

On the Fast Track: New Usability Testing Methods for Web Sites

Laura Downey and Jean Scholtz

National Institute of Standards and Technology
Bldg. 225, A216
Gaithersburg, MD 20899
Tel: (301) 975-4659

E-mail: {Laura.Downey, Jean.Scholtz}@NIST.gov

ABSTRACT

The dynamic nature of the Web poses problems for usability evaluations. Development times are rapid, changes to Web sites occur frequently, often without reevaluating the usability of the entire site, and 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 identify the feasibility of using modified usability testing methods or nontraditional methods that meet our criteria of rapid, remote and automated.

Keywords

Usability testing, Web development, Web site design, Remote usability testing, Web site evaluation.

WEB DEVELOPMENT CONSTRAINTS

Web development places severe time constraints on developers and evaluators. There is a rush to get the Web site developed and released. The tight coupling of content, navigation, and appearance of Web sites means that separate evaluation of any one component is not meaningful.

Web sites change frequently. Hopefully, the portion of the site that is being added or changed is evaluated prior to release. 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.

Users' expectations change. New software and hardware developments are implemented in the Web sites of others. Ensuring that a given Web site is not outdated is also an issue.

The Web is worldwide. Getting representative users into a lab for a usability test is often not feasible. Moreover, users view Web sites using different types of browsers with different preferences and different types of Internet connections. Testing under all these different conditions adds to the time and complexity of setting up and conducting tests.

USABILITY EVALUATION METHODS

Previous research has compared usability evaluation methods and identified advantages and disadvantages of several techniques, including usability testing [1]. John and Marks [2] 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 [3] found laboratory testing of users to be the most effective source of information for identifying usability data. However, user-testing places limits on the type and number of users geographically available and does not allow us to view use in the context of other work activities and hardware/software configurations. In-house user testing is also expensive [1]. Our objective is to identify ways to gather data about usability problems of Web sites from real users without relying solely on in-house usability testing.

METHOD

We carried out an exploratory study to see how different traditional usability testing methods and some nontraditional usability testing methods can be used to evaluate Web sites in keeping with our criteria of rapid, remote and automated methodologies. As there are many diverse types of Web sites, we tested some hypotheses

about the types of testing methodologies that would be appropriate for particular varieties of Web sites.

Web Sites Selected

We selected three different services in the NIST Web site that we felt were representative of three different types of sites: a form fill-in site, a general-purpose library site, and a special purpose 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 work. In the next paragraphs, we briefly explain each site and its use.

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, notices of talks 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.

Previously, articles for inclusion in the technicalendar were faxed, phoned in, or e-mailed to a staff person. This person 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 both professional and administration personnel.

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 NIST phone books, maps to the Gaithersburg and Boulder campuses, and assorted other resources, including a clock and software that can be used to synchronize a PC clock.

The NVL site designers are considering various possibilities for redesign and are interested in any recommendations we can provide. 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 Matrix Market

One specialized service provided by NIST staff is a set of test data for comparing algorithms for numerical linear algebra. One of the mathematics groups has compiled a set of sparse matrices and matrix generators that can be

downloaded from their Web site. These matrices can take considerable time to download so the group is especially interested in ensuring that visitors can quickly and accurately locate the matrix they need. The users are mathematicians from all over the world.

STUDY ONE- THE TECHNICALNDAR WIZARD

Method

We felt that it was necessary to have users actually use this site in order to give useful feedback. We decided to use a modified beta test and determine if we could obtain usability feedback via this method. 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?
- Would users be able to describe usability problems?
- What types of usability problems would be identified this way and what types of usability problems would be missed?

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 baseline.

We then worked with the Web designer to setup a beta test. We constructed an evaluation form for users to fill out after they had used the Technicalendar Wizard. This consisted of nine rating questions, two open-ended questions, and five other questions, including demographic information. In addition, users were given the opportunity to submit a real item or a test item. Announcements about this test were placed in the Technicalendar. The Webmaster setup the site so that all evaluation forms were e-mailed to one of us, as were all test submissions. We were given access to the Technicalendar data repository on the Web in order to view the actual submissions.

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.

Results

Results are discussed in terms of the original three questions posed.

Question 1: Would users be willing to participate in the beta test?

