Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search

THORSTEN JOACHIMS
Dept. of Computer Science, Cornell University
and
LAURA GRANKA
Google Inc.
and
BING PAN
School of Business and Economics, College of Charleston
and
HELENE HEMBROOKE
Dept. of Information Science, Cornell University
and
FILIP RADLINSKI
Dept. of Computer Science, Cornell University
and
GERI GAY
Dept. of Information Science, Cornell University

This paper examines the reliability of implicit feedback generated from clickthrough data and query reformulations in WWW search. Analyzing the users' decision process using eyetracking and comparing implicit feedback against manual relevance judgments, we conclude that clicks are informative but biased. While this makes the interpretation of clicks as absolute relevance judgments difficult, we show that relative preferences derived from clicks are reasonably accurate on average. We find that such relative preferences are accurate not only between results from an individual query, but across multiple sets of results within chains of query reformulations.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Human Factors, Measurement, Reliability, Experimentation
Additional Key Words and Phrases: Clickthrough Data, Eyetracking, Implicit Feedback, Query Reformulations, User Studies, WWW Search

1. INTRODUCTION

The idea of adapting a retrieval system to particular groups of users and particular collections of documents promises further improvements in retrieval quality for at least two reasons. First, a one-size-fits-all retrieval function is necessarily a compromise in environments with heterogeneous users and is therefore likely to act suboptimally for many users [Teevan et al. 2005]. Second, as evident from the TREC evaluations, differences between document collections make it necessary to tune retrieval functions with respect to the collection for optimum retrieval performance. Since manually adapting a retrieval function is time consuming or even impractical, research on automatic adaptation using machine learning is receiving much attention (e.g. [Fuhr 1989, Bartell et al. 1994, Boyan et al. 1996, Freund et al. 1998, Cohen et al. 1999, Herbrich et al. 2000, Crammer and Singer 2001, Kemp and Ramamohanarao 2002, Joachims 2002, Holland et al. 2003, Almeida and Almeida 2004, Radlinski and Joachims 2005, Burges et al. 2005]). However, a great bottleneck in the application of machine learning techniques is the availability of training data.

In this paper we explore and evaluate strategies for how to automatically generate training examples for learning retrieval functions from observed user behavior. In contrast to explicit feedback, such implicit feedback has the advantage that it can be collected at much lower cost, in much larger quantities, and without burden on the user of the retrieval system. However, implicit feedback is more difficult to interpret and potentially noisy. In this paper we analyze which types of implicit feedback can be reliably extracted from observed user behavior, in particular clickthrough data in WWW search. Following and extending prior work reported in [Radlinski and Joachims 2005, Joachims et al. 2005, Granka et al. 2004], we analyze implicit feedback from within individual queries as well as across multiple consecutive queries about the same information need (i.e. query chains). The feedback strategies across query chains exploit that users typically reformulate their query multiple times before their information need is satisfied. We elaborate on the query chain strategies proposed in [Radlinski and Joachims 2005], as well as propose and explore additional strategies.

To evaluate the reliability of these implicit feedback signals, we conducted a user study. The study is designed to analyze how users interact with the list of ranked results (i.e. the "results page" for short) from the Google search engine and how their behavior can be interpreted as relevance judgments. We perform two types of analysis in this study. First, we use eyetracking to understand how users behave on Google's results page. Do users scan the results from top to bottom? How many abstracts do they read before clicking? How does their behavior change, if we artificially manipulate Google's ranking? Answers to these questions give insight into the users' decision process and suggest in how far clicks are the result of an informed decision. Based on these results, we propose several strategies for generating feedback from clicks and query reformulations. To evaluate the degree to which feedback signals indicate relevance, we compare the implicit feedback against explicit feedback we collected manually.

The study presented in this paper is different in at least two respects from previous work assessing the reliability of implicit feedback [Morita and Shinoda

1994, Claypool et al. 2001, White et al. 2002, Kelly and Belkin 2004, Fox et al. 2005]. First, our study provides detailed insight into the users' decision-making process through the use of eyetracking. Second, we evaluate relative preference signals derived from user behavior. This is in contrast to previous studies that primarily evaluated absolute feedback.

Our results show that users make informed decisions among the abstracts they observe and that clicks reflect relevance judgments. However, we show that clicking decisions are biased in at least two ways. First, we show that there is a "trust bias" which leads to more clicks on links ranked highly by Google, even if those abstracts are less relevant than other abstracts the user viewed. Second, there is a "quality-of-context bias": the users' clicking decision is not only influenced by the relevance of the clicked link, but also by the overall quality of the other abstracts in the ranking. This shows that clicks have to be interpreted relative to the order of presentation and relative to the other abstracts. We propose several strategies for extracting such relative relevance judgments from clicks and show that they accurately agree with explicit relevance judgments collected manually.

2. RELATED WORK

The idea of using machine learning to automatically tune retrieval functions has a long history in the retrieval and learning communities. However, most methods assume that explicit relevance judgments are available (e.g. [Fuhr 1989, Bartell et al. 1994]). While Cohen et al. [Cohen et al. 1999] discuss the use of clickthrough data, they derive the data for their experiments from explicit judgments.

Some attempts have been made to use implicit feedback. Browsing assistants observed clicking behavior to highlight and prefetch links during Web browsing [Lieberman 1995, Joachims et al. 1997]. An algorithm that adapts the retrieval function to minimize the rank of the clicked links was proposed in [Boyan et al. 1996]. Extending the ordinal regression method of [Herbrich et al. 2000] to the task of ranking, Joachims proposed a Support Vector Algorithm and showed that it can be trained with pairwise preferences extracted from clicks [Joachims 2002]. A similar approach is followed in [Holland et al. 2003]. Extending the preference elicitation strategies presented in [Joachims 2002, Joachims et al. 2005], Agichtein et al. [2006] show that preferences extracted from aggregate clickthrough statistics can be more accurate than preferences from individual clicks, and they demonstrate how clicks in addition to other implicit feedback signals (e.g. dwell time) can provide significant improvements of retrieval accuracy in real-world search engines [Agichtein et al. 2006]. Kemp and Ramamohanarao [Kemp and Ramamohanarao 2002] used clickthrough data for document expansion by adding the query words to the clicked documents. Using implicit feedback on a retrieval-session level, several studies have explored substituting implicit measures for explicit relevance feedback (see e.g. [White et al. 2002, 2005, Shen et al. 2005]). With a similar goal, session logs from an online bookstore are used in [Almeida and Almeida 2004] to identify communities and personalize search. Among commercial search engines, "Direct Hit" (now part of Teoma) was the first to make use of clickthrough data. The precise mechanism, however, is unpublished.

Beyond clicks and document accesses, as primarily considered in the approaches above, we also explore whether query reformulations give insight into the users information need and the relevance of result sets. Query reformulation chains are used in [Furnas 1985, Radlinski and Joachims 2005] for associating results eventually found at the end of the chain with queries tried earlier in the chain. Alternatively, query reformulations can be used to learn and predict common reformulation patterns (e.g. [Jones and Fain 2003]). Methods for identifying query chains can be found in [Silverstein et al. 1999, Jansen et al. 2000, Radlinski and Joachims 2005].

How reliable are the implicit feedback signals used by these algorithms? Only few studies have addressed this question so far, which motivated the work presented in this paper. The studies in [Morita and Shinoda 1994, Oard and Kim 1998] find that reading time is indicative of interest when reading news stories. Similarly, Claypool et al. [Claypool et al. 2001] find that reading time as well as the amount of scrolling can predict relevance in WWW browsing. However, for the task of retrieval we consider in this paper, Kelly and Belkin [Kelly and Belkin 2004] report that reading time is not indicative of document relevance. They show that reading time varies between subjects and tasks, which makes it difficult to interpret. Nevertheless, Fox et al. [Fox et al. 2005] show in their study that implicit measures like session duration and number of result sets returned, are indicative of user satisfaction with an entire search session. They also show that a combination of several implicit measures including reading time and the way the user exited from the result page, can predict the relevance of an individual result well. See [Kelly and Teevan 2003] for a survey of the state-of-the-art in extracting and using implicit feedback. In the following, we revisit some of the implicit feedback indicators mentioned above, but primarily explore and evaluate implicit feedback in the form of relative preferences, which was not considered by any of the previous studies.

To the best of our knowledge, very few studies have used eye-tracking in the context of online information retrieval, and none have addressed the issues detailed in this present paper. Many of the studies using eye-tracking to study Web-based "information search", use the term loosely, and are actually referencing users' patterns of navigation across general web page content – not the display of search engine results [Goldberg et al. 2002, Pan et al. 2004, Halverson and Hornof 2004]. Furthermore, the questions addressed in these studies are of a much more general nature, depicting general patterns of eye movement and navigation across the page [Goldberg et al. 2002, Pan et al. 2004], and assessing how link color may influence visual search patterns [Halverson and Hornof 2004].

