaining data in the collection. The tree is grown until all of the data has been assigned to a node. Recursive partitioning has an intuitive interpretation as a collection of rules to classify the dataset. An example (illustrative only) might be "The document is useful if dwell time > 20 seconds and the time to the first click < 34 seconds and there are fewer than 5 SERPs in the query interval". This method has not been used much in identifying predictors of document usefulness or user interests.

Recursive partitioning may be superior to logistic regression when the goal is to correctly classify members rather than optimize overall accuracy. Recursive partitioning can take account of the possible interactions between predictor variables natively in the algorithm.

We used recursive partitioning analysis to generate prediction models of document usefulness based on the potentially important behavioral variables listed in Table 2. Decision trees were generated using these predictors, and we identified the cutoff points based on the observed distribution of the variables in the dataset. To better learn the behavioral measures, we made a training collection that balanced the number of saved pages and not-saved pages by sampling the larger not-saved pages pool. In this way, ten balanced training sets were constructed, each sharing the same saved pages pool. The recursive partitioning method was applied to each sample to generate the prediction model. We then compared the variables and cutoff values in these generated models to identify the repeated variables and cutoff values for the decision-tree based prediction model.

4.2 General Prediction Models

We first built the general prediction model when considering all tasks in the experiment, without taking account of the task type. In this model, three behavioral measures were identified as the important factors: *visit_id*, *dwell time* and *time to first click*.

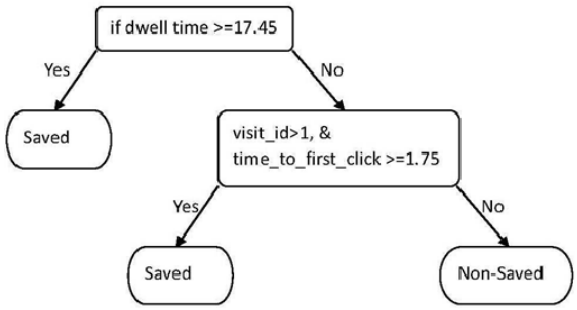

Figure 1 General prediction model

Among these three variables, the *dwell time* on documents is the most important variable in the general prediction model. Examining the distribution of dwell time in the dataset, we found that the model cutoff (17.45 seconds) is the third quarter of the dwell time in all the data. At the same time, we found users saved one quarter of all the visited pages in the assigned task, so it is reasonable to infer that the cutoff point for the dwell time variable depends upon the proportion of saved documents among all visited documents. If the relative frequency of saved documents is not clear in a new dataset, the median dwell time can be adopted as used in White and Kelly [17]. Not many documents were visited more than once by users in one search session, so the model specifies that if a page has been visited more than once, even though the dwell time is shorter than the median, the page is predicted to be useful. Finally, the cutoff time for time to first click is close to the first quarter of the distribution in the dataset. We will apply these cutoff points in our test dataset to evaluate the prediction model.

4.3 Specific Prediction Models

We also conducted 10 random sampling and recursive partitioning analyses on each of the individual tasks to see if tasks of different types would lead to different combinations of important predictors and prediction models. Following the procedure in section 4.1, for each task, we made balanced pools of saved and non-saved pages with the total number of saved pages, and then generated the prediction models using recursive partitioning. Because the CPE task and OBI task in our user experiment differ the most from one another as task types, we present the specific prediction models for CPE (Figure 2) and OBI (Figure 3).

These tree models show that *dwell time* on documents is still the most important predictor. However, the cutoff point for each task is very different, i.e. 30.65 seconds in CPE and 17.5 seconds in OBI. This result indicates that appropriate cutoff points are needed for different types of tasks, as White and Kelly [16] suggested. As for the general model, we suggest the median as the cutoff point for each task. The CPE model also identifies visit time of the page, number of mouse clicks on the page, and the proportion of time spent on content pages during the query interval (*prop_content*) as important variables. Among them, the number of mouse clicks was negatively related to the saved documents, and the *prop_content* was positively related to the saved documents. The OBI model looks very similar to the general model, except that it specifies a range for the *variable time to first click*, which is close to the first quarter to the median value for that variable in the dataset.

