2001

User Modeling for Information Filtering Based on Implicit Feedback

Jinmook Kim
University of South Carolina - Columbia, jinmook@mailbox.sc.edu

Douglas W. Oard

Kathleen Romanik

Publication Info
Published in 2001, pages 25-37.

User Modeling for Information Filtering Based on Implicit Feedback

Jinmook Kim,[*] Douglas W. Oard,[*] and Kathleen Romanik[+]

[*]College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742-4345
{jinmook, oard}@glue.umd.edu
phone: 1-301-405-2033
fax: 1-301-314-9145

[+]LogicTree Corp.
8400 Baltimore Blvd. Suite 301
College Park, MD 20740
kathleen@logictree.com

Abstract

This study reports the results of a pair of user studies that can provide a practical basis for designing an information filtering system that employs implicit feedback for user modeling. In information filtering systems, user modeling can be used to improve the representation of a user's information needs. User models can be constructed by hand, or learned automatically based on feedback provided by the user about the relevance of documents that they have examined. By observing user behavior, it is possible to infer implicit feedback without requiring explicit relevance judgments. Previous studies based on Internet discussion groups (USENET news) have shown reading time to be a useful source of implicit feedback for predicting a user's preferences. Our study extends that work by examining whether reading time is useful for predicting a user's preferences for academic and professional journal articles and by exploring whether printing behavior can usefully augment the information that reading time provides. Two user studies were conducted in which undergraduate students examined articles and abstracts related to the telecommunications and pharmaceutical industries. The new results showed that reading time could still be used to predict the user's assessment of relevance, although reading time for journal articles and technical abstracts are longer than has been reported for USENET news documents. Observation of printing events was found to provide additional useful evidence about relevance beyond that which could be inferred from reading time. The paper concludes with some observations on the limitations of our study, future work that is needed, and the larger implications of work on implicit feedback.

1. Introduction

No one person can read all the information available on the Internet. One must choose a small portion of information from among the panoply of sources available to them. It is the classic needle in a haystack problem. But the haystacks are growing so rapidly that there is continued demand for improved search technology. Information filtering and retrieval systems could provide them with a support for finding information they need. Information retrieval is a *pull* service that users can search for information they need, whereas information filtering is a *push* service that the system finds new information and presents it to the user (Kim et al, 2000). Content-based filtering systems select documents based on the characteristics of the documents such as the words they contain (Sheth, 1994; Oard, 1997). An alternative, now commonly referred to as recommender systems, is to base the search at least in part on annotations made to the documents by other users (CACM, 1997).

A user model that represents some aspect of a user's information needs and/or preferences can be useful in any information system design, and in the case of information filtering it is clearly a central component. User models can be hand-crafted, but machine learning techniques offer the potential to automatically develop or continuously refine a user model. The usual approach in research systems has been to assemble a set of training instances that have been labeled by the user as relevant (either absolutely, or to some degree) or as not relevant. Studies have shown that such explicit feedback from the user is clearly useful (Yan & Garcia-Molina, 1995; Goldberg et al., 1992), but obtaining explicit feedback would likely be problematic in many applications. It is well known that users of commercial information filtering systems make little use of explicit relevance feedback mechanisms when they are provided, at least in part because providing feedback takes time and may increase the cognitive load on the user. Implicit feedback,

in which the system learns by observing the user's behavior, offers an attractive alternative that has received increased attention in recent years (Stevens, 1993; Morita & Shinoda, 1994; Konstan et al., 1997; Nichols, 1997; Oard & Kim, 1998; Kim et al, 2000).

This study seeks to provide a practical basis for user modeling for information filtering based on implicit feedback. In the next section, we briefly describe prior work on the use of implicit feedback in information filtering systems. We then present the results of a pair of user studies that explore how two such sources, observations of reading time and observations of printing behavior, might be used jointly to build a better user model than could be built using either source alone. The paper concludes with some observations on the limitations of our study, future work that is needed, and the larger implications of work on implicit feedback.

2. Implicit Feedback

Implicit feedback may bear only an indirect relationship to the user's assessment of the usefulness of any individual document. But because it can be collected ubiquitously (and thus potentially in great quantities), the potential impact of implicit feedback might ultimately be even greater than that of explicit feedback. InfoScope, a system for filtering Internet discussion groups (USENET), utilized both implicit and explicit feedback for modeling users (Stevens, 1993). Three sources of implicit evidence were used: whether a message was read or ignored, whether it was saved or deleted, and whether or not a follow up message was posted. In summarizing this groundbreaking study, Stevens observed that implicit feedback was effective for tracking long-term interests because it operates constantly without being intrusive.

