Evaluating Implicit Measures to Improve Web Search

STEVE FOX, KULDEEP KARNAWAT, MARK MYDLAND,
SUSAN DUMAIS, and THOMAS WHITE
Microsoft Corp.

Of growing interest in the area of improving the search experience is the collection of implicit user behavior measures (implicit measures) as indications of user interest and user satisfaction. Rather than having to submit explicit user feedback, which can be costly in time and resources and alter the pattern of use within the search experience, some research has explored the collection of implicit measures as an efficient and useful alternative to collecting explicit measure of interest from users.

This research article describes a recent study with two main objectives. The first was to test whether there is an association between explicit ratings of user satisfaction and implicit measures of user interest. The second was to understand what implicit measures were most strongly associated with user satisfaction. The domain of interest was Web search. We developed an instrumented browser to collect a variety of measures of user activity and also to ask for explicit judgments of the relevance of individual pages visited and entire search sessions. The data was collected in a workplace setting to improve the generalizability of the results.

Results were analyzed using traditional methods (e.g., Bayesian modeling and decision trees) as well as a new usage behavior pattern analysis ("gene analysis"). We found that there was an association between implicit measures of user activity and the user's explicit satisfaction ratings. The best models for individual pages combined clickthrough, time spent on the search result page, and how a user exited a result or ended a search session (exit type/end action). Behavioral patterns (through the gene analysis) can also be used to predict user satisfaction for search sessions.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback, search process*

General Terms: Experimentation, measurement

Additional Key Words and Phrases: Implicit measures, search sessions, explicit feedback, explicit ratings, user interest, user satisfaction, prediction model

1. INTRODUCTION

In many real-world information retrieval or filtering applications, it is difficult to obtain explicit feedback from users about the relevance of the results, the

Authors' address: Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-6399; email: {stevef,kuldeep,markmy,sdumais,tomwh}@microsoft.com.

appropriateness of the presentation, and more generally about the quality of their experience. Yet explicit judgments are assumed by researchers for many activities like the tuning and selection of ranking algorithms, information combination, user modeling, information presentation, etc. The focus of our research is to explore how implicit measures of user interest (such as time spent on a page, clickthrough, and user activities like annotation, printing, and purchasing) can be used to develop predictive models for a variety of purposes.

As search becomes more widely used for a broad range of information retrieval tasks (e.g., search for friends, information, help, and shopping), understanding whether the user was satisfied with that information is becoming evermore problematic. Consider a Web search service in which hundreds of millions of queries are issued every day. How does the service know what users want? How does it know when it has returned good results? How does it know when its users are satisfied? One way is to explicitly ask the user. This is often done in Cranfield-style evaluations of information retrieval systems, and has been quite useful in developing and tuning information retrieval algorithms. But this type of data collection is expensive, limited in coverage, and subject to selection biases since users decide whether to participate or not. Explicit feedback can be augmented by other approaches that try to understand the user's needs by collecting and analyzing implicit measures. In short, there may be answers in the way in which people interact with applications; stories if you will that can help application developers improve the user's experience.

Nichols [1997] evaluated the costs and benefits of using implicit measures over explicit ratings. In this study, he asked the question of whether implicit user feedback can substitute for explicit ratings with the end goal of avoiding the difficulties associated with gathering explicit ratings from users" (p. 2). Part of the benefit, Nichols argued, is the fact that collecting implicit ratings "removes the cost to the evaluator of examining and rating the item. Whilst there remains a computational cost in storing and processing the implicit rating data this can be hidden from the system" (p. 2). Nichols suggested that implicit ratings can be combined with existing rating systems to form a hybrid system, that is, using "implicit data as a check on explicit ratings" (p. 5), thus having good potential for being able to predict user satisfaction. Thus, one might argue that implicit measures can provide us with a rich stream of data that cannot only be used to improve the users' experience, but can do so without interrupting their normal workflow.

Our research seeks to develop predictive models of user satisfaction with search results based on implicit measures. We first provide an overview of related research. We then describe an empirical study in which a number of implicit measures were collected along with explicit feedback, and modeled to draw connections between *how* a user interacted with a search engine and their level of satisfaction with the search. Finally, we describe open research issues and directions.

2. RELATED WORK

The *Lumiere* research project [Horvitz et al. 1998] explored the use of probabilistic techniques to improve help and assistance to users while they interacted with Microsoft Office applications. A special event monitoring system, *Eve*, was developed to capture a wide range of user actions. Bayesian models were developed to predict users' goals by considering their background, their interactions with the application, as well as their explicit queries. While the end result of this research was slightly different from replacing explicit ratings and feedback with implicit measures, the underlying goal was similar: to try to understand what users want and what satisfies them without them having to explicitly tell the system what that need is or how satisfied they are.

Morita and Shinoda [1994] and Konstan et al. [1997] evaluated the relationship between reading time as an implicit measure and user interest (which was explicitly measured for experimental purposes). Morita and Shinoda [1994] studied the amount of time that users spent reading Usenet news articles and found that reading time could predict a user's interest levels. Konstan et al.'s [1997] study with their GroupLens system, also showed that reading time was a strong predictor of user interest. By providing a ratings system based on implicit measures (e.g., reading time), GroupLens was able to predict user's interest thus rating a specific Usenet article.

Oard and Kim [1998] studied whether implicit feedback could substitute for explicit ratings in recommender systems and identified three broad categories of "potentially useful observations: examination, retention, and reference" (p. 1). They used these categories to group observable behaviors "in a way that is useful when thinking about how to make predictions" (p. 3). As an extension of the Morita and Shinoda [1994] and Konstan et al. [1997] studies on reading time as an accurate predictor for Usenet, Oard and Kim [1998] also found that reading time and whether a page was printed were useful indicators of user interest. More recently, Oard and Kim [2001] presented a framework for characterizing observable user behaviors using two dimensions—the underlying purpose of the observed behavior (*Behavior Category*—examine, retain, reference, annotate) and the scope of the item being acted upon (*Minimum Scope*—segment, object, class). User behaviors were classified according to these two axes. For example, printing was characterized as retaining a segment; bookmarking as retaining an object, markup as annotating a segment, and so on. Most of the implicit measures we measured in our study involved examining or retaining objects.

Goecks and Shavlik [1999] presented an approach that circumvented "the need for human-labeled pages" with the collection of a specific set of implicit measures while users browsed the World Wide Web. The assumption within their work was that there was a connection between users' clickthrough, scrolling activities, and adding to favorites and their level of interest. In this study, they hypothesized correlations between a high degree of page activity and a user's interest. According to Goecks and Shavlik, "our cross-validation experiment suggests that the agent [that collected the data] can learn to predict, at a high degree of accuracy, the surrogate measurements of user interest." While these results were promising, one drawback Goeck and Shavlik mentioned is that the implicit measures were not tested against explicit judgments of user interest.

Claypool et al. [2001] studied how several implicit measures related to the interests of the user. They developed a custom browser called the *Curious*

Browser to gather data about implicit interest indicators and to probe for explicit judgments of Web pages visited. They then used this browser to collect data from 70 students who used the instrumented browser in a computer lab. Their users browsed over 2000 Web pages, with no particular task context. Claypool et al. found that the time spent on a page, the amount of scrolling on a page, and the combination of time and scrolling have a strong positive relationship with explicit interest, while individual scrolling methods and mouse-clicks were ineffective in predicting explicit interest. Like Nichols [1997], Claypool et al. found that a combination of factors (time and scrolling) led to the most accurate predictions.

