Accurately Interpreting Clickthrough Data as Implicit Feedback

Thorsten Joachims
Dept. of Computer Science
Cornell University
Ithaca, NY, USA
tj@cs.cornell.edu

Laura Granka
Dept. of Communication
Stanford University
Palo Alto, CA, USA
granka@stanford.edu

Bing Pan,
Helene Hembrooke &
Geri Gay
Information Science
Cornell University
Ithaca, NY, USA
{bp58,hah4,gkg1}@cornell.edu

ABSTRACT

This paper examines the reliability of implicit feedback generated from clickthrough data 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.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: User Studies

General Terms

Human Factors, Measurement, Reliability, Experimentation

Keywords

Implicit Feedback, Eyetracking, WWW Search, Clickthrough

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. 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. [9, 2, 4, 17, 14, 13, 1]). 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.

To evaluate the reliability of 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 performed 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. 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 [20, 6, 26, 8, 16]. 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 bias": the users' clicking decision is not only influ-

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

enced 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. IMPLICIT FEEDBACK IN RETRIEVAL

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. [9, 2]). While Cohen et al. [7] 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. An algorithm that adapts the retrieval function to minimize the rank of the clicked links was proposed in [4]. Joachims proposed a Support Vector Algorithm that can be trained with pairwise preferences extracted from clicks [14]. A similar approach is followed in [13]. Kemp and Ramamohanarao [17] use clickthrough data for document expansion by adding the query words to the clicked documents. Session logs from an online bookstore are used in [1] to identify communities and personalize search.

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 study in [20] finds that reading time is indicative of interest when reading newsstories. Similarly, Claypool et al. [6] 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 [16] 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. [8] show in their study that the overall time a user interacts with a search engine, as well as the number of clicks, are indicative of user satisfaction with the search engine. In the following, we explore a new set of strategies for generating implicit feedback from clicks that was not considered by any of these previous studies.

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. To our knowledge, this is the first study of user behavior in retrieval systems of this kind.

3.1 Task, Participants, and Conditions

We designed the study to resemble typical use of a WWW search engine. Participants were asked to answer the same ten questions using Google as a starting point for their search. Half of the searches were navigational [5], asking subjects to find a specific Web page or homepage. The other five tasks were informational [5], 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 1.

Table 1: Questions used in the study.

Navigational
– Find the homepage of Michael Jordan, the statistician.
– Find the page displaying the route map for Greyhound buses.
– Find the homepage of the 1000 Acres Dude Ranch.
– Find the homepage for graduate housing at Carnegie Mellon University.
– Find the homepage of Emeril - the chef who has a television cooking program.
Informational
– Where is the tallest mountain in New York located?
– 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?
– Which actor starred as the main character in the original Time Machine movie?
– 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?
– What is the name of the researcher who discovered the first modern antibiotic?

Users were instructed to start their search with a Google query of their choice and then search for the answer as they normally would. There were no restrictions on what queries users may choose, how and when to reformulate the query, or which links to follow. Users were told that we were studying how people search on the Web, but were not told that we were specifically interested in their behavior on the results page of Google. Users were read each question in turn and answered orally when they found the answer.

We conducted the user study in two phases. In Phase I, we recruited 34 participants, all of which were undergraduate students of various majors at Cornell University. 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. The majority of students were given extra credit in communication courses for their participation. 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, and all subjects indicated at least a general familiarity with the Google interface, as 31 of the subjects reported that Google is their primary search engine.

Phase II of the study was designed to investigate how users react to manipulations of the search results. Using the same ten question and the same instructions to the subjects as in Phase I, each subject was 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 and they did not know that we manipulated the results. When asked after their session, none of the subjects had suspected any manipulation.

22 participants were recruited for Phase II of the study and we were able to record usable eye tracking data for 16 of them. 6 users were in the "normal" condition, 5 in the "swapped" condition, and 5 in the "reversed" condition. Again, the participants were students from various majors with a mean age of 20.4 years.

3.2 Data Capture

The subjects' eye movements were recorded using an ASL 504 commercial eyetracker (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 simultaneous acquisition and analysis of the subject's eye movements [19].

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.

3.3 Eyetracking

We classify eye movements according to the following significant indicators of ocular behaviors, namely fixations, saccades, pupil dilation, and scan paths [23]. 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 [15, 23].

Other indices, such as saccades, are believed to occur too quickly to absorb new information [23]. 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 [23]. While pupil dilation could be interesting in our analysis, we focus on fixations in this paper.

