

The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent™ Simulator

Ed H. Chi, Adam Rosien, Gesara Supattanasiri, Amanda Williams, Christiaan Royer,
Celia Chow, Erica Robles, Brinda Dalal, Julie Chen, Steve Cousins

Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94301
{echi, bloodhound}@parc.com

ABSTRACT

According to usability experts, the top user issue for Web sites is difficult navigation. We have been developing automated usability tools for several years, and here we describe a prototype service called InfoScent™ Bloodhound Simulator, a push-button navigation analysis system, which automatically analyzes the information cues on a Web site to produce a usability report. We further build upon previous algorithms to create a method called Information Scent Absorption Rate, which measures the navigability of a site by computing the probability of users reaching the desired destinations on the site. Lastly, we present a user study involving 244 subjects over 1385 user sessions that shows how Bloodhound correlates with real users surfing for information on four Web sites. The hope is that, by using a simulation of user surfing behavior, we can reduce the need for human labor during usability testing, thus dramatically lowering testing costs, and ultimately improving user experience. The Bloodhound Project is unique in that we apply a concrete HCI theory directly to a real-world problem. The lack of empirically validated HCI theoretical models has plagued the development of our field, and this is a step toward that direction.

Categories & Subject Descriptors: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems--Evaluation/methodology; H.5.2 [Information Interfaces and Presentation]: User Interfaces--Benchmarking, Theory and methods; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—Web-based Interaction; H.3.4 [Information Storage and Retrieval]: System and Software—Performance evaluation; D.2.2 [Software Engineering]: Design Tools and Techniques--User interfaces; I.2.7 [Artificial Intelligence] Natural Language Processing--Text analysis

General Terms: Algorithms, Design, Experimentation, Human Factors, Measurement, Performance, Verification.

Keywords: Information Scent, Information Foraging, Web-based Services, Usability Prediction, User Modeling, User Simulation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA.

Copyright 2003 ACM 1-58113-630-7/03/0004...\$5.00.

INTRODUCTION

There currently exists a major gulf between HCI theories and how they can be applied to the design of Web information architecture. As a scientific field, HCI has almost been sideswiped by the Web, because very few existing HCI theories seem to point the direction for designers to follow. Instead, rules of thumb have been developed from experiences of usability gurus and consultants. Years of development in HCI theory seem not to have prepared us for an answer to how Web sites should be designed. Indeed, designers and researchers of Web interactions have been seeking ways to quantify the quality of user experience for the last five or six years. The lack of applied theory has resulted in the development of ad-hoc methods for designing Web site navigation and content structure. How can we advance the state of art in HCI theory and apply it so that it is more relevant and directly applicable to Web designers?

Recently, we have discovered several surprises in the fundamental theories of how people access information. For example, the “Law of Surfing” shows that users display different but regular patterns of surfing. The equation from this law predicts the limits of how far people will click on sites [10]. Information Foraging Theory predicts the information gathering behavior of users in an information environment [15]. From these user models of how people access and understand information, researchers have discovered that we can predict and simulate how users surf sites with specific tasks, such as how patients access their personal medical records and seek answers to their medical questions [7, 4].

One of these models is based on the notion of “Information Scent,” which is the user’s perception of the value and the cost of accessing a particular piece of information [6]. The idea is that the user decides whether the distal information that lies on the other side of a link is worthwhile to explore and assimilate based on the proximal cues that surround the hyperlink. The theory posits that users decide on their particular courses of action based on these cues, and their behavioral patterns are guided by information scent. One open research question is whether these models can be used directly by practitioners to measure something about the user interaction in their site testing.

To be sure, Information Scent as a concept has been directly utilized in the field to help the design of sites by enabling designers to think about what proximal cues might lead directly to user action [19, 13]. Published research thus far has described the model and how it might be applied [6, 7, 4] as an automated simulation system. The attractiveness of this approach is that Web sites can be measured without employing usage logs, enabling alternative designs of the site to be tested simultaneously. However, what is missing is a method to utilize the model directly, and to embed this method in a system that practitioners can use effortlessly.

In this paper, we describe our effort to directly close this gap. To achieve this, the paper is divided into three related parts. First, we describe the algorithm of Information Scent Absorption Rate (ISAR), which is the needed theory to directly apply Information Scent to measure where users ended up during the simulation. Second, we illustrate the development of a system that directly utilizes Information Scent and ISAR to help designers. We describe our development of this prototype service that enables us to simulate the behavior of users with specific information goals. This service is called the InfoScent™ Bloodhound Simulator. Finally, we present a large user study involving 244 users and some 1385 user sessions over 4 sites and 32 tasks. The user study shows how Bloodhound correlates with real users surfing for information.