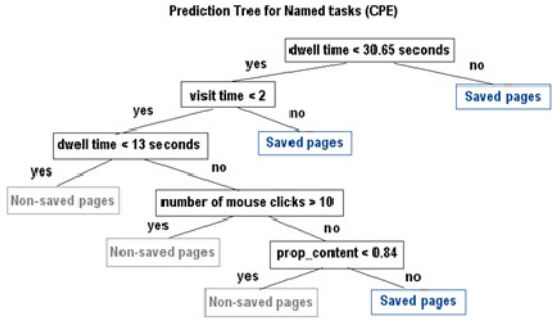

Figure 2. Specific prediction model for CPE

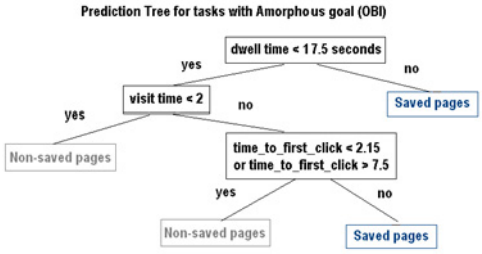

Figure 3. Specific prediction model for OBI

5. IMPLEMENTATION OF PREDICTION MODELS

In order to evaluate the prediction models (both the general model and the specific models) generated from the user experiment, we applied our models to the data collected for the TREC 2011 Session Track. The Session Track [8] aims to provide test collections and evaluation measures for studying information retrieval based on previous user interactions during a search session, rather than one-time queries. The track used the ClueWeb09 collection (English only) [1], and the topics were

[1] lemurproject.org/clueweb09

defined in the usual TREC ad hoc sense, with a title, description, and narrative. Sessions of real user interactions were recorded, including users' queries, clicks on search result pages, and associated time stamps. There were search sessions for 76 topics in total.

5.1 Task Classification

Before implementing the prediction models, we first classified the tasks in the TREC 2011 Session Track using our modification of the Li & Belkin [12] task classification scheme. The 76 sessions were classified into 4 types of tasks according to the task facets. From Table 3 we can see that all of the 76 tasks required searching for factual information, so the product for them is *Factual*. In addition, 66 of the 76 sessions can be classified as searching for information on the *Segment* level, with *Specific* goal, with search targets *Named* (SSN). There are very few examples of three other types of tasks (i.e. SSU, DSN, DAN) in the data set. When we tried to match these task types with the task classification in our user experiment (Table 1), we found that the SSN type of task was mostly similar to CPE, and DSN and DAN types of tasks were very similar to OBI, so we used these specific models for those tasks. We applied the General model for the SSU task type.

Table 3 Task classification of the search tasks in TREC 2011 Session Track

Task type	SSN	SSU	DSN	DAN
Level	**S**egment	**S**egment	**D**ocument	**D**ocument
Goal(quality)	**S**pecific	**S**pecific	**S**pecific	**A**morphous
Naming	**N**amed	**U**nnamed	**N**amed	**N**amed
Product	Factual	Factual	**F**actual	**F**actual
Model applied	CPE	General model	OBI	OBI
number of tasks	66	4	4	2

5.2 Prediction Models on the TREC Dataset

The general model for the TREC dataset was based on the general model we built from the user experiment and the threshold was determined based on the targeted dataset.

The general model (decision tree) is:

if (visit_id > 1),{then it is a useful page};
else if (dwell time > 28.55 seconds), {then it is a useful page};
else if (6.33 seconds **< time-to-first-click <14.55** seconds), {then it is a useful page};
else {non-useful pages}.

The implementation of specific models in the TREC dataset is based on our task classification (Table 3). The CPE specific model is used for SSN type of task, the OBI specific model is used for DSN and DAN types of task, and the other task types use the general model. Since the interaction log for the TREC 2011 Session Track was collected on the server side, and did not include any user interactions on the documents (i.e. content pages), the variable *number of mouse clicks* was not included in the specific models..

