Adaptation of Traditional Usability Testing Methods for Remote Testing

Jean Scholtz
National Institute of Standards and Technology
100 Bureau Dr., Stop 8940
Gaithersburg, MD 20899-8940
Jean.Scholtz@nist.gov

Abstract

Traditional usability testing methods are difficult to use in producing web sites and web applications mainly because of the decreased development times that companies demand for this type of software. Users of web sites have diverse platforms, computer expertise, and expectations. Companies want to use web sites to sell merchandise and provide services to customers. Therefore, it is essential to make usability a high priority in the development of web-based software. How can we resolve this seemingly contradictory situation? We believe that usability testing tools that are remote, rapid, and automated will be helpful in providing more usability information in a shorter time and in a form that can be immediately useful to usability professionals.

In this paper we discuss the approach we have taken to designing such tools. We currently have several tools available for public use which will also be discussed in this paper. Our next steps will be to conduct methodological studies to validate the use of these tools.

1. Introduction

The World Wide Web is expanding rapidly. According to www.netcraft.com/survey/Reports , there were 3,997,270 .COM sites as of September 1999. As of August 2000 there were 11,332,014 .COM sites. As more and more information becomes available, individuals will use the web for many more purposes. Students and scholars use the web for research. Individuals use the web to order books and clothes, make travel arrangements, send flowers, read newspapers, track investments, etc. Companies use the intranet to track orders and to disseminate information to employees. However, many times web users are frustrated because they are unable to locate the sites or the information they want on a particular site.

User-centered design and usability testing of user interfaces are being applied in some traditional software development processes, but are not yet commonplace. Issues of time, resources, and perceived value are obstacles to be overcome in moving usability into the software development cycle. Moving usability engineering procedures to the development of web sites and web applications is even more difficult. The normal software development time has been shortened considerably for web site development. Most companies want a web presence within three months (Vora, 1998). Usability testing is suited to an iterative design environment. Cycles of design, usability testing, and reiterating the design are difficult to carry out in reduced time scales. Sites and applications developed for the web become "unusable" for external reasons. That is, traditional software is tested when changes are made to the software. Web sites may become unusable because of outside forces. For example, a new capability becomes available and other sites that use that capability become more desirable than sites lacking the capability. A site that was "state of the art" yesterday may no longer be appealing. Finally, usability testing is conducted using "typical users" doing "typical tasks" with the software running on "typical platforms." Typical user populations, user tasks, and platforms for web sites constitute a very broad range. There are too many variations of types of users, tasks that can be carried out on a site, and different configurations of machines and browsers to attempt to bring a representative sample to a usability lab for testing. One solution is to recruit volunteers for testing from the current users of the site. As this will likely produce users from many geographical areas, it becomes costly to either bring these users to a central location for testing or to send usability professionals out to conduct usability studies in various geographical regions. Remote testing methods are a possible solution to this problem.

1.1 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 (such as heuristic reviews and cognitive walkthroughs) to laboratory usability tests with actual users 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. However, in-house user testing is expensive and places limits on the type and often, 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' actual hardware and software configurations. This can be especially important when evaluating the usability of web sites. Unlike much traditional software, informational web sites are often used in the context of carrying out another task.

Users of web sites are remote from the site itself. Therefore, remote usability testing quickly comes to mind when thinking of ways to test web sites. Remote testing can be done synchronously or asynchronously. Synchronous testing can be done 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.

Qualitative data is certainly desirable in the redesign of a site. However, quantitative data is also useful in identifying areas where usability problems exist. In our work, we focus on what quantitative data can be collected in a remote, automated, and rapid fashion to identify usability problems. Traditional user-centered design and usability testing methods must be adapted to the web development cycle. We maintain that those methods must be rapid, remote, and automated. The goal of the WebMetrics project (NIST WebMetrics, 1998) at The National Institute of Standards and Technology (NIST) is to produce a methodology, complete with tools and techniques, for producing usable web sites and applications. In this paper we discuss the process we used in producing our first set of tools for remote testing.