More similar to the research presented here, Salogarvi et al. [Salogarvi et al. 2003] used measures of pupil dilation to infer the relevance of online abstracts, and found that pupil dilation increased when fixated on relevant abstracts. However, this study only collected eye movements from three subjects, so the generalizability is unclear, and furthermore, no other measures of searcher performance were addressed. Larger-scale studies were performed by Brumby and Howes [Brumby and Howes 2003, 2004] and by Klöckner et al. [Klöckner et al. 2004]. Both studies analyze how users scan a list of results and how they trade-off exploring the result set with exploring outgoing links. Unlike our work in this paper, however, the motivation and focus of these works was on the use of eye tracking towards understanding

cognitive processes and towards improved interface design, not towards exploiting user behavior as implicit feedback for machine learning.

3. USER STUDY

To gain an understanding of how users interact with the list of ranked results and how their clicking behavior relates to relevance judgments, we conducted two consecutive user studies. Unlike in the majority of the existing user studies, we designed these studies to not only record and evaluate user actions, but also to give insight into the decision process that lead the user to the actions. This is achieved through recording the users' eye movements. Eye tracking provides an account of the users' subconscious behavior and cognitive processing, which is important for interpreting user actions [Rayner 1998].

3.1 Task, Participants, and Conditions

We designed the study to resemble typical use of a WWW search engine. All participants were asked to answer the same ten questions using Google as a starting point for their search. Half of the searches were navigational [Broder 2002], asking subjects to find a specific Web page or homepage. The other five tasks were informational [Broder 2002], asking subjects to find a specific bit of information. The questions vary in difficulty and topic. The complete list of questions is given in Table I.

We conducted the user study in two phases. In Phase I, we recruited 34 participants, all of which were Cornell undergraduate students, mostly from majors in the social sciences and engineering. Students were offered extra class credit for participating in the study and were recruited through announcements in classes. Due to recording difficulties and the inability of some subjects to be precisely calibrated, comprehensive eye movement data was recorded for 29 of the subjects. All subjects were between 18 and 23 years old, with a mean age of 20.3. The gender distribution was split between 19 males and 15 females. In a pre-study questionnaire, all subjects indicated at least a general familiarity with the Google interface and 31 of the subjects reported that Google is their primary search engine. Furthermore, most subjects considered themselves savvy users of internet search engines, using search engines at least several times per week. Overall, these participants appear to be more savvy than average users.

Phase II of the study was designed to investigate how users react to manipulations of the search results. Using the same ten questions, the same recruiting mechanisms, and the same instructions to the subjects as in Phase I, each subject was randomly assigned to one of three experimental conditions.

normal: Subjects in the "normal" condition received Google's original ranking just like in Phase I.

swapped: Subjects assigned to the "swapped" condition received a ranking where the top two results returned by Google were switched in order.

reversed: Subjects in the "reversed" condition received the (typically 10) results from Google in reversed order.

The manipulations to the results page were performed by a proxy that intercepted the HTTP request to Google. None of the changes were detectable by the subjects

Table I. Questions used in the study and the average number of queries and clicks per question and subject.

	No.	Question	Phase I #Queries	Phase I #Clicks	Phase II #Queries	Phase II #Clicks
navigational	1.	Find the homepage of Michael Jordan, the statistician.	2.8	1.6	2.6	1.7
	2.	Find the page displaying the route map for Greyhound buses.	1.3	1.5	1.6	1.6
	3.	Find the homepage of the 1000 Acres Dude Ranch.	2.2	2.6	2.2	1.9
	4.	Find the homepage for graduate housing at Carnegie Mellon University.	2.0	1.7	2.2	0.9
	5.	Find the homepage of Emeril - the chef who has a television cooking program.	1.9	1.6	3.0	1.8
informational	6.	Where is the tallest mountain in New York located?	1.7	2.0	2.0	1.6
	7.	With the heavy coverage of the democratic presidential primaries, you are excited to cast your vote for a candidate. When are democratic presidential primaries in New York?	1.6	1.8	1.6	2.1
	8.	Which actor starred as the main character in the original Time Machine movie?	1.8	1.8	1.9	1.9
	9.	A friend told you that Mr. Cornell used to live close to campus - near University and Steward Ave. Does anybody live in his house now? If so, who?	2.0	1.5	2.9	1.6
	10.	What is the name of the researcher who discovered the first modern antibiotic?	2.0	2.0	2.3	1.6

and they did not know that we manipulated the results. When asked after their session, none of the subjects had suspected any manipulation.

Twenty-two participants were recruited for Phase II of the study using the same recruiting strategies as in Phase I. Again, the participants were offered extra class credit for participating. Their mean age was 20.4 years and their pre-study questionnaires generally showed similar characteristics as those of the subjects in Phase I. We were able to record usable eye tracking data for 16 of the subjects, 11 males and 5 females.

3.2 Data Capture

The subjects' eye movements were recorded using an ASL 504 commercial eye-tracker (Applied Science Technologies, Bedford, MA) which utilizes a CCD camera that employs the Pupil Center and Corneal-Reflection method to reconstruct a subject's eye position. GazeTracker, a software application accompanying the system, was used for the acquisition and analysis of the subject's eye movements [Lankford 2000].

An HTTP-proxy server was established to log all clickstream data and store all Web content that was accessed and viewed. In particular, the proxy cached all pages the user visited, as well as all pages that were linked to in any results page

returned by Google. The proxy did not introduce any noticable delay. In addition to logging all activity, the proxy manipulated the Google results page according to the three conditions, while maintaining the appearance of an authentic Google page. The proxy also automatically eliminated all advertising content, so that the results pages of all subjects would look as uniform as possible, with approximately the same number of results appearing within the first scroll set. With these pre-experimental controls, subjects were able to participate in a live search session, generating unique search queries and results from the questions and instructions presented to them.

Table I gives an overview of the data that was collected for both phases. Overall, the subjects of Phase I issued on average 1.9 queries per question. 70% of all queries issued were unique (i.e. were issued only once throughout Phase I). The subjects clicked on an average of 0.9 results per query. In Phase II, the subjects issued on average 2.2 queries per question, and 81% of all queries were unique. Subjects made on average 0.8 clicks per query in Phase II. While some questions produced more clicks and queries than others as shown in Table I, the data is fairly blanced. The contributions from different subjects are also reasonably balanced. The standard deviation of the number of queries per subject is 3.8 in Phase I and 4.5 in Phase II. Regarding the number of clicks per subject, the standard deviations are 3.6 in Phase I and 4.9 in Phase II.

3.3 Eyetracking

We classify eye movements according to the following significant indicators of ocular behaviors, namely fixations, saccades, pupil dilation, and scan paths [Rayner 1998]. Eye fixations are the most relevant metric for evaluating information processing in online search. Fixations are defined as a spatially stable gaze lasting for approximately 200-300 milliseconds, during which visual attention is directed to a specific area of the visual display. Fixations represent the instances in which most information acquisition and processing occurs [Just and Carpenter 1980, Rayner 1998].

Other indices, such as saccades, are believed to occur too quickly to absorb new information [Rayner 1998]. Saccades, for example, are the continuous and rapid movements of eye gazes between fixation points. Because saccadic eye movements are extremely rapid, within 40-50 milliseconds, it is widely believed that only little information can be acquired during this time.

Pupil dilation is a measure that is typically used to indicate an individual's arousal or interest in the viewed content matter, with a larger diameter reflecting greater arousal [Rayner 1998]. While pupil dilation could be interesting in our analysis, we focus on fixations in this paper.

3.4 Explicit Relevance Judgments

To have a basis for evaluating the quality of implicit relevance judgments, we collected explicit relevance judgments for all queries and results pages encountered by the subjects. With "results page" we refer to the set of typically 10 results returned by Google. For the data from Phase I, the explicit relevance judgments were based on the abstract (i.e. title, query-dependent snippet, URL, and meta-data) presented by Google. For Phase II, we collected two sets of judgments, one based on the abstract as in Phase I, and one based on the actual WWW page.

In particular, for each results page from Phase I, we asked judges to order the results by how promising their abstracts are for leading to information that is relevant to answering the question. We chose this ordinal assessment method, since it was demonstrated that humans can make such relative decisions more reliably than absolute judgments for many tasks (see e.g. [Belew 2000, page 109]). The judges were allowed to give the same rank to different abstracts (therefore producing a weak ordering), if they found both to be equally promising. Five judges (different from the subjects) each assessed all results pages (i.e. all sets of 10 results ever displayed to a subject) for two of the questions, plus ten results pages from two other questions for inter-judge agreement verification. The judgments were collected using a WWW interface that the judges could access remotely. For each question and query, the set of (typically 10) abstracts was displayed to the judge on one WWW page in randomized order. Next to each abstract was a text box into which the judge had to enter a numerical value indicating relevance on an ordinal scale. The system required that all text boxes had to be filled. Judges were instructed that only the ordering induced by the value was important, but not their absolute value. Giving the same value to multiple results meant indistinguishable relevance. All queries and all results pages encountered during Phase I were judged in this fashion. The judges received detailed instructions and examples of how to judge relevance. However, we intentionally did not use specially trained relevance assessors, since the explicit judgments will serve as an estimate of the data quality we could expect when asking regular users for explicit feedback. The agreement between judges is reasonably high. Whenever two judges expressed a strict preference between two abstracts, they agree in the direction of preference in 89.5% of the cases.