The paper is organized as follows. First we describe related work and our past work on the Information Scent simulation algorithm called WUFIS. Next we describe the modifications to the WUFIS simulation to give the Information Scent Absorption Rate (ISAR) algorithm. We then describe how ISAR is used in practice in a prototype service code-named Bloodhound. We describe the capability of the system, and the contents of its usability report. Lastly, we describe the large user study, and present some concluding remarks.

RELATED WORK

Automated usability tools can be broken up into two different types. First, there are a wide variety of systems for making sure Web sites conform to Web accessibility standards for users with disabilities. There are numerous systems for measuring the accessibility of a Web site based on U.S. government regulation Section 508 (section508.gov) and W3C's Web Content Accessibility guidelines (w3c.org/wai/). Example systems include AccessEnable (www.retroaccess.com) and LIFT (www.usablenet.com).

The second type of automated usability tools are systems that try to predict Web site usage patterns or usability based on site designs. Previously, we reported on the precise algorithm we employed to predict and simulate web traffic [6,7]. Some recent systems include CWW [4], WebTango [11], and WebCriteria SiteProfile [22]. WebTango is a system that uses empirically validated data to correlate against design elements such as text placement, color, and

other design features of a page. Then design features that correlate with successful sites as measured by judges' ratings are then used as measuring sticks for future web site page designs. WebTango focuses on individual page design issues rather than information architecture and navigation, so it is not directly relevant to this paper.

WebCriteria SiteProfile [22] employs software agents as surrogate users to traverse a Web site and derives various usability metrics from simulated surfing. The simulated browsing agents perform a random walk of the Web site. It neither simulates users with specific information needs, nor users who can perceive navigational choices and make navigational decisions. There is some well-known controversy surrounding the validity of this system [16, 20].

CWW, on the other hand, uses Latent Semantic Indexing techniques to estimate the degree of semantic similarity in the calculation of information scent for each link [4]. However, this technique has not yet been completely automated for the analysis of all of the pages for an entire site. It is applied manually to each page of a site selectively, creating a rather cumbersome process.

Our work in Information Scent simulation [7] is also similar to several information retrieval algorithms based on network inferences. Turtle and Croft proposed the use of Bayesian networks to model information retrieval problems [21]. More recently, a number of efforts in the Web research community have concentrated on combining linkage information with user queries in order to rank search results. Most similar to the Information Scent approach, Chakrabarti et. al. [5] and Silva et. al. [18] proposed combining link-based and keyword-based pieces of evidence in a single information retrieval model. Chakrabarti's system uses the text surrounding a link as keyword-based evidences to determine a weight for each link analyzed. These evidences are similar to the ideas of proximal cues. This weighting is then used to compute rankings of the retrieval results using a modified version of the Kleinberg authority algorithm [12].

Fundamentally, the new development of an automated usability service using Information Scent simulation seems unique in its approach. While past information retrieval techniques are interested in using Bayesian networks and linkage information to re-rank search results, researchers are not interested in using these algorithms to measure how users might reach these destinations.

PAST WORK ON INFOSCENT SIMULATIONS

Here we present the necessary summary of past work for understanding how the Bloodhound service works and the modifications to the simulation algorithm necessary for calculating the Information Scent Absorption Rate.

Our system is based on the theoretical notion of *information scent* [6, 7] developed in the context of *information foraging theory* [15]. Information Foraging is related to other research, such as Berrypicking [2] and ASK [3], of how

users optimize behavior to seek information both in directed structured and opportunistic unstructured ways. We have found that users commonly have some *information goal* – some specific information they are seeking – when they visit a Web site. Users typically forage for information by navigating from page to page along hyperlinks. The content of pages associated with these links is presented to the user by some snippets of text or graphic called *proximal cues*. Foragers use these browsing proximal cues to access the *distal content*: the page at the other end of the link. *Information Scent* is the imperfect, subjective perception of the value, cost, or access path of information sources obtained from proximal cues.

During information seeking, when choosing from a set of outgoing links on a page, the user examines some of the links and compares the cue (*i.e.*, link anchor and/or surrounding text) with her information goal. The user takes the degree of similarity as an approximation to how much the content reachable via that link coincides with the information goal. Using this concept, *Web User Flow by Information Scent* (WUFIS) [7] is a predictive simulation technique based on a combination of information retrieval techniques and spreading activation [1].

