# An Empirical Approach to Automated Web Site Evaluation

Melody Y. Ivory-Ndiaye
The Information School
University of Washington
myivory@ii.washirigton.edu
Received 18 March 2003; Accepted 18 April 2003

*Abstract: The Web enables broad dissemination of information and services, yet most sites have in-adequate usability and accessibility. Numerous automated evaluation methodologies and tools have been developed to help designers to improves their sites. We describe the state-of-the-art in automated web site evaluation and then elaborate on our WebTango approach, which entails deriving design guidelines by mining empirical data. We compute over 157 measures, which assess many web interface aspects, and then use these measures along with expert ratings from Internet professionals as input to data mining algorithms. This mining process enables us to derive statistical models of highly rated web interfaces, such that the models reflect effective design patterns that are used on them. We then deploy the models so that designers can use them in the automated analysis of their sites.*

**Keywords** : World Wide Web, WWW - Empirical Studies, Automated Usability Evaluation, Web Site Design

Received 18 March; Accepted 18 April 2003

## 1 Introduction

The World Wide Web plays an important role in our society-enabling broader dissemina-tion of information and services than was previously available. However, many sites provide inadequate usability and are inaccessible to users with disabilities [Forrester Research, 1999; Jackson-Sanborn *et al.,* 2002]. Much has been said about the way in which to design web sites so that they can be used and accessed by users who have a broad range of abilities, backgrounds, and skills, but building a high-quality site is a challenging task. Furthermore, it is a challenge that many non-professional or occasional designers need to undertake.

There are numerous approaches to address the current state of the Web:

- Design guidelines, patterns, and methods to help designers to structure their design pro-cesses and to inform their design decisions (e.g., [Comber, 1995; Computer Science and Telecommunications Board, 1997; Detweiler and Omanson, 1996; Farkas and Farkas, 2000; Flanders and Willis, 1998; Fleming, 1998; Levine, 1996; Lyardet *et al.,* 1999; Lynch and Horton, 1999; National Cancer Institute, 2001; Nielsen, 1998; Nielsen, 1999; Nielsen, 2000; Paciello and Paciello, 2000; Rosenfeld and Morville, 1998; Sano, 1996; Schriver, 1997; Shedroff, 2001; Shneiderman, 1997; Spool *et al,* 1999; Thatcher *et al.,* 2002; Turns and Wagner, 2002; van Duyne *et al.,* 2002; Williams, 2000; World Wide Web Consortium, 1999]).
- Web site templates to help designers to implement sites within HTML authoring environments.
- Form-based site creation services to build sites for designers (e.g., autowebmaker [autowebmaker, 2003] and easiwebmaker [Easiwebmaker, 2003] for e-commerce sites, Yahoo! GeoCities [Yahoo! Inc., 2003] and Terra Lycos Tripod [Lycos, Inc., 2003] for personal sites,and WebCT [WebCT, Inc., ] and Blackboard [Blackboard, Inc., ] for course sites).
- Automated tools to critique designers' web pages in the late design stages [ivory, 2003a; Ivory and Hearst, 2002b].
- Sketch-based tools to help professional designers to draw site maps, storyboards, and schematics in the early design stages [Lin and Landay, 2002; Lin *et al.,* 1999; Newman and Landay, 2000].
- Automated tools to enable users to transform web pages to fit their needs [ivory *et al.,* 2003].
- Each of these approaches contribute to helping designers to build better sites and to help-ing users to have better experiences on-line. In this article, we examine the contribution of automated web site evaluation tools and methodologies (fourth item above). Our research on the development, use, and role of automated evaluation tools has shown that

they are important for helping designers to learn about effective design practices, to apply evaluation criteria consistently and broadly across entire sites, and to reduce the cost of non-automated evaluation methods like usability testing [ivory, 2001; Ivory, 2003a; Ivory and Hearst, 2001; Ivory and Hearst, 2002b; Ivory *et al,* 2003].