Morita and Shinoda (1994) introduced another source, proposing an information filtering technique based on observations of reading time. They conducted user study over a six-week period with eight users to determine whether preference for Internet discussion group USENET messages was reflected in the time spent reading those messages. The results showed a strong positive correlation between reading time and explicit feedback provided by those users. They also discovered that treating messages that the user read for more than 20 seconds as relevant actually produced better recall and precision in an information filtering simulation than using the messages explicitly rated by the user as relevant would. Konstan et al. (1997) repeated this study in a more natural setting, distributing modified software that allowed volunteers to participate in a recommender system trial in which both explicit feedback and reading time were recorded for a small set of USENET discussion groups. Their results indicated that recommendations based on reading time can be nearly as accurate as recommendations based on explicit feedback. They also suggested some additional observable behaviors, including printing, forwarding, and replying privately to a message, as sources for implicit ratings.

Nichols (1997) began the effort to develop a comprehensive view of implicit feedback, with a focus on its use in information filtering systems. He presented a list of potentially observable behaviors; adding purchase, assess, repeated use, refer, mark, glimpse, associate, and query to those mentioned above. Oard & Kim (1998) extended that work, organizing the behaviors into three broad categories (examination, retention, and reference). They also presented examples from related fields, for example, using Web link analysis (Brin & Page, 1998) and indexing based on bibliographic citations (Garfield, 1979) to illustrate the potential of implicit feedback based on reference behavior.

3. Experiment Design

Previous studies on implicit feedback have found that predictions based on reading time can be about as accurate for USENET as those based on explicit ratings (Morita & Shinoda, 1994; Konstan et al., 1997). Evidences from practice clearly indicates that some types of reference behavior are valuable as well (Brin & Page, 1998; Garfield, 1979). We know less, however, about the utility of many other types of observable user behaviors, such as print, save, or purchase. We thus chose to focus on printing behavior, because our intuition suggested that users might spend less time reading a document in cases in which they decided to print it for later use. The system that we used was designed to provide access to scientific and professional journal articles (both full text and abstracts), which might exhibit different characteristics than USENET news. So, we were also interested how reading time and explicit ratings were related in this case.

3.1 Hypotheses

Our hypotheses to be tested were:

a. On average, users spend more time reading relevant full-text journal articles than non-relevant articles.
b. On average, users spend more time reading abstracts of relevant journal articles than abstracts of non-relevant articles.
c. The combination of reading time and printing behavior would be more useful for predicting explicit ratings than using reading time alone.

3.2 Experimental System

Powerize Server,™ developed by Powerize.com, is a Windows NT text retrieval and filtering system that searches multiple internal and external information sources simultaneously and presents the retrieved documents to the user in a customized manner that can be viewed with a Web browser. It presently uses a manually constructed user model known as a search profile. Once a user sets up a search profile, she can choose to save the profile and have it re-executed on a regular schedule. Our experiments were done using the Powerize Server 1.0. A custom version of Powerize Server 1.0 was created for our experiments by Powerize.com. It was instrumented to measure reading time and printing behavior and to record user-entered ratings for individual documents.

On the Powerize Server 1.0, users interact with the system through two principal interfaces: Publications and Studio. The Studio interface allows users to select and manage profiles based on their interests and includes five collections of profiles known as "wizard packs:" General, Pharmaceutical, Aerospace, Telecommunications, and Energy. Each wizard pack is designed to serve the needs of a group of users. For example, the Pharmaceutical wizard pack is intended for users in the pharmaceutical industry. The Pharmaceutical and Telecommunications wizard packs were used in our experiments. Each wizard pack consists of several "wizards," and each wizard is designed to help the user complete a particular task. For example, there is a competitive intelligence wizard to help users find information about a competitor. Each wizard is further divided into "topics," which are collections of profile templates designed to retrieve information about a particular subject. For example the competitive intelligence wizard contains topics such as " Mergers and Acquisitions" and "Financial Information." Each profile template encodes the structure of a query for a set of information sources. Users create actual profiles by selecting templates and providing search terms such as a drug or company name. By using templates, users can create queries without being familiar with the individual information sources or their query interfaces. Once, users construct their profiles through the Studio interface, they can browse documents retrieved by the system using the Publications interface.