The prediction model employs a simulation of an arbitrary number of users traversing the links and content of a Web site. The users have information goals that are represented by vectors of content words. At each page visit, the model assesses the information scent associated with each hyperlink. The scent of a link is calculated as a degree of similarity between the proximal cues and the information need. It then computes a probabilistic network of the likelihood of one user moving from one page to another page along hyperlinks that may or may not match the user’s information goal. This probabilistic network is then used to simulate the user flow throughout the site based on that information goal. Figure 1 summarizes this simulation.

Here is a mathematical sketch of the simulation algorithm. Readers may refer to past paper for more details [7]. First, we extract the content and linkages of a Web site. We obtain the hyperlink topology as an adjacency matrix T . We also obtain the content (word x document) W matrix. An entry in the W matrix specifies how important a word is in that document according to TF.IDF, a well-known information retrieval algorithm. A user’s information need is expressed as a keyword query vector Q .

For each link $E(i,j)$, we obtain the proximal cue words that are associated with that link, and insert this information into a matrix K . K is a three dimensional matrix, with an entry $K(i,j,k)$ specifying that link $E(i,j)$ contains the keyword k . There are a variety of ways to obtain proximal cues. For example, we may look at (1) the words in the link itself, (2) the text surrounding a link, (3) the graphics related to a link, (4) the position of the link on the page, etc.

We look up the weighting of each keyword in K in the matrix W to measure the importance of each keyword. Finally, we multiply the link cues in K with Q to obtain the Proximal Scent matrix PS . Thus, for each link $E(i,j)$, we find the corresponding proximal cue words from K , obtaining a vector $K(i,j,*)$. $PS(i,j) = K(i,j,*) * Q$. PS is then normalized so that each column sums to 1.0. This Proximal Scent matrix specifies the probabilities of users following each particular link.

At this point, we can then use the scent matrix to simulate users flowing through various links of a site, giving each link a different proportion of the users relative to the strength of the scent. The probability associated with each link essentially specifies the proportion of users that will flow down various link choices.

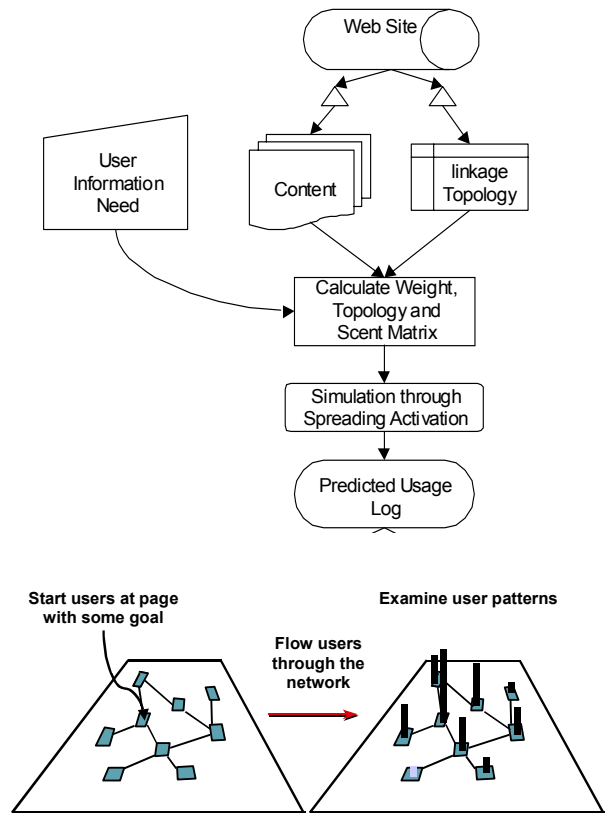


Figure 1: (top) Flow chart of the Web User Flow by Information Scent (WUFIS) algorithm. (bottom) Flowing users through the network using Spreading Activation

We use a network flow algorithm called *spreading activation* [1], depicted pictorially in Figure 1 (bottom). We take an entry page, and construct an entry vector E . We set the initial activation vector, $A(1) = E$. The algorithm goes through $t=1..n$ number of clicks:

$$A(t) = \alpha S A(t-1) + E.$$

The parameter α simulates the proportion of users that do not go from step $t-1$ to step t (e.g. Law of Surfing estimation could be used here). This process generates a predicted user flow, which can be used to extract simulated user paths and infer the usability of a Web site.

The Bloodhound Project

Using this simulation algorithm, the novel idea behind the Bloodhound project is to create a service that would automatically infer the usability of a site. The application scenario is that a customer of the service would specify the site to be analyzed, and the information goal to be simulated in the analysis. Then the Bloodhound service would return usability metrics that tell the customer how easy it is to accomplish the information goal that was given. Figure 2 describes the conceptual idea of the service pictorially.

