Methods for Identifying Usability Problems with Web Sites

Jean Scholtz

National Institute of Standards and Technology, Gaithersburg, MD USA, jean.scholtz@nist.gov

Laura Downey

Vignette Corporation, Austin, TX USA, laura@vignette.com

Abstract: The dynamic nature of the Web poses problems for usability evaluations. Development times are rapid and changes to Web sites occur frequently, often without a chance to re-evaluate the usability of the entire site. New advances in Web developments change user expectations. In order to incorporate usability evaluations into such an environment, we must produce methods that are compatible with the development constraints. We believe that rapid, remote, and automated evaluation techniques are key to ensuring usable Web sites. In this paper, we describe three studies we carried out to explore the feasibility of using modified usability testing methods or non-traditional methods of obtaining information about usability to satisfy our criteria of rapid, remote, and automated evaluation. Based on lessons learned in these case studies, we are developing tools for rapid, remote, and automated usability evaluations. Our future work includes using these tools on a variety of Web sites to determine 1) their effectiveness compared to traditional evaluation methods, 2) the optimal types of sites and stages of development for each tool, and 3) tool enhancements.

Key words: Usability testing, Web development, Web site design, Remote usability testing, Web site evaluation

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35349-4_22](https://doi.org/10.1007/978-0-387-35349-4_22)

S. Chatty et al. (eds.), *Engineering for Human-Computer Interaction*

© IFIP International Federation for Information Processing 1999

1. WEB DEVELOPMENT CONSTRAINTS ON USABILITY EVALUATION

Web development places severe time constraints on developers and evaluators because of the rapid development and release cycle. The tight coupling of content, navigation, and appearance of Web sites means that separate evaluation of any one component is not meaningful. Web sites also change frequently. We feel it is safe to assume that in the majority of cases, there is no testing of the entire site to see how the new portion fits in, even if the new portion has been evaluated. Web sites must evolve as users' expectations change and as new software and hardware developments are implemented in the Web sites of others and content becomes outdated. This implies that usability evaluations for Web sites should be a continuous effort. But the traditional usability tests, and even Nielsen's (1989) "discount usability testing," take time, especially if one has to test an entire Web site.

Web sites reach a diverse audience. Getting representative users to come to a usability laboratory to participate in an evaluation is often not feasible. More importantly, users view Web sites with different types of browsers, preferences, and internet connections. Testing under all these different conditions adds to the time and complexity of setting up and conducting laboratory tests. Therefore, tools to monitor existing sites for potential usability problems should be very useful to Web site developers.

While traditional software has the same constraints, we maintain that these constraints are more severe when developing Web sites and Web-based applications. However, the Web facilitates quick and widespread delivery of information. Moreover, Web server logs record much information about user interactions with Web sites. We wanted to see if we could take advantage of these two properties in developing some methods for obtaining usability information for Web-based software.

2. TRADITIONAL USABILITY EVALUATION METHODS

Jeffries et al. (1991) compared usability evaluation methods and identified advantages and disadvantages of several techniques, including usability testing. John and Marks (1997) compared the effectiveness of several usability evaluation methods to laboratory usability tests and found that less than half of the problems predicted were observed in usability tests. Nielsen (1993) found laboratory testing of users to be the most effective source of information for identifying usability data. In-house user testing is expensive and places limits on the type and number of users geographically available. Moreover, in-house testing does not allow evaluators to view use in the context of other work activities and the users' hardware and software configurations.

Remote testing has been getting increased attention in the evaluation community. Remote testing can be done synchronously by using software tools that allow the evaluator to view the remote user's screen. Audio connections may be provided by

the software or by using additional phone lines. Asynchronous tests can be done by electronically distributing the software and the test procedures and providing a way for the results to be captured and returned to the evaluator.

Hartson et al. (1997) discussed advantages and disadvantages of several types of remote evaluations and presented two case studies: one using teleconferencing and the second using a semi-instrumented method of evaluation. The focus in these two cases was to obtain qualitative information to be used in formative evaluations.

