Effect of task on time spent reading as an implicit measure of interest — **Source link** ⧉

Melanie Kellar, Carolyn Watters, Jack Duffy, Michael Shepherd

Institutions: Dalhousie University

Related papers:

- Implicit interest indicators

- Implicit feedback for inferring user preference: a bibliography

- GroupLens: applying collaborative filtering to Usenet news

- Information filtering based on user behavior analysis and best match text retrieval

- Display time as implicit feedback: understanding task effects

Effect of Task on Time Spent Reading as an Implicit Measure of Interest

Melanie Kellar

Dalhousie University, Halifax, Nova Scotia. melanie@cs.dal.ca

Carolyn Watters

Dalhousie University, Halifax, Nova Scotia. watters@cs.dal.ca

Jack Duffy

Dalhousie University, Halifax, Nova Scotia. Jack.duffy@dal.ca.

Michael Shepherd

Dalhousie University, Halifax, Nova Scotia. shepherd@cs.dal.ca

Abstract

Information Filtering systems learn user preferences either through explicit or implicit feedback. However, requiring users to explicitly rate items as part of the interface interaction can place a large burden on the user. Implicit feedback removes the burden of explicit user ratings by transparently monitoring user behavior such as time spent reading, mouse movement and scrolling behavior. Previous research has shown that task may affect the effectiveness of some implicit measures. In this work we report both qualitative and quantitative results of an initial study examining the relationship between user time spent reading and relevance for three web search tasks: relevance judgment, simple question answering and complex question answering. This study indicates that the usefulness of time spent as a measure of user interest is related to task and becomes more useful as the task becomes more complex. Future directions for this research are presented.

Introduction

The rapid growth of the World Wide Web has highlighted the need for systems that help users sift through the wealth of information available online in order to find documents relevant to their needs. Information Filtering (IF) systems select documents for users based on their previous preferences. In order to learn user preferences, IF systems must determine user interest through either explicit or implicit user feedback. The next generation of IF systems must respond to the context of past user preferences as well as their current task, and it must do so without incurring additional overhead from the user.

Explicit feedback may come in the form of user defined rules or more commonly as user ratings. Users are often required to give binary judgments after reading each document, such as "relevant" or "not relevant", or rate items on a five point Likert scale. Implicit feedback transparently monitors user behavior in order to determine user preferences. Common user behaviors that have been found to indicate interest include time spent reading, mouse movements, scrolling behavior, and interactions with a document, such as saving, forwarding and printing.

Implicit feedback has a number of advantages over explicit feedback. Unobtrusively monitoring user behavior allows users to focus on the task at hand without the interruption of having to give ratings for documents. Requiring users to stop and explicitly rate items before moving on to the next document disrupts a user's normal reading and browsing behavior (Middleton et al., 2003; Nichols, 1997) and complicates the design. It is difficult to motivate users to continuously give explicit ratings (Kim et al., 2002; Konstan et al., 1997) even when the benefits are obvious (i.e., personalized recommendations). Implicit feedback may also help to eliminate inconsistencies in explicit user ratings.

A hurdle for IF systems is user attitude towards having their behavior monitored. If users are not comfortable with having some behaviors monitored in order to make recommendations then an IF system with implicit feedback may be unproductive.

IF systems can be used to make recommendations for users engaged in a variety of tasks, such as news reading, web browsing and web retrieval. Previous research has

shown that the effectiveness of some implicit measures of interest may be dependent on the task at hand. For instance, time spent reading has been found to be a good indicator of interest for news reading (Konstan et al., 1997; Morita & Shinoda, 1994) and web browsing (Claypool et al., 2001; Seo & Zhang, 2000), but contradictory results have been found for Information Retrieval tasks (Kelly & Belkin, 2001; White et al., 2002). Based on this previous work, we have chosen to concentrate on evaluating the effectiveness of time spent reading for tasks related to web searching.