The specific model (decision tree) is:

```
if (visit_id > 1), then useful pages
if (task type =  SSN)
        {if (dwell time >=28.84 #Median dwell time for this type of task)
                {then useful pages }
        else if (prop_content>=0.6 #Median for this type of task)
                {then useful pages }
        else { then non-useful pages }}

else if (task type = SSU)
        {if (dwell time >=17  #Median dwell time for this type of task)
        {then useful pages }
else if (task type = DSN or DAN,
        {if (dwell time >=32.82  #Median dwell time for these types of task)
        {then useful pages}
else if (time_to_first_click > 6.33 & time_to_first_click < 14.55)
        {then useful pages }
        else {non-useful pages }}}
```

5.3 Relevance Feedback Technique

For evaluation of the various prediction models in this paper, we use only one of the three types of query modification specified in the TREC 2011 Session Track, which uses the maximum amount of behavioral data obtained during the search session [8]. Since the Session Track's basic concept was to compare the search results of the final query in a search session using a baseline system with no behavioral data, to the results of a final query using such data, we used the results of our document usefulness prediction models to modify the last query but one (*last -1*) in the session, using both positive and (positive + negative) relevance feedback (RF). The TREC 2011 Session Track database, Clueweb09 English only [8] was the searched collection, and was the source for documents and terms for RF.

From the prediction of document usefulness for each of our models, we generated a pool of "relevant" documents that occurred during each search session, and calculated the term frequency for each term in the pool, and in the equivalent corpus of non-useful documents. The observed term frequency was then discounted by the prior of the expectation of appearance in a random document in the English language using the Brown corpus [15].

With respect to the number of useful and non-useful terms for query expansion, we used the approach described in Belkin et al. [2], in which a negative RF system was implemented. The number of suggested feedback terms was determined by the formula:

> 5n + 5, where n = number of judged documents to a maximum of 25 suggested terms.

The query was parsed as a weighted sum, using the default Lemur[2] weighting for RF term addition for positive terms, and adding the negative terms under the InQuery "NOT" operator, with 0.6 weight.

[2] lemurproject.org

Two relevance feedback methods were implemented:

Positive relevance feedback only: In the runs with positive RF only, the predicted "useful" documents were used to calculate the term frequency, the top 25 terms were selected to expand the *last -1* query in the session.

Both positive and negative relevance feedback: In the runs with both positive and negative RF, the predicted "useful" documents were used to calculate the term frequency for "useful" terms and the top 15 terms were selected to be "useful" terms; the predicted "non-useful" documents were used to calculate the term frequency for "non-useful" terms and the top 10 terms were selected to be "non-useful" terms. We then combined the *last-1* queries with the 15 "useful" terms (with positive weight 1.0), and the 10 "non-useful" terms (with negative weight 0.6) using the Indri query language.

Therefore, we have four prediction models of document usefulness generated and evaluated in this study:

- General model with Positive RF only (GP)
- General model with Positive and Negative RF (GPN)
- Specific model with Positive RF only (SP)
- Specific model with Positive and Negative RF (SPN)

5.4 Baseline model

Our baseline model used Pseudo Relevance Feedback on the last queries issued by the users in each session. The default parameters in the Indri Retrieval System were used as follows:

Parameters for Pseudo Relevance Feedback:

```
int fbDocs = _param.get( "fbDocs" , 10 );
int fbTerms = _param.get( "fbTerms" , 10 );
double fbOrigWt = _param.get( "fbOrigWeight", 0.5 );
double mu = _param.get( "fbMu", 0 );
```

We compare the retrieval performance of our personalized models with the baseline system, to evaluate whether the personalized models based on users' interactions in the search session improve results over pseudo relevance feedback that does not take account of users' interaction behaviors.