3.3 Pilot Study

A pilot study was conducted to validate the experimental procedures. Special consideration was given to data collection procedures in order to determine whether the system could collect and process the required information. The pilot study was done using only "Pharmaceutical Wizards," with 4 students who were taking a microbiology course on Drug Action and Design at the University of Maryland. A total of 21 instances of reading time and rating were gathered, which showed the expected pattern of increasing reading time with increasing rating. The data collected from the pilot study also suggested that printing behavior might prove useful. Every one of the 9 cases in which printing was requested was rated as relevant, and any obvious way of using reading time alone to make predictions would have missed some of those cases.

4. Data Collection

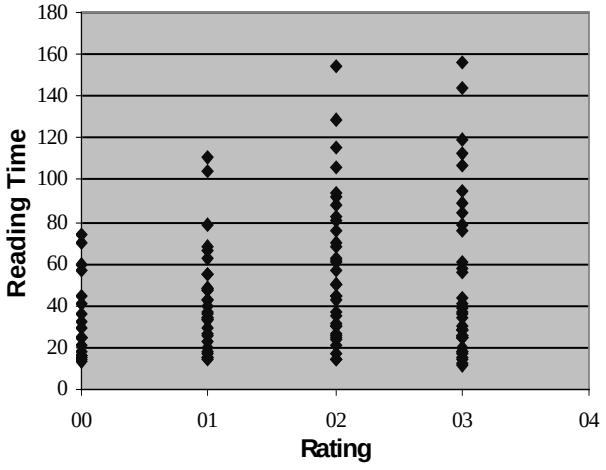
Two experiments were conducted with two different user groups: The Telecommunications user group and the Pharmaceutical user group. Eight undergraduate students taking an honors research seminar at the University of Maryland participated in the first experiment. The students were engaged in research for a group project that required examining new products, services, and technologies for wireless Personal Communications Systems (PCS). After conversations with both the students and their instructor to define their information needs, search topics were created by the authors using the "Telecommunications

Wizards." A total of 97 full-text articles were retrieved using 5 topics: digital PCS, Iridium, Teledesic, Nextel i1000, and Ricochet. All of the selected information sources were from Dialog,™ a provider of professional journal content. The experiment with the Telecommunications user group took place in a single one-hour session. A total of 130 ratings (explicit relevance judgments), with associated reading time and printing behavior observations, were collected. Explicit ratings were collected on a four point scale: "00"for no interest, "01" for low interest, "02" for moderate interest, and "03" for high interest. A rating of "NA" for no comment was also allowed.

The second experiment was done with 85 senior or advanced junior students attending laboratory sessions for a zoology course on Mammalian Physiology at the University of Maryland. Search topics were created by the authors using the "Pharmaceutical Wizards" after interviewing the instructor. A total of 96 articles were returned using 5 topics: beta-blockers, antihypertensives, ACE inhibitors, positive intropic agents, and cardiac sympathomimetics. Again, all of the selected information sources were from Dialog.™ This experiment was conducted in seven sessions during a single week. Sessions 1 and 2 were administered following the same procedure that the Telecommunications user group used. There were 18 subjects in each session, and we discovered that with that many simultaneous users our server's hardware configuration was unacceptably slow, resulting in what we assessed to be unreliable measurements of reading time. To minimize the impact of this problem, students were paired in groups of two for sessions 3 through 7. One student in each group was assigned to examine the documents, while the other observed the session. In this way, all of the students in each lab period were able to participate in some way, but our measurements would (hopefully) still reflect the reactions of a single student. To minimize the potential effect on reading time caused by having two subjects on a machine, students were asked not to talk to each other during the experiment. A total of 698 ratings were collected during the seven sessions.

5. Data Analysis

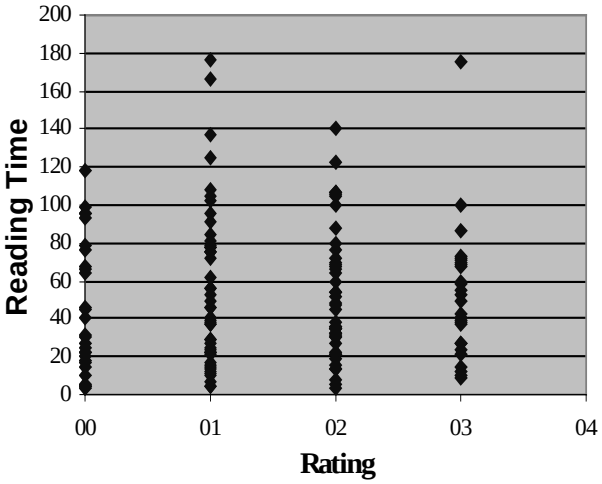
A total of 122 cases out of 130 ratings collected from the eight subjects in the first experiment were considered valid for purposes of data analysis. All five cases collected from one subject were excluded from the data analysis because that student missed the first half of the experiment. Two other cases (exceeding a Z score of ±3) were excluded because they were detected as outliers based on the standardized residual scores for reading time. One case was excluded because it had a rating of "no comments." Figure 1 shows the descriptive data analysis for the Telecommunications user group. An increase in reading time, in general, can be observed as the value of the rating gets higher on the scatterplot. The rating of "00," indicating "no interest," had the lowest mean reading time, and "02," representing "moderate interests," had the highest mean reading time. It seemed that subjects were able to identify highly relevant articles more quickly than those that they rated moderately relevant.