We conducted a study examining the use of time spent reading as an implicit indicator of interest for three web search tasks: relevance judgment, simple question answering (fact finding), and complex question answering. We will present quantitative results from this study as well as qualitative results from user questionnaires exploring user attitudes towards IF systems that use implicit feedback.

Following, we present previous findings with implicit measures and the affect of task on the effectiveness of implicit measures. We then present a detailed methodology of the study we performed followed by the results of the study. Lastly, we end with a discussion and future direction for this work.

Background

Implicit Measures

Behaviors such as time spent reading (Morita & Shinoda, 1994; Konstan et al., 1997), mouse activity (Goecks & Shavlik, 2000; Hijikata, 2004), scrolling behavior (Claypool et al., 2001), items bookmarked (Seo & Zhang, 2000) and interactions with a document (Kim et al., 2001) have been examined as implicit measures of user interest. These behaviors can be used to indicate interest for a variety of systems, such as IF systems, recommender systems and query expansion systems (Golovchinsky, 1999). In this work we are concentrating on the use of implicit measures for IF systems.

Oard and Kim (2001) developed a classification scheme of observable behaviors, shown in Table 1. This table continually evolves as new behaviors are added, with the most recent addition being the *create* behavior added by Kelly and Teevan (2003). Much of the current research has concentrated on the *examine* and *retain* categories.

Time Spent Reading as an Implicit Measure

In this research, we concentrate on time spent reading as an indicator of user interest. This implicit measure has been examined for a variety of user tasks, such as news reading, web browsing, reading journal articles and web searching.

Table 1. Classification scheme of observable behaviors used as implicit measures of interest (Oard & Kim, 2001).

		Minimum Scope		
		Segment	Object	Class
Behavior Category	**Examine**	View Listen Scroll Find Query	Select	Browse
	Retain	Print	Bookmark Save Delete Purchase Email	Subscribe
	Reference	Copy-and-paste Quote	Forward Reply Link Cite	
	Annotate	Mark up	Rate Publish	Organize
	Create	Type Edit	Author	

Previous research has shown that time spent reading can be an effective measure of user interest when users are engaged in a news reading task or web browsing. Morita & Shinoda (1994) conducted an experiment where users explicitly rated news articles and the time spent reading was recorded. They found a strong tendency to spend a long time reading articles they rated interesting as compared to articles they rated not interesting. A later experiment conducted by Konstan et al. (1997) required users to explicitly rate UseNet articles while the time users spent on a page was recorded. The results showed a relationship between time spent reading and explicit ratings.

Claypool et al. (2001) examined the correlation between time spent reading and user interest for user directed web browsing. Users explicitly rated web pages and the time spent reading was recorded. The time spent reading was found to be a good indicator of interest. Kim et al. (2001) evaluated time spent reading for users reading academic journal articles. Users explicitly rated the articles and the time spent reading was recorded. They also found that time spent reading could be used to predict interest. Users tended to spend longer amounts of time reading relevant articles than non-relevant articles.

The effectiveness of time spent reading for Information Retrieval tasks has also been examined but with a variance of findings. White et al. (2002) recorded the time users spent reading while users judged the relevance of a query to a document summary. They reported that the difference between time spent reading relevant documents and non-relevant documents was statistically significant. However, when Kelly & Belkin (2001) attempted to replicate the results of Morita & Shinoda (1984) for web search tasks, they reported that they found no significant difference in the time spent reading relevant and non-relevant documents. These findings suggest that although time spent reading may be good indicator of interest for some tasks, such as news reading and web browsing, it may not be a good indicator for all tasks. Our study extends the work of Kelly & Belkin (2001) by examining the correlation between time spent reading and user interest for three web search tasks: relevance judgment, simple question answering, and complex question answering.

Methodology

The goal of this research was to determine for which web search tasks (relevance judgment, simple question answering, and complex question answering) time spent reading is an effective measure of user interest.