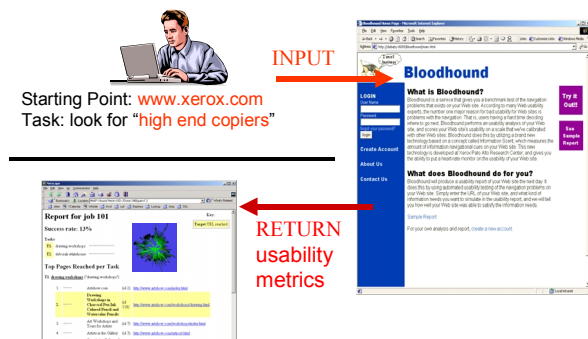


Figure 2: Conceptual idea in the Bloodhound Project

By itself, the simulation algorithm reported earlier is not enough to measure this navigability. We need a method to directly measure how easily can users reach the targeted destinations using the information goal given. To do this, we developed the following new method called Information Scent Absorption Rate (ISAR).

Information Scent Absorption Rate Method

The intuition is that as users discover information items their needs have been satisfied and the simulated users should settle and terminate at a set of documents. These target documents are the documents that satisfy their information needs. The rate in which people finish is a measurement of the navigability of a site.

We first compute the Scent matrix as specified above. Each entry $S(i, j)$ in the Scent matrix is the calculated probability that a user will surf from page i to page j , given that this user has the given information goal. However, the scent matrix that describes the surfing graph has leaves (nodes without connections to further nodes). The spreading activation algorithm does no backtracking, as simulations only move forward on the network. One way to fix this is to tie leaf nodes back to the starting point. So if node j is a leaf, we set $S(j, \text{starting page}) = 1.0$. Any user reaching node j would start over at the initial page.

Now, we need to make the actual destination pages the absorption states. Users reaching these absorption states do not leave these documents. To do this, we turn the destinations into nodes that do not have any children (that is, turn them into leaf nodes of the graph.) So we take the Scent matrix S and zero out the entire column of the target documents. So if target document is t , then the t -th column of the S matrix should be zeroed out. Let's call this new scent matrix S' .

We now do spreading activation user flow simulation using this updated S' scent matrix and sum up the amount of activation still left in the activation vector at the last click of the simulation. Let's call this value β , which can be thought of as the probability that someone would still be searching for the destination page. Then the probability of success is $(1-\beta)$. Figure 3 describes this idea pictorially.

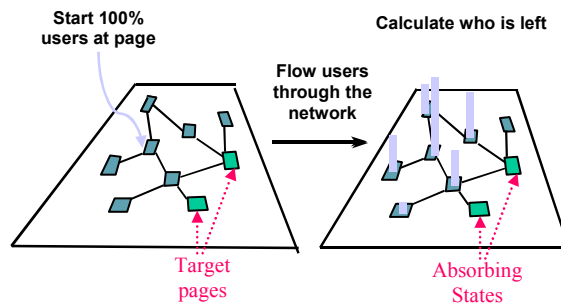


Figure 3: The user absorption model depicted

InfoScent™ Bloodhound Simulator

We want to use the Information Scent algorithm to generate automated usability reports. In this section, we describe the InfoScent™ Bloodhound Simulator, which is a service built using these algorithms. It identifies and tracks navigation problems that exist on the site. Since the analysis is automated, it can be performed over and over again, tracking changes on a Web site and how it affects the usability of the site, thus allowing an analyst to put a heart-rate monitor on the site usability.

Figure 4 shows the input screen that allows analysts to specify the site to be analyzed and a set of user tasks (specified using information keywords related to the goal) and the associated destinations to retrieve the correct information.

Figure 5 shows the result of one of these analyses. It shows that the average success rate of the tasks is 37%, which we consider that to be a “fair” rating. While this seems low, a recent study on user success rate on performing tasks on e-commerce sites showed that an average of 56% was successful [14]. Looking for “demonstrations” succeeded 49% of the time, while looking for “training fleet” material only succeeded 23% of the time. Furthermore, the report shows that several high traffic pages are used as intermediate navigational pages, including pages that may be bottleneck pages.

The report further shows where users are likely to end-up, (i.e. the likely destinations based on their information need.) For example, in the first “demonstration” task, the first highest likely destination is not the target destination page but instead is a page entitled “ICAI Demonstration Projects”. This is certainly potentially confusing to the “demonstration” task.