Rating	Telecommunications Wizards	
	# of Cases	Mean Reading Time (sec.)
00	20	32.85
01	31	42.84
02	34	57.74
03	37	50.24
Total	122	47.60

Figure 1. Descriptive data analysis for the Telecommunications user group.

In the second experiment, there were 7 sessions. In sessions 1 and 2, 36 subjects provided 166 ratings, but data from those two sessions were not used in this study because of the slow system response time described in the previous section. A total of 532 ratings were gathered from 49 subjects that participated in sessions 3, 4, 5, 6, and 7. A total of 153 cases out of the 532 ratings gathered were considered as valid for data analysis in this study, in part because it was discovered after the experiments that only 25 of the 96 articles that had been automatically assembled for presentation to the subjects had abstracts (none had full-text). The 363 ratings that were given for the 71 bibliographic citations that lacked abstracts were excluded from the data analysis because we did not feel that the bibliographic citations alone could provide an adequate basis for assessment by the users. Three cases that were detected as outliers and 13 cases with "no comments" were also excluded from the data analysis. The scatterplot in Figure 2 presents the distribution of 153 valid cases, and the associated table shows both the number of cases and the mean reading time for each rating.



Rating	Pharmaceutical Wizards	
	# of Cases	Mean Reading Time (sec.)
00	29	42.97
01	51	55.14
02	45	52.00
03	28	51.50
Total	153	51.25

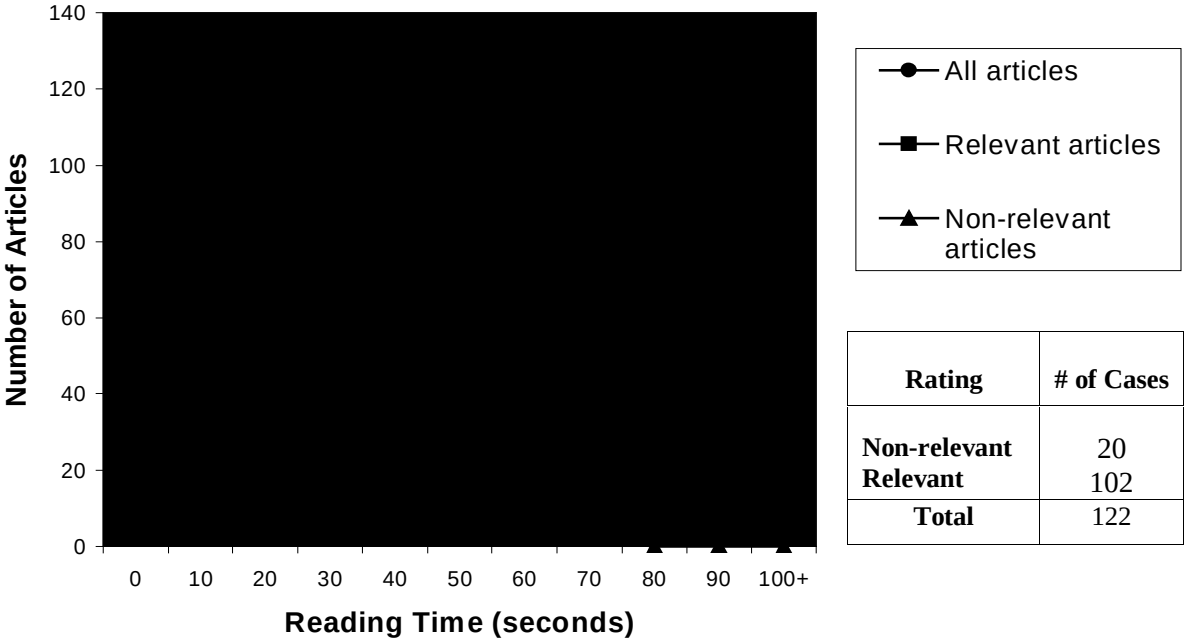
Figure 2. Descriptive data analysis for the Pharmaceutical user group.