- We discuss automated evaluation methods in general and then elaborate on one approach- the WebTango method. The WebTango method identifies design patterns that exist in empirical data. It entails computing 157 quantitative page-level and site-level measures, which assess many aspects of web interfaces (e.g., the amount of text on a page, color usage, and consistency). We use these measures along with expert ratings from Internet professionals as input to data mining algorithms (e.g., decision trees and clustering) [Cooley *et al.,* 1997], which enable us to derive statistical models of highly rated web interfaces; the models reflect effective design patterns that are used within these interfaces. We then deploy the models so that designers can use them in the automated analysis of their sites.
- We begin this article by presenting a web site design and evaluation scenario, which motivates our discussion. We then summarize the state-of-the-art in automated evaluation approaches in Section 3. Section 4 provides an overview of the WebTango methodology and prototype and subsequent sections describe the five steps that we follow in the approach, how we use the approach to derive concrete thresholds for existing design guidelines, and our future research directions. We elaborate on the WebTango approach for several reasons: (1) it represents a new methodology for the human-computer interaction field and, at the time of this publication, is still the only approach of its kind; (2) it has broad implications for other evaluation approaches, due to the extensive set of web interface aspects that it assesses and the quantitative thresholds that it educes for these aspects; (3) it demonstrates the application of data mining; and (4) it enables context-sensitive evaluation of web interfaces.

## 2 Motivating Scenario

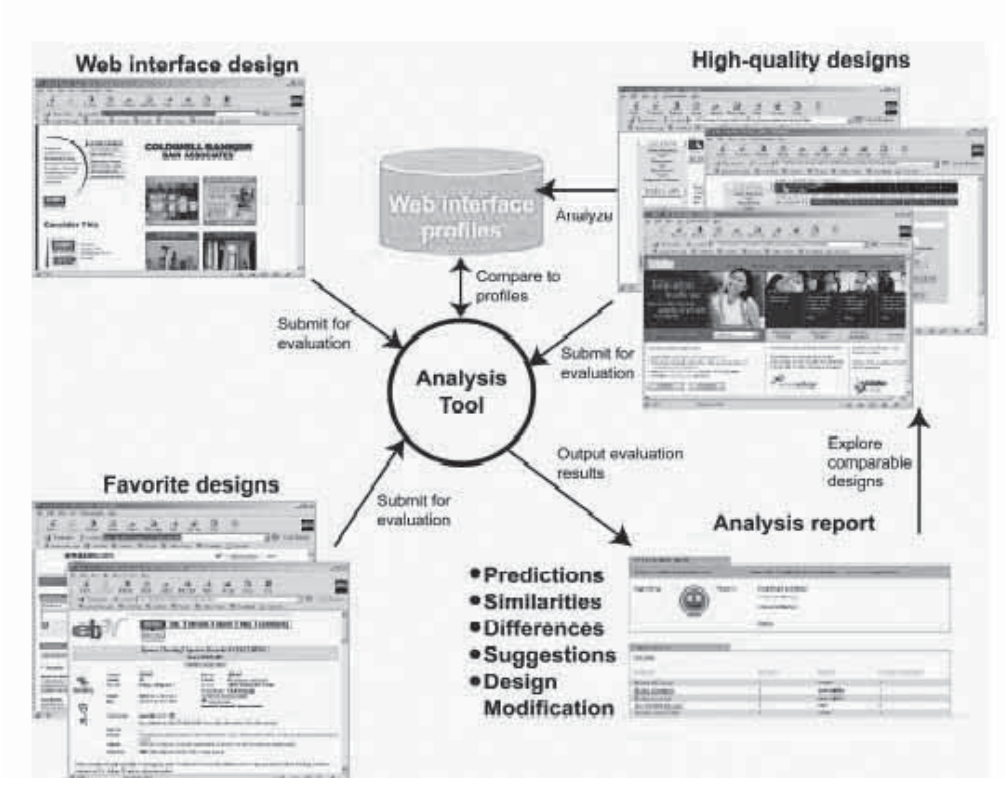


Figure 1: A web interface evaluation scenario. The analysis tool compares features of a submitted design to features of highly rated designs. It then provides a detailed analysis report, which the designer can explore to improve her design.