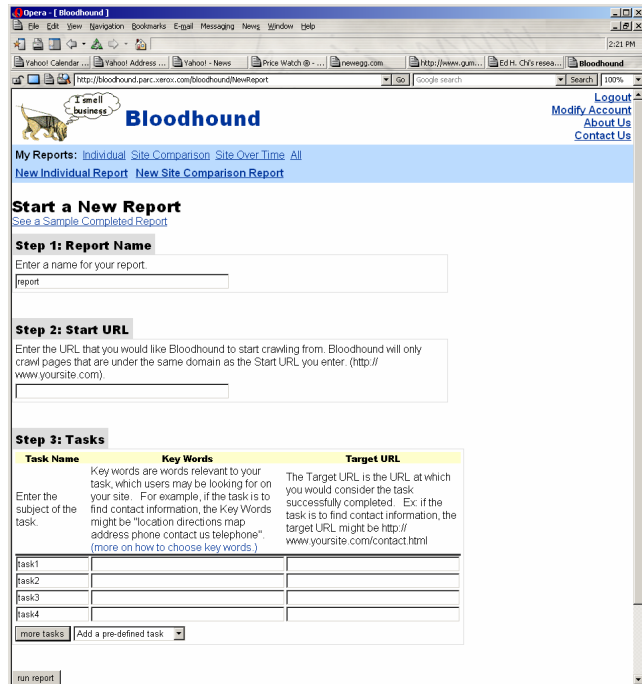


Figure 4: Input tasks into Bloodhound to be simulated.

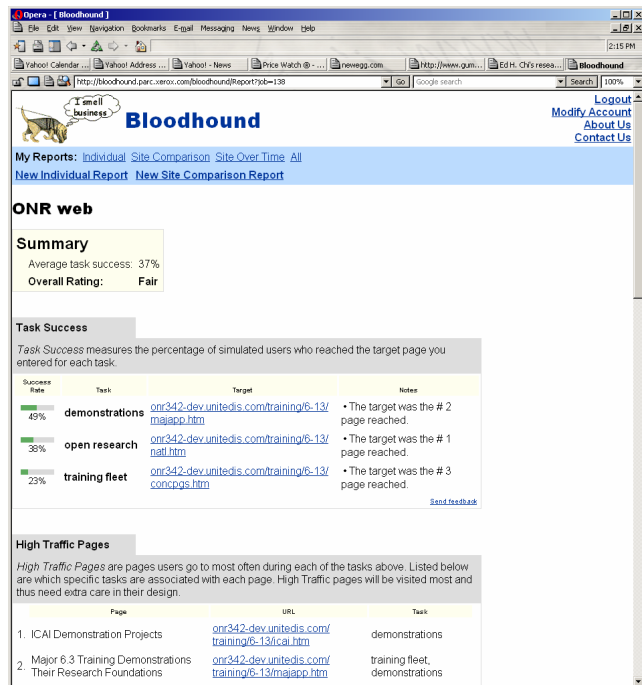


Figure 5: Part of a sample report for the Office of Naval Research Website

Figure 6 shows another example for Inxight.com, a start-up company in information visualization and linguistic technology. The first task is searching for “research papers” that describe the technology that are available, and has a probability of success of 33%. The second task is searching for “events” in which Inxight will be demonstrating their technology, with a task success rate of 52%. Both tasks require clicking twice from the home page to succeed, with the same intermediate pages (Home Page → news and events → [Events, Research Papers]). The difference between the success rates comes from the fact that one task has much better proximal cues that lead the users to the goal. The “Events” link is directly visible from the home page, while the “research paper” link requires some hunt and peck before it can be found. This highlights a case in which usability can be improved by observing the simulated surfing behaviors.

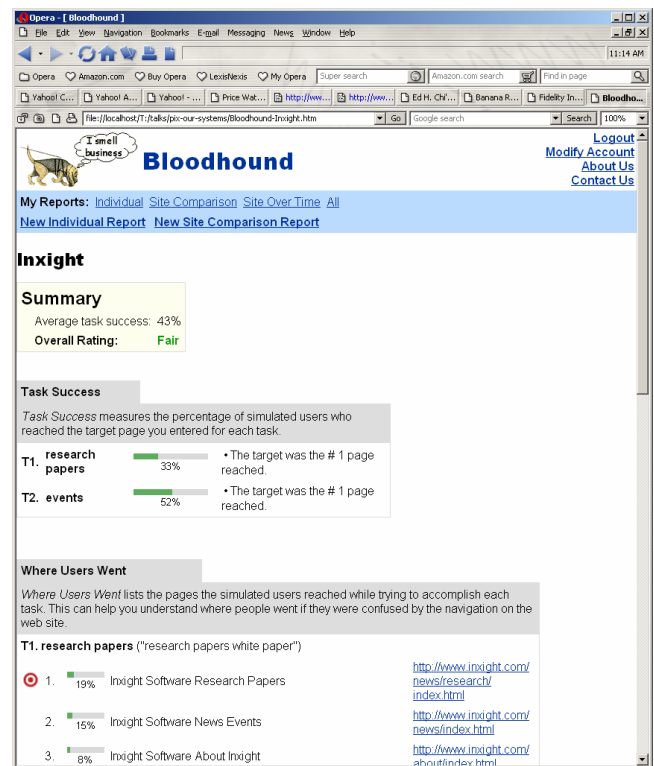


Figure 6: A sample report for Inxight.com.