For the result pages from Phase II we collected explicit relevance assessments for abstracts in a similar manner. However, the set of abstracts we asked judges to weakly order was not limited to the (typically 10) hits from a single results page, but the set included the results from all queries for a particular question and subject. Again, all queries of all subjects were judged. The inter-judge agreement on the abstracts is 82.5%. We conjecture that this lower agreement is due to the less concise judgment setup and the larger sets that had to be ordered.

To address the question of how implicit feedback relates to an explicit relevance assessment of the actual *Web page*, we collected relevance judgments for the pages from Phase II following the setup already described for the abstracts. However, instead of showing the abstracts, the judges were given only a hyperlink. The hyperlink pointed to a locally cached copy of the page that was recorded by the proxy when the subject originally issued the query. Using the browser back-button, the judges would return to fill-in the text box. Again, all queries of all subjects were judged. The inter-judge agreement on the relevance assessment of the pages is 86.4%.

4. ANALYSIS OF USER BEHAVIOR

In our study we focus on the list of ranked results returned by Google in response to a query. Note that clickthrough data on this results page can easily be recorded by the retrieval system, which makes implicit feedback based on this page particularly attractive. In most cases, the results page contains links to 10 pages. Each link is

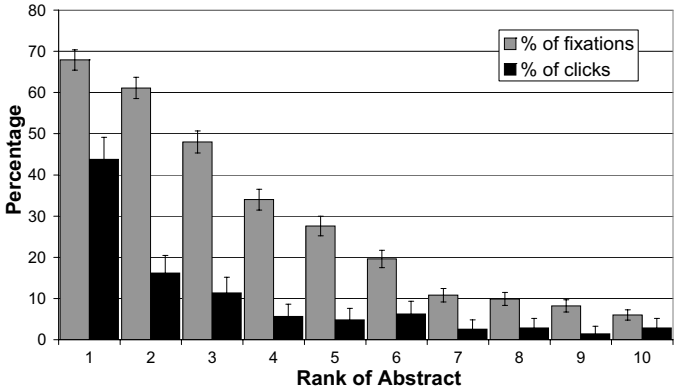

Fig. 1. Percentage of times an abstract was viewed/clicked depending on the rank of the result.

described by an abstract that consists of the title of the page, a query-dependent snippet extracted from the page, the URL of the page, and varying amounts of meta-data.

Before we start analyzing particular strategies for generating implicit feedback from clicks on the Google results page, we first analyze how users scan the results page. Knowing which abstracts the user evaluates is important, since clicks can only be interpreted with respect to the parts of the results that the user actually observed and evaluated. The following results are based on the data from Phase I.

4.1 Which links do users view and click?

One of the valuable aspects of eye-tracking is that we can determine how the displayed results are actually viewed. The light bars in Figure 1 show the percentage of results pages where the user viewed the abstract at the indicated rank. We consider an abstract as "viewed" by the user, if there is at least one fixation within a heuristically defined look-zone covering the abstract. The abstracts ranked 1 and 2 receive most attention. After that, attention drops faster. The dark bars in Figure 1 show the percentage of times a user's first click falls on a particular rank. It is very interesting that users click substantially more often on the first than on the second link, while they view the corresponding abstract with almost equal frequency.

There is an interesting change around rank 6/7, both in the viewing behavior as well as in the number of clicks. First, links listed below this rank receive substantially less attention than those presented higher. Second, unlike for ranks 2 to 5, the abstracts ranked 6 to 10 receive more equal attention. This can be explained by the fact that typically only the first 5-6 links were visible without scrolling. Once the user has started scrolling, rank appears to becomes less of an influence for attention. A sharp drop occurs after link 10, as ten results are displayed per page.

4.2 Do users scan links from top to bottom?

While the linear ordering of the results suggest reading from top to bottom, it is not clear whether users actually behave this way. Figure 2 depicts the instance of first arrival to each abstract in the ranking. The arrival time is measured by fixations;

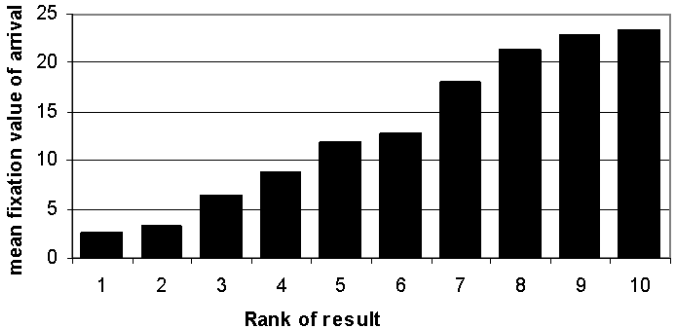

Fig. 2. Mean time of arrival (in number of previous fixations) depending on the rank of the result.

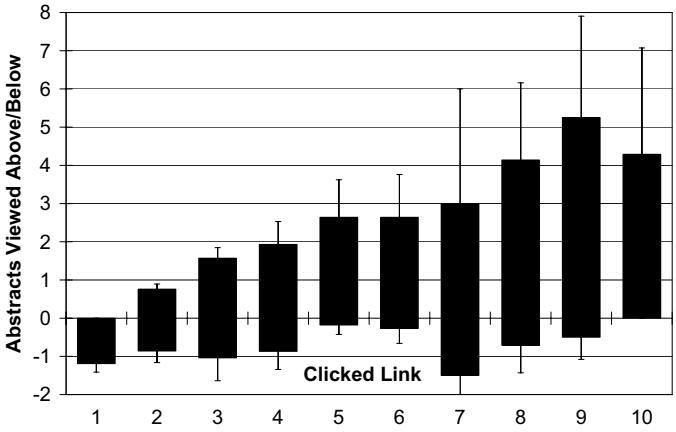

Fig. 3. Mean number of abstracts viewed above and below a clicked link depending on its rank.

i.e., at what fixation did a searcher first view the nth-ranked abstract. The graph indicates that on average users tend to read the results from top to bottom. In addition, the graph shows interesting patterns. First, individuals tend to view the first and second-ranked results right away, within the second or third fixation, and there is a big gap before viewing the third-ranked abstract. Second, the page break also manifests itself in this graph, as the instance of arrival to results seven through ten is much higher than the other six. It appears that users first scan the viewable results quite thoroughly before resorting to scrolling.

4.3 Which links do users evaluate before clicking?

Figure 3 depicts how many abstracts above and below the clicked document users view on average. With "above" and "below" we refer to the layout on the results page. The graph shows that the lower the click in the ranking, the more abstracts are viewed above the click. While users do not neccessarily view all abstracts above a click, they view substantially more abstracts above than below the click. This is consistent with conclusions from the eye-tracking study in [Klöckner et al. 2004], namely that most users follow a depth-first search strategy. In particular, they find that most users tend to scan the list from top to bottom, and click on sufficiently

Table II. Percentage of times the user viewed an abstract at a particular rank before he clicked on a link at a particular rank.

Viewed	Clicked Rank					
Rank	1	2	3	4	5	6
1	90.6%	76.2%	73.9%	60.0%	54.5%	45.5%
2	56.8%	90.5%	82.6%	53.3%	63.6%	54.5%
3	30.2%	47.6%	95.7%	80.0%	81.8%	45.5%
4	17.3%	19.0%	47.8%	93.3%	63.6%	45.5%
5	8.6%	14.3%	21.7%	53.3%	100.0%	72.7%
6	4.3%	4.8%	8.7%	33.3%	18.2%	81.8%

promising abstracts without first exploring many links below that result.

Table II augments the information in Figure 3 by showing which particular abstracts users view (rows) before making a click at a particular rank (columns). For example, the elements in the first two rows of the third data column show that before a click on link three, the user has viewed abstract two 82.6% of the times and abstract one 73.9% of the times. In general, it appears that abstracts closer above the clicked link are more likely to be viewed than abstracts further above. Another pattern is that the abstract right below a click is viewed roughly 50% of the times (except at the page break). Finally, note that the lower-than-100% values on the diagonal indicate some accuracy limitations of the eye-tracker, as well as potentially some higher degree of peripheral vision than incorporated into our definition of look-zones.

5. ANALYSIS OF IMPLICIT FEEDBACK

The previous section explored how users scan the results page and how their scanning behavior relates to the decision of clicking on a link. We will now explore how relevance of the document to the query influences clicking decisions, and vice versa, what clicks tell us about the relevance of a document. After determining that user behavior depends on relevance in the next section, we will explore how closely implicit feedback signals from observed user behavior agree with the explicit relevance judgments.