2. Our Approach to Tool Adaptation

We have been fortunate to have access at NIST to many different types of web sites and applications. The groups at NIST are extremely eager to obtain whatever information we can provide them about the usability of their web site in exchange for allowing us to experiment with different testing methods on their site. In each case study, we have taken a traditional method from user-centered design or usability testing and modified it for remote testing. Then we applied the modified method in a usability study of a particular web site. Based on the lessons learned, we developed a tool to capture usability data when using the method remotely.

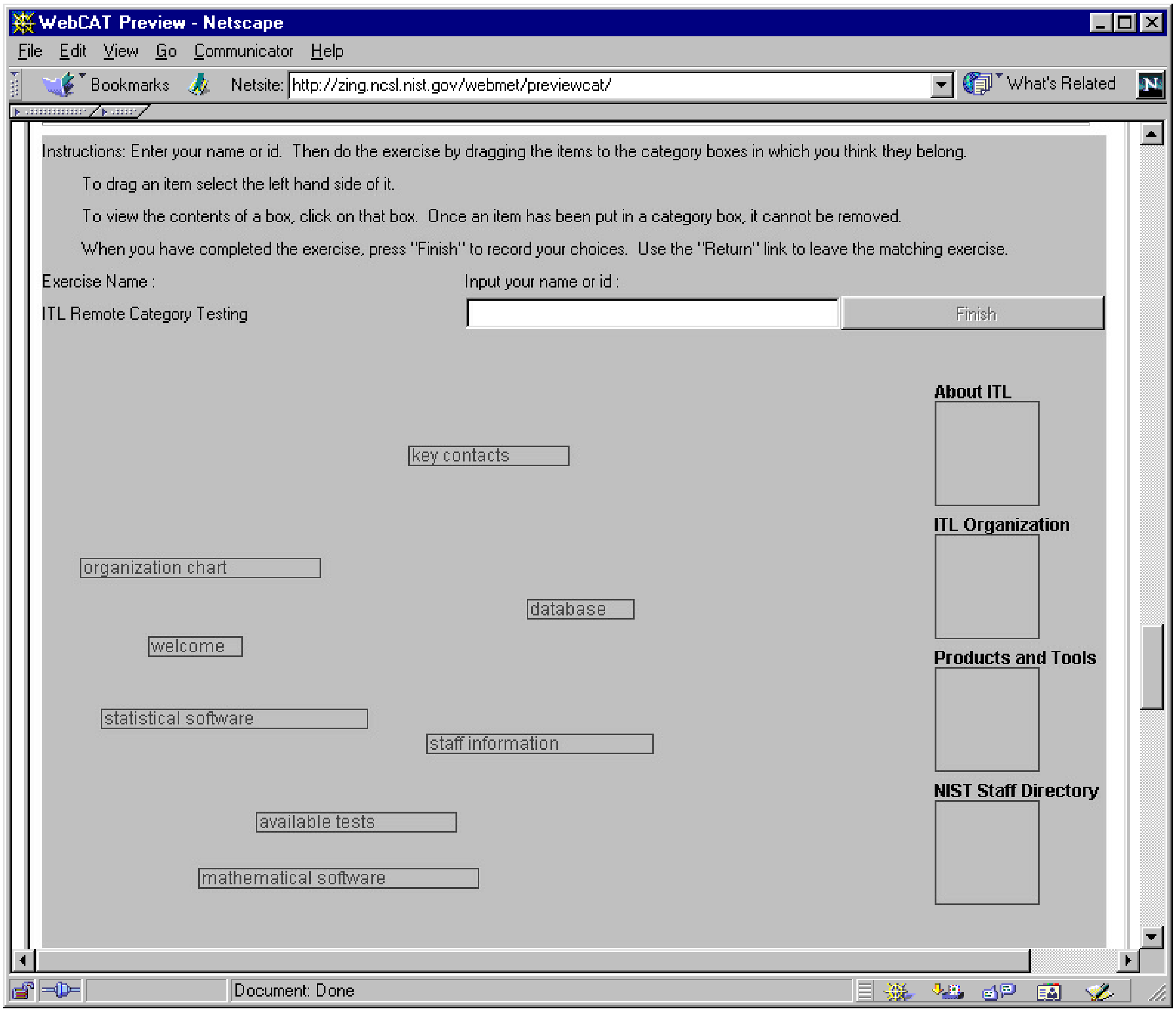
3. The Tools

In this section, we describe three tools we have produced so far and released to the general public. We also describe a small case study using the tool or the case study that led to the development of the tool. In the case study we focus, not on the results of the case study, but on the process we used in the case study.