6. EVALUATION OF PREDICTION MODELS

6.1 Prediction Accuracy of Document Usefulness

We evaluated the prediction accuracy of document relevance against the TREC assessors' relevance judgments [8]. The relevance judgments ranged from -2 to 3 (-2 stands for spam document, 0 stands for not relevant, 1 for relevant, 2 for highly relevant, and 3 means the document is a key to the information need). Since we used binary relevance ratings in our prediction models, we grouped TREC assessors' judgments into two groups: ratings that were -2 or 0 were grouped as "not-relevant", and ratings of 1 to 3 were grouped as "relevant".

Table 4 and Table 5 show the prediction accuracy of the general model and the specific models on the document relevance. They demonstrate that the general model achieved better prediction performance than the specific model for the TREC 2011 Session Track data.

Table 4 Prediction accuracy of the general model

		TREC assessment		
		not- relevant	relevant	total
General model predicted	not-relevant	29 (53%)	45(Type I error)	74
	relevant	26 (Type II error)	83 (65%)	109
	total	55	128	183
Overall accuracy				61%

Table 5 Prediction accuracy of the specific model

		TREC assessment		
		not- relevant	relevant	total
Specific model predicted	not-relevant	24 (44%)	55 (Type I error)	69
	relevant	31(Type II error)	73 (57%)	102
	total	55	128	183
Overall accuracy				53%

6.2 Retrieval Performance Comparison

6.2.1 Overall performance

We first calculated the mean of the retrieval performance over 76 sessions by each of the models (Table 6). Our results demonstrate that all of the four models improve over the baseline, on almost all the measures provided by TREC 2011 Session Track [8].

Table 6 Overall retrieval performance

	Baseline	GPN	GP	SPN	SP
ERR[3]	0.20	0.28	0.29	0.27	0.28
ERR@10	0.19	0.28	0.29	0.26	0.28
nERR[4]	0.29	0.44	0.45	0.41	0.42
nERR@10	0.27	0.43	0.45	0.40	0.42
nDCG[5]	0.30	0.24	0.25	0.24	0.25
nDCG@10	0.20	0.34	0.35	0.32	0.33
AP[6]	0.09	0.09	0.09	0.08	0.09
GAP[7]	0.08	0.09	0.09	0.09	0.09

6.2.2 Retrieval performance by sessions

Expected Reciprocal Rank (ERR) [6] was chosen for the following analysis because the user model underlying the ERR measure matches the user model underlying our prediction models. When examining results by search sessions, we found, as usual,

[3] ERR: Expected Reciprocal Rank;

[4] nERR: ERR normalized by the maximum ERR per query;

[5] nDCG: normalized discounted cumulative gain;

[6] AP: Average Precision

[7] GAP: Graded Average Precision;

that there is great variety in evaluation results for any single technique between sessions. For example, the GP model had better retrieval performance on ERR in 42 sessions, but had worse performance on ERR in 32 sessions.

Table 7 Retrieval performance change over baseline on ERR

	All sessions			
Run Measures	GP	GPN	SP	SPN
Mean ERR	0.29	0.28	0.28	0.27
Percent improvement over baseline (ERR=0.20)	44%	38%	38%	32%
Number of sessions that improve the baseline	40	40	36	36
Number of session that decrease from the baseline	29	32	32	34
Number of session that did not change the baseline	7	4	8	6

6.2.3 The relationship between prediction accuracy and retrieval performance

Among the 76 sessions, there were 12 sessions in which users did not click on any documents. We first compare the retrieval performance and the change over baseline between the sessions that did not contain any clicked documents (N=12) with the sessions that contained at least one clicked document (N=64).