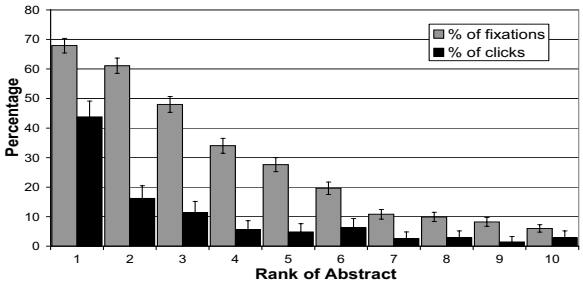

Figure 1: Percentage of times an abstract was viewed/clicked depending on the rank of the result.

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 users.

For each results page from Phase I, we randomized the order of the abstracts and asked judges to (weakly) order the *abstracts* by how promising they look for leading to relevant information. 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. [3, Page 109]). Five judges (different from subjects) each assessed the results pages for two of the questions, plus ten results pages from two other questions for inter-judge agreement verification. The judges received detailed instructions and examples of how to judge relevance. However, we explicitly 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 were 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. 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. 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 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.

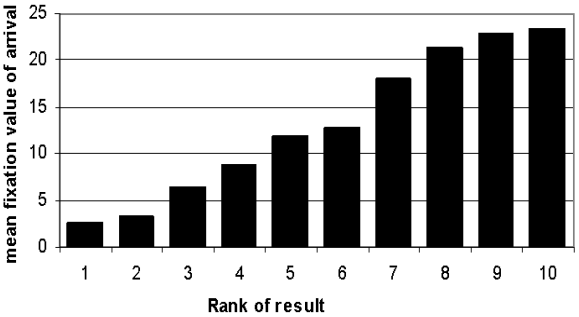

Figure 2: Mean time of arrival (in number of previous fixations) depending on the rank of the result.

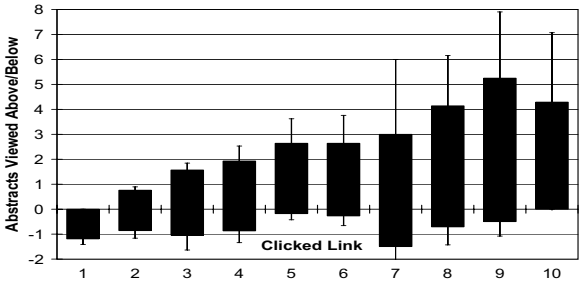

Figure 3: Mean number of abstracts viewed above and below a clicked link depending on its rank.

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. The abstracts ranked 1 and 2 receive most attention. After that, attention drops faster. The dark bars im 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 below this rank receive substantially less attention than those earlier. 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; 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.

Table 2: Percentage of times the user viewed an abstract at a particular rank before he clicked on a link at a particular rank.

Viewed Rank	Clicked 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%

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. 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.

Table 2 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.

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. Users react to the degraded ranking in two ways. First, they view lower ranked links more frequently. In the "reversed" condition subjects scan significantly more abstracts than in the "normal" condition. All significance tests reported in this paper are two-tailed tests at a 95% confidence level. Second, subjects are much less likely to click on the first link, but more likely to click on a lower ranked link. 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 to the Wilcoxon test. Furthermore, the average number of clicks per query decreases from 0.80 in the "normal" condition to 0.64 in the "reversed" condition.

This shows that users 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. [4, 17, 8]). 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 3. The experiment focuses on the top two links, since these two links are scanned relatively equally. Table 3 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 19/20 is significantly different from 2/7 assuming a binomial distribution. 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 3 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.

Table 3: Number of clicks on the top two links depending on relevance of the abstracts for the normal and the swapped condition for Phase II. In the column headings, +/- indicates whether or not the user clicked 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

We conclude that users have substantial trust in the search engine's ability to estimate the relevance of a page, which influences their clicking behavior.

5.2.2 Quality 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" and "swapped" 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 (i.e. 1 to 10 for a typical results pages) as assigned by the relevance judges. We call this number the relevance rank of an abstract. To focus on results pages where the users in the "reversed" condition saw less relevant abstracts, we only consider those cases where the clicks are not below rank 5. For theses cases, the average relevance rank of clicks in the "normal" or "swapped" condition is 2.67 compared to 3.27 in the "reversed" condition. The difference is significant according to the Wilcoxon test.

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.

5.3 Are clicks relative relevance judgments?

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

Table 4: Accuracy of several strategies for generating pairwise preferences from clicks. The base of comparison are either the explicit judgments of the abstracts, or the explicit judgments of the page itself. Error bars are the larger of the two sides of the 95% binomial confidence interval around the mean.