Our work focuses on using remote testing to obtain quantitative information to supplement work done in the laboratory or to identify potential usability problems in existing software. While we believe that qualitative data is needed in order to produce better designs, we are focusing on what quantitative data can be collected in a remote, automated, and rapid fashion to identify usability problems.

3. OBJECTIVES AND METHODOLOGY

3.1 Objectives

Our objective is to develop tools and techniques to facilitate evaluation of Web sites and Web site designs. Our long-term plan is to:

- Carry out case studies to determine what useful quantitative data can be collected in a remote, automated, and rapid fashion;
- Develop tools based on what we learn from the case studies;
- Use these tools on a wide variety of sites, comparing the effectiveness of individual tools as well as combinations of tools and remote testing;
- Redesign tools as needed and develop new tools suggested by the effectiveness comparisons; and
- Generate guidelines about which tools should be used under which conditions.

This paper reports on the first set of case studies and the lessons learned from them. Our first three tools have just been released and two more are in the design stage.

3.2 Methodology for Case Studies

In order to design effective tools, we carried out three initial case studies to investigate the usefulness of quantitative information that can be collected remotely. In these case studies, we looked at the effectiveness of evaluations based on:

- User satisfaction ratings on usability questions
- The time and success users have in carrying out tasks on Web sites
- Usage patterns

3.3 Web Sites Selected

We selected three different services in the NIST Web site representing three different types of sites. The sites selected were a form fill-in application, a general-purpose library site, and a special purpose technical site. The owners of these sites were enthusiastic about our experiments and were grateful for any recommendations that we could provide them in the course of our research. In the next paragraphs, we briefly explain each site and its use.

3.3.1 The NIST Technicalendar Wizard

A printed calendar is published every week at NIST. It contains notices about meetings and talks to be held at NIST or given by NIST employees at other locations, as well as meetings elsewhere that might be of interest to NIST scientists. The calendar is distributed to NIST personnel in hardcopy. It is also viewable on the Web and e-mailed to others outside of the agency. The URL for the calendar is at <http://nvl.nist.gov/pub/nistpubs/calendars/techcal/techcal.htm>.

Previously, articles for inclusion in the technicalendar were faxed, phoned in, or e-mailed to a staff person who spent at least one day per week collecting any missing information for items submitted and formatting them correctly. To streamline this activity, an on-line wizard was developed so that submissions could be made via the Web. It was hoped that this would considerably reduce the time spent in publishing the technicalendar and make the submission process easier for NIST staff.

3.3.2 The NIST Virtual Library

The NIST Virtual Library (NVL) is a service available to both NIST staff and the public from the NIST home page. The NVL gives users access to an online catalog, assorted electronic journals, various databases (some of which are limited to NIST personnel), and NIST publications. Access is also provided to several other resources, including NIST maps and phonebooks, yellow pages, other government agency information, and weather forecasts.

This site supports both NIST personnel conducting scientific research as well as outside visitors who range from school children writing reports to researchers in business and universities. The NVL site designers are considering various possibilities for redesign and are interested in any recommendations we can provide. The current version of this site can be viewed at <http://www.nist.gov/>.

3.3.3 The Matrix Market

The Matrix Market is a specialized service provided by NIST staff in the form of test data for comparing algorithms for numerical linear algebra. The mathematics group has compiled a set of sparse matrices and matrix generators that can be downloaded from their Web site. The users are mathematicians from all over the world. Viewers of these pages can get information about the individual matrices and

can download any matrix test data. This data can take considerable time to download so the group is especially interested in ensuring that visitors can quickly and accurately locate the test set they need. The Web site used in the case study was similar to the current page at <http://math.nist.gov/MatrixMarket/>.

4. STUDY ONE: THE TECHNICALNDAR WIZARD

4.1 Methodology

We used this case study to determine the usefulness of collecting subjective satisfaction ratings. As the wizard had not yet been released, we incorporated this data collection into a beta test. While the beta test method is often used to collect bug information, it is usually not used specifically to collect data about usability problems. We had questions about how useful beta testing would be as a substitute for usability testing. In particular, would users be willing to participate in the beta test and what types of usability problems would be identified this way?