Table 8 Comparison between two types of sessions

Models	Sessions that do not contain clicked documents (N=12)	Sessions that contain clicked documents (N=64)	Mann-Whitney test
	Mean (SD) of retrieval performance evaluated using ERR measure	Mean (SD) of retrieval performance evaluated using ERR measure	p
GPN	0.08 (0.15)	0.32 (0.25)	<0.05
GPN change over baseline	-0.07 (0.3)	0.11 (0.31)	<0.05
GP	0.15 (0.28)	0.32 (0.25)	<0.05
GP change over baseline	0.01 (0.13)	0.11 (0.31)	<0.05
SPN	0.08 (0.15)	0.31 (0.27)	<0.05
SPN change over baseline	-0.07 (0.3)	0.09 (0.32)	<0.05
SP	0.15 (0.28)	0.31 (0.27)	<0.05
SP change over baseline	0.01 (0.13)	0.09 (0.32)	0.095

Table 8 shows that the retrieval performance of our four models and the improvement of our models over baseline (except SP) is significantly better in the sessions that contain clicked documents than in sessions that did not contain any clicked documents. This result shows that our models need clicked document data to improve search results.

Since we are using users' interactions during sessions to build the prediction models of document usefulness, in the following analysis, we present results for sessions that contain at least one clicked document (N=64) to evaluate the performance of our prediction models. We then examine the relationship between prediction accuracy and retrieval performance, by three types of change (decrease, no change, and increase) in retrieval performance compared with the baseline.

We compare the amount of user interaction in each session, and the prediction accuracy of document usefulness by the *general model* among the three groups (Table 9). The amount of user interactions is defined as the sum of the number of queries, the number of search result pages, and the number of content pages visited. Our results show that the prediction accuracy of document usefulness in the sessions for which use of the general model increased the retrieval performance over the baseline was as high as 75%. In contrast, the prediction accuracy was only 51% in the sessions that where personalization decreased the retrieval performance. The Kruskal Wallis test revealed that the differences in the overall accuracy was significant ($p<.05$). Neither the number of clicked documents nor the number of interactions was significantly different among three groups.

Table 9 Comparison of behaviors and prediction in the general model

	Sessions that decrease from the baseline	Sessions that did not change the baseline	Sessions that increase from the baseline	Kruskal-Wallis Test
	Mean	Mean	Mean	p
Number of clicked documents in the session	2.87	2.33	2.84	0.82
Number of interactions in the session	12.57	9.67	12.39	0.70
Overall accuracy of the general model	**51%**	**53%**	**75%**	**<.05**

We also compared the amount of user interaction in each session, and the prediction accuracy of document usefulness by the *specific model* among the three groups (Table 10). As with the general model, we found that the sessions that increased retrieval performance over the baseline by the specific model had significantly higher prediction accuracy of document usefulness (64%) with the specific model as compared to the sessions that decreased the retrieval performance (34%). Again, neither the

number of clicked documents nor the number of interactions was significantly different among three groups.

Table 10 Comparison of behaviors and prediction in the specific model

	Sessions that decrease from the baseline	Sessions that did not change the baseline	Sessions that increase from the baseline	Kruskal-Wallis Test
	Mean	Mean	Mean	p
Number of clicked documents in the session	2.4	2.4	3.21	0.70
Number of interactions in the session	11.44	9.8	13.35	0.45
Overall accuracy of the general model	**0.34**	**0.41**	**0.64**	**.009**

In general, when applying our prediction models on the TREC Session Track 2011 dataset, our results indicate that the better the prediction accuracy, the greater the improvement of retrieval performance over the baseline.

6.2.3.1 When to personalize?

Figure 4 and Figure 5 show the ERR values for each search session for the baseline and for each of our personalization methods. It is clear there are specific search sessions which will benefit from our personalization methods, and others for which such personalization is either not worth the effort, or harmful.