Explicit Feedback	Abstracts					Pages
Data	Phase I	Phase II				Phase II
Strategy	"normal"	"normal"	"swapped"	"reversed"	all	all
Inter-Judge Agreement	89.5	N/A	N/A	N/A	82.5	86.4
Click > Skip Above	80.8 ± 3.6	88.0 ± 9.5	79.6 ± 8.9	83.0 ± 6.7	83.1 ± 4.4	78.2 ± 5.6
Last Click > Skip Above	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	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	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	84.1 ± 4.9	75.6 ± 14.5	66.7 ± 13.1	70.0 ± 15.7	70.4 ± 8.0	67.4 ± 8.2

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.

We will now explore implicit feedback measures that respect these dependencies by interpreting clicks not as absolute relevance feedback, but as pairwise preference statements. Such an interpretation is supported by research in marketing, which has shown that humans tend to make pairwise comparisons among options [24]. 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 [14, 7]. 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.

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

While it is difficult to infer whether the links l_1, l_3, and l_5 are relevant on an *absolute* scale, it is 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

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

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 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.

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 4 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 (binomial distribution) better than the random baseline of 50%. 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 (binomial test) w.r.t. degradations in ranking quality in the "swapped" and "reversed" condition. As expected, trust bias and quality 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 that 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

$$\text{rel}(l_5) > \text{rel}(l_2), \qquad \text{rel}(l_5) > \text{rel}(l_4).$$

Table 4 shows that this strategy is slightly more accurate than "Click > Skip Above". The difference is significant in Phase I, but not Phase II (binomial test).

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

$$\text{rel}(l_1) > \text{rel}(l_3), \quad \text{rel}(l_5) > \text{rel}(l_3), \quad \text{rel}(l_5) > \text{rel}(l_1).$$

The validity of this strategy is not supported by the data. The accuracy is 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 test) difference between "normal" and "reversed" condition. 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. (Last 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 $i - 1 \notin C$.

The accuracy is given in Table 4. This strategy shows no significant (binomial test) differences compared to "Click > Skip Above".

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 $(i + 1) \notin C$.

Table 4 shows that this strategy appears highly 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.

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

The previous section 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 4 shows the agreement with the explicit relevance judgments of the pages for the different 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"). While none of the individual differences is significant (binomial test), 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. While the difference is not significant (binomial test), 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.

We conclude that the implicit feedback generated from clicks shows reasonable agreement with the explicit judgments of the pages. While 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.

6. RELATED WORK ON EYETRACKING IN INFORMATION SEARCH

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 [10, 21, 12]. 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 [10, 21], and assessing how link color may influence visual search patterns [12].

More similar to the research presented here, Salogarvi et al. [25] 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 a bit weak, and furthermore, no other measures of searcher performance were addressed. The most similar published research [11] reported descriptive eye movement analyses to depict the overall user pattern of evaluation on results from a search engine, but did not yet correlate this with implicit relevance evaluations.

One other study used eye-tracking in online search to assess the manner in which users evaluate search results [18]. They conducted two experiments to determine whether users engaged in a more exhaustive "breadth-first" search (meaning that users will look over a number of the results before clicking any), or a "depth-first" search. In both studies, users were significantly more likely to engage in the depth-first strategy, clicking on a promising link before continuing to view other abstracts within the results set.

7. CONCLUSIONS

We presented the first 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 trust they have in the retrieval function, 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. 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. [14]).

In future work, we plan to continue to build adaptive retrieval systems that use implicit feedback signals. We also plan to extend our analysis in several ways. For example, we will explore additional feedback strategies that include timing information and the behavior on pages downstream from the results page. Including such additional information could lead to more accurate implicit feedback. Furthermore, we are exploring relative feedback from clicks not only for results within a single query, but spanning a chain of related queries. Initial findings are reported in [22].

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.

8. REFERENCES

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

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

[3] R. Belew. *Finding Out About*. Cambridge, 2000.

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

[5] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

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

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

[8] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve the search experiences. In *Talk presented at SIGIR03 Workshop on Implicit Measures of User Interests and Preferences*, 2003.

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

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

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

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

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

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

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

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

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

[18] K. Kloeckner, N. Wirschum, and A. Jameson. Depth- and breadth-first processing of search result lists. In *ACM Conference on Computer-Human Interaction*, 2004.

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

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

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

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

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

[24] E. Russo and F. LeClerc. An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*, 21(2):274–290, 1994.

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

[26] R. White, J. Jose, and I. Ruthven. Comparing explicit and implicit feedback techniques for web retrieval: Trec-10 interactive track report. In *Text Retrieval Conference (TREC)*, 2001.