3.1 WebCAT (The Web Category Analyzer Tool)

Often web sites are organized by defining a number of subcategories of information. One usability test is to determine if the names for these subcategories are intuitive to users. That is, can users correctly identify the subcategory in which various types of information would be located. In the laboratory, we might accomplish this by giving users a list of subcategories and another list of information needs and asking them to match these – either verbally or on paper.

To do this type of testing remotely, we developed WebCAT. Using this tool, a usability engineer can construct a short exercise to test out existing or proposed categories. The engineer specifies the categories and the subcategories. Then he/she performs the exercise (much like a matching exercise) by dragging the subcategories to the properly labeled category box. The usability engineer's results are used as the baseline with which to compare subjects' results. After the baseline has been established, the usability engineer can send out the URL where the created exercise is stored to individuals who have agreed to participate in the study. Results are automatically collected as individuals perform the exercises. The usability engineer can view the collated results at anytime. The exercises created using WebCAT can be run remotely or can be used to automate the test and data collection in the usability lab. Figure 1 shows the user's view of a typical WebCAT exercise. Category boxes appear on the right of the screen. Items appear on the left side of the screen. A user simply drags and drops items in category boxes to indicate where he/she believes this information would be located. The usability engineer completes the exercise to generate a baseline set of data to use in the automatic analysis. Raw data is also provided to the usability engineer in spreadsheet form.



Instructions: Enter your name or id. Then do the exercise by dragging the items to the category boxes in which you think they belong.

To drag an item select the left hand side of it.

To view the contents of a box, click on that box. Once an item has been put in a category box, it cannot be removed.

When you have completed the exercise, press "Finish" to record your choices. Use the "Return" link to leave the matching exercise.

Exercise Name : ITL Remote Category Testing

Input your name or id :

Figure 1 An exercise created by a usability engineer to test categories on a NIST web site

3.1.1 The NIST Virtual Library: A Case Study for Category Matching

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. In this section, we focus on the matching exercise as this is what led us to build WebCAT.

We recruited five subjects from different scientific disciplines who worked at the NIST site in Gaithersburg, MD. Although we did NOT conduct this test remotely, we designed the test so that, given the appropriate software, it could be conducted remotely.

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 collected the results of this variation of a card-sorting task (Nielsen, 1993). As a benchmark for comparing the results, we had two experts complete the matching exercise: a reference librarian at NIST who was very familiar with the NVL site and the designer of the NVL Web site.

Table 1 shows some of the categories and items from the NVL which were used to devise our matching test. Category matching can easily show categories that are too broad or confusing for users. This is an easy test to conduct and could even be linked to from a current web site if the designers want to obtain data to use in a redesign.

Category	**Items**
Subject Guides	Biotechnology, chemistry, diversity, engineering
Visiting NIST	NIST Tour Information, Street map of NIST
Web Resources	List of Federal Library Web sites, Weather, Phone/ email directories
E-Journals	Elsevier Science Tables of Context, Proceedings of the National Academy of Sciences Online
NIST Publications	NIST Research Library Online
Hints and Help	Goals and Mission of NVL, about Frames
Databases	CD-ROM Databases, Library of Congress