More recently, Joachims [2002] provided some interesting insight into the collection of implicit measures in place of explicit measures. In his study, Joachims proposed a technique based entirely on clickthrough data. The goal of his work was to develop a method for learning a ranking function based on clickthrough, rather than more costly explicit judgments. The results of Joachim's study indicate that clickthrough data was found to closely follow the relevance judgments, and was useful in learning a ranking function using a Ranked SVM algorithm. While Joachims indicated that clickthrough was a significant predictor of user interest, other studies reviewed earlier indicated that there is the potential for augmenting clickthrough with other implicit measures as well.

While a review of these studies hardly gives justice to many of the research efforts underway (see Kelly and Teevan [2003] for an annotated bibliography of studies on implicit measures), it does provide a representation of the work that is going on and highlights three significant points. First, there is good potential for implicit measures to either replace or act in conjunction with explicit ratings or feedback. Second, there is some disagreement in the existing research on *exactly* what implicit measures are useful—at least within the domain of search engines. Finally, most of the studies have been conducted in laboratory settings. In these situations, experimenters can exercise careful control over the content and guarantee that subjects are only focusing on the task. While laboratory studies reduce noise, the extent to which they generalize to real-world situations in which users are doing many things at once and are frequently interrupted is unclear.

Our study tried to cast some additional light on these points, as well as extend previous research in a number of different ways. First, we used a non-laboratory setting to collect the data from a sample of 146 people over a 6-week period of time. This meant a relatively normal user search environment and an abundance of rich implicit and explicit data. Second, we focused on a Web search scenario, looking at how users interact with the results of search engines. Within this task we looked at satisfaction with individual pages visited and also at satisfaction with an entire search session. Third, our analysis recorded more than 20 implicit measures, which provided us with a rich set of inputs for modeling. Last, we used Bayesian modeling techniques to develop predictive models, and also developed a novel pattern analysis technique (which we call *gene analysis*) to describe user behavior patterns within the search sessions. Previous work (e.g., Morita & Shinoda [1994]; Claypool et al. [2001]) reported simple descriptive correlations between implicit measures and explicit user satisfaction. Our approach was to learn models based on a subset of the data and apply them to a hold-out set, to get an estimate of the predictive accuracy of the models.

3. APPROACH

3.1 Browser Instrumentation

To collect the data required for this research, we developed an Internet Explorer (IE) add-in within a client-server architecture, a technique similar to that used by Claypool et al. [2001]. The IE add-in was a browser helper object that was developed using C-Sharp and installed on a client machine. It monitored a user's search session for several user behaviors (described in more detail in the next two sections). The user behaviors included explicit judgments of satisfaction with the search results and implicit measures of interest collected from mouse and keyboard activities. The IE add-in collected implicit measures and explicit feedback on the client and communicated the data back to an SQL Server database where it was stored and analyzed. The data was sent from the client via different types of XML envelopes (e.g., one for explicit user feedback and another for implicit measures). The IE add-in worked for search results from MSN Search or Google. The user could turn the add-in on or off at any time.

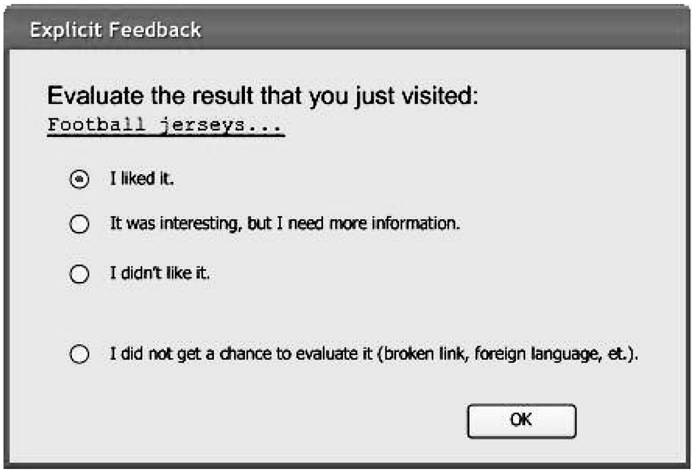
Collecting both implicit measures of user activity and explicit judgments allowed us to model which implicit measures best predicted user satisfaction within the search experience. Details about what explicit feedback and implicit measures we collected are described below.

3.2 Explicit Feedback

Explicit feedback was collected at two levels of detail. First, feedback was collected for individual result visits (i.e., all the pages in the list of search results that the user visited). Second, feedback was collected for the overall search session, which could involve several result visits and/or several queries. A state machine was developed to prompt the user for feedback at appropriate times.

Figure 1 illustrates the feedback dialog for evaluating individual result visits. This dialog was triggered when a user left a search result he/she was visiting by using the Back button to return to the list of search results, closing the IE window, issuing a new query, navigating using history or favorites, typing a new URL in the address bar, or after being inactive for 10 min. For the purpose of analysis, *I liked it* was coded as satisfied with the result (SAT), *It was interesting, but I need more information* was coded as partially satisfied with the result (PSAT), *I didn't like it* was coded as dissatisfied with the result (DSAT), and *I did not get a chance to evaluate it* was ignored except to record how often this happened.

Figures 2 and 3 illustrate dialogs for obtaining session-level feedback. When the user issued a new query, the dialog shown in Figure 2 was used to probe



Explicit Feedback

Evaluate the result that you just visited:
Football jerseys...

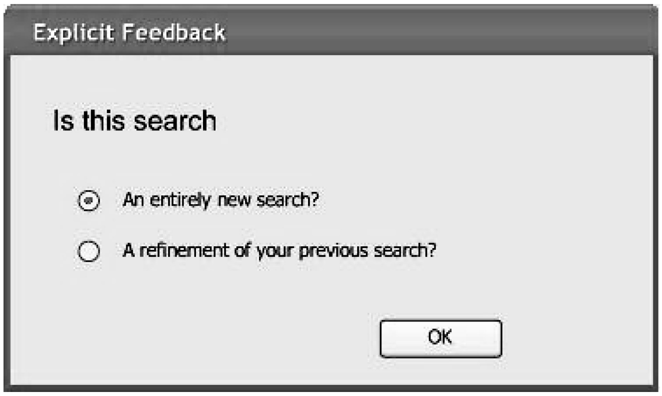
● I liked it.

○ It was interesting, but I need more information.

○ I didn't like it.

○ I did not get a chance to evaluate it (broken link, foreign language, et.).

Fig. 1. Result-level evaluation.



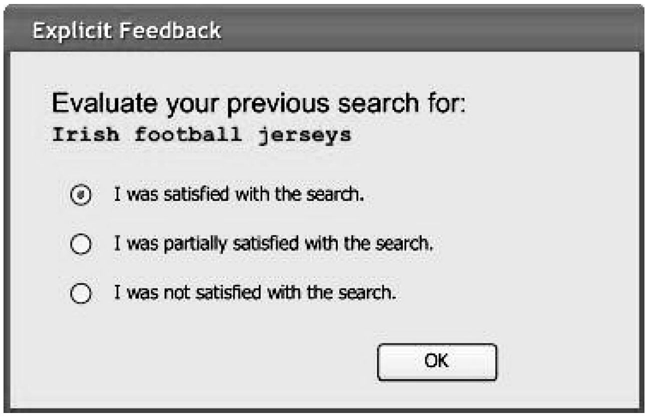
Explicit Feedback

Is this search

● An entirely new search?