5.1 Does relevance influence user decisions?

Before exploring particular strategies for generating relevance judgments from observed user behavior, we first verify that users react to the relevance of the presented links. We use the "reversed" condition as an intervention that controllably decreases the quality of the retrieval function and the relevance of the highly ranked abstracts. Figure 4 includes the same type of graph as Figure 1 for the "normal" and the "reversed" condition for the data from Phase II. The graphs show that the users react to the degraded ranking in two ways. First, they view lower ranked links more frequently. In particular, in the "reversed" condition the average position of the viewed links within a results page is significantly further down in the ranking than in the "normal" condition (Wilcoxon, $p = 0.03$). All significance tests reported in this paper are two-tailed tests at a 95% confidence level. Second, subjects are less likely to click on the first link, but more likely to click on a lower ranked link. More generally, the average rank of a clicked document in the "normal" condition is 2.66 and 4.03 in the "reversed" condition. The difference is significant according

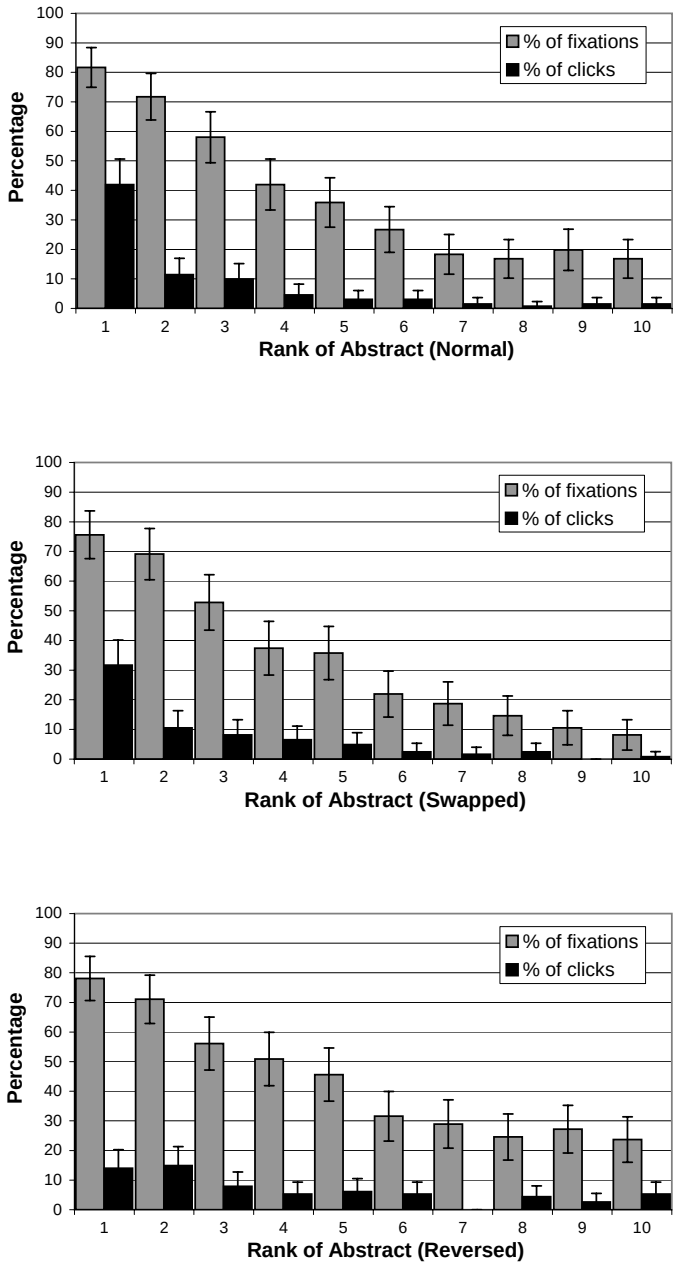

Fig. 4. Percentage of abstracts viewed and clicked depending on the rank of the result for the "normal", the "swapped", and the "reversed" condition in Phase II. Due to differences in data cleaning and the definition of look-zones, the fixation percentages are slightly higher than in Figure 1.

Table III. Number of clicks on the top two presented links depending on relevance of the abstracts for the normal and the swapped condition for Phase II. In the column headings, +/- indicates whether the user clicked (+) or did not click (-) on link l_1 or l_2 in the ranking. rel() indicates manually judged relevance of the abstract.

"normal"	l_1^-, l_2^-	l_1^+, l_2^-	l_1^-, l_2^+	l_1^+, l_2^+	total
$rel(l_1) > rel(l_2)$	15	19	1	1	36
$rel(l_1) < rel(l_2)$	11	5	2	2	20
$rel(l_1) = rel(l_2)$	19	9	1	0	29
total	45	33	4	3	85
"swapped"	l_1^-, l_2^-	l_1^+, l_2^-	l_1^-, l_2^+	l_1^+, l_2^+	total
$rel(l_1) > rel(l_2)$	11	15	1	1	28
$rel(l_1) < rel(l_2)$	17	10	7	2	36
$rel(l_1) = rel(l_2)$	36	11	3	0	50
total	64	36	11	3	114

to the Wilcoxon test with $p = 0.01$.

This shows that user behavior does depend on the quality of the presented ranking and that individual clicking decisions are influenced by the relevance of the abstracts. It is therefore possible that, vice versa, observed user behavior can be used to assess the overall quality of a ranking, as well as the relevance of individual documents. In the following, we will explore the reliability of several strategies for extracting implicit feedback from observed user behavior.

5.2 Are clicks absolute relevance judgments?

One frequently used interpretation of clickthrough data as implicit feedback is that each click represents an endorsement of that page (e.g. [Boyan et al. 1996, Kemp and Ramamohanarao 2002, Fox et al. 2005]). In this interpretation, a click indicates a relevance assessment on an absolute scale: clicked documents are relevant. In the following we will show that such an interpretation is problematic for two reasons.

5.2.1 *Trust Bias.* Figure 1 shows that the abstract ranked first receives many more clicks than the second abstract, despite the fact that both abstracts are viewed much more equally. This could be due to two reasons. The first explanation is that Google typically returns rankings where the first link is more relevant than the second link, and users merely click on the abstract that is more promising. In this explanation users are not influenced by the order of presentation, but decide based on their relevance assessment of the abstract. The second explanation is that users prefer the first link due to some level of trust in the search engine. In this explanation users are influenced by the order of presentation. If this was the case, the interpretation of a click would need to be relative to the strength of this influence.

We address the question of whether the users' evaluation depends on the order of presentation using the data from Table III. The experiment focuses on the top two links, since these two links are scanned relatively equally. Table III shows how often a user clicks on either link 1 or link 2, on both links, or on none of the two depending on the manually judged relevance of the abstract. If users were not influenced in their relevance assessment by the order of presentation, the number of clicks on link 1 and link 2 should only depend on the judged relevance of the

abstract. This hypothesis entails that the fraction of clicks on the more relevant abstract should be the same independent of whether link 1 or link 2 is more relevant. The table shows that we can reject this hypothesis with high probablility, since 19 vs. 1 is significantly different from 2 vs. 5 (Fisher Exact Test, $p = 0.003$). To make sure that the difference is not due to a dependence between rank and magnitude of difference in relevance, we also analyze the data from the swapped condition. Table III shows that also under the swapped condition, there is still a strong bias to click on link one even if the second abstract is more relevant. With 15 vs. 1 compared to 7 vs. 10, subjects fail to click on the more relevant link significantly more frequently when it is presented in the second position (Fisher Exact Test, $p = 0.003$). Note that the number of subjects in the "normal" and the "swapped" condition is different, so that comparing total counts between the two conditions is not meaningful.

We conclude that the position of a result has substantial influence on the users' decision to click. We conjecture that users trust in the search engine's ability to estimate the relevance of a page, which influences their clicking behavior. Related ordering effects were also found in other decision settings (see e.g. [Mantell and Kardes 1999]).

5.2.2 *Quality-of-Context Bias.* We now study whether the clicking behavior depends on the overall quality of the retrieval system, or only on the relevance of the clicked link. If there is a dependency on overall retrieval quality, any interpretation of clicks as implicit relevance feedback would need to be relative to the quality of the retrieval system.

To address this question, we control the quality of the retrieval function using the "reversed" condition and compare the clicking behavior against the "normal" condition. In particular, we investigate whether the links users click on in the "reversed" condition are less relevant on average. We measure the relevance of an abstract in terms of its rank as assigned by the human relevance judges (i.e. the abstracts judged most relevant have rank 1, the next most relevant links have rank 2, etc.). We call this number the human relevance rank of an abstract. The average human relevance rank of clicks in the "normal" condition is 2.10 compared to 2.45 in the "reversed" condition. The difference is significant according to the two-tailed Wilcoxon test with $p = 0.02$.

We conclude that the quality of the ranking influences the user's clicking behavior. If the relevance of the retrieved results decreases, users click on abstracts that are on average less relevant[1].

5.3 Are clicks relative relevance judgments within one results page?

Interpreting clicks as relevance judgments on an absolute scale is difficult due to the two effects described above. An accurate interpretation would need to take into account the user's trust into the quality of the search engine, as well as the quality of the retrieval function itself. Unfortunately, trust and retrieval quality are two quantities that are difficult to measure explicitly.

[1]The study of Brumby and Howes [Brumby and Howes 2003, 2004] shows that a similar dependence on the overall quality of the results set also exists for the users exploration and satisficing behavior.

We will now explore implicit feedback measures that respect these dependencies by interpreting clicks not as absolute relevance feedback, but as preference statements among the available options. This interpretation is motivated by the theory of revealed preferences [Samuelson 1948, Varian 1992] from economics. It states that consumer behavior (i.e. purchasing decision from a set of options) can be used to reveal the (typically unobservable) utility function that governs the consumer's decision process. Analogously, in the retrieval setting we would like to identify the utility (i.e. relevance) of a result from the user's choice among the available options (e.g. click on result). Motivation for interpreting user actions as a choice among a limited set of options (and not as an absolute statement of utility or relevance) comes from models of bounded rationality that have been found to explain information-navigation behavior (see e.g. [Brumby and Howes 2003, 2004, Klöckner et al. 2004]). In particular, users were found to often act before they had evaluated all options, therefore trading-off quality against exploration effort.