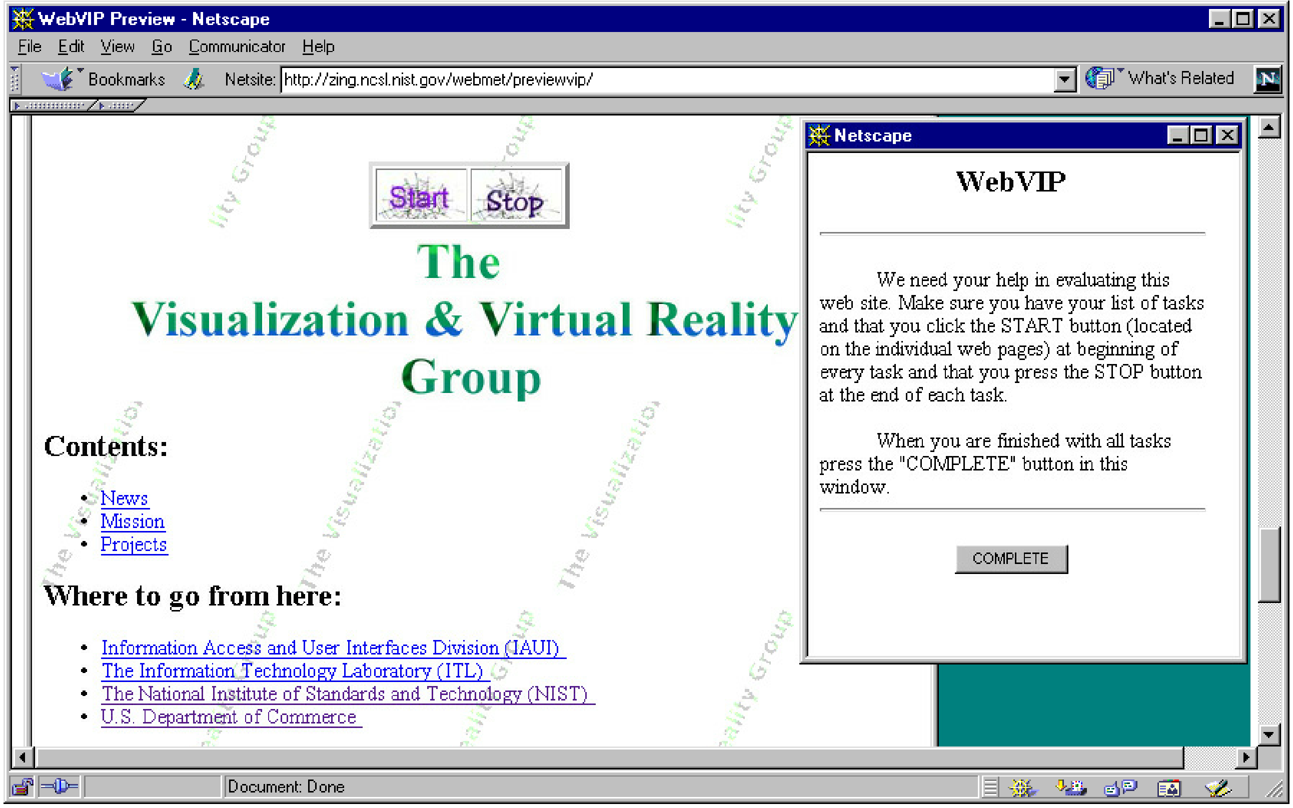
Table 1 A Sample of categories and items in the matching exercise

In a subsequent experiment, we conducted the matching exercise remotely using e-mail. This experiment produced similar results and verified that remote experiments were feasible.

3.2 WebVIP (The Web Visual Instrumenter Program)

Many usability professionals use video data to record the computer screens in a usability test. Suppose we want to see how users navigate to a particular section of a given website. A typical procedure would be to give users a task that necessitates getting to a particular portion of the web site. Their navigation path could be recorded either on videotape or by an observer who would record the links selected by the user. To implement this type of test remotely, some way of capturing the navigation paths (including the use of the back button) is needed.

WebVIP automates data collection of user paths during usability testing. This tool allows a usability engineer to specify which links in a web site should be recorded and time stamped as the user goes about doing specified tasks. This instrumentation is performed on a copy of the web site that is given only to individuals who have agreed to participate in a usability session. WebVIP inserts visual symbols on all the links on a web site and the usability engineer specifies, by clicking, the links to be logged. The instrumented site is then saved and the usability engineer can send the URL of this site to users who have agreed to participate. As users carry out specific tasks with the web site, their paths, including the use of the Back button, are saved in log files. Each link that is visited is saved in the file with a time stamp. A new file is created for each user. These files can then be analyzed by the usability engineer to determine whether users were able to navigate successfully within the site. Again, usability testing using WebVIP can be conducted remotely or used in the usability lab to facilitate data collection. Figure 2 shows the user's view of an instrumented site. Start and stop buttons are located on the top and bottom of each web page. A separate browser window contains the WebVIP instructions and a complete button for users to select when they are finished with the entire set of tasks.