○ A refinement of your previous search?

Fig. 2. Requery dialog.



Explicit Feedback

Evaluate your previous search for:
Irish football jerseys

● I was satisfied with the search.

○ I was partially satisfied with the search.

○ I was not satisfied with the search.

Fig. 3. Session-level evaluation.

Table I. Example Search Session

User Behavior	Description of User Behavior
Query 1 = "information retrieval"	The user submits the query "information retrieval" to the search engine.
Result list returned	A result list is returned to the user in response to the query submission.
Result 1 clicked	User clicks the first result in the result list.
Back button clicked	User clicks the Back button and returns to the result list.
Result-level feedback prompt	*A dialog box prompts the user for level of satisfaction with Result 1 (Figure 1).*
Result 4 clicked	User clicks the fourth result in the result list.
Query 2 = "information retrieval, TREC"[a]	The user submits a second query to narrow the focus of the search.
Result-level feedback prompt	*A dialog box prompts the user for level of satisfaction with Result 4 (Figure 1).*
Requery prompt	*A dialog box asks the user if this is a new search or a continuation of their original search (Figure 2).*
Result 1 clicked	User clicks the first result in the result list.
Navigate to another URL	The user types a new URL in the address bar.
Result-level feedback prompt	*A dialog box prompts the user for level of satisfaction with Result 1 (Figure 1).*
Session-level feedback prompt	*A dialog box prompts the user for level of satisfaction with the entire session (Figure 3).*

[a] Acronym stands for *T*ext *RE*treival *C*onference. Website `http://trec.nist.gov/>`.

whether the search intent had changed; that is, whether the user was moving on to a new search task or continuing a previous search. We could have tried to infer a change in intent from temporal patterns and query string overlap, but thought it was safer to ask participants. When the user indicated she/he were continuing the previous search, no additional feedback was requested. When the user indicated she/he had a new search intent, the dialog shown in Figure 3 was presented for evaluating the quality of the previous search session. Figure 3 was also presented to the user when the user closed the IE window, navigated using history or favorites, typed a URL in the address bar, opened another instance of IE, or was inactive for longer than 10 min. Note that search session judgments were collected even when no results were clicked. For the purpose of analysis, *I was satisfied with the search* was coded as satisfied with the search session (SAT), *I was partially satisfied with the search* was coded as partially satisfied with the search session (PSAT), and *I was not satisfied with the search* was coded as dissatisfied with the search session (DSAT).

Table I illustrates an example search session, showing sample User Actions (e.g., Query, Result Clicks, etc.) and the corresponding dialog prompts. The dialog prompts are shown in italics.

3.3 Implicit Measures

Implicit measures were also gathered while the users were conducting their searches and viewing results. Mouse and keyboard actions were recorded and time-stamped by the IE add-in. Table II provides an overview of the main implicit measures we collected. For each page visited, several time and scrolling

Table II. Result-Level Implicit Measures

Result-Level Measure	Description
Time Difference in seconds Duration in seconds	Time spent on a page is represented with two different measures. *Difference in seconds*: time from when the user left the results list to the time he/she returned. *Duration in seconds*: subset of the above time during which the page was in focus.
Scrolled, scrolling count, average seconds between scroll, total scroll time, maximum scroll	Each time a user scrolled down the page a "scrolled" event was logged, along with the percentage of the page that the user moved within that scroll and a timestamp.
Time to first click, time to first scroll	Initial activity times. Time to first click and first scroll.
Page, page position, absolute position	Position of page in results list. The number of the search results page, the search result position on the page, and the absolute search result position.
Visits	Number of visits to a result page.
Exit type	End of page visit. The way in which the user exited the result—kill browser window; new query; navigate using history, favorites, or URL entry; or time out.
Image count, page size, script count	Characteristics of the page. Count of image, size of page, and number of scripts on page.
Added to favorites, printed	Other user actions with page. Whether the user added the search result to his/her favorites or printed the search result page.

Table III. Session-Level Implicit Measures

Session Level Measure	Description
Query count	Number of queries.
Results set count	Number of result sets that were returned.
Results visited	Number of results visited.
End action	The way in which the user exited the session—kill browser window; navigate using history, favorites, or URL entry; open another instance of IE; or time out
Average result duration seconds	Average of duration in seconds.
Average maximum scroll	Average of maximum scroll.
Average page, average page position, average absolute position	Averages of page, page position, and absolute position of result.
Average printed, added to favorites	Average of printed and added to favorites.

activities, whether the user added it to favorites or printed it, and how he/she left the page (Exit type) were recorded. In addition, characteristics of the page (its position in the results list, the number of images, scripts and its size) were recorded. Table III shows additional measures computed for each session. Session-level measures include averages of all result-level measures as well as the number of queries, results lists, results visited, and how the session ended (End action).

The general method of analysis, as we describe in more detail below, was to build models that predicted the explicit judgments of satisfaction (at both the page and session levels) using the implicit measures. We also explored methods for describing sequences of user actions and correlated those with explicit judgments of satisfaction.

3.4 Participants

We collected data from 146 internal Microsoft employees who volunteered for the experiment. The employees were asked to deploy the IE add-in and then respond to the dialogs requesting explicit feedback whenever they conducted Web searches. There were no special laboratory or data collection sessions; data was collected constantly as people searched the Web in the course of their daily work activities. The IE add-in was only available for use on the internal corporate network, so access speeds were fairly constant across the query sessions. The collection of data spanned approximately 6 weeks.

4. DATA ANALYSIS

We first report a few summary statistics for our searches, then provide a brief introduction to two of the analysis techniques we used, and finally describe the main findings for individual result views and entire search sessions.

4.1 Summary Data Characteristics

Data was collected from 146 participants over a span of approximately 6 weeks. During this time, explicit feedback was collected for 2560 sessions and 3659 page visits. Although our user population was more computer savvy than the general Web population, characteristics of their searches were generally consistent with what others have reported. The average query length was 2.99 words, which is somewhat longer than the value of 2.35 reported by Silverstein et al. [1998] or the value of 2.40 reported by Spink et al. [2001]. The average number of queries per session was 2.50, which was very close to the value of 2.52 reported by Spink et al. [2001]. The choice of a work setting may influence the information needs that users seek to address. However, an informal examination of the queries suggests that a range of search intents from specific questions (e.g., *C# XML editor, mappoint*) to general informational and browsing tasks (e.g., *activism online communities, windows scripting*) was represented. More detailed characterizations of the participants and their tasks are beyond the scope of this article, although an interesting problem for future research.