USER EXPERIMENT

The above section described the example uses of the Bloodhound Simulator. We are interested in how well Bloodhound simulates real users in surfing tasks. To this end, we conducted an extensive user study involving 244 users and collected 1386 user sessions.

Subjects: We recruited subjects from a wide-variety of sources, using several large mailing lists. A popular yellow-pages style site called Craig’s List was used to recruit the majority of the users. For the intranet site we tested, we recruited employees of the company to participate. Sub-

jects were paid a \$20 Amazon gift certificate. A total of 244 users participated in this study.

Material: Four different types of web sites were used: help.yahoo.com (the help system section of Yahoo!), www.rei.com (a camping/outdoor online store), hivin-site.ucsf.edu (AIDS and HIV medical information site), and parcweb.parc.com (an intranet of company internal information). A remote version of the usability testing tool based on WebLogger is used to conduct the tests [17]. Users downloaded this testing apparatus and went thru the test at their leisure.

Tasks: Each site had a set of eight tasks, for a total of $8 \times 4 = 32$ tasks. For each site, the eight tasks were grouped into 4 categories of similar types. For each category, one task was considered to be hard and the other to be easy. For each task, the user was given an information goal in the form of a question. The tasks were chosen after we spent some time getting familiar with the sites. We wanted to be sure that the tasks were somewhat representative of the users of the site. Generally, at least 50 users were assigned to each task. Here is a sample of the information goals given:

Help.yahoo.com (7484 documents):

- You want Yahoo! to add your site to the Yahoo! Directory. Find some guidelines for writing a description of your site.
- When is the playing season for Fantasy Football?
- You want to get driving directions to the airport, but you don't know the street address. How else can you get accurate directions there?

REI (36422 documents):

- You are planning a week-long hiking trip for this summer, and you're on a budget. Find a single person tent for less than \$120.
- Find the location of the REI store nearest you.
- Find yourself some warm, fairly heavy long underwear for the upcoming ski season.

HivInSite (8308 documents):

- What are the impacts of the AIDS epidemic on Latin America?
- How often should HIV infected women have pap smears to check for the HIV-Associated malignancy, cervical cancer?
- What are some safer sex resources that are targeted towards teens?

Parcweb (19227 documents):

- Suppose this is your first time using Amberweb. Find some documentation that will help you figure out how to use it.
- Find the 2002 Holiday Schedule
- Find out where you can download the latest DataGlyph Toolkit.

Procedure: We sent each subject a URL link designed specifically for the subject. The link contained an online con-

sent form and instructions for the study. They then downloaded the WebLogger and then the subjects were asked to perform the study in the comfort of their office or anywhere else they chose. Subjects could abandon a task if they felt frustrated, and they were also told that they could stop and continue the study at a later time. The idea was to have them work on these tasks as naturally as possible.

Users were explicitly asked not to use the search feature of the site, since we are only interested in navigation data. Each subject was assigned a total of eight tasks (four easy, and four hard tasks) from across different sites. We made sure to counter-balance the task assignments for difficulty. In the end, each task is assigned roughly the same number of times. We recorded the time of each page access. Whenever the user wanted to abandon a task, or if they felt they had achieved the goal, the user clicked on a button signifying the end of the task. Subjects were then taken to a form, where they could give feedback on any usability problems they might have encountered. We recorded the time they took to handle each task, the pages they accessed, and the keystrokes they entered (if any).

We also ran Bloodhound reports on each of these tasks and recorded the activation vectors and the success values.

USER STUDY RESULTS

User Sessions Obtained

Usable Sessions	Task 1a	1b	2a	2b	3a	3b	4a	4b	Sum
Yahoo	44	46	46	43	44	47	44	44	384
REI	24	37	49	27	44	35	47	38	392
HivIn-Site	29	35	30	20	25	31	28	22	332
Parcweb	28	31	29	27	25	30	30	33	304

Table 1: Summary of usable user sessions.