During the first month of testing, there were 24 electronic submissions. Of these, 16 were real submissions and eight

were test submissions. Any given Technicalendar contains between 25- 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 about 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, five from the test submissions and 10 from the real submissions.

Question 2: Would users be able to describe usability problems?

Twenty-one usability problems were identified in the heuristic evaluation. Of these, five were fixed prior to the start of the actual test. The types of problems fixed were:

- How and where users were able to access help
- Supplying different types of help (calendar policy help as well as wizard help)
- How and where users were able to view information submitted so far
- Terminology

Of the sixteen remaining problems, beta users commented about six issues that were identified in the heuristic review. These comments were collected from the open-ended section of the questionnaire that we provided. The types of issues that users described were:

- Missing text fields
- Directions were included in the pull down lists, leading users to think a selection had been made when it had not
- Having to supply a field that users did not see as applicable
- Feedback about what would appear in response to a user selection
- Placement of item summary information

Users also 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. This problem was not uncovered during our heuristic evaluation.

Question 3: What types of usability problems would be identified through beta testing and what types of usability problems would be missed?

The heuristic review identified an issue with knowing which fields were required and which were optional. Users

did not comment on this but a question about identifying optional fields in the usability questionnaire received a lower rating than the other questions. Therefore, we could assume this was a problem for users.

A problem identified in the heuristic review was that the text field for phone numbers did not specify a format. While users did not comment on this, we found many variations in the input by looking at the data. Users supplied anything from ten digit numbers to a four-digit extension. As our area has recently gone to ten-digit dialing and the technicalendar is distributed to non-NIST employees, the phone number definitely needs to consist of all ten digits.

Another problem noted in the heuristic was that users were asked to make a selection that seemed to be out of order. They were asked about the type of their submission and when they wanted their item to appear in the Technicalendar in one of the initial steps. However, they were not asked the date of the event until much later in the process. While users did not comment on the order of this selection, they did note problems because they did not understand that there were interactions between the types of items submitted and when those items could be published.

This leaves seven problems noted in the heuristic that did not show up as errors in the user data and were not commented on by users. Five of these issues were cosmetic.

- Alignment of text fields
- Better grouping of text fields
- Consistent labeling on buttons in form
- Allowing keyboard navigation to move around in form
- Change the label from "reset" to "clear"

Two other issues noted in the heuristic had to do with user feedback and use of interface controls:

- Number steps on Wizard
- Inappropriate use of interface controls

These did not appear to cause problems for users.

Discussion

Beta testing was quite successful in this case. The best information came from the open-ended comments. Users commented directly on six of the issues identified in the heuristic and indirectly on a seventh. We identified one more problem by analyzing the submission. Another problem was indicated by the low ratings on a question in the usability questionnaire.

Our rating questions addressed the following issues:

1. Overall use of the wizard
2. Navigating between fields
3. Navigating between steps
4. Knowing what information to enter
5. Changing previously entered information
6. Knowing what fields are required
7. Understanding the terminology
8. The order in which information needed to be supplied
9. The value of the item summary

With the exception of the last question, these questions are general enough to apply to most wizard applications.

The difficulty in using the wizard to input information for exceptional types of items was identified in comments and also by analyzing the actual items submitted.

The response from users was good. We did not uncover any new problems in the second round of testing. And in this case, we were actually able to use the ratings in the usability questionnaire to verify that our redesigns resulted in improvements. A side benefit of the beta testing and formal evaluation was that exactly one complaint was received when users were required to use the wizard for all Technicalendar submissions.

STUDY TWO: THE NIST VIRTUAL LIBRARY

Method

The NIST Virtual Library (NVL) is a scientific library accessible to the public from the NIST Web site. While some of the data are restricted to NIST personnel, most of the library resources are open to the general public. The NVL is accessible via the Web. Therefore, we were interested in designing testing that had the possibility of being carried out remotely. We were also interested in ways to decrease the time needed to design, run, and analyze the test. So, we were looking for test methodologies that would lend themselves to automation.

The usability test consisted of three parts: a matching exercise to test existing categorization, 10 representative tasks, and a short demographic and satisfaction questionnaire. We recruited five subjects from different scientific disciplines who worked at site in Gaithersburg, MD.

We needed a benchmark to compare the results of our subjects. We had two experts complete the matching exercise and the ten tasks. 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.