4.2 Bayesian Modeling

To construct predictive Bayesian models for inferring the relationships between implicit measures and explicit satisfaction judgments, we used Bayesian model-structure learning. This approach generates a Bayesian network, which highlights dependencies among variables and influences on a dependent variable of interest (in our case explicit judgments of satisfaction with characteristics of individual results visited or an entire search session). Methods for inferring Bayesian networks from data have been developed and refined over the last decade (e.g., Cooper and Herskovits [1992]; Heckerman et al. [1995]). Given a set of variables, Bayesian-network learning methods perform heuristic searches over a space of dependency models and use a Bayesian model score to identify models that best predict the data. The Bayesian model score estimates the likelihood of a model given data, *p(model|data)*, by approximating the

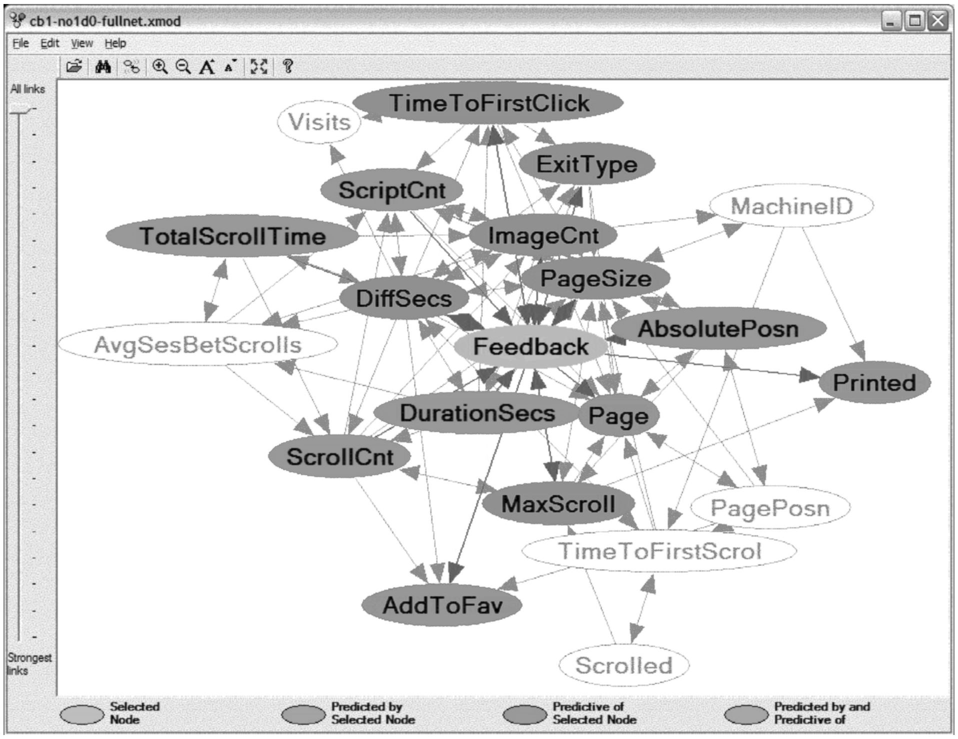


Fig. 4. Bayesian model of the influence of implicit measures on the dependent variable feedback.

quantity, *p(data|model) * p(model)*. Chickering [1997] have developed methods for representing conditional probability distributions encoded within the variables (nodes) of Bayesian networks as decision graphs. These decision graphs represent the conditional probability distributions of each variable, and are a generalization of decision trees in which nonroot nodes may have multiple parents.

This Bayesian approach provides several advantages which we found useful in our analyses. In general, it provides a flexible framework for understanding the relationships between implicit measures and explicit satisfaction, and for predictive modeling. The use of a dependency network (as illustrated in Figure 4) allows us to explore the relationships among variables graphically, as we describe in more detail below. Both continuous and discrete variables are represented in the same model. For purposes of prediction, however, we are interested in the probability distributions for individual output variables, so we learn a decision tree for each output variable. Thus, this approach is similar to other techniques for building classifiers using decision trees. The main difference is the use of Bayesian scoring and pruning techniques for learning decision trees (see Chickering et al. [1997] for details). We could have used alternative learning machine learning techniques to develop predictive models (e.g,, SVMs, linear or nonlinear regression). However, since our main goal was to understand which implicit measures were most predictive of explicit ratings, we were more interested in comparative performance using different combinations of variables (e.g., clickthrough alone vs. clickthrough plus other variables) than in the comparative performance of different learning techniques. Bayesian networks and decision trees have been used in other user modeling work (e.g., Horvitz et al. [1998]), and we believe that they provide a good starting place for our evaluations.

The WinMine tool-kit 1.0 was used for our analyses [Chickering, 2002]. The data was split into training and test sets. Eighty percent of the data was used as the training set to build a predictive model, and the remaining 20% of the data was used to evaluate the accuracy of the model in predicting new data. We explored splits based on users (use 80% of the users to predict the remainder) and time (use the first 80% of the data to predict the last 20% of the data). The results were very similar, so only results from the temporal splits are reported in this article. The complexity of the learned models can be controlled using a kappa parameter to penalize more complex models, and by setting a minimum number of cases represented in a leaf node. For the experiments reported below, we set kappa to 0.90 and required a minimum of 50 observations per leaf node.

We built two Bayesian network models, one for predicting satisfaction for individual page visits (using the variables in Table II) and one for predicting satisfaction for entire search sessions (using the variables in Table III). The variables listed in the tables were used as input to predict users' explicit feedback of SAT (Satisfied), PSAT (Partially Satisfied), and DSAT (Dissatisfied). Figure 4 provides a snap-shot of a Bayesian network that was built for page visits using this technique. Nodes correspond to variables and arcs represent the statistical dependencies between variables. Selecting a node shows the other variables which predict it (arcs pointing in) and which it predicts (arcs going out). For example, in Figure 4, the node Feedback (i.e., user rating of SAT, PSAT, DSAT) has been selected and the statistical dependencies with this variable are shown by the nodes which are shaded darker and their associated arcs. Feedback is predicted by several variables including time (e.g., Duration in Seconds, Time to First Click), scrolling (e.g., Total Scroll Time, Maximum Scroll Extent), and page variables (e.g., Page Size, Image Count).

A decision tree can be used to summarize the model for any node. Figure 5 shows a portion of the probabilistic decision tree for the dependent variable Feedback. Nodes correspond to variables, and each leaf node stores a probability distribution for the dependent variable. The dependent variable Feedback can take on three possible categorical values (DSAT, PSAT, SAT), and these are shown as histograms.

Figure 6(a) provides a drill-down view into one of the nodes where the probability for satisfaction (*p(Satisfied)*) was 88% (second from top node in Figure 5). Note that the model predicts a probability distribution over the three possible outcomes (*p(Satisfied)* = 88%, *p(Partially Satisfied)* = 8%, and *p(Dissatisfied)* = 4%), so a single model is used to predict the different values of the dependent variable. This node was reached when the difference in seconds was greater than 58.4 s, the exit type was not back to the result list, the absolute position was less than 3.45, and the image count was greater than or