We collected a total of 1386 sessions for all of the tasks. This is smaller than $244 \text{ subjects} \times 8 \text{ tasks} = 1952$ sessions because some subjects did not participate in all of the tasks in all of the sites. Of these, we cleaned the data to throw out any sessions that employed the site's search engine as well as any sessions that did not go beyond the starting home page. We were not interested in sessions that involved the search engine because we wanted users to find the information using only navigation. In the end, **1112** user sessions were usable (Yahoo=358, REI=301, HivIn-Site=220, Parcweb=233). Table 1 summarizes the number of usable sessions that were collected for each task.

Data Analysis

For each of the user sessions, we tallied the frequency of accesses for each document on the site. We then took these tallies and generated a single frequency distribution over the document space for all of the user sessions in that task.

This frequency vector is the “user summary vector”. This gave us 32 user summary vectors.

For each of the tasks, we also had data from the Bloodhound service that specified the probability of each user ending up at each document during each step of the simulation. We produced a sum of these probabilities across each step and scaled the activation values so that it matched the scale of the user summary vector. This generated a Bloodhound frequency vector that specified how many times Bloodhound predicted the users should have visited each document. This frequency vector is the “Bloodhound summary vector”.

We then computed the correlation coefficients between the two corresponding summary vectors for each task. The following table gives the correlation coefficients for each of the 32 tasks.

<i>Corr. Coeff.</i>	Yahoo	REI	HivIn-Site	Parc-web
task 1a	0.7528	0.4701	0.6811	0.7394
task 1b	0.7218	0.4763	0.7885	0.8756
task 2a	0.7489	0.9892	0.6671	0.8930
task 2b	0.8840	0.7073	0.6880	0.8573
task 3a	0.7768	0.7321	0.8835	0.7197
task 3b	0.6973	0.6979	0.5660	0.7123
task 4a	0.9022	0.9415	0.8407	0.8340
task 4b	0.9052	0.7600	0.4634	0.9344

Table 2: correlation coefficients for frequency distribution comparisons between Bloodhound generated frequency vector versus user study data.

Statistically, a correlation coefficient above 0.8 is generally considered to be strong correlation, and between 0.5 and 0.8 is considered moderate, while below 0.5 is considered weak correlation [8, p.486]. Accordingly, three cases have a weak correlation. Twelve correlated strongly, and seventeen of the 32 tasks correlated moderately.

This is a reasonable result. The user study shows that in nearly all of the cases, Bloodhound was able to produce click streams that moderately correlate with user data, and in a third of the time, Bloodhound actually produced click streams that correlate strongly with user streams.

Our goal in using Bloodhound is to reduce the cost of conducting usability testing for Web sites. From this study, we can be reasonably confident in Bloodhound creating moderately reasonable approximations in nearly all cases. It gives us slight comfort in knowing that nearly a third of the cases are likely to be fairly accurately simulated, even though we do not know *a priori* which third.

In the course of the study, we noticed that Bloodhound appears to be sensitive to the task query keywords. In future

work, careful understanding of how the query keywords capture domain knowledge of the task is essential in improving the accuracy of Bloodhound.

CONCLUSION

Practitioners have widely deployed conventional usability evaluation techniques, such as card sorting, cognitive walk-throughs, accessibility guidelines, and direct user testing. Recent work on automated usability techniques has generally emphasized the continual need for such conventional techniques, as it is difficult, if not impossible, to completely replace these techniques with automated tests. Automated tools are intended to be used as a component in the comprehensive evaluation of site usability.

As a step in that direction, in this article, we have described an automated tool for analyzing the usability of a Web site. InfoScent™ Bloodhound Simulator uses Web agents to predict the user traffic flow through a Web site by examining the information scent surrounding every hyperlink on the site with respect to some given information need. The resulting simulation produces a report that specifies the probability of success for each individual task. The hope is that by employing Web agents to discover usability problems, we can dramatically reduce the cost of searching and fixing Web site navigation problems.

We presented a user study involving 244 subjects producing 1385 user sessions over 32 tasks for 4 sites. The results show that Bloodhound strongly correlates with real user data in a third of the tested 32 tasks. In the other roughly two third of the cases, Bloodhound moderately correlates with real user data. The user study showed that Bloodhound gave measurements that reasonably approximate real users, giving designers a way to measure how well users might perform on tasks. They could consider possible alternative designs that can be tested immediately again using Bloodhound.