5.1 Reading Time as a Source for Implicit Feedback

In both experiments, we noted a decline in mean reading time between articles rated as moderate interest and those rated as high interest. In fact, a consistent decline in reading time in the second experiment was evident as interest increased. This suggests that we will likely not be able to reliably distinguish between degrees of interest using reading time, so we converted the ratings to a binary scale: "00" to "non-relevant" and "01, 02 and 03" to "relevant" for our subsequent analysis in both experiments. Figure 3 presents the descriptive data analysis on reading time with this binary rating scale for data collected from the Telecommunications user group. An increase in mean reading time was observed from non-relevant to relevant documents on the graph. Ratings made on non-relevant documents and on relevant documents were approximately normally distributed below and above the mean reading times of 32.85 and 50.49 seconds, respectively.

An independent-samples t-test, comparing the mean reading time on relevant documents with non-relevant ones, was done to test our first hypothesis. A statistically significant difference between the two mean reading times was found at $\alpha = 0.05$. We therefore conclude that users tend to spend a longer time reading relevant articles than non-relevant articles, which is a consistent result with the two previous studies by Morita and Shinoda (1994) and by Konstan et al. (1997). Morita and Shinoda, in their study in 1994, concluded that preference of a user for an article was the dominating factor that affected time spent reading it, and they suggested using a threshold on reading time to detect relevant articles. Their results showed that 30 % of interesting articles could be retrieved with precision of 70 % by using a threshold of 20 seconds. A much higher threshold would be required in our first experiment to reach a similar recall level. This comports with our intuition, since Morita and Shinoda used USENET messages, while our first experiment was conducted with full-text professional journal articles. Several factors, such as the length of the article, levels of difficulty for understanding the contents, and differences in language skills, could

affect the reading time. Subjects in our study might also require longer reading time to understand the content of an article because none of them were experts in the field. Figure 4 shows the recall and precision for different ranges of reading time. For example, the recall and precision that would result from treating articles with reading time of at least 40 seconds as relevant were 0.418 and 0.894, respectively. The horizontal line at a precision of 0.836 shows the value that would be achieved if the user selected articles randomly, since 102 of the 122 articles were judged as relevant.



Rating	# of Cases
Non-relevant	20
Relevant	102
Total	122

Figure 3. Number of articles read for at least the given duration (Telecommunications user group).