We hoped to address the first question by advertising the availability of the online wizard in the paper bulletin. NIST personnel were told that after the trial period the online submission procedure would be the sole submission method. They were informed about the usability study and were given the opportunity to “test” the wizard if they did not have actual calendar items to submit.

One problem with using beta testing to uncover usability problems is the difficulty of correlating user reports of problems with the task the user was doing when the problem occurred. As this Web application is quite simple, containing only a few high-level tasks, we hypothesized that, by collecting the submitted calendar item, we could identify the user task that was being done when the usability problem was encountered.

We were interested to see what types of usability problems could be identified using this approach. The authors conducted independent heuristic evaluations of the Technicalendar Wizard first. We listed the issues that at least one of us had identified as a problem. We used this list of problems as a comparison for the actual problems identified during the beta test. While we realize the limitations of heuristic reviews, we were faced with some real-world constraints. First, we needed some input for constructing our rating questionnaire. Secondly, we wanted a baseline to compare the usability problems identified during the beta test. A comparison based on an actual user test would have been more desirable but there are issues about charging administrative staff time for usability tests in government institutions which raise the cost of in-house user testing.

We constructed an evaluation form for users to fill out and e-mail to us after they had used the Technicalendar Wizard. The evaluation form included six questions for rating usability and an open-ended comment field.

After a month of use, we compiled the data collected from the user test and looked to see what, if any, overlap we had with the problems identified in the

heuristic evaluation. We reviewed the user data along with the problems identified in the heuristic review and fixed a number of problems. We continued collecting user data for the next six weeks to see if our redesigns were construed as better.

4.2 Results

During the first month of testing, there were 24 electronic submissions. Of these, 16 were real submissions and 8 were test submissions. Any given Technicalendar contains between 25 and 40 items. Some of the items are published in more than one Technicalendar, so a very rough guess is that the 16 real submissions constituted between 15% -25% of the total submissions for the month. Of the 24 electronic submissions, 13 filled out evaluation sheets. Eight questionnaires were from real submissions and five were from test submissions.

The second phase of testing lasted six weeks. During this time there were 59 electronic submissions. Of these, 43 were real submissions and 16 were test submissions. We received 15 evaluation questionnaires, 10 from the real submissions and 5 from the test submissions.

Twenty usability problems were identified in the heuristic evaluation. Of these 20, 4 were fixed prior to the beta test. We wanted to see if and how the remaining problems were identified during beta testing and what problems were discovered in beta testing that were not identified during the heuristic evaluation. Usability problems discovered during beta testing could be identified in one of three ways: a low rating in the rating section of the questionnaire, a calendar submission with missing or incorrectly formatted data, or user comments. As shown in Table 1, the most problems (including two we thought we had fixed prior to beta) were identified through user comments. However, the calendar submissions and the ratings yielded three more problems. In addition, all these methods helped us identify usability issues.

What problems were noted in the heuristic review that were not identified as problems during the beta test? Of the eight problems identified by the heuristic review that were not identified during beta testing, two had to do with alignment of fields and grouping of fields. Two others were terminology and inconsistent labeling problems. No keyboard navigation was provided to move between steps and no numbers were provided on the wizard steps. No field was provided for a title for a speaker. And finally, directions for selecting an item in a drop-down list appeared as the first item in the list.

What problems were identified by users that were not identified in the heuristic review? Users described difficulties in submitting some unusual items. For example, a user had difficulty using the wizard to fill in the proper information for a panel with six speakers. There was no way to specify that this was a panel, and it was difficult to list the names of all six speakers nicely formatted. This individual usually wrote nicely formatted descriptions for the calendar items and then submitted them. The wizard did not support her formatting preferences. The open-ended comments were especially useful for identifying unusual problems.