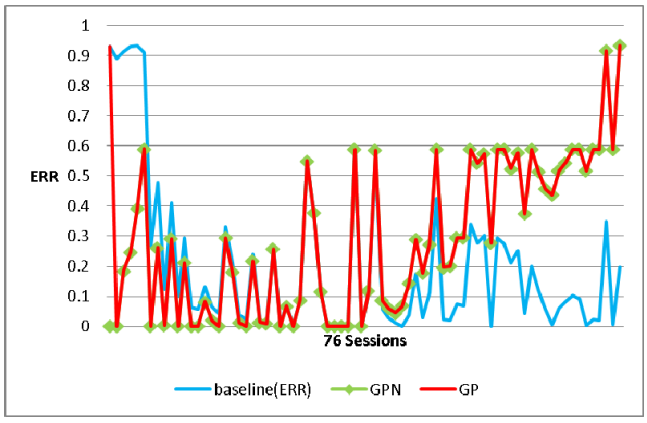

Figure 4 When to personalize using the general model (sessions ordered according to increasing positive difference between GPN and baseline)

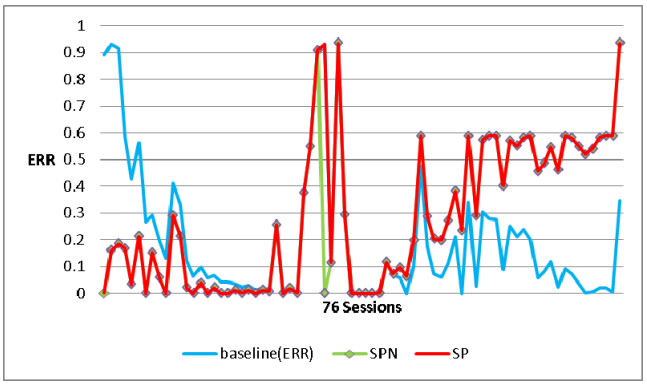

Figure 5 When to personalize using the specific model (sessions ordered according to increasing positive difference between SPN and baseline)

The question is when should one apply personalized models? (cf. [16]) In practice we need to identify these sessions from the observable user interactions during the search session. To accomplish this we focus on interaction behaviors that are associated with sessions where our models were particularly good and particularly bad. Specifically, we look at the behaviors in sessions where our models improved retrieval performance by greater than 0.20 on ERR and for sessions where the models did worse than-0.05 on ERR. We compare the behavioral variables in the search sessions between the two extremes in retrieval performance defined by the GP model in Table 11. For the sessions in which retrieval performance benefited a lot from the personalization model, users spent significantly more dwell time on each content page (51.25 seconds) than in the other group of sessions where the performance was decreased (34.61 seconds). They also spent a relatively greater proportion of task time on reading content pages (43%) than in the other group of sessions (29%). These results, in particular the former, suggest that it is possible to identify search sessions which would benefit from personalization relatively early in the course of the search session.

Table 11 Comparison of behavioral measures between two groups defined by GP model

Behavioral measures	Decreased group (N=13)	Benefit group (N=24)	Mann-Whitney test
	Mean	Mean	p
Total time on task (seconds)	354.27	317.35	1
Total time on content pages (seconds)	102.78	141.71	.19
Total time on SERPs	223.03	141.73	.18
Mean dwell time on each content page (seconds)	**34.61**	**51.25**	**<.05**
Mean dwell time on each SERPs (seconds)	37.43	24.00	.31
Number of query	4.23	3.96	.37
Time to first click (seconds)	33.47	22.44	.19
Query interval time (seconds)	86.97	86.64	.52
Proportion of time spent on content pages in the session	**29%**	**43%**	**<.05**

7. DISCUSSION

In this study, we generated document usefulness prediction models based on a laboratory experiment of the influence of different types of tasks on task session behaviors. We constructed a general prediction model and several specific prediction models tailored to different types of tasks. Recursive partitioning analysis was used to generate decision tree models based on users' interaction behaviors during search sessions. Our models identified three main behavioral measures as important predictors of document usefulness: *dwell time* on document, the number of times a page has been visited in the session (*visit_id*), and the time before first click after issuing a query (*time to first click*). Besides the positive relationship between dwell time and document usefulness, we found that, in all but one case, the *visit_id* and *time_to_first_click* to be positively related to document usefulness. The specific models for different types of tasks show major differences by task type for both the predictors and the rules. For example, in the specific prediction model for CPE, *number of mouse clicks* on the document was negatively related to document usefulness, and the proportion of time spent on content pages during query interval (*prop_content*) was positively related to document usefulness. However, neither of these two measures and associated rules was useful in the specific model for OBI. Furthermore, although all models included dwell time as a significant factor, cut-off points for dwell time to predict usefulness differed between the different tasks, and from that of the general model.