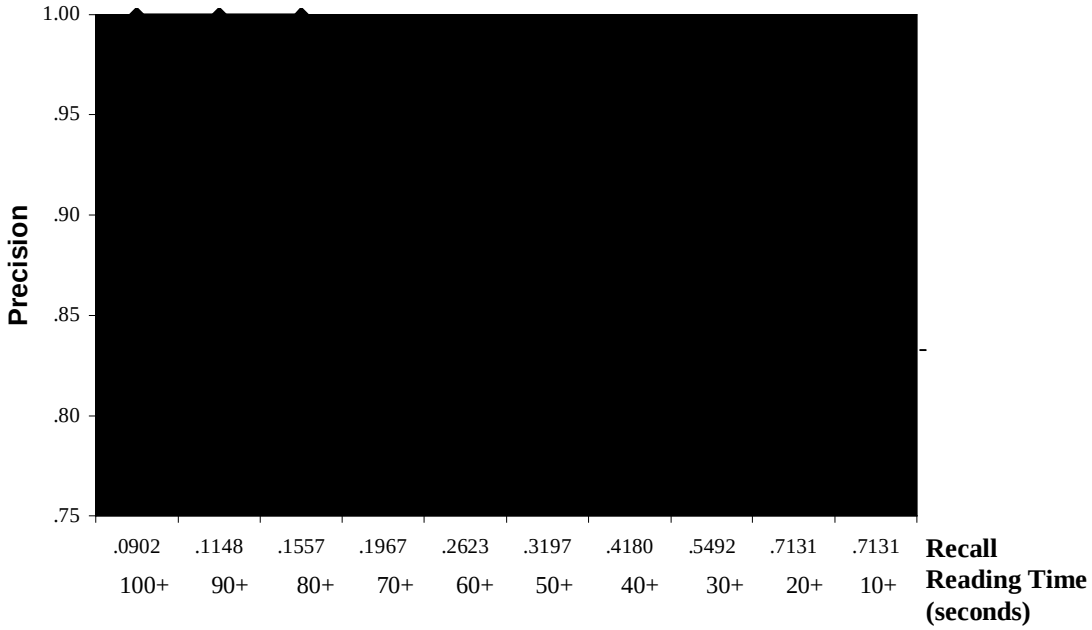
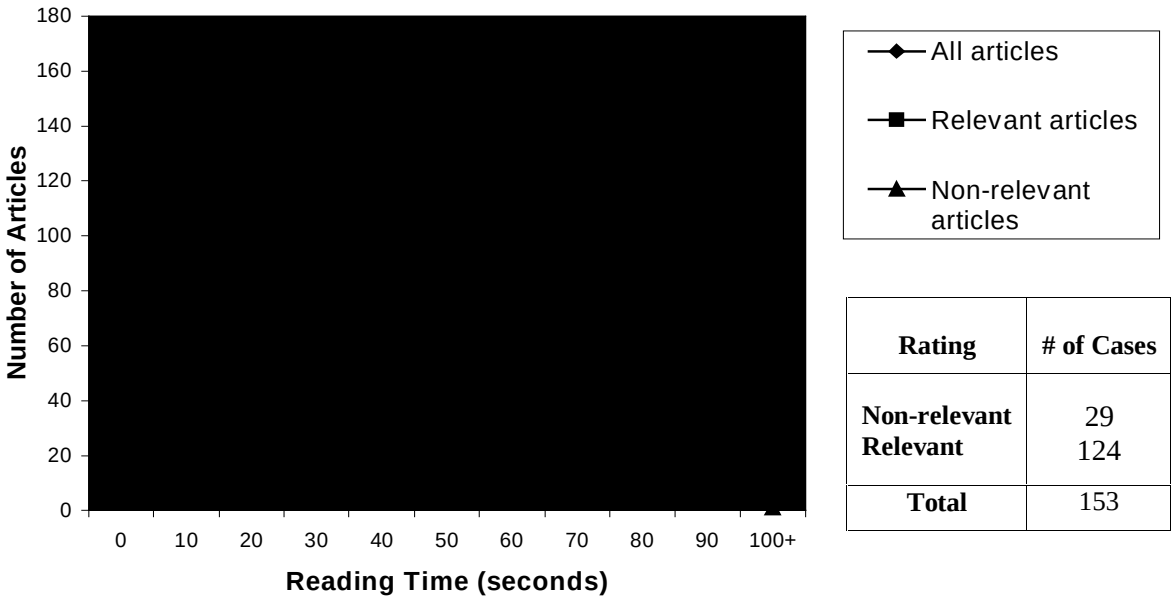


Figure 4. Precision vs. reading time (Telecommunications user group).



Rating	# of Cases
Non-relevant	29
Relevant	124
Total	153

Figure 5. Number of articles read for at least the given duration (Pharmaceutical user group).