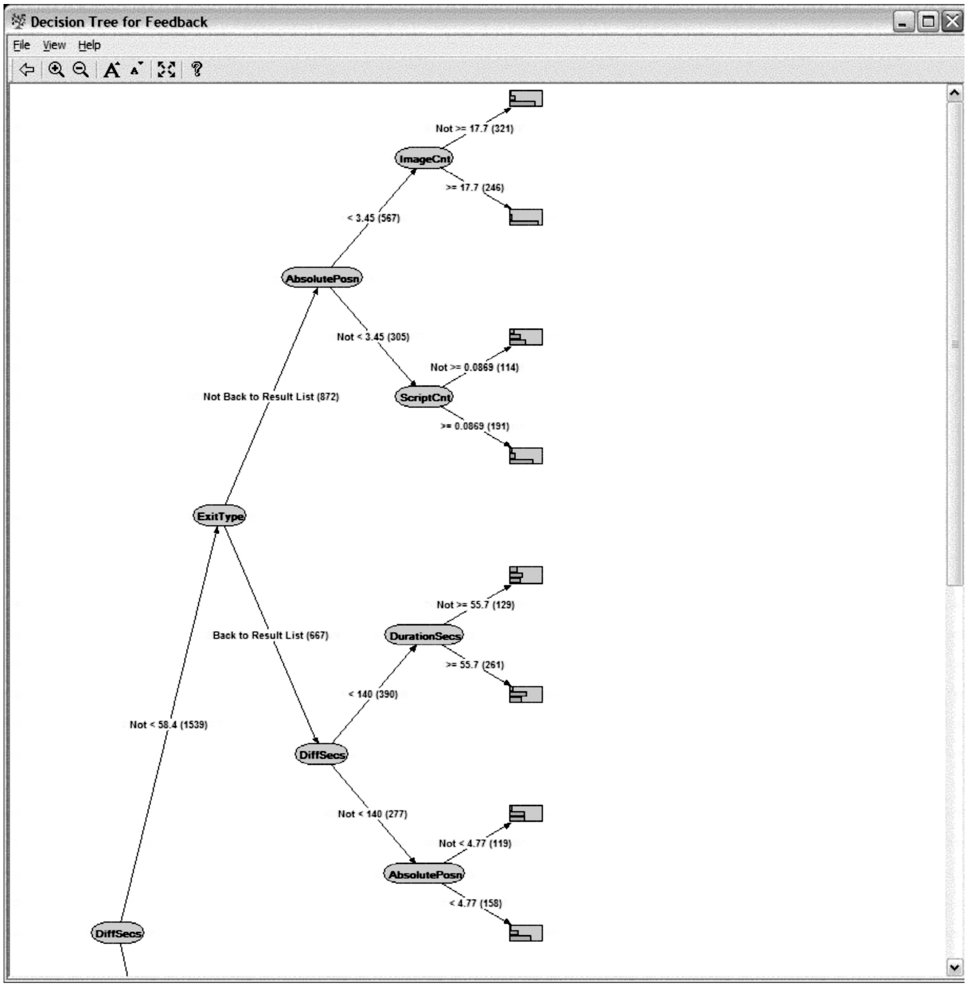
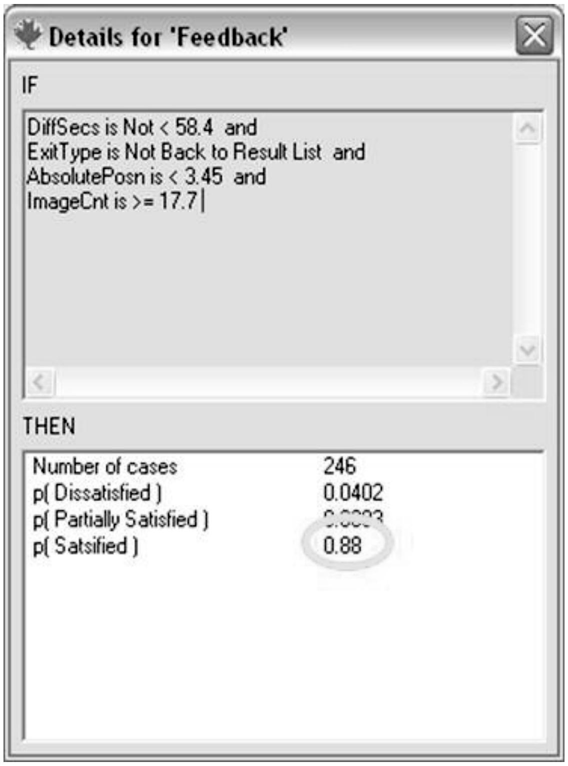


Fig. 5. Decision tree for the dependent variable feedback.

equal to 17.7. Roughly speaking, this meant that when users spent more than 58 s on a page (which had lots of images and was in the top three results) and did not go back to the results list, they were satisfied with the page 88% of the time.

Figure 6(b) provides a drill-down view of another of the decision tree nodes, one which is highly predictive of dissatisfaction. In this node, the probability that the user was dissatisfied (*p(Dissatisfied)*) was 73.4% when the difference in seconds was less than 58.4 s, the exit type was going back the result list, the difference in seconds was less than 27.1, the absolute position of the result was greater than 5.04, and the duration in seconds was less than 9.93. Roughly speaking, this says that when users spent very little time on a page and they did go back to the results list, they were likely to be dissatisfied (with a probability of 73.4%).



IF

```
DiffSecs is Not < 58.4  and
ExitType is Not Back to Result List  and
AbsolutePosn is < 3.45  and
ImageCnt is >= 17.7
```

THEN

```
Number of cases              246
p( Dissatisfied )            0.0402
p( Partially Satisfied )     0.0803
p( Satsified )               0.88
```

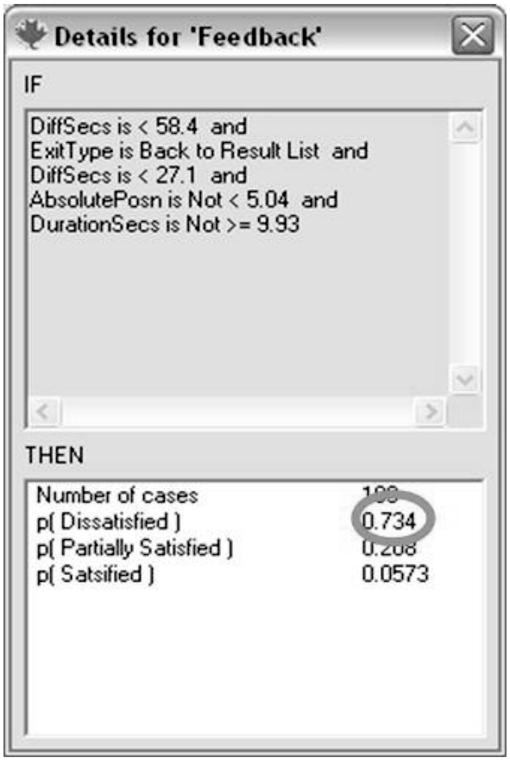
Fig. 6(a). Detailed satisfaction prediction model information.

More formal evaluations of the predictive accuracy of the Bayesian models are described below in the Results-Level and Session-Level findings.

4.3 Gene Analysis

The gene analysis was a descriptive and innovative technique that allowed us to look at *patterns* of user behavior within a search session in the context of what happened around a user's interaction with a result. In this sense, the search session provided the scope for the analysis and the behaviors in and around the result interaction provided the context for the analysis. The gene analysis represented another, more descriptive way in which we could look at the data.