Netsite: http://zing.ncsl.nist.gov/webmet/previewvip/

The
Visualization & Virtual Reality
Group

Contents:

- News
- Mission
- Projects

Where to go from here:

- Information Access and User Interfaces Division (IAUI)
- The Information Technology Laboratory (ITL)
- The National Institute of Standards and Technology (NIST)
- U.S. Department of Commerce

WebVIP

We need your help in evaluating this web site. Make sure you have your list of tasks and that you click the START button (located on the individual web pages) at beginning of every task and that you press the STOP button at the end of each task.

When you are finished with all tasks press the "COMPLETE" button in this window.

Document: Done

Figure 2 An instrumented web site

3.2.1 A Case Study for the Development of WebVIP

In our case study for the NIST Virtual Library, we conducted a performance experiment as well. We concentrated on tasks that required users to locate specific information. Our goal was to see if we could collect a bare minimum of data and still identify usability problems. We collected the following information as the users performed representative tasks:

- Whether users found the answer (yes/no)
- Elapsed time
- Users' perceived difficulty

- Users’ perception of the time for completing the task
- Users’ navigation paths

We simulated a remote test by limiting the interaction with the experimenter. We collected the users' paths by video taping the screen and replaying the tape to obtain the links users followed. We gave users a rating sheet to mark if they had found the answer and to rate the difficulty of obtaining the answer. The time was obtained by the user specifying "ready" and "stop."

Ratings were not reliable. Users often did not rate the tasks they missed as the most difficult, probably because many of them thought they had located the answer. This indicated to us that we needed to collect the answers to information seeking tasks to ensure that users were successful.

Looking at the paths that users took to locate information gave us quantitative data about different strategies that users took. We were also able to identify category names that were misleading to users. This data was the most useful to us in deciding what portions of the web site needed to be redesigned. We decided from this study to concentrate on a tool that logged user paths automatically. Time data is collected as each link is time stamped. Overall task time is also collected using "start" and "stop" buttons incorporated into WebVIP. Presently the answer must be collected separately from WebVIP but we plan to incorporate the collection of this information at a later date.

3.3 VISVIP (A Visualization Tool for Analyzing WebVIP Data)

VISVIP (Cugini and Scholtz, 1999) is a visualization developed to help usability engineers analyze the data from WebVIP. VISVIP generates a 3D visualization that the usability engineer can manipulate using a mouse or space ball. In addition to the visualization window, a control window allows the user to refine the visualization in several ways.

To generate the visualization, we use the free software Linklint (Linklint, 1997) to determine the static structure of the web site. The web site is then depicted as a directed graph, with pages as nodes, and hyperlinks as edges. VISVIP automatically generates a 2D layout of the graph, using a force-directed algorithm. In our model, adjacent nodes exert a spring-like force on each other - they "try" to set themselves apart at a fixed distance. Non-adjacent nodes repel each other with a force inversely proportional to the distance between them. Finally there is a third force that weakly attracts all nodes towards the origin; this serves to keep parts of the graph that become unconnected from flying off to infinity.