Our matching task was a variation of a traditional card sorting task [3]. In a card-sorting task, users supply names for actions and objects in the interface, group them into categories and then label the categories. As we were not starting from scratch in our design, we used a matching task. Our goal was not to redo all the categories and labels but to determine if any categories were troublesome.

In the matching task users were asked to match 29 items to one of 10 choices, nine categories from the NVL home page plus a "none" category. We scored users' responses according to:

- The number of items misidentified
- The number of times a category was misidentified
- The number of users per category who misidentified that category.

We wanted to determine if a matching task could be used to indicate the usability of categories in Web site.

In the task performance section, we concentrated on tasks that required users to locate information. Our goal was to see if we could collect a bare minimum of data and still identify usability problems. For the 10 representative tasks users were asked to do, we collected:

- Whether users found the answer (yes/no)
- The time it took
- Users' perceived difficulty
- Users' perception of the time for completing the task

We wanted to see if this simplification of usability testing could give us information about the usability of the Web site. This basic information could be obtained using remote, asynchronous testing methods. That is, we could send an instrumented version of the Web site along with instructions for doing the tasks to participants and have the results transferred back to us electronically.

We also had users complete a satisfaction questionnaire consisting of 10 questions. Of these 10 questions, six were rating scale questions, 3 were demographic questions, and the last was an open-ended question for users to comment about the interface and tasks. Again, this questionnaire could have been given remotely, using an electronic delivery method.

Results

The Matching Task

Our baseline users misidentified 2/29 items. Our non-expert subjects misidentified 13/29 items, which clearly indicated that the categories were a problem. Out of the

nine categories, all of the subjects had problems with three of them. The troublesome categories were: databases, hints and help, and NIST resources. Four subjects had problems with two other categories. Three subjects had problems with two more categories.

This clearly indicated to us that the categories used in the Web site were not clear to our users. Moreover, this type of test could be automated quite easily and could be administered remotely also. What we miss, of course, is getting input concerning possibilities for category names that are more in line with users' expectations.

The Performance Test

Our expert users were able to do 9/10 tasks. However, each expert user missed a different task. Our five non-expert users were able to complete between six and seven of the ten tasks. Three users completed six tasks and the other two users completed seven 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 some interesting issues. 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. There were three other tasks that only two users were successful in doing.

Users rated the difficulty and time factors for the tasks quite high given the success and time they needed to complete these tasks. A seven-point scale was used, with 1 being an unacceptable rating and 7 being an excellent rating. Experts gave a difficulty rating of 5.7/7 compared to the non-experts 4.8/7. Ratings of the time it took to accomplish the task were 5.8/7 for the experts compared with 4.8/7 for the non-experts.

Originally, we had just intended to use success or failure in completing the task. However, we found instances where users thought they had located information but had not. Therefore, recording the actual answer for comparison is necessary.

The user ratings of task difficulty and time were very closely correlated. There was only one instance where the two ratings differed by more than one point. And in that case, the ratings differed by two points. That user found the time needed to complete the task more acceptable than the perceived difficulty of the task. Thus it seems that a perceived difficulty rating for the task alone is sufficient. However, the ratings for difficulty don't necessarily reflect success in the task as users sometimes did not realize that they had not been successful.

Discussion

Because we were not actually conducting this test remotely, we were able to observe users and interview them after the test. We asked them retrospectively to talk to us about what they did in some of the problematic tasks. We video taped the screen while users were carrying out the tasks and we used this videotape to help users remember what it was they did. 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., which category to begin looking under. If they did know the category, then they preferred to use that.

We also noted that users used the category icons in the menu frame to jump to those pages, rather than using the links within the home page. We observed that users often missed seeing links within pages that they should have located. In the interviews, users suggested that it was helpful to have alphabetical ways of viewing pages that were not familiar. When the material was familiar, then grouping was useful; assuming it was done correctly.

We feel that remote and automated test procedures are, not only feasible, but will yield valuable information. The matching task is easily automated and this gave us information about confusion with the current category groupings. Automating the usability test for remote users is also feasible. However, it is necessary to collect the answer if users are asked to locate information, as they are not always aware of whether they have done the task correctly. Being able to collect the path that users take to locate the information is also extremely beneficial. If we use programs that allow the tester to view the participant's screen and maintain an audio connection, we can also collect information about why users employ the strategies we observe.