Figure 1 depicts an analysis scenario: a web designer seeking to determine the quality (i.e., usability, accessibility, performance, etc.) of an interface design. If the site has already been designed and implemented, the designer could use the site as input to an analysis tool. The analysis tool (or a benchmark program) would then sample pages within the site and generate a number of quantitative measures pertaining to all aspects of the interface. A key component of benchmarking is the ability to determine how well benchmark results compare to other systems or a best case [Jain, 1991; Sauer and Chandy, 1981]. For this scenario, designs that have been rated favorably by expert reviewers or users could be used for comparison purposes. Hence, the analysis tool could compare the input design's quantitative measures to those for highly rated designs.

Ideally, the analysis tool goes beyond traditional benchmarking and generates a report that contains: (1) an interface quality or usability prediction, (2) links to similar highly rated designs from the comparison sample, (3) differences and similarities to the highly rated designs, and (4) specific suggestions for improvements. The designer could use these results to choose between alternative designs as well as to inform design improvements. The tool could also perform some simple design modifications (e.g., changing font types or sizes or changing color combinations) automatically. This analysis process could be iterated as necessary.

Similarly, the designer could use the analysis tool to explore results for other interface designs, such as favorite sites. This process may help to educate the designer on subtle aspects of design that may not be apparent from simply inspecting interfaces.

One can also imagine that a designer would want to obtain feedback on interface designs earlier during the design phase as opposed to after implementation. If the tool supported analysis of designs represented as images or templates, then it would be possible to support this evaluation. In particular, the tool needs to use image processing during analysis.

## 3 Automated Evaluation Methodologies

We have conducted several studies and assessments of automated evaluation methodologies and tools (see [Ivory, 2001; Ivory, 2003a; Ivory and Hearst, 2001; Ivory *et al.,* 2003] for more details). Our analysis of 84 evaluation methods revealed two important ways to automate web site evaluation:

- **Analysis:** Software automatically identifies potential problems.
- **Critique:** Software automatically identifies potential problems and suggests improvements.

Of the 84 web evaluation methods surveyed, thirteen percent support analysis and five percent support critique. Furthermore, web site aspects evaluated by most approaches are inadequate [Brajnik, 2000] and are based on guidelines that have not been empirically validated. We summarize the predominate evaluation approaches in the remainder of this section.

### 3.1 Analysis of Web Server Performance

Performance monitoring and stress-checking approaches measure server response time for actual or simulated requests and enable designers to pinpoint performance bottlenecks, such as slow server response time, that negatively affect the usability and accessibility of a web site [BMC Software, 2002; Exodus, 2002; Bacheldor, 1999; Freshwater Software, 2002; Wilson, 1999; Zona Research, Inc., 1999]. In general, these approaches provide little insight into the quality of the web site itself.

### 3.2 Analysis of Site Usage Data

Given that web servers automatically log client requests, web log analysis is a heavily used methodology [Drott, 1998; Fuller and de Graaff, 1996; Hochheiser and Shneiderman, 2001; Sullivan, 1997], despite the fact that server log data is unreliable, inadequate, and missing valu-able information about information-seeking tasks [Byrne *et al.,* 1999; Etgen and Cantor, 1999; Fuller and de Graaff, 1996; Schwartz, 2000]. Task-based approaches aggregate traces of multi-ple user interactions and produce visualizations or reports, which compare users' task flows to the designer's task flow [Beirekdar *et al.,* 2002; Helfrich and Landay, 1999; Hong *et al.,* 2001; Paganelli and Paterno, 2002; Tiedtke *et al.,* 2002]. Inferential approaches include statistical anal-ysis of site traffic and user interactions [Drott, 1998; Fuller and de Graaff, 1996; NetIQ, 2002; Sullivan, 1997; Theng and Marsden, 1998], as well as on-line analytical processing and min-ing of usage data [Beirekdar *et al.,* 2002; Biichner and Mulvenna, 1998; Chi *et al.,* 2002; Cooley *et al.,* 1997; Dyreson, 1997; Etzioni, 1996; Heer and Chi, 2002; Kosala and Blocked, 2000; Spiliopoulou *et al.,* 1999; Spiliopoulou, 2000; Zaiane *et al.,* 1998], and three-dimensional visual-izations of web sites and traversed paths [Chi *et al.,* 2000; Chi, 2002; Cugini and Scholtz, 1999]. Starfield visualization [Hochheiser and Shneiderman, 2001] enables designers to interactively explore usage patterns via an interface that supports zooming, filtering, and dynamic querying [Ahlberg and Shneiderman, 1994].