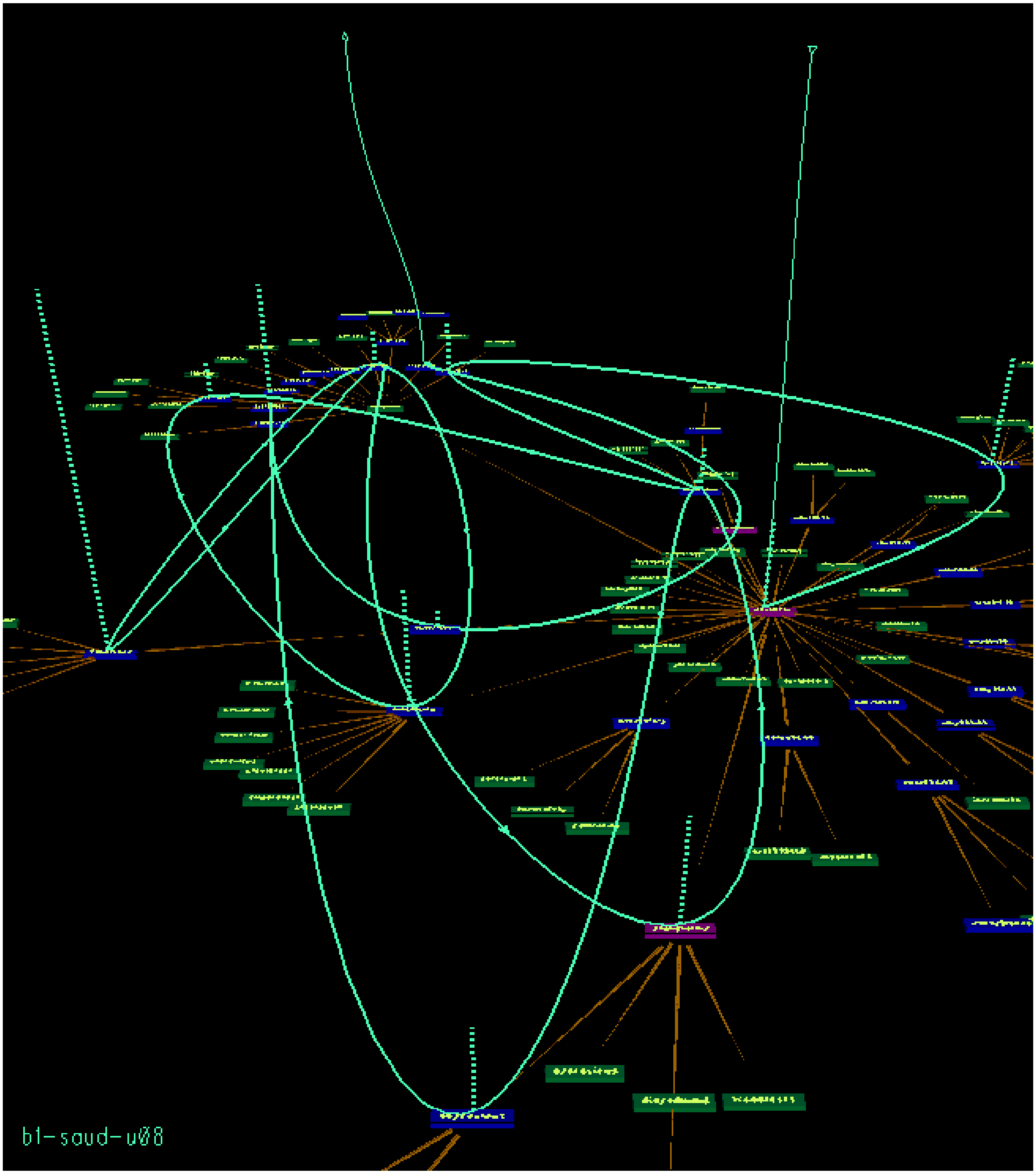


Figure 4 A user path displayed on a partial view of nodes of the web site

The usability engineer can dynamically adjust the constants of the underlying force model so as to control the overall density and appearance of the graph. Also, for finer control, the usability engineer can drag and drop individual nodes into precise locations as desired.

After conducting an experiment with an instrumented web site, the usability engineer uses VISVIP to display the paths of the subjects. Each subject path is depicted as a spline overlaid on top of the directed graph. Each subject is coded with a distinct name and color. VISVIP also allows the usability engineer to specify the colors manually so as to examine groups of subjects (e.g. novices vs. experts). The spline curves are decorated with arrowheads to indicate direction, and a special curvy arrow into and out of the plane of the graph highlights the starting and ending point of the path. The usability engineer may dynamically select any number of user paths to display simultaneously. Figure 3 shows a typical VISVIP visualization of one subject's path through a web site. The dotted vertical bars on the graph indicate the time spent at those nodes. You can also view this at the NIST web site (http://www.itl.nist.gov/iaui/vvrg/cugini/webmet/visvip/vv-home.html).