In the gene analysis, the search session behavior was encoded as a string. There were five primary strings used to demarcate user actions: (1) S represented the start of the session; (2) q represented the submission of a query; (3) L represented a result list being displayed to the user; (4) r represented a user clicking on a result; and (5) Z represented the end of the user's search session. The sequence SqLrZ, then, represented a simple session in which a session started; the user issued a query, was presented a result list, visited one result; and then the session ended. We use an asterisk (*) to indicate that there were additional user behaviors before or after a pattern of interest. Thus, SqLr* represents a session beginning with the start of a search session, the



```
Details for 'Feedback'

IF
DiffSecs is < 58.4  and
ExitType is Back to Result List  and
DiffSecs is < 27.1  and
AbsolutePosn is Not < 5.04  and
DurationSecs is Not >= 9.93

THEN
Number of cases              108
p( Dissatisfied )            0.734
p( Partially Satisfied )     0.208
p( Satsified )               0.0573
```

Fig. 6(b). Detailed disatisfaction prediction model information.

submission of a query, a result list being returned to the user, the user clicking on a result, and this then being followed by any other activities. Similarly, the pattern *qLrLrLr* means that, somewhere inside a session, three results were visited after a query.

As mentioned earlier, the gene analysis technique was a secondary descriptive analysis, and, admittedly, it requires additional exploration. Nonetheless, the patterns identified by gene analysis can be used on their own to provide insights about user interaction patterns, or they can be used as additional input variables into the Bayesian models we described earlier. More detail is discussed in the Result-Level and Session-Level findings.

5. RESULT-LEVEL FINDINGS

The result-level analyses explore how accurately we can predict explicit judgments of satisfaction with individual results that were visited. Table IV shows the extent to which clickthrough alone could be used to predict user satisfaction. This table summarizes the number of pages that users clicked on for which they were Satisfied, Partially Satisfied, Dissatisfied, and Could Not Evaluate. When users clicked on a page, they were Satisfied 39% of the time. This baseline model is to predict Satisfied whenever a result is visited.

As described above, we learned a Bayesian model to predict Feedback, using the nineteen variables shown in Table II. The data was split by time,

Table IV. Result-Level Clickthrough Satisfaction

	Training		Testing		Total	
Feedback from User	Number	Percent	Number	Percent	Number	Percent
Satisfied (SAT)	1164	0.42	278	0.40	1442	0.39
Partially Satisfied (PSAT)	843	0.30	230	0.33	1073	0.29
Dissatisfied (DSAT)	782	0.28	190	0.27	972	0.27
Could Not Evaluate	0		0		172	0.05
Total	2789		698		3659	

Table V. Result-Level Predictions Using Bayesian Model

Levels	SAT	PSAT	DSATs	Accuracy
Predict SAT	172	53	20	70%
Predict PSAT	67	91	36	47%
Predict DSAT	39	86	134	52%

with the first 80% of the data used to build the model (2789 judgments) and the remaining 20% used to evaluate the model (698 ratings). Figures 4 and 5 show the learned Bayesian model and a portion of the decision tree. Using the baseline model of always predicting Satisfied for clicked results gives an accuracy of 40% for the test data. (There was a slightly different distribution of Satisfied ratings for the full and test data, 39% vs. 40%.) Table V shows how accurately the learned Bayesian model could predict users' feedback. The rows show the predictions of the learned model and the columns show the actual user judgments. The learned model, using a combination of many implicit measures, was able to predict Satisfaction 70% of the time, which represented a large increase over the baseline accuracy of 40% when clickthrough alone was used to predict Satisfaction. Overall predictive accuracy on all of the test cases (for the three different judgments) was 57%. A nonparametric McNemar test for paired observations showed that the accuracy for the Bayesian model was higher than that for the baseline clickthrough only model ($\chi^2(1) = 42.8$, $p < 0.001$). For some leaf nodes, the probability distribution was highly skewed toward one outcome (as in Figures 6(a), and 6(b)), but in other cases the distribution was more uniform so there was less confidence in the outcome. If one looks at only the cases for which model confidence was high (i.e., the score for the most probable outcome was > 50%), the predictive accuracy increased to 77% for SAT and 66% overall, although this covered fewer of the test cases (407 vs. 698). A nonparametric McNemar test for paired observations showed that predictive accuracy for this subset of cases was higher than the baseline clickthrough model ($\chi^2(1) = 44.0$, $p < 0.001$).

The two most important variables in the Bayesian model were Difference in Seconds and Exit Type, as shown in the decision tree in Figure 6(a). Using just these two variables in the Bayesian model, accuracy for predicting SAT was 66% and 56% for all three judgments overall, both of which were very close to the model using the full set of 19 predictor variables. The difference from a baseline clickthrough model is again significant using the McNemar test for paired observations ($\chi^2(1) = 43.3$, $p < 0.001$). Difference in Seconds represented the total time spent on a clicked result, that is, starting from the instant when

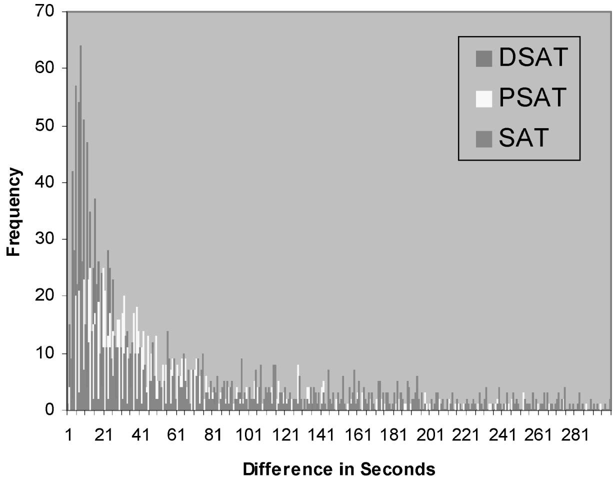


Fig. 7. Distributions of Difference in Seconds.

a result was clicked to the time when the user either came back to the results description list or closed the search application in some other way. Figure 7 illustrates the distribution of values for difference in seconds broken down by whether the user judged the result she/he visited as SAT, PSAT, or DSAT.

For ease of presentation, only values between 1 and 300 seconds are shown in Figure 7. Even for this truncated range of durations, it is evident that, for shorter times, the user was more likely to be dissatisfied with the result, and with longer times the user was more likely to be satisfied. The variance of the distributions was larger when users were satisfied. We used a one way analysis of variance (ANOVA) to compare the distributions of Difference in Seconds for SAT, PSAT, and DSAT ratings using the full set of values. We used two common techniques to normalize the common skew associated with response time data. For the first analysis, we truncated the maximum time to 5000 s. Means for the time distributions were 108 s, 244 s, and 658 s for the SAT, PSAT, and DSAT, respectively. The one-way ANOVA was highly significant: $F(2, 3484) = 108.2$, $p \ll 0.001$. All pair-wise t-tests were significant ($t(1927) = 4.3$, $p \ll 0.001$; $t(2412) = 12.5$, $p \ll 0.001$; $t(2513) = 9.54$, $p \ll 0.001$). For the second analysis, we used the log (difference in seconds + 1). Again, the one-way ANOVA was highly significant ($F(2, 3484) = 410.9$, $p \ll 0.001$), and all pair-wise t-tests were significant ($t(1927) = 15.1$, $p \ll 0.001$; $t(2412) = 27.5$, $p \ll 0.001$; $t(2513) = 14.04$, $p \ll 0.001$). It is important to note that all of our participants were using a high-speed corporate LAN, so bandwidth remained fairly constant throughout the data collection period. (It would, however, be interesting to explore this relationship in environments with a larger range of connectivity parameters.) In addition to time, Exit Type was also one of the top predictors for satisfaction. When a user went back to the results list, he/she was more likely to be dissatisfied than satisfied ($n = 779$ DSAT, $n = 586$ SAT, $\chi^2(1) = 13.7$, $p \ll 0.001$). Conversely, when a user closed the browser on a result page, he/she was more likely to be satisfied than dissatisfied ($n = 347$ SAT, $n = 51$ DSAT, $\chi^2(1) = 127.7$, $p \ll 0.001$). These two variables (Difference in Seconds

Table VI. Result-Level Clickthrough Satisfaction Model

	Training		Testing		Total	
Feedback from Users	Number	Percent	Number	Percent	Number	Percent
Satisfied (SAT)	1175	0.57	289	0.56	1464	0.57
Partially satisfied (PSAT)	403	0.20	102	0.20	505	0.20
Dissatisfied (DSAT)	470	0.23	121	0.24	591	0.23
Total	2048		512		2560	

and Exit Type) in combination with clickthrough represented a strong combination of implicit measures when building our own prediction models within the Web search scenario.