Client data captured via an instrumented web site [Etgen and Cantor, 1999; Scholtz and Laskowski, 1998], an instrumented browser [Enviz, Inc., 2001; Vividence Corporation, 2002], or a proxy server [Hong *et al.,* 2001] is one way to address some of the limitations of server log data. However, web log analysis approaches require that the site is already built and in use, thus they are more useful for assessing navigation patterns than design elements.

### 3.3 Assessment of Guideline Conformance

There are over 30 tools, including the one discussed in this article, for assessing whether a web site conforms to usability, accessibility, HTML coding, or browser-compatibility guide-lines. This type of assessment is referred to as guideline review, and guideline review tools include: Bobby [WatchFire, 2002], WAVE [Pennsylvania's Initiative on Assistive Technology, 2001], A-Prompt [Adaptive Technology Research Center, 2002], LIFT [UsableNet, 2000; Us-ableNet, 2002], WebSAT [Scholtz and Laskowski, 1998], W3C HTML Validator [World Wide Web Consortium, 2001], LinkScan [Electronic Software Publishing Corporation, 2002], Any-Browser [AnyBrowser.com, 2002], and others [Faraday, 2000; World Wide Web Consortium, 2002]. Bobby, LIFT, A-Prompt and a few others provide critique support (i.e., they recommend design changes). The LIFT tools assist designers with making recommended changes and in-cludes a tool, LIFT - Nielsen Norman Group Edition [UsableNet, 2002], that assesses a site's conformance to accessibility guidelines developed from studies of users with disabilities [Coyne and Nielsen, 2001]. Conforming to the guidelines embedded in these tools can potentially elim-inate problems that arise due to poor HTML syntax (e.g., missing page elements), but in many cases the guidelines have not been empirically validated.

Other techniques compare quantitative web page measures-such as the number of links or graphics-to thresholds [Stein, 1997; Theng and Marsden, 1998]. However, concrete thresholds for a wider class of quantitative web page and site measures still remain to be established. The methodology presented in this article makes it possible to derive quantitative thresholds for design aspects, which we discuss in Section 9.

### 3.4 Analysis of Navigation Text

Cognitive Walkthrough for the Web (CWW) [Blackmon *et al,* 2002; Blackmon *et al,* 2003] is an automated critique approach that can identify potential navigation problems and provide guid-ance for correcting them. The approach entails the use of Latent Semantic Analysis [Landauer and Dumais, 1997] to contrast web page text (headings or links) to a specified information-seeking goal (100-200 word narrative and correct link selection on the page). The LSA system computes several similarity measures for the two text inputs, based on the semantic space that the practitioner specifies (e.g., for grade nine reading ability). The authors provide LSA thresh-olds for three navigation problems: confusable heading or link text, unfamiliar heading or link text, and goal-specific competing heading or link text. Empirical studies have demonstrated that modifying page text to fit within these thresholds improves usability. This tool is under development.

### 3.5 Simulation of Hypothetical Users Navigating a Site

WebCriteria's Site Profile[1] [Lynch *et al.,* 1999; Web Criteria, 1999] uses an idealized user model that follows an explicit, pre-specified navigation path through a web site and estimates several metrics, such as page load and optimal navigation times. Several researchers-Chi, Pirolli, and Pitkow [Card *et al,* 2001; Chi *et al,* 2000; Chi *et al,* 2001; Chi *et al,* 2003], Blackmon and colleagues [Blackmon *et al,* 2002; Kitajima *et al,* 2000; Blackmon *et al,* 2003], and Miller and Remington [Miller and Remington, 2000; Miller and Remington, 2002]-simulate information-seeking behavior by modeling hypothetical users traversing the site from specified start pages, making use of information "scent" (i.e., common keywords between the user's goal and content on linked pages) to make navigation decisions. None of these approaches account for the effects of various web page attributes, such as the amount of text or layout of links, on navigation behavior.

This tool is no longer available for use.