3.3.1 VISVIP: Anecdotal Use

We conducted a usability study of two web sites and we used this data to explore working with our visualization tool. We intend to design some formal studies later to compare traditional methods of analysis with a visual analysis. These exploratory studies will be used to refine the visualization tool.

In our first efforts to use VISVIP, we found it useful to initially view user paths against all nodes of the web site. This gives the usability engineer a good indication of the portion of the site that was actually used. After isolating the portion of the site used, visualizing only those nodes visited by at least one user makes the visualization less complex and easier to locate patterns of use. At this point, circular paths are easily seen. It is more useful to start by viewing the paths of all users than by viewing one path at a time. By seeing all paths, the usability engineer is able to see individual deviations easily. These can be explored by restricting the view to only those paths. VISVIP supports playback of user paths, in real time or in a speeded up time frame selected by the usability engineer.

By zooming in on the 3D view, the usability engineer can read the URLs of the various nodes. When the node is clicked on, statistics about the users who visited are displayed in a browser window. Double clicking on the node displays that web page in the browser. To make it easier for the usability engineer to see if particular nodes are visited, VISVIP allows a node to be "marked." Marking highlights the node so it is not necessary to zoom into the display to check the URL.

By titling the entire visualization it is easy to view times spent at nodes. This gives a good indication of nodes where all users spent a lot of time. This might either indicate that they found useful information there or they were confused about what to do.

3.4 Requirements Collection: A Virtual Participatory Design Meeting

Design meetings in which users participate are often used to gather requirements for a project. Design meetings take time to run and time to prepare for. Holding asynchronous virtual design meetings will allow users to participate when they have time and when they think of a requirement. Virtual meetings allow us to expand the number of users we can include. This could be extremely beneficial in designing a site for use within a large company.

We used a virtual design meeting to gather for the redesign of a major NIST on-line resource – the NIST Virtual Library (NVL).

One of the interesting issues with a library site is that there are two very different categories of people who need to be considered when redesigning such a site. While we are, of course, interested in the users of the library user and how well the site meets their needs, we also need to consider the impact of redesign on the library staff. Much behind the scenes work is still needed to make a virtual library "virtual." We wanted to ensure that we considered first of all, input from library staff for the redesign. The library staff already has a tremendous amount of work and due to the various hours they work, scheduling meetings is difficult. But it was important to make sure that we gathered as much input from them as possible. We also felt that people are much more likely to think of issues affecting the design one issue at a time and these ideas arise because of something that is happening at work that moment. How could we make sure that we captured that information in a timely way?

We decided to collect requirements in the form of tasks or scenarios and to collect these scenarios from the library workers via the intranet as all of the staff has easy access to this during their work. We scheduled an initial meeting with the staff and explained what we wanted to do. We then had a period of several weeks during which the staff contributed their scenarios using a prototype tool accessible from the intranet. We supplied a template for the scenarios. We allowed the staff to comment on others scenarios anonymously and also made it easy to see which scenarios others had commented on. We supplied several example scenarios to help everyone get started. Automatic e-mail notification was used to alert staff when a new scenario was posted.

For each scenario, the submitter was asked to include a description, identify the benefits (speed, accuracy, not currently possible, etc.), and who would benefit (library reference staff, end users, other library staff, etc.). We asked the submitter how frequently this scenario would happen and its importance on a 7-point scale. We also asked for any negative aspects to this scenario.

3.4.1 Results

We received 28 scenarios. Of these, 18 also included comments from one or more participants. After the collection period was over, we classified the scenarios into basic categories, allowing scenarios to be in more than one category. We used these categories and scenarios to construct initial requirements for the revised NVL. Unfortunately, the redesign of the NVL was postponed to a later date. When the project is restarted, we will use these scenarios in the design of our usability evaluations. We also plan to collect scenarios from end users as we redesign the end user portion of the virtual library.