We evaluated the following three hypotheses:

H₁: Users will spend more time reading relevant documents than non-relevant documents for relevance judgment tasks.

H₂: Users will spend more time reading documents that contain the answer for simple question answering tasks then those that do not contain the answer.

H₃: Users will spend more time reading documents that contain the answer for complex question answering tasks then those that do not contain the answer.

Every participant first took part in a practice trial for each of the three conditions and then completed two trials as part of the study. The documents used in this study were news articles from the Reuters News Corpus, Volume 1, a collection of 810,000 news stories.

In addition to logging their time spent reading each document, participants responded to a summary questionnaire after the session. The questionnaire explored perceptions the participants had regarding how much time they thought they spent reading relevant and non-relevant documents. Participants were also questioned about their feelings towards a system that would monitor their behavior and for which tasks they would find recommendations useful.

Ten undergraduate science students, four female and six male, were recruited to participate in this study. Undergraduate students were selected as most graduate students possess advanced research skills that may not reflect the behavior of an average user. Participants were not paid for their participation in the study and each session lasted approximately one hour.

Relevance Judgment

The relevance judgment task required participants to judge five news documents as either relevant or non relevant to a supplied topic phrase. Table 2 shows the two topics used and titles of the documents belonging to that document set. The order of documents in each trial was randomly generated. Figure 1 shows the system during a relevance judgment task, with the topic "Protest held against Walt Disney Studios". The document set included non-relevant documents, documents that may have appeared relevant at first glance but in actuality were not relevant, and one relevant document. This condition was a loose replication of the experiment done by Kelly and Belkin (2001).

Table 2. Topic phrases and documents used for the relevance judgment task

Relevance Judgment Task
Topic Phrase: Protest held against Walt Disney Studios
Titles from document set:
• "The English Patient" early winner at Oscars.
• Catastrophes cost insurers 22 percent less in 1996.
• Hollywood readies for biggest night of year.
• Race on to translate Oscar wins into dollars.
• Elton John wants to smell the roses on his 50th.
Topic Phrase: Conspiracy theories surrounding President Kennedy's assassination in 1963
Titles from document set:
• Drug cuts heart deaths by 60 pct after surgery.
• Mystery light may be meteor - TWA probers.
• Sudanese activist wins human rights award.
• Followers, theorists mourn Kennedy 33 years on.
• FEATURE-33 years after, morbid tour recreates JFK murder

Figure 1.Example of a relevance judgment task. The topic is repeated at the top of every article and participants make their judgment using the *Relevant and Not Relevant* buttons at the bottom of the page.

Figure 2. Example of a complex question answering task. The question is repeated at the top of every article and participants submit their answers using *Answer* button at the bottom of the page.

Simple Question Answering

The next condition was a simple question answering task. This was a basic fact finding task in which participants were given five documents and a question, for example "As of July 1996, how many members did CompuServe have worldwide?" Again, documents were presented to the users one at a time. Table 3 displays the questions and document sets used for the simple question answering task. We were careful to choose questions that were not obvious enough the participants could answer them without reading the articles. For each document, the participants were required to either submit the answer or state that the article did not contain the answer. Each set consisted of non-relevant documents, documents that were somewhat relevant but did not contain the answer, and one document that contained the answer.

Table 3. Questions and documents used for simple question answering task

Simple Question Answering Task
Question: How much does a player on Namibia's national soccer team make per appearance?
Titles from document set: • Fed seen holding U.S. interest rates steady.   • Soccer-Aberdeen Struggle into UEF A Cup First Round after Scare.   • Southern Africa summit to tackle trade conflicts.   • Soccer-Namibia out to Enhance Reputation.   • Soccer-Keegan Despairs as Money Fails to Buy Results
Question: As of July 1996, how many members did CompuServe have worldwide?
Titles from document set: • Decision nears on Indiana tobacco lawsuit.   • Europe Online logs off for good after talks fail.   • CompuServe reports loss, cutting work force.   • CompuServe sees Q2 loss of $0.10-$0.15.   • Sprint says Internet can top long distance share.