Table 1. Ways in which Problems were Identified during Beta Testing

<i>Identification Method</i>	<i>Number</i>	<i>Type of problem</i>
Calendar submission	2	Text field formatting
Low ratings	1	Determining optional fields
User comments	5	Access to help
		Relationship between fields
		Terminology
		Layout
		Missing defaults

4.3 Discussion

The response rate from the users was good. We collected more input than we would have been able to during a typical laboratory usability test. We worked with the Web master to correct the problems identified and a second version was installed on the Web site. The second round of beta testing uncovered no new problems and allowed us to verify our redesign by comparing the usability ratings.

The case study suggested that a useful tool would include functionality to automatically generate satisfaction questionnaires along with an analysis capability. In the case of transaction-based Web applications, it may be feasible to generate rating questions in response to the completion of a checklist of the components of the site.

5. STUDY TWO: THE NIST VIRTUAL LIBRARY

5.1 Methodology

The NIST Virtual Library (NVL) is a scientific library accessible to the public from the NIST Web site. While some of the databases are restricted to NIST personnel, most of the library resources are open to the general public. The NVL staff was considering a redesign of the web interface and was very interested in obtaining data that would help them focus on specific areas to target.

The usability test consisted of three parts: a matching exercise to test existing categorization, ten representative tasks, and a short demographic and satisfaction questionnaire. We recruited five subjects from different scientific disciplines who worked at the NIST site in Gaithersburg, MD. It is important to note that we did NOT conduct this test remotely. We designed the test so that, given the appropriate software, it could be conducted remotely. In simulating remote testing, we kept the experimenter interaction with the users during the test to a minimum.

In the matching task, users were asked to match 29 items to one of ten choices, nine categories from the NVL home page plus a “none” category. We collected the results of this variation of a card-sorting task (Nielsen, 1993). In the performance task we collected the time it took users to complete each of ten tasks and their answers for each task. We also collected users' perceived difficulty ratings for each task. After the test was over, the experimenter conducted a retrospective interview with the users to identify qualitative information in order to determine what kinds of information we would miss in a purely automated asynchronous remote test.

We needed a benchmark to compare the results of our subjects. We had two experts complete the matching exercise, the ten tasks, and the satisfaction questionnaire. One expert was a reference librarian at NIST who was very familiar with the NVL site. The second expert was the designer of the NVL Web site.

5.2 Results

5.2.1 The Matching Task

Our baseline users misidentified two items out of the 29 total. Our non-expert subjects misidentified 13 items. Out of the nine categories, two of them, Databases and Hints & Help, were misidentified the most times. Figure 1 shows some of the categories and items in the matching task.

Category	Items
Subject Guides	Weather forecasts
Visiting NIST	CD-ROM databases
Hints and Help	Street map of NIST
Web Resources	Online Commerce Business Daily
E-Journals	Britannica Online
NIST Publications	List of Federal Library Web sites
Databases	NIST Tour Information
	NIST index to technical activities

Figure 1. A Sample of Categories and Items in the Matching Exercise

5.2.2 The Performance Test

Tasks

Find 6 computer science journals.

Find a list of the periodic tables.

Find at least one NIST person to contact on the subject of visualization.

Find a link in the NVL site that lets you look up U.S. area code information.

Find the link to physics dissertation abstracts.

Figure 2. A Sample of the Tasks used in the Performance Test

Figure 2 shows a sample of the tasks users were asked to do. Our expert users were able to do nine of the ten tasks. However, each expert user missed a different task. Our five non-expert users were able to successfully complete between six and seven of the ten tasks.

The expert users took just over eight minutes to complete the ten tasks. The non-expert users needed over 31 minutes to complete the same tasks. Looking at individual tasks, we find an interesting issue. All the non-expert users missed one task. However, the users did not rate this task as the most difficult. This is probably because many of them thought they had located the answer.

5.2.3 The Satisfaction Questionnaire

Users rated the difficulty for the tasks quite high given their success and the time they needed to complete these tasks. A seven-point scale was used, with one being an unacceptable rating and seven being an excellent rating. Experts gave an average difficulty rating of 5.7 compared to an average of 4.8 for the non-experts.