## 4 Overview of the WebTango Method

The WebTango methodology entails deriving design guidance (i.e., prevalent design patterns) by examining well-designed web interfaces. More specifically, the idea is to enable designers to compare their designs to the well-designed ones to determine whether their designs exhibit similar properties, and if not, how they differ. The general approach involves the following steps, which we elaborate on in the remainder of this article.

1. Identifying an exhaustive set of quantitative interface measures
2. Computing measures for a large sample of rated interfaces
3. Deriving statistical models from the measures and ratings
4. Using the models to predict ratings for new inter-

faces

5. Validating model predictions

The analysis methodology consists of two distinct but related phases (Figure 2): (1) establish-ing an interface quality baseline, and (2) analyzing interface quality. Both phases share common activities-crawling web sites to download pages and associated elements (Site Crawling) and computing page-level and site-level quantitative measures (Metrics Computation). During the first phase, the page- and site-level measures coupled with expert ratings of sites are analyzed to determine profiles (predictive models) of highly rated interfaces. These profiles encapsulate key quantitative measures, thresholds for these measures, and effective relationships among key measures; thus, they can be considered as an interface quality baseline.

During the second phase, an design's page- and site-level measures are compared to the developed profiles to assess its quality. Sites analyzed in the latter phase are usually not the same as the sites used to develop profiles. Once profiles are developed, the analysis phase can be used continuously. However, the interface quality baseline phase needs to be repeated periodically (annually or semi-annually) to ensure that profiles reflect current web design practices.

This analysis methodology is consistent with other guideline review methods (discussed in the preceding section) and benchmarking methods. What distinguishes this analysis approach from other guideline review methods is: (1) the use of quantitative measures, (2) the use of empirical data to develop guidelines, and (3) the use of profiles as a comparison basis. The methodology and tools are designed to support many aspects of the web interface evaluation scenario that we described in Section 2. Currently, recommending design improvements, presenting comparable designs, and automated design modification are not supported; future work will focus on these aspects. All other aspects of the scenario are fully supported.

## 4.1 WebTango Prototype

We have developed a rudimentary prototype to support profile development and interface evalu-ation (Figure 3). This prototype is available for public use via separate form-fillin interfaces for the site crawling and analysis steps; future work will integrate these interfaces to support the entire process. The interface for each tool routes requests to a server daemon (the *Tool Server)* for processing; the daemon in turn forks new processes to forward requests to the appropriate backend tool *(Site Crawler, Metrics Computation,* or *Analysis Tool).* Both the Site Crawler and Metrics Computation Tools interact with the *HTML Parser and Browser Emulator;* this component creates a detailed representation of a web page, including the screen coordinates (x and y location) for each page element, the width and height of each element, the font used, foreground and background color, and other attributes. The browser emulator determines many of the details, such as the height and width of text, by querying the graphics environment via the X Server running on the system console.

Currently, each tool sends an email notification to the client when the request completes; this notification includes a link to an archive file (in gzipped tar format) containing the output

of running the tool. The output of running each tool is used as input to the subsequent step. For example, pages downloaded by the Site Crawler Tool are then processed by the Metrics Computation Tool to output page- and site-level measures. The Viewer Tool (Figure 4) enables designers to interactively explore results from the Analysis and Metrics Computation Tools. Future work will focus on developing an interactive, integrated tool to support the entire process. All tools are available for online use via the WebTango Project's web site (http: //webtango . ischool. Washington. edu/tools/). More details about each component can be found in [ivory, 2001, Chapter 4].

## 5 Identification of Web Interface Measures

The first step of the WebTango process entails identifying an exhaustive set of quantitative measures to assess many aspects of web interfaces. We developed a set of 157 page-level and site-level measures based on an extensive survey of design recommendations from recognized experts and usability studies (e.g., [Flanders and Willis, 1998; Fleming, 1998; Nielsen, 1998; Nielsen, 1999; Nielsen, 2000; Rosenfeld and Morville, 1998; Sano, 1996; Schriver, 1997; Shedroff, 1999; Shneiderman, 1997; Spool *et al.,* 1999]). Our intention was to first quantify features discussed in the literature and then to determine their importance in producing high-quality designs. We begin with a view of a web interface's structure and then summarize the 157 mea-