A primary focus of this study was to evaluate whether the prediction models generated from our user experiment could improve retrieval performance, so we applied our models to the TREC 2011 Session Track data. The comparison of the prediction by the different models against TREC assessors' relevance judgments showed the general model had higher accuracy than the specific models. One reason for this is that there is a great imbalance in task type in the TREC 2011 Session Track data. The vast majority of the tasks were of a task type similar to the CPE task in our experiment. Our specific model for CPE contains *number of mouse clicks* on the documents as one of the predictors, but such data was not available in the server-side logs in the TREC 2011 Session Track. The task type imbalance also meant that not enough data was available (10 sessions in total) for three of the four types of tasks (SSU, DSN, DAN). Consequently, one can expect the specific models will fail when they try to adapt to the individual task types. In addition, other facets of the task types may present in the TREC Session Track that differ from those in the user experiment in our lab. Such facets could influence users' behaviors in different ways. For example, users were given 20 minutes for searching in our experiment, whereas in the crowdsourced search sessions in the TREC Session Track, users were given only 5 minutes for searching. Users were likely to have searched with an awareness of time-pressure which might not have been the case during our experiment.

With respect to retrieval performance, all of our models improved a great deal from the baseline which used pseudo relevance feedback on the last queries users issued in each session. Our general model had somewhat better retrieval performance than the specific model on both of the two relevance feedback methods. The models with positive relevance feedback performed only slightly better than the models with both positive and negative relevance feedback. This suggests that using negative IRF may not be worth the effort, although perhaps the limited size of the data set for the TREC Session Track mitigates this conclusion. In addition, we found that the better the prediction accuracy of document usefulness, the more likely that the model could improve retrieval performance.

Teevan, Dumais & Horvitz [16] examined users' clicks and explicit judgments for the same queries and found they differed greatly from one another. They proposed that the *potential for personalization* could be defined by the gap between how well search engines could perform if they were to tailor results to the individual, and how well they currently perform by returning results designed to satisfy everyone. In our study, we also addressed the question of *when to apply personalization models* by presenting the retrieval difference between our models and the baseline. We found our prediction models consistently improved retrieval performance over the baseline in some sessions, but consistently decreased the performance of the baseline line in some other sessions. We compared some users' behavioral measures during these sessions, and found significant differences between the two extreme groups. In particular, users spent significantly more dwell time on each content page and a greater proportion of task time on content pages for the sessions where the personalization models we applied were beneficial. Dwell time on content pages has been proven to be positively related to document relevance, and one explanation for our results is that users were visiting more useful documents in the sessions where the models performed well and that query expansion based on the useful documents helped users to articulate their information needs. Therefore, IRF based on the predicted useful documents achieved better results than using the top *n* returned documents, as in pseudo relevance feedback. Further work is needed to investigate when the system should apply personalization models to improve retrieval performance.

Because the data that we used for our personalization models should be applicable to any general retrieval system, one might expect that our levels of improvement would be applicable to techniques with much higher baseline performance, resulting in higher absolute performance levels. It is also the case that our usefulness prediction models were used as input to quite standard, and rather simple relevance feedback techniques. More sophisticated use of the models could result in better overall performance improvement.