There were also some low-frequency events that were highly predictive of user interest. These did not enter into the Bayesian model because of the choice of a minimum of 50 cases per node. For optimizing overall predictive accuracy, it was important to have broad coverage so we set a minimum number of cases per node. For other applications, it may be useful to know that some activities are highly predictive of user satisfaction; printing and adding to favorites were two such actions. From the full sample of 3659 result visits, 44 resulted in a page being printed and 47 resulted in a page being added to Internet Explorer favorites. When users printed a page, they were satisfied with that page 73% of the time, and partially satisfied 23% of the time—only once was a user dissatisfied with a page he or she printed. When users added a page to their favorites, they were satisfied 81% of the time, and partially satisfied 19% of the time—thus no one added a page to his or her favorites when dissatisfied. Although these behaviors were infrequent, they were highly correlated with user satisfaction, and might provide useful diagnostics for some applications.

To summarize, at the result level, we describe three major findings. First, we found that by using a combination of implicit measures we could better predict user satisfaction than by just using clickthrough. Second, we found that time and exit type were the two best predictors of satisfaction. And third, we found that some actions like printing and adding to favorites were highly correlated with satisfaction, but infrequent.

6. SESSION-LEVEL FINDINGS

The session-level analyses explore how accurately we can predict explicit judgments of satisfaction for entire search sessions which may consist of multiple queries and/or result visits. Table VI summarizes overall session-level satisfaction. Users were satisfied with 57% of their search sessions. This is higher than the 39% satisfaction observed at the result level. This is because users often viewed several result pages or issued several queries in a query session. They could be satisfied with the entire search session even though some of the results examined or queries issued were not satisfactory. The baseline model is to predict Satisfied for every session.

We learned a Bayesian model to predict Feedback for a session, using the variables shown in Table III. Again, the data was split by time, with the first 80% of the data used to build the model (2024 session judgments) and the remaining 20% used to evaluate the model (512 session judgments). The baseline model

Table VII. Prediction of Session-Level Satisfaction Using Implicit Measures

Levels	SAT	PSAT	DSAT	Accuracy
Predict SAT	267	44	50	74%
Predict PSAT	9	42	23	57%
Predict DSAT	13	16	48	62%

of always predicting Satisfied for a session gave an accuracy of 56% for the test data. (There was a slightly different distribution of Satisfied ratings for sessions in the test data than in the full data, 56% vs. 57%.) Table VII shows how accurately this learned Bayesian model could predict users' feedback. The rows show the predictions of the learned model and the columns show the actual user judgments. Using a learned combination of implicit measures, we were able to predict SAT 74% of the time, and overall predictive accuracy for the three judgments was 70%. This is higher than the baseline model, and a nonparametric McNemar test for paired observations showed that the overall accuracy for the model was significantly higher than the baseline ($\chi^2(1) = 41.3$, $p \ll 0.001$). If we look only at the cases for which model confidence was high (i.e., the score for the most probable feedback outcome in a leaf node is > 50%), the predictive accuracy of the learned model increased to 86% for SAT and 77% overall, and this was reliably different from the baseline ($\chi^2(1) = 41.3$, $p < 0.001$).

The most useful individual variables were the number of individual result visits which the user judged to be Satisfied, Partially Satisfied, and Dissatisfied, the Number of Pages Visited, and End Action (that is, how the search session was terminated—typing in a new URL address, closing the browser, etc.). In practical applications, however, one would not have explicit judgments of user satisfaction. So, we also constructed Bayesian models in which we did not include the explicit judgments for individual results. This model was able to predict a user rating of SAT 60% of the time, and the overall predictive accuracy was 60%. Although the predictive accuracy is not as high as for the full model, it was reliably better than baseline ($\chi^2(1) = 7.0$ $p = 0.008$). We believe that this accuracy could be further increased by using predictions of result-level satisfaction (rather than the actual judgments), and we will explore this in future work. For this model, the most important variables were Average Duration on Results, Number of Results Sets, and End Action. As we found with the results-level analyses, time on page and end action were the most important implicit measures in predicting user satisfaction.

It is worth noting that in both the result-level and session-level analyses we included the Search Engine used (MSN Search vs. Google) as a variable in the Bayesian model. This variable was never an important predictor of satisfaction, at either the result level or session level. Thus, the two search engines produced similar user behavior patterns in our study ($\chi^2(2) < 1$, $p > 0.05$ for both result-level and session-level judgments).

6.1 Gene Analysis

As part of the session-level analysis, we used the gene analysis technique to extract user behavior *patterns* that we could then map to user satisfaction.

Table VIII. Gene Analysis Behavior Patterns

Pattern	Freq.	%SAT	%PSAT	%DSAT	Avg. SAT Duration (s)	Avg. PSAT Duration (s)	Avg. DSAT Duration (s)
SqLrZ	509	81	10	7	4599	610	211
SqLrLr*	362	23	39	36	30	71	14
SqLrLrLr*	129	20	37	42	16	26	10
SqLrLZ	117	75	15	9	64	6	11
SqLrLrLrLr*	114	13	35	51	12	46	25
SqLrLrZ	82	73	13	13	4819	349	72
SqLrqLr*	70	64	25	10	2002	49	20
SqLrLrLrZ	61	57	22	19	2178	209	53

In a gene analysis, the search session behavior was encoded as a string of actions, and session behavior sequences themselves were composed of smaller constituent patterns that could also be mapped to user satisfaction. That is, a pattern could be found for each interacted result within the session by extracting the substring in the behavior sequence around the result with which the user interacted. For example, the pattern *qLrLr* meant that, anywhere inside a session, two results were visited after a query. Table VIII summarizes the most common gene patterns and subpatterns for search sessions. For each pattern, the table shows the frequency of occurrence and the satisfaction ratings, along with the dwell time for each satisfaction rating.

Some of these patterns, like SqLrZ (first row), were highly associated with Satisfaction. For example, participants were satisfied 81% of the time when their session was characterized by the pattern SqLrZ and 73% of the time when it was characterized by SqLrLrZ. These sequences can be used on their own in an exploratory fashion to suggest patterns of interaction associated with user satisfaction. Or interesting gene patterns can be entered into a Bayesian model and used as additional input variables. We conducted some preliminary analyses using genes selected from tables like this as inputs to the Bayesian model (along with the other independent variables) and found, for example, that with the gene variable *qLrZ (i.e., sessions that ended with a query, result list presentation, and result visit) was predictive of a judgment of SAT. There is clearly much more work to do in combining behavioral patterns into the Bayesian models, but the technique offers some promise as a way of understanding richer patterns of user interactions with search results.