Following these motivations, the strategies we explore are based on the idea that not only clicks should be used as feedback signals, but also the fact that some links were *not* clicked on [Joachims 2002, Cohen et al. 1999]. Consider the example ranking of links l_1 to l_7 below and assume that the user clicked on links l_1, l_3, and l_5.

$$l_1^* \;\; l_2 \;\; l_3^* \;\; l_4 \;\; l_5^* \;\; l_6 \;\; l_7 \tag{1}$$

While it is difficult to infer whether the links l_1, l_3, and l_5 are relevant on an *absolute* scale, it seems much more plausible to infer that link l_3 is more relevant than link l_2. As we have already established in Sections 4.2 and 4.3, users scan the list from top to bottom in a reasonably exhaustive fashion. Therefore, it is reasonable to assume that the user has observed link l_2 before clicking on l_3, making a decision to *not* click on it. This gives an indication of the user's preferences between link l_3 and link l_2. Similarly, it is possible to infer that link l_5 is more relevant than links l_2 and l_4. This means that clickthrough data does not convey *absolute* relevance judgments, but partial *relative* relevance judgments for the links the user evaluated. A search engine ranking the returned links according to their relevance should have ranked link l_3 ahead of l_2, and link l_5 ahead of l_2 and l_4. Denoting the user's relevance assessment with rel(), we get partial (and potentially noisy) information of the form

$$\text{rel}(l_3) > \text{rel}(l_2),\;\; \text{rel}(l_5) > \text{rel}(l_2),\;\; \text{rel}(l_5) > \text{rel}(l_4)$$

This strategy for extracting preference feedback is summarized as follows.

Strategy 1. (Click > Skip Above)
For a ranking $(l_1, l_2, l_3, ...)$ and a set C containing the ranks of the clicked-on links, extract a preference example $rel(l_i) > rel(l_j)$ for all pairs $1 \leq j < i$, with $i \in C$ and $j \notin C$.

Note that this strategy takes trust bias and quality-of-context bias into account. First, it only generates a preference when the user explicitly decides to not trust the search engine and skip over a higher ranked link. Second, since it generates pairwise preferences only between the documents that the user evaluated, all feedback is relative to the quality of the retrieved set.

Table IV. Accuracy of several strategies for generating pairwise preferences from clicks within one result set. The base of comparison are either the explicit judgments of the abstracts, or the explicit judgments of the page itself. Behind each ± we show the larger of the two sides of the 95% binomial confidence interval around the sample mean. The column "p/q" shows the average number of preferences generated per query by this strategy.

Explicit Feedback		Abstracts					Pages
Data		Phase I	Phase II				Phase II
Strategy	p/q	"normal"	"normal"	"swapped"	"reversed"	all	all
Inter-Judge Agreem.	N/A	89.5	N/A	N/A	N/A	82.5	86.4
Click > Skip Above	1.37	80.8±3.6	88.0±9.5	79.6±8.9	83.0±6.7	83.1±4.4	78.2±5.6
LastClick > SkipAbove	1.18	83.1±3.8	89.7±9.8	77.9±9.9	84.6±6.9	83.8±4.6	80.9±5.1
Click > Earlier Click	0.20	67.2±12.3	75.0±25.8	36.8±22.9	28.6±27.5	46.9±13.9	64.3±15.4
Click > Skip Previous	0.37	82.3±7.3	88.9±24.1	80.0±18.0	79.5±15.4	81.6±9.5	80.7±9.6
Click > No Click Next	0.68	84.1±4.9	75.6±14.5	66.7±13.1	70.0±15.7	70.4±8.0	67.4±8.2

How accurate is this implicit feedback compared to the explicit feedback? To address this question, we compare the pairwise preferences generated from the clicks to the explicit relevance judgments. Table IV shows the percentage of times the preferences generated from clicks agree with the direction of a strict preference of a relevance judge. On the data from Phase I, the preferences are 80.8% correct, which is substantially and significantly better than the random baseline of 50% ($p < 10^{-6}$, testing against a Binomial distribution with mean 0.5)[2]. Furthermore, it is fairly close in accuracy to the agreement of 89.5% between the explicit judgments from different judges, which can serve as an upper bound for the accuracy we could ideally expect even from explicit user feedback.

The data from Phase II shows that the accuracy of the "Click > Skip Above" strategy does not change significantly with respect to degradations in ranking quality in the "swapped" (Binomial Proportion Test, $p = 0.14$) and "reversed" condition (Binomial Proportion Test, $p = 0.32$). As expected, trust bias and quality-of-context bias have no significant effect.

We next explore a variant of "Click > Skip Above", which follows the intuition that earlier clicks might be less informed than later clicks (i. e. after a click, the user returns to the search page and selects another link). This lead us to the following strategy, which considers only the last click for generating preferences.

Strategy 2. (Last Click > Skip Above)
For a ranking $(l_1, l_2, l_3, ...)$ *and a set* C *containing the ranks of the clicked-on links, let* $i \in C$ *be the rank of the link that was clicked temporally last. Extract a preference example* $rel(l_i) > rel(l_j)$ *for all pairs* $1 \leq j < i$, *with* $j \notin C$.

Assuming that l_5 was the last click in the example from above, this strategy would produce the preferences

$$\mathrm{rel}(l_5) > \mathrm{rel}(l_2), \quad \mathrm{rel}(l_5) > \mathrm{rel}(l_4).$$

Table IV shows that this strategy is slightly more accurate than "Click > Skip Above". To analyze this difference, we analyze the accuracy of the preferences that

[2]Note that assuming a Binomial distribution has to be taken with a grain of salt. It assumes independence between preferences statements, which is not necessarily the case for statements derived from the same query.

are produced by "Click > Skip Above" but not by "Last Click > Skip Above" (i.e. preference not resulting from the last click on a results page). The accuracy of these preferences is only 67.1% in Phase I and 77.5% over all pages in Phase II. For the data from Phase I, the difference in accuracy between preferences resulting from last clicks to those not from last clicks is significant (Binomial Proportion Test, $p = 0.001$).

The next strategy we investigate also follows the idea that later clicks are more informed decisions than earlier clicks. But, stronger than the "Last Click > Skip Above", we now assume that clicks later in time are on more relevant abstracts than earlier clicks.

Strategy 3. (Click > Earlier Click)
For a ranking $(l_1, l_2, l_3, ...)$ and a set C containing the ranks of the clicked-on links, let $t(i)$ with $i \in C$ be the time when the link was clicked. We extract a preference example $rel(l_i) > rel(l_j)$ for all pairs j and i, with $i, j \in C$ and $t(i) > t(j)$.

Assuming that the order of clicks is 3, 1, 5 in the example ranking from above, this strategy would generate the preferences

$$\mathrm{rel}(\mathrm{l}_1) > \mathrm{rel}(\mathrm{l}_3),\ \mathrm{rel}(\mathrm{l}_5) > \mathrm{rel}(\mathrm{l}_3),\ \mathrm{rel}(\mathrm{l}_5) > \mathrm{rel}(\mathrm{l}_1).$$

The validity of this strategy is not supported by the data. The accuracy is substantially worse than for the "Click > Skip Above" strategy. It also appears that the ranking quality has an influence on the accuracy of the strategy, since there is a significant (Binomial Proportion, $p = 0.01$) difference between "normal" and "reversed" condition in Phase II. We conjecture that the increased amount of scanning (see Section 5.1) before making a selection in the "reversed" condition leads to a very well informed choice already for the early clicks.

As found in the behavioral data from Section 4.3, the abstracts that are most reliably evaluated are those immediately above the clicked link. This lead us to the following strategy, which generates constraints only between a clicked link and a not-clicked link immediately above.

Strategy 4. (Click > Skip Previous)
For a ranking $(l_1, l_2, l_3, ...)$ and a set C containing the ranks of the clicked-on links, extract a preference example $rel(l_i) > rel(l_{i-1})$ for all pairs $i \geq 2$, with $i \in C$ and $i - 1 \notin C$.

For the example, this strategy generates the preferences $\mathrm{rel}(\mathrm{l}_5) > \mathrm{rel}(\mathrm{l}_4)$ and $\mathrm{rel}(\mathrm{l}_3) > \mathrm{rel}(\mathrm{l}_2)$. The accuracy is given in Table IV. This strategy does not show a consistent gain in accuracy compared to "Click > Skip Above". There is no significant difference between the accuracy of the preferences generated by "Click > Skip Previous" and the 80.2% accuracy of the additional preferences generated by "Click > Skip Above" (Binomial Sign Test, $p = 0.59$ for Phase I).

Finally, we explore another strategy that is motivated by the findings in Section 4.3. While Section 4.3 showed that users do not scan much below a click, the data suggests that they view the immediately following abstract in many cases. This leads us to the following strategy, where we generate a preference constraint between a clicked link and an immediately following link that was not clicked.