Originally, we had intended to use only success or failure in completing the task. However, we found instances where users thought they had located information but had not. Therefore, recording the users' answers was necessary.

5.3 Discussion

Because we were not actually conducting this test remotely, we were able to observe users and interview them after the test. We did this to get an idea of the data we would not be able to collect remotely. Our observations of users' strategies and retrospective interviews gave us some insights into user search strategies. We found that users tended to use a search engine if they didn't know where to start a search, i.e., under which category to begin searching. If they did know the category, they preferred to use that.

We also noted that users preferred the category icons in the menu frame to jump to those pages, rather than the links within the home page. Alphabetical listings of

links were more helpful than other groupings when the material was unfamiliar.

This qualitative information could not have been easily obtained using asynchronous remote testing, although a comparison of the paths users take to optimal paths could be used to identify critical decision points. Was the quantitative information we collected useful? The results of the matching tasks pointed to two category names that were difficult for users to understand. The performance test identified four tasks that were difficult for users. Looking at the paths these users took compared to the paths of users who obtained the correct answers could have helped isolate where the confusion occurred.

What lessons did we learn in order to design successful, remote testing tools? We need to collect users' answer to specific tasks to ensure the task really was successfully completed. Additionally, collecting the paths that the users take in information seeking tasks allows us to determine how well our Web site organization matches users' mental models. An automatic way to compare these paths to ideal paths would also be useful. Quantitative measures of time, success, and perceived difficulty can be obtained remotely and automatically. The matching task can be easily automated and used for remote testing.

6. STUDY THREE: THE MATRIX MARKET

6.1 Methodology

As we noted earlier, the Matrix Market is a very specialized site used primarily by mathematicians in testing numerical linear algebra algorithms. The developers of this site told us that visitors to the site would primarily be interested in 1) finding information about a particular test set or 2) downloading a particular test set. While the information contained in the site made it quite large, the use of the site was limited to two primary uses. We used this case study to determine if usage patterns derived from server log data would be effective in identifying usability problems. We recognize the numerous problems with using server log data as the sole source of information (Stout, 1997). However, server log data can still be used to determine overall patterns of traffic, changes in traffic patterns and dead areas in a site. Sullivan (1997) describes the use of server logs to provide inferential statistics about Web site usability.

6.2 Heuristic Review

We first did a heuristic review of the site to use as an indicator of potential problems. We used this data to get an indication of the type of information to look for in the server log data. In the review, we identified 17 problems that we grouped into eight basic kinds of problems. In Table 2, we list the eight categories of problems identified in the review and our hypotheses about what information from the server log might be used to confirm or deny these potential problems. We also included a result column to show our conclusions based on examination of one month of server log data.. We plan to extend this study to use more data in the

future. Please note that the web page of the Matrix Market (<http://math.nist.gov/MatrixMarket>) has changed somewhat since the time of our case study and some problems discussed here are no longer present.

6.3 Results

Our analysis was done mostly by “brute force”; that is, we used scripts to filter and sort the data. Our long-term goal is to develop queries and visualizations that usability professionals can use to analyze traffic on Web sites, with an emphasis on uncovering usability problems.

For each of the potential problems identified by our heuristic, we hypothesized what data in the server log might be used to determine if the problem actually existed in real use. We simplified the access log file by removing all references to graphics and to scripts. We built paths of user visits each day, recognizing that caching prevents us from seeing the complete picture. We placed a time limit on visits and discarded visits lasting longer than 30 minutes.

6.3.1 Overall Results

In one month, we counted 1199 visits and 1010 unique IP addresses. To see whether users were having any major problems with the site, we looked at the percent of visits where help was accessed at least once. Just over 5% of the visits used help.

The home page provided six ways for users to browse through the matrices. We found that for this month, the percentage of visits using each access method ranged from 4% to 19%. This gave us an indicator of the top two or three access methods. We also found that 40% of the visits started from the home page, while 24% of the visits started from a page explaining one of the matrices. However, almost 70% of the visits requested the home page at some time.