As a field, very little HCI theory has been able to inform designers how to architect their Web sites. HCI needs theories that have been validated and that can be applied again and again in practice to reduce costs and point the direction for future design and future research. Web information access is fundamentally about two things: user interfaces and cognition. Automated ways to analyze information access interfaces such as a Web browsing *must* therefore utilize *cognitive theories* of how people surf for information. Ideally, the cognitive predictive model should model user’s context. The development of these theories is what enables the field to develop and prosper, because it encodes what we have learned as a field. We hope that our work in this area is a step toward that direction, yet we know that the difficulty in understanding user contexts makes Web usability a significant challenge for years to come.

ACKNOWLEDGEMENT

The research and development described in this article could not have happened without the collaboration of Peter Pirolli. Pam

Desmond helped with proof-reading of an earlier draft. This work was supported in part by Office of Naval Research grant No. N00014-96-C-0097 to Peter Pirolli and Stuart Card.

REFERENCES

- [1] Anderson, J. R., Pirolli, P. L. (1984) Spread of Activation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10: pp. 791-798.
- [2] Bates, M.J. (1989). The design of browsing and berypicking techniques for the on-line search interface. *Online Review*, 13(5): pp. 407-431.
- [3] Belkin, N. J. (1980). Anomalous states of knowledge as the basis for information retrieval. *Canadian Journal of Information Science*, 5, May 1980 (pp. 133-143).
- [4] Blackmon, M. H., Polson, P.G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. CHI 2002 Conference on Human Factors in Computing Systems. ACM Press, pp. 463-470.
- [5] Chakrabarti, S., B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. (1998) Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. Of the 7th International World Wide Web Conference (WWW7)* (pp. 65-74), Brisbane, Australia
- [6] Chi, E., P. Pirolli, and J. Pitkow. (2000) The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a Web site. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI 2000*, pp.161-168. Hague, Netherlands.
- [7] Chi, E.H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using information scent to model user information needs and actions on the Web. *Proc. of the ACM Conference on Human Factors in Computing Systems, CHI 2001* (pp. 490-497), Seattle, WA.
- [8] Devore, Jay. (1987) Probability & Statistics for Engineering and the Sciences. 2nd Ed. Brooks/Cole Publishing: Monterey, CA.
- [9] Heer, J., Chi, E.H. Separating the Swarm: Categorization Methods for User Access Sessions on the Web. In *Proc. of ACM CHI 2002 Conference on Human Factors in Computing Systems*, pp. 243--250. ACM Press, April 2002. Minneapolis, MN.
- [10] Huberman, B.A., Pirolli, P., Pitkow, J.E., and Lukose, R.M. (1998). Strong regularities in World Wide Web surfing. *Science*, April 3, 1998, vol. 280, num. 5360 (pp. 95-97).
- [11] Ivory, M. Y., R. R. Sinha, and M. A. Hearst, (2001) "Empirically Validated Web Page Design Metrics," *Proc. Human Factors in Computing Systems (CHI2001)*, pp. 53-60.
- [12] Kleinberg, J. M. (1998) Authoritative sources in a hyperlinked environment. In *Proc. Of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, (pp. 668-677), San Francisco, CA.
- [13] Koman, Richard. (1998) The Scent of Information. *WebReview*, May, 1998. http://www.webreview.com/1998/05_15/strategists/05_15_98_1.shtml
- [14] Nielsen, J. <http://www.useit.com/alertbox/20010819.html>, August 19, 2001
- [15] Pirolli, P. and S.K. Card. (1999) Information foraging. *Psychological Review*. 106: p. 643-675.
- [16] Pirolli, Peter. (2000) A Web Site User Model Should at Least Model Something About Users. *Internetworking 3:1*, Mar 2000. ITG Publications. http://www.internetg.org/newsletter/mar00/critique_max.html
- [17] Reeder, R. W., Pirolli, P. and Card, S. K. (2001). WebEyeMapper and WebLogger: Tools for Analyzing Eye Tracking Data Collected in Web-use Studies. *Proceedings of CHI 2001*, Seattle.
- [18] Silva, I., B. Ribeiro-Neto, P. Calado, E. Moura, N. Ziviani. (2000) Link-based and Content-based Evidential Information in a Belief Network Model. In *Proc. of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.96-103). Athens, Greece.
- [19] Spool, J.M., Scanlon, T., Snyder, C., and Schroeder, W. (1998). Measuring Website usability. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '98* (pp. 390), Los Angeles, CA.
- [20] Tilt, Chris. (2000) Response to Pirolli's Critique of MAX model. *Internetworking 3:1*, Mar 2000. ITG Publications. http://www.internetg.org/newsletter/mar00/response_critique_max.html
- [21] Turtle, H., Croft, W. (1991) Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187-222
- [22] WebCriteria SiteProfile. (2002) <http://www.webcriteria.com>