Strategy 5. (Click > No-Click Next)
For a ranking $(l_1, l_2, l_3, ...)$ *and a set* C *containing the ranks of the clicked-on links, extract a preference example* $rel(l_i) > rel(l_{i+1})$ *for all* $i \in C$ *and* $(i+1) \notin C$.

For the example, this strategy generates the preferences

$$\mathrm{rel}(\mathrm{l}_1) > \mathrm{rel}(\mathrm{l}_2),\ \mathrm{rel}(\mathrm{l}_3) > \mathrm{rel}(\mathrm{l}_4),\ \mathrm{rel}(\mathrm{l}_5) > \mathrm{rel}(\mathrm{l}_5).$$

Table IV shows that this strategy appears very accurate in the "normal" condition. However, this number is somewhat misleading. Unlike e.g. "Click > Skip Above", the "Click > No-Click Next" strategy generates preferences aligned with the estimated relevance ordering of Google. First, since aligned preferences only confirm the current ranking, they are probably less valuable for learning. Second, generating preferences that follow Google's ordering leads to better than random accuracy even if the user behaved randomly. For example, if the user just blindly clicked on the first link for every query, the accuracy of "Click > No-Click Next" would be 62.4%. More convincing and conservative support for this strategy comes from the "reversed" condition. While the confidence intervals are large, the strategy appears to be less accurate than "Click > Skip Above". However, the results confirm that the strategy is more accurate than random even in the reversed condition ($Binomial(0.5)$, $p = 0.01$).

5.4 Are clicks relative relevance judgments within a query chain?

All strategies from the previous section generate preferences only between results from the same query. As already argued in [Radlinski and Joachims 2005], restricting ourselves to such within-query preferences is likely to be suboptimal for at least two reasons. First, such strategies only ever produce preferences between the top few results presented to the user. If no highly relevant link is among those results, there will never be a preference directly indicating that such a link should be ranked higher. Second, these strategies do not exploit that users typically run not only a single query, but reformulate the query to refine and improve the results if necessary. So, later queries might help disambiguate earlier queries. In our study, we found that users ran on average 2.2 queries for answering a single question in Phase II[3].

In the following we investigate whether it is possible to generate sufficiently accurate relative preference judgments between results from different queries within a chain of query reformulations relating to the same information need. We will not argue that these preferences are necessarily more accurate than those from the within-query strategies, but that they reveal information that is not available from the within query strategies (see [Radlinski and Joachims 2005]) and that they are accurate enough to be potentially useful. For example, a query chain with no clicks after the query "oed", then a reformulation to "oxford english dictionary" with a click, might indicate that the clicked results is relevant to the query "oed" even if the document does not contain this string. It is unlikely that this connection could ever be inferred from within-query preferences. See [Radlinski and Joachims 2005]

[3]Studies of Web search engine logs report between 2.8 and 1.6 queries on average [Silverstein et al. 1999, Jansen et al. 2000], depending on their definition of what constitutes a session.

for learning experiments that support the value of preferences from query chains compared to learning from within-query preferences alone. Note that in our study we know the proper segmentation into query chains by construction (i.e. queries for the same question). For practical applications, we briefly discuss the automatic detection of query chains at the end of this section.

The first strategy we propose is an analogous extension of "Click > Skip Above" to multiple result sets. A preference is generated between two links from different result sets within the same query chain, if a link in an earlier result set was skipped and a link in a later result set was clicked.

STRATEGY 6. (CLICK > SKIP EARLIER QC)
For a ranking $(l_1, l_2, l_3, ...)$ followed (not necessarily immediately) by ranking $(l'_1, l'_2, l'_3, ...)$ within the same query chain and sets C and C' containing the ranks of the clicked-on links in either ranking, extract a preference example $rel(l'_i) > rel(l_j)$ for all pairs $i \in C'$ and $j < \max(C)$, with $j \notin C$.

To illustrate this strategy, consider the following example of a query chain with four queries. As in the previous section, clicks are indicated with a "*".

$$\begin{array}{l} q_1 : \mathrm{l}_{11}\ \ \mathrm{l}_{12}\ \ \mathrm{l}_{13}\ \ \mathrm{l}_{14}\ \ \mathrm{l}_{15}\ \ \mathrm{l}_{16}\ \ \mathrm{l}_{17} \\ q_2 : \mathrm{l}^*_{21}\ \ \mathrm{l}_{22}\ \ \mathrm{l}^*_{23}\ \ \mathrm{l}_{24}\ \ \mathrm{l}^*_{25}\ \ \mathrm{l}_{26}\ \ \mathrm{l}_{27} \\ q_3 : \mathrm{l}_{31}\ \ \mathrm{l}^*_{32}\ \ \mathrm{l}_{33}\ \ \mathrm{l}_{34}\ \ \mathrm{l}_{35}\ \ \mathrm{l}_{36}\ \ \mathrm{l}_{37} \\ q_4 : \mathrm{l}^*_{41}\ \ \mathrm{l}_{42}\ \ \mathrm{l}_{43}\ \ \mathrm{l}_{44}\ \ \mathrm{l}_{45}\ \ \mathrm{l}_{46}\ \ \mathrm{l}_{47} \end{array} \tag{2}$$

For this example, strategy "Click > Skip Earlier QC" will generate the preferences

$$\begin{array}{l} \mathrm{rel}(\mathrm{l}_{32}) > \mathrm{rel}(\mathrm{l}_{22}),\ \ \mathrm{rel}(\mathrm{l}_{32}) > \mathrm{rel}(\mathrm{l}_{24}), \\ \mathrm{rel}(\mathrm{l}_{41}) > \mathrm{rel}(\mathrm{l}_{22}),\ \ \mathrm{rel}(\mathrm{l}_{41}) > \mathrm{rel}(\mathrm{l}_{24}),\ \ \mathrm{rel}(\mathrm{l}_{41}) > \mathrm{rel}(\mathrm{l}_{31}). \end{array}$$

The accuracy of this strategy is shown in the first row of Table V. While the accuracy of this strategy as evaluated against the explicit judgments of the abstracts is significantly different from random in the "normal" ($Binomial(0.5)$, $p < 10^{-4}$) and "swapped" condition ($Binomial(0.5)$, $p = 0.01$), the table shows a significant influence of the presentation. In particular, the accuracy is not significantly different from the random baseline in the "reversed" condition ($Binomial(0.5)$, $p = 0.61$), and the accuracy in the "normal" condition is significantly different from the accuracy in the "reversed" condition (Binomial Proportion Test, $p = 0.007$). Our conjecture is that this results from the sequential way in which the user evaluates the options. When deciding to click in the later query, the links from the previous query are no longer visible to the user, so that a direct comparison is limited by memory. Such memory-based decision making is generally believed to be less accurate (see e.g. [Hutchinson and Alba 1991]).

To improve the accuracy of the preferences, we tested the subset of preferences generated only by the last click in a query chain. Otherwise, the following strategy is equivalent to "Click > Skip Earlier QC".

STRATEGY 7. (LAST CLICK > SKIP EARLIER QC)
For a ranking $(l_1, l_2, l_3, ...)$, let C contain the ranks of the clicked-on links. If the last ranking $(l'_1, l'_2, l'_3, ...)$ within the same query chain received a click, then let i be the temporally last click in this ranking and extract a preference example $rel(l'_i) > rel(l_j)$ for all pairs $j < \max(C)$, with $j \notin C$.

Table V. Accuracy of several strategies for generating pairwise preferences from clicks between multiple queries within a query chain (some results already appeared in [Radlinski and Joachims 2005]). The base of comparison are either the explicit judgments of the abstracts, or the explicit judgments of the page itself. Behind each ± we show the larger of the two sides of the 95% binomial confidence interval around the sample mean. The column "p/q" shows the average number of preferences generated per query by this strategy.

Explicit Feedback		Abstracts				Pages
Data		Phase II				Phase II
Strategy	p/q	"normal"	"swapped"	"reversed"	all	all
Click > Skip Earlier QC	0.49	84.5±16.4	71.1±17.0	54.6±18.1	70.2±9.7	68.0±8.4
Last Click > Skip Earlier QC	0.33	77.3±20.6	80.8±20.2	42.1±24.4	68.7±12.6	66.2±12.2
Click > Click Earlier QC	0.30	61.9±23.5	51.2±17.1	35.3±26.4	50.6±11.4	65.8±11.8
Click > TopOne NoClickEarl. QC	0.35	86.4±21.2	77.3±15.1	92.6±16.9	83.9±9.1	85.4±8.7
Click > TopTwo NoClickEarl. QC	0.70	88.9±12.9	80.0±10.1	86.8±12.1	84.2±6.1	84.5±6.1
TopOne > TopOne Earlier QC	0.84	65.3±15.2	68.2±12.7	75.6±15.1	69.4±7.8	69.4±7.9

For the example, this strategy will generate the preferences

$$\mathrm{rel}(l_{41}) > \mathrm{rel}(l_{22}),\ \mathrm{rel}(l_{41}) > \mathrm{rel}(l_{24}),\ \mathrm{rel}(l_{41}) > \mathrm{rel}(l_{31}).$$

Row two of Table V shows that there is no evidence that preferences derived from the last click are more accurate. The accuracies are well within the confidence intervals of the preferences generated from the more general strategy "Click > Skip Earlier QC" accross all conditions.