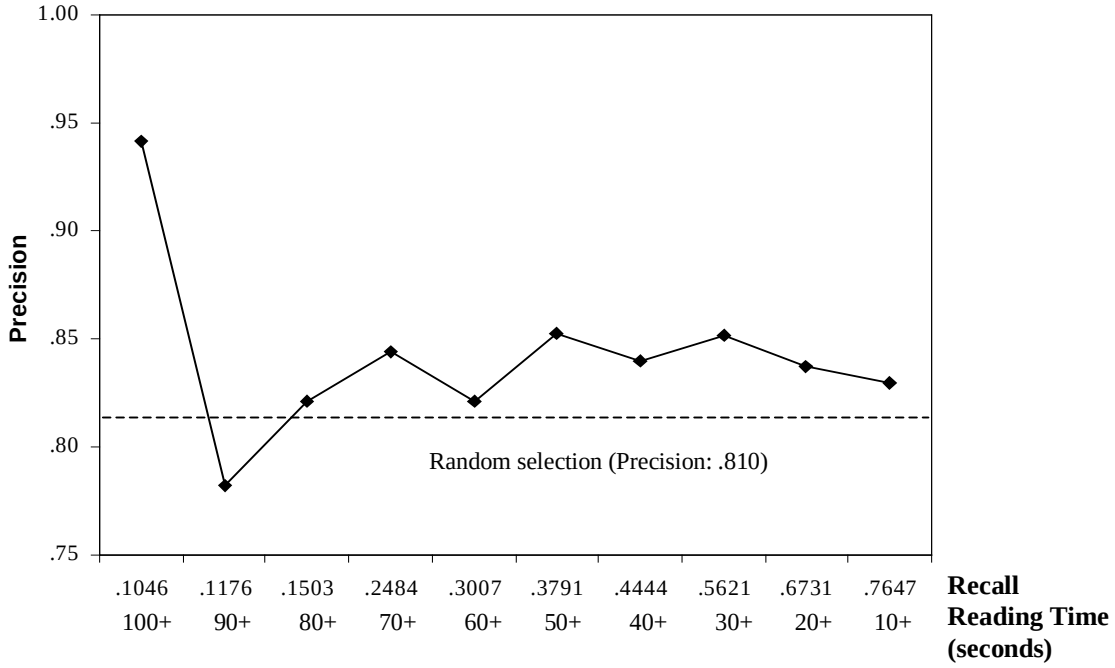


Figure 6. Precision vs. reading time (Pharmaceutical user group).

Figure 5 shows the descriptive data analysis for our experiment with the Pharmaceutical user group. There was a 10.22 second difference between the mean reading times on relevant and non-relevant documents, but no statistical significance was found at $\alpha = 0.05$, based on the independent-samples t-test. The mean reading time on relevant documents was 53.19 seconds, which was close to the one (50.49 seconds) for the Telecommunications user group in our first experiment. The mean reading time on non-relevant documents, however, was 42.97 seconds, which was 10.12 seconds more than was observed with the Telecommunications user group. We suspect that this unexpected outcome resulted at least in part from the different setting in which we paired two students together. As we mentioned in Section 4, one student

in each group was observing the session, while the other was browsing retrieved articles. In this case, the student doing the browsing might have sometimes chosen to wait until the other student had also examined the article before clicking on the feedback button. Figure 6 presents the observed recall and precision for different ranges of reading time. Only for extremely long times (over 100 seconds) does reading time provide any clear improvement over random selection (shown by the horizontal line at a precision of 0.810).

5.2 Printing Behavior as Evidence of Interest

Printing behavior was examined in this study with the hope that it may provide us with clues that can predict explicit ratings beyond those clues given by reading time. There were a number of relevant documents that could not be discriminated from non-relevant ones using only reading time in Figures 3 and 5. For example, using 47.60 and 51.25 seconds as thresholds for cutting off non-relevant documents in Figures 3 and 5 will also throw 61 out of 102 (59.80 %) and 68 out of 124 (54.84 %) relevant documents away, respectively. Can printing behavior provide a clue for detecting those relevant documents that would have been thrown away using reading time alone?

	Telecommunications User Group		**Pharmaceutical User Group**	
	Reading Time	**Rating**	**Reading Time**	**Rating**
	156	03	100	03
	81	02	58	03
			53	03
			43	03
			38	03
			12	03
			100	02
			67	02
			66	02
			48	02
			36	02
			35	02
			32	02
			8	02
			17	01
			11	01
Mean	118.50		45.25	

Table 1. Reading time and ratings for printed articles.

Unfortunately, only two cases of printing behavior were available from the data collected from the experiment with the Telecommunications user group, as shown in Table 1. No meaningful interpretation on the data collected could be made with only two cases. We believe that the low frequency of the printing behavior may have resulted from a disparity of goals among the subjects. The members of that undergraduate research team had previously assigned responsibility for technology research to a few of the team members. As a result, the other members of the team may have treated this session more as a familiarization opportunity than as a directed search for information.

There were 16 cases of printing behavior, which accounted for more than 10% of the 153 cases for the Pharmaceutical user group. As Table 1 shows, users expressed some degree of interest in every printed document. Eleven of the 16 cases had reading times less than the mean for relevant documents (53.19

seconds), suggesting that users were sometimes able to rapidly identify and print interesting documents. Printing behavior thus provides a useful cue for predicting explicit ratings because it makes it possible to reliably detect relevant documents in some cases. Reading time, by contrast, provides noisier estimates of relevance, but balances that limitation by providing a larger number of observations. Our experiment suggests that the combination of reading time and printing behavior can provide more useful evidence for predicting explicit ratings than could be obtained from either source alone.

6. Conclusion

We have shown that reading time can be a useful source of implicit feedback for systems that search academic and professional journal articles in full text, but we were not able to demonstrate a similar effect for abstracts of similar materials. When retention behavior (printing, in this case) was observed, it was found to contribute complementary information, suggesting that systems which couple both types of observations may be able to better model a user's information seeking behavior than those that rely on reading time alone. Oard and Kim (1998) suggested additional behaviors that might be observed, which could help system designers recognize useful sources of implicit feedback that would be practical to obtain in their application. Implicit feedback could be useful in a broad array of information access applications, including filtering or retrieval using content-based and/or annotation-based techniques. Annotation-based techniques stand to benefit in two ways – by using implicit feedback to develop better user models and by sharing with other users the annotations derived from implicit feedback. Annotation-based techniques that can exploit large sets of simple (and noisy) observations could see the greatest impact, perhaps significantly accelerating the deployment of large-scale recommender systems.