STUDY THREE: THE MATRIX MARKET

Methodology

The Matrix Market is a very specialized site used primarily by mathematicians using or developing algorithms in numerical linear algebra. The site contains sparse matrices to use in testing algorithms in addition to a procedure for submitting matrices for inclusion in the site. Information about the NIST staff and other contributors is also available at the site.

We felt that finding domain experienced users to test would be difficult. Because of the specialized nature of this site, we felt that constructing scenarios and conducting usability testing would result in artificial results. In this instance we felt that trying to obtain as much information from server logs as possible would produce the most *inexpensive results*. This assumes, of course, that easy ways could be built to obtain information from server logs.

We recognize the numerous problems with using server log data as the sole source of information [4]. 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 [5] describes the use of server logs to provide inferential statistics about Web site usability.

Heuristic

We first did a heuristic review of the site to use as an indicator of potential problems. In the heuristic, we identified 17 problems that we classified in five categories. These categories are described below.

Consistency

1. *Behavior problem*: Users could browse for matrices by name, a collection type, by application or by contributor. However, the browse functionality differed depending on what the user was browsing by.
2. *Terminology problem*: Inconsistent uses of terminology occurred between link names and text in general information pages.

Navigation

3. *Scrolling problem*: There were several instances of long lists of matrices in which users had to scroll to reach either the top or bottom of the page.

Easy Access to Information

4. *Discrimination problem*: One potential problem was a page of around 500 matrix names, each of which was a link, arranged in 6 columns. These names ranged from two alphanumeric characters to eight alphanumeric characters. Often names differed only by one digit from the name above or below it.
5. *Arrangement of groups*: The home page of the site had nine groupings of information plus five graphic buttons embedded in a large graphic. Therefore, not all the groupings were visible if users had all the toolbars on the browsers turned on.
6. *Extra step to access information*: Users were able to have matrices created dynamically. However, accessing any of the pages for dynamically created matrices caused the display of a page explaining which versions of browsers supported access to these matrices. This page was displayed each time the user tried to access a dynamically generated matrix, although the same information was given for each.

Prevent User Errors

7. *Long search form problem*: The entire search form could not be viewed without scrolling. It also contained groupings separated by horizontal bars. We felt that users might not see the entire form and have to make several attempts at searching, having missing some of the options that could be filled in. A "submit query" button was located at the top of the form as well as at the bottom of the form so users did not have to scroll to the bottom of the form in order to submit the query.

User Feedback

8. *Download information problem*: Users were told the number of bytes for each matrix. We were interested if users would be able to translate this into download times or if they would become impatient and stop downloading the files.

Results

Our intent is to see if we can make use of server log data to indicate usability problems. For the purposes of this study, we analyzed only one month of server log data. 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 also tried to see if the questions we were asking could be generalized to apply to other Web sites.

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.

Overall Use

First we wanted to get an idea about traffic and the number of users visiting the site. 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, leading us to conclude that users did not have major problems with the site.

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 was 19%, 14%, 9%, 6%, 5%, and 4% respectively. 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.

Comparing server log data with heuristic results

In this section, we'll discuss the server log data that we felt could be used to indicate if the problem identified in the heuristic evaluation was actually a problem encountered by users.

Consistency

1. *Behavior problem:* The two instances where the "browse" worked differently were used in 9% and 4% of the visits. We found that fewer than 8% of the visits used both types of browse methods. Therefore we can assume that this inconsistency is not a large problem.
2. *Terminology problem:* We found that only 5% of the visits accessed the help functionality so we decided that this was not a severe problem. We could also have looked to see if the percentage of users accessing help first did so after accessing a general reference page or a technical page.

Navigation

3. *Scrolling required problem:* We decided that this was indeed a problem that needed further investigation as we found 20% of the visits accessed this page. We can now look further to see if information accessed within that page is currently arranged at the beginning or end of the page.

Easy Access to Information

4. *Discrimination problem:* Users use this page to access data pages about matrices. We found that 52% of visits accessed at least one data file and 29% of visits accessed more than one data file. We speculated that if the percentage of users accessing multiple data files from this page were much higher than the average, this would be a problem to investigate further.