Complex Question Answering

The third condition used more complex question answering tasks requiring participants to make a comparison within a document or complete a small calculation using data from within the document. Figure 2 shows the system during a complex question answering task. An example of a complex question is "Who are the three party leaders running for Prime Minister during Britain's 1997 elections?" Table 4 shows the questions and documents used for this condition. The names of the three leaders were found in three different places within one and only one article. Again, for each of the five documents, participants were required to either submit the answer or

state the article did not contain the answer. Each set consisted of non-relevant documents, documents that were somewhat relevant but did not contain the answer and one document that contained the answer.

Table 4. Questions and documents used for the complex question answering task

Complex Question Answering Task
Question: Al Gore witnessed the signing to two multi-million dollar contracts for US business in China. What are the names of those two businesses?
Titles from document set: • Drive opens to cleanse British politics of sleaze.   • China says did not aid U.S. Democrats.   • U.N. chief off to Zaire crisis summit in Togo.   • Gore meets China Premier, to witness business deals.   • Gore says no China policy change on fund raising.
Question: Who are the three party leaders running for Prime Minister during Britain's 1997 elections?
Titles from document set: • UK's Labour saps Tory support from business.   • Women could hold key to victory in UK election.   • UK poll battle heats up as ex-minister stands down.   • Suspected IRA blasts damage UK rail links.   • Rwandan strongman wishes Kabila success in Zaire.

Data Collection

A system was built for this study that presented the news articles to the participants, administered the topics and questions, and recorded the time spent on each article. All judgments and answers were submitted by the participants through the system. The time spent was determined by recording the time when participants loaded a new document and also when they submitted their judgment or answer. The order of documents in each set was randomly generated for each participant.

The summary questionnaire was administered through an oral interview lasting approximately ten to fifteen minutes after the session. A researcher asked participants questions from the questionnaire and recorded the participants' answers. The oral nature of the questionnaire allowed the researcher to explore further statements made by the participants.

Results

Quantitative Results

H_1: Users will spend more time reading relevant documents than non-relevant documents for relevance judgment tasks.

Using a between subjects analysis, there was no significant difference between time spent on relevant and non-relevant articles in either trial of the relevance

judgment task. The results for this condition are shown in Table 5.

Table 5. Results of relevance judgment task

	Trial 1 time(sec)	Trial 2 time(sec)
mean time of reading for relevant docs	34.37	33.75
mean time of reading for non-relevant docs	27.09	23.83
sd of relevant docs	19.87	23.14
sd of non-relevant docs	20.21	16.49
Calculated t-value	1.03	1.28

H_2: Users will spend more time reading documents that contain the answer for simple question answering tasks then those that do not contain the answer.

The results for the simple answer task are shown in Table 6. The first trial showed no significant difference but the second one did ($t=4.33$, $p < 0.005$, $df = 48$).

Table 6. Results of simple question answering task

	Trial 1 time(sec)	Trial 2 time(sec)
mean time of reading for relevant docs	36.94	61.14
mean time of reading for non-relevant docs	26.57	23.96
sd of relevant docs	19.86	26.07
sd of non-relevant docs	23.50	15.04
Calculated t-value	1.42	4.33

H_3: Users will spend more time reading documents that contain the answer for complex question answering tasks then those that do not contain the answer.

For the complex answering task both trials were significant ($t=3.28$, $p < 0.005$, $df=48$, $t=2.56$, $p < 0.005$, $df=48$, respectfully). The results for this condition are shown in table 7.

The non-significant results of the first condition, the relevance judgment task, indicate that when users read an article to judge how relevant it is, they do not spend more time on the article that they ultimately judged as relevant. This finding replicates those of Kelly and Belkin (2001).