Several important research issues remain, however, if we are to fully capitalize on the potential of implicit feedback to support information filtering. Our approach leverages prior work on information filtering using explicit feedback by predicting the feedback that a user would have provided. It remains to be seen whether greater effectiveness could be achieved using more closely coupled techniques. The development of explainable systems is another topic that merits increased effort. Ultimately, the systems we build will be tools in the hands of their users. If we provide users with tools they understand, that may use them to accomplish things that the tools' developers never envisioned. If we are to exploit this potential, we will need to give serious thought to how users will understand what their systems are doing for them so that they can make most of their potential for intentional action. Our work also suggests specific technical questions that now need to be addressed. Perhaps the most urgent is the question of how to accommodate the uncertainty inherent in implicit feedback. We have, for example, shown the precision improvement that can be achieved at various reading time thresholds, but it is not clear that applying a sharp threshold would be the best approach. And if a threshold does turn out to be about as good as any more nuanced strategy, some guidance on how to select that threshold for particular applications will be needed.

Finally, it is important to realize that our work was conducted in a controlled environment. There is now considerable evidence from practice that implicit feedback from situated users is of value, particularly for examination and reference behavior (e.g., doubleclick.com and Google, respectively). The experiment reported in this paper is a first step towards gaining similar experience with printing behavior as well, but evidence based on observations of situated users will be needed before we can fully understand the potential impact of any combination of techniques in a specific application.

Acknowledgements

The authors wish to thank Nick Carmello of Powerize.com for modifying Powerize Server 1.0, Professors William Higgins and Carol Pontzer at the University of Maryland for working closely with us to find subjects for our experiments and to craft meaningful tasks for them to perform, and our volunteer participants, without whom our research would not have been possible. This work has been supported in part by the Maryland Industrial Partnerships program and Powerize.com.

References

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. Dept. of Computer Science, Stanford Univ. http://dbpubs.stanford.edu:8090/pub/1998-8

CACM (1997) Special Issue on Recommender Systems, *Communications of the ACM*, 40(3), March.

Garfield, E. (1979*) Citation indexing: Its theory and application in science, technology, and humanities*. New York: Wiley-Interscience.

Goldberg, D., Nichols, D., Oki, B. M, and Terry, D. (1992) Using collaborative filtering to weave an information tapestry. *Communication of the ACM*, December, 35(12): 61-70.

Kim, J., Oard, D. W., and Romanik, K. (2000) Using implicit feedback for user modeling in Internet and Intranet searching. Technical Report, College of Library and Information Services, University of Maryland at College Park. http://www.clis.umd.edu/research/reports/

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. (1997) GroupLens: Applying collaborative filtering to Usenet News. *Communication of the ACM*, March, 40(3), 77-87. http://www.acm.org/pubs/articles/journals/cacm/1997-40-3/p77-konstan/p77-konstan.pdf

Morita, M and Shinoda, Y. (1994) Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 272-281.

Nichols, D. M. (1997) Implicit ratings and filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapaest, Hungary 10-12, ERCIM. http://www.ercim.org/publication/ws-proceedings/DELOS5/index.html

Oard, D. W. (1997) The state of the art in text filtering. *User Modeling and User-Adapted Interaction*, 7(3), 141-178.

Oard, D.W., and Kim, J. (1998) Implicit Feedback for Recommender System. In *AAAI Workshop on Recommender Systems*, Madison, WI: 81-83. http://www.glue.umd.edu/~oard/research.html

Sheth, B. D. (1994) "A learning approach to personalized information filtering." Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science. http://lcs.www.media.mit.edu/groups/agents/publications/

Stevens, C. (1993) *Knowledge-based assistance for accessing large, poorly structured information spaces*. Ph.D. thesis, University of Colorado, Department of Computer Science, Boulder. http://www.holodeck.com/curt/mypapers.html

Yan, T.W. and Garcia-Molina, H. (1995) SIFT – A tool for wide-area information dissemination. In *Proceedings of the 1995 USENIX Technical Conference*, pp.177-186. ftp://db.stanford.edu/pub/yan/1994/sift.ps.