We should add that one of the problems with gene analysis is that several patterns occur with very low frequency, making them unreliable for purposes of prediction. We believe that using subpatterns and abstractions (e.g., those containing more than three qLr actions) can help mitigate this problem. In addition, we believe that patterns are attractive because they are likely to be consistent across search applications. A pattern may be thought of as a superset of several implicit measures and can be potentially improved and extended to represent those component measures correctly.

To summarize, at the session level, we describe four major findings. First, we found that by using a combination of implicit measures we could better predict user satisfaction than by just using the base rate of satisfaction with sessions.

Second, we found that duration in seconds, clickthrough, and end action were the strong predictors of user satisfaction. Third, we found that result-level satisfaction was associated with session-level satisfaction. And fourth, we found that exploring behavior patterns provided some insights about sequences of user activity and that these patterns can be incorporated into Bayesian models.

7. DISCUSSION AND CONCLUSION

The goal of this research was to understand the relationship between implicit and explicit measures of user satisfaction. We focused on Web search applications, and collected more than 30 implicit measures (along with explicit judgments) from 146 people over a 6-week period of time in their normal work context. We used Bayesian modeling techniques and found that a *combination* of the *right* implicit measures can provide good predictions of explicit judgments of user satisfaction. At the result level, large and statistically significant improvements over a baseline model were observed, and clickthrough, time, and exit type proved to be the best predictors of satisfaction. At the search session level, smaller, but still significant, improvements over a baseline model were observed, and again clickthrough, time, and end action were the best predictors of satisfaction.

We also explored the use of usage patterns (which we call *gene sequences*) for characterizing sequences of user behavior patterns and predicting user satisfaction. For example, users were more satisfied in sessions consisting of the pattern with SqLrZ than in those starting with SqLrLrLrL*. This suggests that the longer users search through a particular result list, the less likely they are to be satisfied with the search session. Analyzing user interaction logs may also reveal patterns unique to a specific search application (e.g., Internet search engine vs. Intranet search engine; one user interface vs. a different user interface).

We believe that this study resulted in useful and significant evidence on the importance of combining implicit measures using probabilistic formalisms for predicting user satisfaction. We believe this study helped elucidate that there is great potential in using the right combination of implicit measures to augment and extend explicit ratings or feedback. Explicit feedback should not necessarily be used *in lieu* of implicit measures; rather, it might be considered one more measure in the combination of all measures (and appropriately weighted). The descriptive analysis of user behavioral patterns within search sessions offers another interesting way to look at the implicit measures in the context of how the user interacted with results. Further, consistent with Nichols' [1997] conclusion, careful analysis of both implicit and explicit measures should be considered with the appropriate weighting of all measures in mind, but great potential exists for the employment of implicit measures in real-world search applications.

Last, as echoed in some of the previous studies discussed in this article, security and privacy must be primary concerns when delving into the realm of implicit measures. During this study, we respected the privacy of all those involved in the study and were able to collect data where the queries and results were distinct from the users. Extending this respect for privacy and security would have to be a key consideration in further research in this area.

8. FUTURE WORK

Having discovered that implicit measures can be used to build accurate predictive models of user satisfaction, especially at the individual result level, we would like to be able to use the predicted judgments as a cost-effective way to augment and extend explicit judgments. One approach would be to use the predictions to prioritize queries requiring a more detailed human analysis. A more interesting alternative would be to substitute predictions for explicit judgments. This is challenging for at least two reasons. First, the predicted satisfaction estimates are not 100% accurate. Then again, neither is the consistency of human relevance assessments, so perhaps the level of predicted accuracy is sufficient to support the comparison of ranking algorithms. Second, the judgments are probabilistic and most relevance judgments are binary. There has been some work on using graded relevance judgments (e.g., Voorhees [2001]) and we believe that the probabilistic outputs of our learned models fit nicely into this framework. Understanding the extent to which predicted satisfaction can be used instead of or in combination with explicit judgments is an important next step in our research.

Another direction for future work is to collect additional data in a wider variety of natural settings. We used a work environment with high-speed network connections and participants who were reasonably savvy technically. We would like to extend our data collection to other settings like the home, to explore different connection speeds and different user populations, and to explore the consistency of the learned models in these different environments. And we would like to explore a wider range of implicit measures and techniques like gene analysis for exploring patterns of interactions.

Another direction would be to explore different learning methods. We used Bayesian modeling techniques because they have sound probabilistic foundations, allow both continuous and discrete variables to be combined into a single model, and allow dependencies among variables to be represented. Many other approaches could be used to model the relationships between implicit measures and explicit judgments (e.g., linear and nonlinear regression, alternative classification algorithms such as SVMs or kNN). In this work, we were more interested in comparative performance using different combinations of variables (e.g., clickthrough alone vs. clickthrough plus other variables) than in the comparative performance of different learning techniques. However, better understanding the most useful models is an important direction both practically and theoretically.

We believe that the results and techniques presented in this article are a promising start in understanding how implicit measures of user activity relate to explicit judgments of user satisfaction. Fully understanding the best modeling techniques, the consistency of models derived in different usage contexts, and the situation in which implicit measure can complement explicit judgments will require more detailed analysis and investigation.

REFERENCES

Chickering, D. M. 2002. The WinMine Tookit. Microsoft Research Tech. Rep. MSR-TR-2002-102. Microsoft Research, Redmond, WA. Go online to `http://research.microsoft.com/~dmax/WinMine/tooldoc.htm`.

Chickering, D. M., Heckerman, D., and Meek, C. 1997. A Bayesian approach to learning bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. 80–89.

Claypool, M., Brown, D., Le, P., and Waseda, M. 2001. Inferring user interest. *IEEE Internet Comput. 5*, 6 (Nov.-Dec.), 32–39.

Cooper, G. and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn. 9*, 309–347.

Goecks, J. and Shavlik, J. 1999. Learning users' interests by unobtrusively observing their normal behavior. In *Proceedings of the IJCAI Workshop on Machine Learning for Information Filtering*. 129–132.

Heckerman, D., Geiger, D., and Chickering, D. M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn. 20*. 197–243.

Horvitz, E., Breese, J., Heckerman, D., Hovel, D., and Rommelse, K. 1998. The Lumiere Project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (July). 256–265.

Joachims, T. 2002. Optimizing search engines using clickthrough data. *In Proceedings of KDD* 2004. 133–142.

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

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. 1997. GroupLens: Applying collaborative filtering to usenet news. *Commun. ACM 40*, 3, 77–87.

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

Nichols, D. M. 1997. Implicit ratings and filtering. In *Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering* (Nov.). 221–228.

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

Oard, D. W. and Kim, J. 2001. Modeling information content using observable behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*. 38–45.

Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. 1998. Analysis of a very large AltaVista query log. SRC Tech. Note 1998-014, Compaq Systems Research Center, Palo Alto, CA. Website: `http://www.research.compaq.com/SRC/publications`.

Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T. 2001. Searching the Web: The public and their queries. *J. Amer. Soci. Informat. Sci. 52*, 3, 226–234.

Voorhees, E. 2001. Evaluation by highly relevant documents. In *Proceedings of the Twenty-Fourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 74–82.

Received November 2003; revised June 2004, October 2004; accepted October 2004