Perhaps the user has to read all the way through an article before judging its relevance. Thus in the worst case every article would have to be completely read and there is no

significant difference in reading time between the two types of articles.

Table 7. Results of complex question answering task

	Trial 1 time(sec)	Trial 2 time(sec)
mean time of reading for relevant docs	57.15	72.10
mean time of reading for non-relevant docs	30.23	41.70
sd of relevant docs	23.67	34.57
sd of non-relevant docs	21.10	29.11
Calculated t-value	3.28	2.56

The results of the simple and complex question answering task when taken together suggest that when users have to find a specific piece of information, as opposed to judging relevance, they do in fact spend more time on the article that has that information. This effect seems to be more powerful as the complexity of the task increases (the simple question answering task had one trial reach significance but the complex question answering task had both trials reach significance).

These results indicate that implicit collection of time spent on articles is useful as a measure of user interest as the task becomes more complex, such as when a user wants to find a specific answer, and not as useful when the user is merely reading the article to judge general relevance.

Qualitative Results

Qualitative results were collected through a summary questionnaire. We examined participant perceptions regarding how much time they thought they spent reading news articles and then using within subjects results, we compared the results. Potential privacy concerns were explored as well. Participants were also asked for which types of information personalized recommendations would be useful.

For the first condition, the relevance judgment task, participants were asked if they thought they spent more time reading: relevant document, non-relevant documents, or spent equal time reading all documents. Eight of the ten participants stated that they thought they spent more time reading relevant documents than non-relevant documents. Of these eight participants, only four actually spent a significant amount of time reading relevant documents compared with the non-relevant documents in either trial.

Participants were asked for the second condition, the simple question answering task, if they thought they spent more time reading: documents that contained the answer, documents that didn't contain the answer, or spent equal

time reading all documents. Seven of the ten participants stated that they thought they spent more time reading documents that contained the answer. Of these seven participants, four actually spent a significant amount of time reading documents that contained the answer when compared with the documents that did not contain the answer.

For the third condition, the complex question answering task, users were asked whether they though they spent more time reading: documents that contained the answer, documents that didn't contain the answer, or spent equal time reading all the documents. Eight of the ten participants stated that they thought they had spent more time reading the documents that contained the answer. Of these eight participants, six actually spent a significant amount of time reading documents that contained the answer when compared with the documents that did not contain the answer.

It is interesting that the users' perception and their performance differ for the simple tasks but are more in line as the task became more complex.

User privacy and comfort is an important consideration when monitoring user behavior. In order for a system to track implicit feedback, users have to be willing to have certain behaviors monitored. In our study, 8 out of the 10 users said they would be comfortable with a system that monitored behaviors *if* it was in order to make personalized recommendations. Many participants stated they would want the option to turn the monitoring on or off depending on what they were working on. Participants were also worried that their information would be misused for other purposes, such as marketing.

Participants were also asked for types of information they would find personalized recommendations useful. Table 8 summarizes these findings. Participants were unanimous in that they would find recommendations of news articles useful. Nine of the 10 participants found recommendations of scholarly articles while doing research would be useful. Eight of the 10 participants would find recommendations useful for web browsing while seven out of 10 stated they would find recommendations useful while online shopping. One concern participants had with online shopping was that the system would only recommend items similar to those they had already purchased. One participant stated that he would not like recommendations while shopping online because "I like to find things I wasn't looking for, otherwise you won't find new things". Another was concerned that his varied shopping history would result in inaccurate recommendations: "my history is very varied so there would be no way to make good recommendations. Would just get in the way". Only three of the ten participants thought that it would be useful to recommend documents while web browsing. Participants worried that it would be too overbearing and would hinder their web browsing experience.