The site developers told us that they expected two types of users. Users might come to the site, having read a research publication about an algorithm for numerical linear algebra, to read a description of the matrix that was referenced. Users would also come to the site to determine if the supplied matrices would be useful in testing their algorithms and if so, download the appropriate file. We found that 52% of the visits looked at the matrix descriptions. However, only 6% of the visits downloaded a file.

6.3.2 Comparing Server Log Data with Heuristic Results

Of the eight potential problems we investigated using server log data, we verified that one (the scrolling problem) was a significant problem. Two remain to be verified (the long search form and the download problem) and five others were not verified as significant problems. Table 2 shows the results of looking at usage patterns to see if the potential usability problems affected users doing their work.

The next step is to do actual user testing on this site to determine how accurately the server log data reflects these usability problems. We must also determine other

usability problems not identified in the heuristic and look for indications of these in the server log data.

Table 2. Usage Pattern Results

<i>Problem</i>	<i>Usage Data Examined</i>	<i>Results</i>
Behavior inconsistency	% of users using inconsistent behavior	< 8% of users used inconsistent behavior
Terminology inconsistency	% of users using help	Only 5% of users used help
Need to scroll long lists	% of users viewing page	20% - potential problem
Discriminating between link names for data files	Average number of links followed from this page to data files would be greater than from other pages to access data files	29% of users accessed more than one data file but only 7% accessed more than one data file from this page
Scrolling was most likely needed to view several groups of links on home page	Most frequently accessed pages by visits	Most frequently accessed pages were in easily visible area
Extra step needed to access some information	% of users viewing this data	Only 5% of visits went to this data
Need to scroll to view entire search form	Search followed immediately by another search	Unable to validate at this time
Estimates for download times are not given	% of users stopping transfer of data downloads	Unable to validate at this time

6.4 Discussion

We believe that for specialized Web sites with limited uses, the following usage questions can be answered through server log data:

- From the home page, which links are most frequently used?
- Do users have a difficult time discriminating between names of links from a given page?

- Do users have difficulty locating information via searching and need to make multiple attempts?
- Do visitors use help frequently?
- What pages are used most frequently as entry points by users?

Server log analysis allows us to estimate the percentage of users that a potential usability problem affects. We concluded from this case study that a useful tool would construct user paths from server log data and display the appropriate visualizations in response to usability questions such as those listed above.

7. CONCLUSIONS

We believe that usability evaluation techniques that will prove effective for the Web must be rapid, remote, and automated. We have investigated the data that could be collected in such a fashion and shown the usefulness of that data in identifying usability problems through three case studies of different types of web sites. We have also gathered some requirements for tools that could provide much of this information.

7.1 Gamma Testing

We suggest the term “gamma testing” for a variation of beta testing focusing on identifying usability problems. We showed that this type of testing is useful for Web applications consisting of forms. A tool to support this type of data collection would generate rating questions based on information supplied by the developer about the components of the form. While we have not yet developed this tool, we have written a document about using this methodology.

7.2 Automated Testing

We conducted a usability test to see if specific information-seeking tasks could be automated. We found that a category-matching exercise was quite useful and could easily be automated. Our abbreviated usability test, collecting whether or not users were successful in carrying out a task and the time they needed to complete the task, can also be automated. We have released the first version of the NIST WebMetrics tool suite (<http://zing.ncsl.nist.gov/~webmet/>) which contains a category matching tool (WebCAT) and an automated path collection tool (WebVIP). We are currently working on visualizations to facilitate the analysis of the user paths obtained through WebVIP.

7.3 Server Log Analysis

For specialized Web sites, using server logs to obtain more information about the use and usability of the site is an excellent starting point. We found server log data

useful in giving indications of the relative amount of use of various portions of the site and in judging the possible effect of potential usability problems. We are currently designing a tool to construct approximate paths and provide appropriate visualizations for investigating potential usability problems.