However, we found that only 7% of the visits that accessed this page accessed more than one data file. This leads us to believe that the users who use this page to access information about matrices are highly skilled and have few problems in discriminating between the matrix names.

5. *Arrangement of groups:* We were investigating the arrangement of the groups of information on the home page. We found that almost 70% of the visits accessed the home page at least once. Over 50% of the visits accessed at least one matrix data page. We also found that visitors most frequently accessed the matrices using the two pages listed first in that group. This group was in the center of the groupings on the home page. So we concluded that the arrangement provided sufficiently easy access to information for the majority of users.
6. *Extra step to access information:* We found that only 5% of the visits accessed matrices this way. Therefore, we concluded that the problem was not severe.

Prevent User Errors

7. *Long search form problem:* We speculated that users who were confused about the search form might perform another search immediately after one that did not yield useful information. However, due to caching we are unable to see a return to the search page itself. To investigate this problem, we need to include the search scripts that we had eliminated to simply the log data. We have not yet completed this analysis but we intend to look at successive accesses to search scripts to determine if users actually have problems with the search procedure.

User Feedback

8. *Download information problem:* In order to note the magnitude of this problem we could look at the percentage of times that a user stopped the transfer of a matrix download. We would then compare this to the percentage of downloads requested. In order to obtain this data, the server access log file and the error log file need to be merged by time stamp, as the error log does not indicate the name of the file being transferred when the stop transfer request is received. We have not yet completed this analysis.

Discussion

We believe that for specialized Web sites, server log data can be used effectively to gain insights about potential problems of use. The questions we researched in the data can be generalized to provide the following information.

- From the home page, which links are most frequently used?
- Do users have a difficult time discriminating between names of links?
- Do users have difficulty locating information via searching and need to make multiple attempts?
- Is the time it takes to download information on a site acceptable to users?
- Do visitors use help frequently?

Server log analysis also allows us to estimate the percentage of users that a potential usability problem affects.

CONCLUSION

We believe that usability evaluation techniques that will prove effective for the Web must be rapid, remote and automated. We have investigated different types of usability evaluation techniques that meet these criteria. We have used three case studies of different types of Web sites to show the different techniques that can provide useful usability information for each.

Gamma Testing

We suggest the term "Gamma testing" for a variation of beta testing focusing on identifying usability problems. This type of testing is useful for Web applications consisting of forms. The open-ended question provided us with the best information in this case. A short usability questionnaire and analysis of the submitted data was also useful. Letting users submit "test" data to tryout the form is a way to increase participation.

Automated Testing

We conducted usability testing to see what 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. A backend to automatically analyze the resulting data can also be developed. This type of automated testing can be done remotely using software that allows the tester to view participants' screens and to maintain audio connections. However, even asynchronous testing can provide useful data from a more diverse set of users than it would be feasible to test in a lab setting.

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 to give us indications of the relative amount of use of various

portions of the site. This can also be used to judge the possible effect of potential usability problems. By constructing an approximate path for users, we were able to determine if users had discrimination problems with links and identify instances where behavioral inconsistency was or was not a problem. We also speculated that we would be able to obtain information in the user path about possible confusion with a search form.

Future Work

The rapid, remote and automated methods for usability evaluation that we have espoused here are currently in construction. We are continuing to conduct case studies to investigate other types of usability evaluations that are possible, as well as trying out methods on different types of Web sites. We are developing tools to instrument Web sites and provide automated data collection and analysis. We are developing queries and visualizations for Web sever access log data that can be used as indicators of possible problems. These tools will exist in a suite of tools that also includes an html analyzer for locating potential usability and accessibility problems. Once working prototypes are developed, we will continue to conduct case studies to determine the usefulness of rapid, remote and automated testing tools.

ACKNOWLEDGMENTS

Special thanks to Charles Sheppard for all his efforts in obtaining the server log data used in our analysis.

REFERENCES

1. Jeffries, R., Miller, J., Wharton, C. and Uyeda, K. 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.
2. 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.
3. Nielsen, J. 1993, *Usability Engineering*, Boston, MA : Academic Press:
4. Stout, Rick. 1997. *Web Site Stats: Tracking Hits and Analyzing Traffic*. McGraw-Hill, Berkeley, Ca.
5. Sullivan, Terry, 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) <http://www.uswest.com/web-conference/>.