Table 8. Summary of participant responses

Would you find recommendations useful for these types of information?	Participants who responded yes
News	10 out of 10
Search Result Ranking	8 out of 10
Web Browsing	3 out of 10
Scholarly articles	9 out of 10
Online Shopping	7 out of 10

Future Work

This study indicates that the usefulness of using time spent as a measure of user interest is related to task and becomes more useful as the task becomes more complex. The *post hoc* analysis of users' perceptions revealed that users assessed how they spent their time more accurately as the task became more complex. Further investigation to replicate this finding and to establish when user perceptions are valid will be done.

Based on the results of this work, further studies are necessary to determine if there are more nuances within the relevance judgment tasks or if this finding holds in general. The development and evaluation of a prototype recommender system will be the focus of future research. This system will use implicit feedback in two ways. First, it will incorporate implicit feedback in order to learn user preferences, eliminating the overhead imposed by explicit feedback. Second, we will use user behavior to identify the current task, or context, in which a user is engaged. Knowledge of the type of task a user is engaged in can be used to make context-sensitive recommendations. It is our hope that this work will result in an IF system that uses implicit feedback to make better recommendations and improve satisfaction without increasing the overhead on the part of the user.

ACKNOWLEDGMENTS

REFERENCES

Claypool, M., Le, P., Waseda, M., & Brown, D. (2001). Implicit Interest Indicators. In *Proceedings of ACM Intelligent User Interfaces Conference (IUI '01)*, Santa Fe, NM. 33-40.

Goecks, J., & Shavlik, J. (2000). Learning users' interests by unobtrusively observing their normal behavior. In *Proceedings of International Conference on Intelligent User Interfaces (IUI'00)*, New Orleans, LA. 129-132.

Golovchinsky, G., Price, M. N., & Schilit, B. N. (1999). From Reading to Retrieval: Freeform Ink Annotations as Queries. In *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR '99)*, Berkeley, CA. 19-25.

Hijikata, Y. (2004). Implicit User Profiling for On Demand Relevance Feedback. In *Proceedings of ACM Intelligent User Interfaces Conference (IUI '04)*, Madeira, Portugal. 198-205.Kelly, D., & Belkin, N. J. (2001). Reading Time, Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevance Feedback. In *Proceedings of ACM Conference on Research and Development in Information Retrieval (SIGIR '01)*, New Orleans, LA 408-409.

Kelly, D. & Teevan, J. (2003). Implicit Feedback for Inferring User Preference: A bibliography. *SIGIR Forum*, 37(2), 18-28.

Kim, K., Carroll, J. M., & Rosson, M. (2002). An Empirical Study of Web Personalization Assistants: Supporting End-Users in Web Information Systems. In *Proceedings of the IEEE 2002 Symposia on Human Centric Computing Languages and Environments*. Arlington, VA. 60-62.

Kim, J., Oard, D. W., & Romanik, K. (2001). User Modeling for Information Access Based on Implicit Feedback. In *Proceedings of ISKO-France 2001*, Nanterre, France.

Konstan, J., Miller, B., Maltz, M., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3), 77-87.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2003). Capturing Interest through Inference and Visualization: Ontological User Profiling in Recommender Systems. In *Proceedings of the Second Annual Conference on Knowledge Capture 2003*. 62-69.Morita, M., & Shinoda, Y. (1994). Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In Proceedings of ACM *Conference on Research and Development in Information Retrieval (SIGIR '94)*, Dublin, Ireland. 272-281.

Nichols, D. (1997). Implicit Rating and Filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary. 31-36.

Oard, D. W., & Kim, J. (2001). Modeling Information Content Using Observable Behavior. In *Proceedings of ASIST 2001 Annual Meeting*, Washington, DC.

Seo, Y.-W., & Zhang, B.-T. (2000). Learning User's Preferences by Analyzing Web-Browsing Behaviors. In *Proceedings of International Conference on Autonomous Agents 2000*, Barcelona, Spain. 381-387.

White, R. W., Ruthven, I., & Jose, J. M. (2002). The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Proceedings of 24th BCS-IRSG European Colloquium on IR*.