8. CONCLUSION

In this study, we used searchers' interaction behavioral measures derived from a laboratory experiment to build prediction models of document usefulness. We then evaluated the prediction accuracy and retrieval performance of these models by applying them to the TREC 2011 Session Track data set, where users search interactions were generated in radically different search sessions. We found several behavioral measures, e.g. dwell time, visit time, time to first click, proportion of time on content pages, and others could be potential predictors of document usefulness. We also found different combinations of variables and rules of prediction models based on these measures for different types of tasks. The results demonstrate that the prediction models we generated from our user experiment improve the performance over a baseline that did not take account of user interaction information in search sessions with quite different characteristics than those from which the prediction models were developed. The positive "transfer" effect leads us to believe the models we have developed may be used for personalization of retrieval in a variety of searching circumstances, and that we could expect even greater performance benefit when richer, client-side data that our prediction models depend upon are applied.

ACKNOWLEDGMENTS

The research that led to this work was funded by the IMLS, under grant number LG-06-07-0105-07. We thank all of the members of the Personalization of the Digital Library Experience (PoODLE) research team at Rutgers University, without whose efforts this work could not have been accomplished. Thanks to Si Sun for assistance in TREC 2011 Session Track. Thanks to Jingjing Liu for validating task classification as the third coder. Thanks to David Pane at CMU, who helped us greatly and generously in performing the Indri runs.

REFERENCES

[1] Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. (pp. 3-10). Seattle, Washington, USA.

[2] Belkin, N. J., Carballo, J. P., Cool, C., Lin, S., Park, S. Y., Rieh, S. Y., et al. (1998). Rutgers' TREC-6 interactive track experience. *Proceedings of the Sixth Text REtrieval Conference*, 597-610.

[3] Bierig, R., Cole., M.J., & Gwizdka, J. (2009). A user centered experiment and logging framework for interactive information retrieval. In N.J. Belkin, R. Bierig, G. Buscher, L. van Elst, J. Gwizdka, J. Jose, et al. (Eds), CEUR Workshop Proceedings: 512. Proceedings of the SIGIR 2009 Workshop on Understanding the User: Logging and interpreting user interactions in information search and retrieval, UIIR'2009 (pp. 8-11). Aachen, Germany: CEUR Workshop Proceedings.

[4] Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), paper no. 152. Retrieved from http://informationr.net/ir/8-3/paper152.html.

[5] Breiman, L. (1984). *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC.

[6] Chapelle, O., Metzler, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, 2009, 621-630.

[7] Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 147-168.

[8] Kanoulas, E., Hall, M. Clough, P. Carterette, B., & Sanderson, M. (2012) Overview of the TREC Session Track. In: *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*. Gaithersburg, MD: National Institute of Standards and Technology. Retrieved on 20 May 2012 at http://trec.nist.gov/pubs/trec20/t20.proceedings.html.

[9] Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In A. Spink and C. Cole (Eds.) *New Directions in Cognitive Information Retrieval* (pp.169-186). Netherlands: Springer Publishing.

[10] Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04*). ACM, New York, NY, USA, 377-384. DOI= http://doi.acm.org/10.1145/1008992.1009057

[11] Kelly, D. & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2), 18-28.

[12] Li, Y. & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage*. 44, 6 (November 2008), 1822-1837. DOI=10.1016/j.ipm.2008.07.005 http://dx.doi.org/10.1016/j.ipm.2008.07.005

[13] Liu, C., Gwizdka, J., & Liu, J. (2010). Helping identify when users find useful documents: examination of query reformulation intervals. In *Proceeding of the third symposium on Information interaction in context (IIiX '10)*. ACM, New York, NY, USA, 215-224. DOI= http://doi.acm.org/10.1145/1840784.1840816

[14] Liu, J. & Belkin, N.J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '10)*. Geneva, Switzland, July 19-23, 2010.

[15] Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics,* Volume 1, July, 2002 Association for Computational Linguistics.

[16] Teevan, J., Dumais, S.T., and Horvitz, E. (2010). Potential for personalization. *ACM Trans. Comput.-Hum. Interact.* 17, 1, 1-31.

[17] White, R. W. & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 297-306). Arlington, Virginia, USA.

[18] White, R.W., Ruthven, I., and Jose, J.M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*. ACM, New York, NY, USA, 35-42.