8. FUTURE WORK

We based our designs for the tools on a few case studies with specific types of Web sites. As our first set of tools is now available, we are requesting feedback from the hundreds of developers who have downloaded the tools. We plan to document their experiences using the tools on various types of Web sites. Based on this feedback, as well as our continuing case studies, we plan to revise the existing tools and to design new tools to facilitate rapid, remote, and automated usability evaluations of Web sites.

9. REFERENCES

- Hartson, H., Castillo, J., Kelso, J., and Neale, W. (1996) Remote Evaluation: The Network as an Extension of the Usability Laboratory. *Proceedings ACM CHI'96 Conference*, (Denver, CO, April 13-18), 228-235.
- Jeffries, R., Miller, J., Wharton, C. and Uyeda, K. (1991) User Interface Evaluation in the real world: A comparison of four techniques. *Proceedings ACM CHI'91 Conference*, (New Orleans, LA, April 28-May 2), 119-124.
- John, B.E. and Marks, S.J. (1997) Tracking the effectiveness of usability evaluation methods. *Behaviour and Information Technology*, Vol. 16, no. 4/5, 188-203.
- Nielsen, J. (1993) *Usability Engineering*, Academic Press, Boston.
- Nielsen, J. (1989) Usability engineering at a discount. In *Designing and Using Human-Computer Interfaces and Knowledge Based Systems*, (ed. G. Salvendy and M.J. Smith) Elsevier Science Publishers, Amsterdam, 394-401.
- Stout, R. (1997) *Web Site Stats: Tracking Hits and Analyzing Traffic*. Mc-Graw-Hill, Berkeley, CA.
- Sullivan, T. (1997). *Reading Reader Reaction: A Proposal for Inferential Analysis of Web Server Log Files*. 3rd Annual Conference on Human Factors and the Web, (Denver, CO. June 12). For proceedings see:
http://www.research.att.com/conf/hfweb/conferences/prev_conferences.en.html

10. BIOGRAPHY

Dr. Jean Scholtz is currently a researcher in the Visualization and Virtual Reality Group at NIST. Her interests are in tools for evaluating software from the user perspective, primarily CSCW systems and Web applications. Dr. Scholtz has a PhD in computer science.

As a researcher at NIST in the Natural Language Processing and Information Retrieval Group, Laura Downey focused on evaluation and analysis techniques. She is the designer of several of the tools in the NIST WebMetrics tool suite.

Discussion

Kenneth Fishkin: The usability of a Web site can be affected drastically by the browser people are using. Can WebSAT accommodate this?

Jean Scholtz: Yes. Firstly, some of the guidelines are browser-specific, warning the user about features which may work on some browsers and not others. Secondly, the trace log contains browser information. So you can tell what browser was used if you use Webvip to log links.

Morten Harning: Do your tools support dynamically generated HTML pages?

Jean Scholtz: No, but they are next on the list.

Prasun Dewan: If you allow applets in your pages, then Web applications would not be a special case of interactive applications, since Java lets you write arbitrary applications.

Jean Scholtz: Yes! However, we would still need to modify our work as we currently apply only guidelines about informational-type sites.

Fabio Paterno: How do you plan to use information on usage of a Web site to improve its design?

Jean Scholtz: We plan to develop some models of web usage and compare the actual usage to what the designers intended to support.

Christian Gram: A pessimistic comment to user testing of Web sites based on an experiment where many students made "thinking aloud" tests on commercial Web sites: They agreed on a few (serious) errors, but all in all there was only little overlap. That means, each Web site contains a lot of errors, and it is impossible to find all - or just the major ones.

Michael Freed: Responding to Christians comment, I just want to point out that tools for detecting usability problems in a design have not to detect all such problems in order to be useful. Such tools can be very useful if, e.g., they sometimes draw attention to problems that would be obvious from hindsight, but are unlikely to be considered from advance. Also, such tools can complement designers' intuition by considering usability problems that cannot be predicted on the basis of common sense.