In analogy to "Click > Earlier Click" for within query preferences, we designed the following strategy to explore the relationship between pairs of clicked links between queries. In particular, we generate a preference between a clicked link of an earlier query and a clicked link of a later query in the same query chain.

Strategy 8. (Click > Click Earlier QC)
For a ranking $(l_1, l_2, l_3, ...)$ *followed by ranking* $(l'_1, l'_2, l'_3, ...)$ *within the same query chain and sets* C *and* C' *containing the ranks of the clicked-on links in either ranking, extract a preference example* $rel(l'_i) > rel(l_j)$ *for all pairs* $i \in C'$ *and* $j \in C$.

Applied to the example query chain, this strategy will generate the preferences

$$\begin{aligned}&\mathrm{rel}(l_{32}) > \mathrm{rel}(l_{21}),\ \mathrm{rel}(l_{32}) > \mathrm{rel}(l_{23}),\ \mathrm{rel}(l_{32}) > \mathrm{rel}(l_{25}),\\ &\mathrm{rel}(l_{41}) > \mathrm{rel}(l_{21}),\ \mathrm{rel}(l_{41}) > \mathrm{rel}(l_{23}),\ \mathrm{rel}(l_{41}) > \mathrm{rel}(l_{23}),\\ &\mathrm{rel}(l_{41}) > \mathrm{rel}(l_{32}).\end{aligned}$$

The strategy "Click > Click Earlier QC" shows results that are qualitatively similar to those of "Click > Earlier Click". The accuracy is not significantly different from random ($Binomial(0.5)$, $p = 0.91$), and there appear to be strong biases resulting from the presentation.

One shortcoming of the two strategies "Click > Skip Earlier QC" and "Last Click > Skip Earlier QC" is that they generate preferences only if an earlier query within the chain drew a click. However, about 40% of all queries in Phase II did not receive any clicks. For such queries without clicks, we rely on our eye-tracking results that show that users typically view the top links. For queries without clicks, we therefore assume that the user evaluated the top two links and decided to not click on them, but rather to reformulate the query. This leads to the following two

strategies, where we generate a preference between a clicked link in a later query, and the first (or second) link in a earlier query that received no clicks.

STRATEGY 9. (CLICK > TOPONE NOCLICKEARLIER QC)
For a ranking $(l_1, l_2, l_3, ...)$ *that received no clicks followed by ranking* $(l'_1, l'_2, l'_3, ...)$ *within the same query chain having clicks on ranks* C'*, extract a preference example* $rel(l'_i) > rel(l_1)$ *for all* $i \in C'$*.*

STRATEGY 10. (CLICK > TOPTWO NOCLICKEARLIER QC)
For a ranking $(l_1, l_2, l_3, ...)$ *that received no clicks followed by ranking* $(l'_1, l'_2, l'_3, ...)$ *within the same query chain having clicks on ranks* C'*, extract preference examples* $rel(l'_i) > rel(l_1)$ *and* $rel(l'_i) > rel(l_2)$ *for all* $i \in C'$*.*

Applying strategy "Click > TopOne NoClickEarlier QC" to the example generate the preferences

$$\text{rel}(\text{l}_{21}) > \text{rel}(\text{l}_{11}),\ \text{rel}(\text{l}_{23}) > \text{rel}(\text{l}_{11}),\ \text{rel}(\text{l}_{25}) > \text{rel}(\text{l}_{11}),$$
$$\text{rel}(\text{l}_{32}) > \text{rel}(\text{l}_{11}),\ \text{rel}(\text{l}_{41}) > \text{rel}(\text{l}_{11}).$$

For "Click > TopTwo NoClickEarlier QC" the analogous preferences for l_{12} would be added as well. Table V shows that the preferences from these two strategies are highly accurate across all conditions. The fact that the user decided to not click on any link but rather to reformulate the query appears to give particularly strong evidence.

The accuracy of the previous strategies "Click > TopOne NoClickEarlier QC" and "Click > TopTwo NoClickEarlier QC" suggests that users not only give negative feedback about the result set by not clicking on any link, but also that they learn from the result set how to formulate a better query. In particular, a user might discover an unanticipated ambiguity of the orignial query, which is avoided in a query reformulation. To see whether users manage to improve their queries within a chain of reformulations, we evaluated how often the top result of a later query is more relevant than the top result of an earlier query.

STRATEGY 11. (TOPONE > TOPONE EARLIER QC)
For a ranking $(l_1, l_2, l_3, ...)$ *followed by ranking* $(l'_1, l'_2, l'_3, ...)$*, extract the preference example* $rel(l'_1) > rel(l_1)$ *(i.e. the links top-ranked by Google).*

To illustrate, this strategy would generate the preferences

$$\text{rel}(\text{l}_{21}) > \text{rel}(\text{l}_{11}),\ \text{rel}(\text{l}_{31}) > \text{rel}(\text{l}_{11}),\ \text{rel}(\text{l}_{41}) > \text{rel}(\text{l}_{11}),$$
$$\text{rel}(\text{l}_{31}) > \text{rel}(\text{l}_{21}),\ \text{rel}(\text{l}_{41}) > \text{rel}(\text{l}_{21}),\ \text{rel}(\text{l}_{41}) > \text{rel}(\text{l}_{31}),$$

for the example. Table V shows that query reformulations generally improve the top result. For 69.4% of the query pairs, Google's top result of the later query is more relevant than its top result of an earlier query. The difference from the random baseline of 50% is significant ($Binomial(0.5)$, $p < 10^{-5}$). This shows that users learn how to formulate better queries during the search process. While this is an interesting insight, we do not advocate using "TopOne > TopOne Earlier QC" to generate training examples for learning improved retrieval functions, since it's preferences strongly depend on the quality of the current retrieval function.

Finally, to make any of the strategies introduced above applicable in practice, it will be necessary to detect query chains automatically. Initial experiments with an intranet search engine for the Cornell Library Web pages indicate that segmenting a sequence of queries into intervals of approximately constant information need is tractable. Using machine learning based on features like the overlap of query words, overlap and similarity of the retrieved results, and time between queries, it was possible to learn an accurate segmentation rule [Radlinski and Joachims 2005]. Despite this success, it remains an open question whether this segmentation can be done equally accurately in a web search setting, and in how far the information need "drifts" within long query chains.

5.5 How accurately do clicks correspond to explicit judgment of a document?

Sections 5.3 and 5.4 showed that certain types of preference statements derived from clicks correspond well with explicit relevance judgments of the abstract. This means that implicit and explicit feedback based on the same (limited) amount of information, namely the abstract, are reasonably consistent. However, it is not clear whether users make reliable relevance judgments of the actual pages based on the abstract alone. We will now use the explicit judgments we collected for the data from Phase II to investigate in how far the preference statements derived from clicks agree with the explicit relevance judgments of the pages.

The last colum of Table IV shows the agreement with the explicit relevance judgments of the pages for the different within-query strategies. We compare this column to the neighboring column that shows the agreement with the explicit judgments of the abstract on the same data. For most strategies, the agreement with the explicit page judgement is slightly lower than the agreement with the abstract judgments ("Click > Skip Above", "Last Click > Skip Above", "Click > Skip Previous", "Click > No Click Next"). On average there seems to be a drop in agreement of around 3%. The only exception is the strategy "Click > Earlier Click", where there is an increase in agreement. Such an increase is plausible: a misleadingly promising abstract might attract the click of a user, but the user returns to the results page and selects another link.

The agreement with the explicit relevance judgments of the pages for the query-chain strategies is shown in the last column of Table V. The results are similar to those of the within-query strategies. In general, the preferences agree well with the explicit judgments of the pages. For all but one strategy, the accuracies are not substantially different from those for the judgments of the abstracts. The strategy "Click > Click Earlier QC" that generates preferences between multiple clicks is the exception, showing an increase in agreement similar to "Click > Earlier Click".

We conclude that the implicit feedback generated from clicks both within result sets and between result sets in a query chain shows reasonable agreement with the explicit judgments of the pages. While for most strategies the agreement between implicit and explicit judgments is lower than the average agreement of 86.4% between two explicit judgments, the implicit judgments are still reasonably accurate. For the two strategies "Click > TopOne NoClickEarlier QC" and "Click > TopTwo NoClickEarlier QC" that exploit the lack of any click in a previous query, the accuracy is particularly high. These two strategies are likely to be particularly valuable not only for their high accuracy, but also for the kind of preferences they

produce. Their preferences are very informative, since they provide information about relevant results potentially deep down in the earlier ranking.

6. DISCUSSION AND LIMITATIONS

While a key motivation for the work in this paper is the use of implicit feedback as training data for automatically learning improved retrieval functions, the paper only addresses the first question towards such a system. It evaluates different strategies for generating pairwise preferences against human relevance judgments and analyses their accuracy. It does not adress the question of how these preferences can be used in a learning algorithm, how to represent a retrieval function, how to combine the strategies, or how "informative" the preferences from different strategies are. For example, the strategy "Click > No Click Next" always generates preferences that confirm the current ordering, so that these preferences by themselves would never encourage a learning algorithm to change the current ordering. Furthermore, none of the strategies generates independently identically distributed training data as assumed by most machine learning algorithms. Some of these issues are discussed in [Joachims 2002, Radlinski and Joachims 2005].