We felt it was helpful to enter some data in the tool initially. This served as an example so that users could easily recall what type of information to enter in the template. Automatic e-mail notification was helpful as it reminded users to enter their own requirements. We pilot tested the template with several participants to make sure it was easily used and understood.

We intend to try this method for several other redesigns and then we will design a tool that usability professionals can use to generate remote requirements collection.

4. Conclusions and Future Work

We have used these tools in several different case studies (Scholtz, Laskowski, and Downey, 1998; Scholtz and Downey, 1999). We are currently working on modifications to our current tools. For example, WebCAT will be modified to include the typical card sorting method where users name the groupings of the categories. Cluster analysis will be provided for this type of test.

WebVIP will be modified to present user tasks, collect demographic questionnaires, user comments, and satisfaction ratings. In addition, the latest version of WebVIP will allow typing, mouse movements, and scrolling to be logged. The tool will contain templates for these that usability professionals can easily customize.

Currently, VISVIP is being modified and will include new views and functionality for usability engineers. We have only a few uses of VISVIP but we are designing formal studies of this tool as well as using it for analyzing our current case studies to determine its usefulness.

We have not yet implemented a generic requirements collection tool. We plan to experiment with other web site designs to understand the functionality needed in such a tool.

The next step is to demonstrate the validity of these tools. We will approach this in two steps. First, we have placed the finished tools in the public domain. We invite other usability professionals to use them and to provide us with feedback and data about their usefulness. We will also design and conduct a series of methodological studies to assess the usefulness of these tools over traditional usability methods for web evaluation.

While there is no substitute for laboratory user studies to provide qualitative data for usability assessments, we are convinced that remote testing methods provide valuable additional data. Remote testing can be used to identify areas of web sites that are problematic, difficult user tasks, or even user populations that should be considered for detailed usability testing.

Remote testing can be used to assess a current site or it can be used when a new site or a new version of a site is being developed. A link from the current site to a remote evaluation test might be setup to solicit input from current users. This is valuable in tow ways. First, it provided information about the usability of the new site. Second, it helps prepare users for a change. It is difficult to drastically change a heavily users web site as users are often frustrated when navigation paths and shortcuts they have finally figured out are changed. For web sites that have a global user population, remote assessment and analysis methods are necessary to continually evaluation global usability.

5. Acknowledgements

The author would like to thank the many people who helped on these various projects. Sharon Laskowski had many helpful comments and as manager of the group provided the resources to do this work. John Cugini designed and implemented VISVIP. Laura Downey (now at Vignette Software) designed WebVIP and WebCAT and Paul Hsiao implemented them. Bill Trefzger helped with the case studies of the NIST Virtual Library and the requirements data collection.

6. References

1. Cugini, J. and Scholtz, J. (1999). VISVIP: 3D Visualization of Paths through Web Sites. The International Workshop on Web-Based Information Visualization (WebVis'99).
2. Hartson, H., Castillo, J. and Kelso, J., (1996). Remote Evaluation: The Network as an Extension of the Usability Laboratory, Proceedings *of CHI 96*, pp. 228-235. New York: ACM Press.
3. 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.
4. 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.
5. Linklint: http://www.goldwarp.com/bowlin/linklint/, 1997.
6. Nielsen, J. (1993) *Usability Engineering*, Academic Press, Boston.
7. The NIST WebMetrics Tool Suite, http://www.nist.gov/webmetrics
8. Scholtz, J., Downey, L. (1999) Methods for Identifying Usability Problems with Web Sites. Engineering for Human-computer Interaction. (Eds. S. Chaty and P. Dewan). Boston: Kluver. pp. 191-206.
9. Scholtz, J., Laskowski, S, and Downey, L. (1998). Developing Usability Tools and Techniques for Designing and Testing Web Sites, *4th Annual Human Factors and the Web Conference*, June 5, ATT, Basking Ridge, NJ.
10. Vora, P. (1998). Designing for the Web: A survey, *Interactions*, Vol. 5 (3), pp.13-30.