It is important to keep in mind that the results we report are obtained for one particular search engine and one particular group of users. The participants in our study were young, well educated, and internet savvy search-engine users. The behavior of less proficient users might be substantially different. Furthermore, the way the search engine presents results and generates abstracts is likely to have an influence on user behavior. It would be interesting to see whether the results change substantially for other search engines and other search tasks (e.g. intranet search, desktop search).

Another limitation is that all feedback strategies we analyzed only use clicks, but do not consider other signals like timing information and the behavior on pages downstream from the results page. Including such additional information could lead to more accurate implicit feedback. Furthermore, while click-spam is not an issue in intranet and desktop applications of search, one would likely incur adversarial users in web search. It would be interesting to investigate the robustness of learning algorithms to a bounded fraction of preferences from malicious clicks.

7. CONCLUSIONS

We presented a comprehensive study addressing the reliability of implicit feedback for WWW search engines that combines detailed evidence about the users' decision process as derived from eyetracking, with a comparison against explicit relevance judgments. Our results indicate that users' clicking decisions are influenced by the relevance of the results, but that they are biased by the order in which they are presented, and by the overall quality of the result set. This makes it difficult to interpret clicks as *absolute* feedback. However, we examine several strategies for generating *relative* feedback signals from clicks, which are shown to correspond well with explicit judgments. In particular, we define a set of strategies for extracting pairwise preferences within results sets and between different result sets within a chain of query reformulations. While the implicit relevance signals are less consistent with the explicit judgments than the explicit judgments among each other, the

difference is encouragingly small. The fact that implicit feedback from clicks is readily available in virtually unlimited quantity might more than overcome this quality gap, if implicit feedback is properly interpreted using machine learning methods for pairwise preferences (e.g. [Joachims 2002]). In future work, we plan to continue to build adaptive retrieval systems that explore the use of machine learning from implicit feedback signals.

ACKNOWLEDGMENTS

We thank the subjects and the relevance judges for their help with this project. This work was funded in part through NSF CAREER Award IIS-0237381 and through a gift from Google.

REFERENCES

Agichtein, E., Brill, E., and Dumais, S. 2006. Improving web search ranking by incorporating user behavior. In *Conference on Research and Development in Information Retrieval (SIGIR)*. 19–26.

Agichtein, E., Brill, E., Dumais, S., and Ragno, R. 2006. Learning user interaction models for predicting web search preferences. In *Conference on Research and Development in Information Retrieval (SIGIR)*. 3–10.

Almeida, R. and Almeida, V. 2004. A community-aware search engine. In *Proceedings of the World Wide Web Conference (WWW)*.

Bartell, B., Cottrell, G., and Belew, R. 1994. Automatic combination of multiple ranked retrieval systems. In *Annual ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*. 173–181.

Belew, R. 2000. *Finding Out About*. Cambridge.

Boyan, J., Freitag, D., and Joachims, T. 1996. A machine learning architecture for optimizing web search engines. In *AAAI Workshop on Internet Based Information Systems*. 1 – 8.

Broder, A. 2002. A taxonomy of web search. *SIGIR Forum 36,* 2, 3–10.

Brumby, D. and Howes, A. 2003. Interdependence and past-experience in menu choice assessment. In *Poster presented at the 25th Annual Meeting of the Cognitive Science Society*.

Brumby, D. and Howes, A. 2004. Good enough but i'll just check: Web-page search as attentional refocusing. In *International Conference on Cognitive Modeling*.

Burges, C., Renshaw, S., Renshaw, E., Ari, L., Deeds, M., Hamilton, N., and Hullender, G. 2005. Learning to rank using gradient descent. In *International Conference on Machine Learning (ICML)*. Bonn, Germany.

Claypool, M., Le, P., Waseda, M., and Brown, D. 2001. Implicit interest indicators. In *International Conference on Intelligent User Interfaces (IUI)*. 33–40.

Cohen, W., Shapire, R., and Singer, Y. 1999. Learning to order things. *Journal of Artificial Intelligence Research 10*, 243–270.

Crammer, K. and Singer, Y. 2001. Pranking with ranking. In *Advances in Neural Information Processing Systems (NIPS)*.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems 23,* 2, 147–168.

Freund, Y., Iyer, R., Schapire, R., and Singer, Y. 1998. An efficient boosting algorithm for combining preferences. In *International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers Inc., 170–178.

Fuhr, N. 1989. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems 7,* 3, 183–204.

Furnas, G. 1985. Experience with an adaptive indexing scheme. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM Press, New York, NY, USA, 131–135.

Goldberg, J., Stimson, M., Lewenstein, M., Scott, M., and Wichansky, A. 2002. Eye-tracking in web search tasks: design implications. In *Proceedings of the Eye tracking Research and Applications Symposium (ETRA)*. 51–58.

Granka, L., Joachims, T., and Gay, G. 2004. Eye-tracking analysis of user behavior in www search. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

Halverson, T. and Hornof, A. 2004. Link colors guide a search. In *ACM Conference on Computer-Human Interaction (CHI)*.

Herbrich, R., Graepel, T., and Obermayer, K. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 115–132.

Holland, S., Ester, M., and Kieling, W. 2003. Preference mining: A novel approach on mining user preferences for personalized applications. In *Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. 204 – 216.

Hutchinson, W. and Alba, J. 1991. Ignoring irrelevant information: Situational determinats of consumer learning. *Journal of Consumer Research 18*, 325–345.

Jansen, B., Spink, A., and Saracevic, T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage. 36,* 2, 207–227.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Joachims, T., Freitag, D., and Mitchell, T. 1997. WebWatcher: a tour guide for the world wide web. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 1. Morgan Kaufmann, 770 – 777.

Joachims, T., Granka, L., Pang, B., Hembrooke, H., and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*. 154–161.

Jones, R. and Fain, D. 2003. Query word deletion prediction. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM Press, New York, NY, USA, 435–436.

Just, M. and Carpenter, P. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review 87*, 329–354.

Kelly, D. and Belkin, N. 2004. Display time as implicit feedback: Understanding task effects. In *ACM International Conference on Research and Development in Information Retrieval (SIGIR)*. 377–384.

Kelly, D. and Teevan, J. 2003. Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum 37,* 2, 18–28.

Kemp, D. and Ramamohanarao, K. 2002. Long-term learning for web search engines. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. 263–274.

Klöckner, K., Wirschum, N., and Jameson, A. 2004. Depth- and breadth-first processing of search result lists. In *Extended Abstract, ACM Conference on Computer-Human Interaction*.

Lankford, C. 2000. Gazetracker: software designed to facilitate eye movement analysis. In *Proceedings of Eye Tracking Research & Applications*. 51–55.

Lieberman, H. 1995. Letizia: An agent that assists Web browsing. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI '95)*. Morgan Kaufmann, Montreal, Canada, 924–929.

Mantell, S. and Kardes, F. 1999. The role of direction of comparison, attribute-based processing, and attitude-based processing in consumer preference. *Journal of Consumer Research 25*, 335–352.

Morita, M. and Shinoda, Y. 1994. Information filtering based on user behavior analysis and best match text retrieval. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*. 272–281.

Oard, D. and Kim, J. 1998. Implicit feedback for recommender systems. In *Proceedings of the AAAI Workshop on Recommender Systems*. 81–83.

Pan, B., Hembrooke, H., Gay, G., Granka, L., Feusner, M., and Newman, J. 2004. The determinants of web page viewing behavior: An eye tracking study. In *Proceedings of Eye Tracking Research & Applications*, S. Spencer, Ed. ACM, New York.

Radlinski, F. and Joachims, T. 2005. Query chains: Learning to rank from implicit feedback. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*.

Rayner, K. 1998. Eye movements in reading and information processing. *Psychological Bulletin 124*, 372–252.

Salogarvi, J., Kojo, I., Jaana, S., and Kaski, S. 2003. Can relevance be inferred from eye movements in information retrieval? In *Proceedings of the Workshop on Self-Organizing Maps*. 261–266.

Samuelson, P. 1948. Consumption theory in terms of revealed preferences. *Econometrica 15*, 243–253.

Shen, X., Tan, B., and Zhai, C. 2005. Context-sensitive information retrieval using implicit feedback. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM Press, New York, NY, USA, 43–50.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. *SIGIR Forum 33,* 1, 6–12.

Teevan, J., Dumais, S., and Horvitz, E. 2005. Beyond the commons: Investigating the value of personalizing web search. In *Workshop on New Technologies for Personalized Information Access (PIA)*. 84–92.

Varian, H. 1992. *Microeconomic Analysis*. Noron, New York.

White, R., Ruthven, I., and Jose, J. 2002. The use of implicit evidence for relevance feedback in web retrieval. In *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*. Springer-Verlag, London, UK, 93–109.

White, R., Ruthven, I., and Jose, J. 2005. A study of factors affecting the utility of implicit relevance feedback. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. ACM Press, New York, NY, USA, 35–42.

Received Oct. 2005; revised Aug. 2006; accepted Oct. 2006