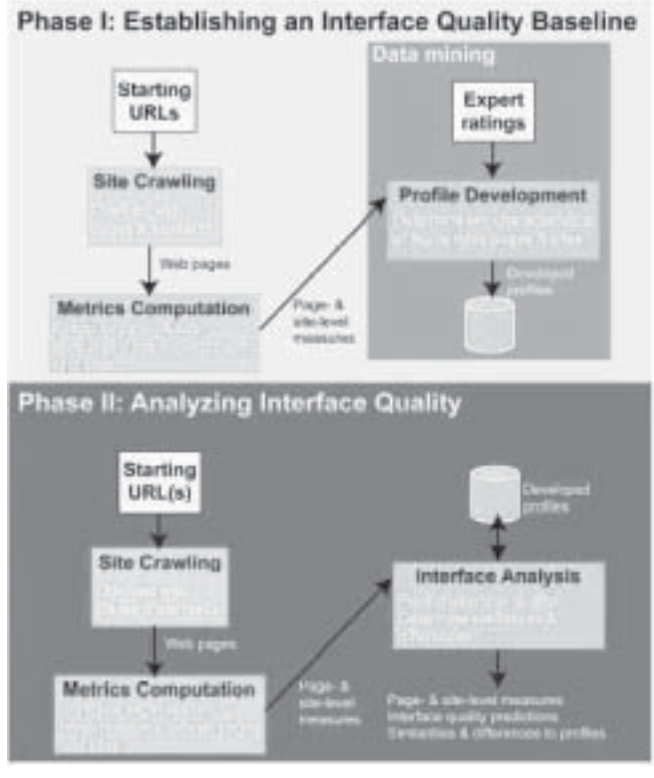


Figure 2: Profile development and interface evaluation phases of the WebTango process. All steps in th< WebTango process are completed to develop profiles of highly rated interfaces. Only two of the steps an completed to assess the quality of a web interface.

sures.

An in-depth discussion of the measures can be found in [Ivory, 2001, Chapter 5]. We also developed an interactive appendix to illustrate all the measures. HTML and PowerPoint versions are available on the WebTango Project's web site (http://webtango. ischool. Washington, edu/).

### 5.1 Web Interface Structure

A web interface is a mix of many elements (text, links, and graphics), formatting of these elements, and other aspects that affect its overall usability, accessibility, and quality. Web interface design entails a complex set of activities for addressing these diverse aspects. To gain insight into web design practices, Newman and Landay [2000] conducted an ethnographic study wherein they observed and interviewed eleven professional web designers. One important finding was that most designers viewed web interface design as being comprised of three components- information design, navigation design, and graphic design.

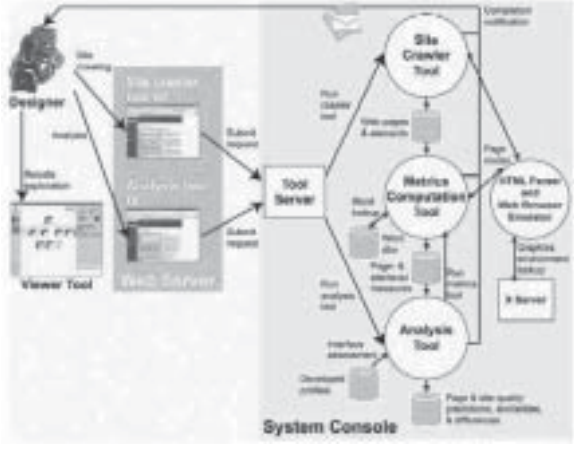


Figure 3: WebTango proto type architecture. The tools support the analysis methodology

- *Information design* focuses on determining an information structure (i.e., identifying and grouping content items) and developing category labels to reflect the information structure.
- *Navigation design* focuses on developing navigation mechanisms (e.g., navigation bars and links) to facilitate interaction with the information structure.
- *Graphic design* focuses on visual presentation and layout.
- *Experience design* is an overarching aspect and encompasses all three of these categories, as well as properties that affect the user experience, such as download time, the presence of graphical ads, popup windows, etc. [Creative Good, 1999; Shedroff, 2001].

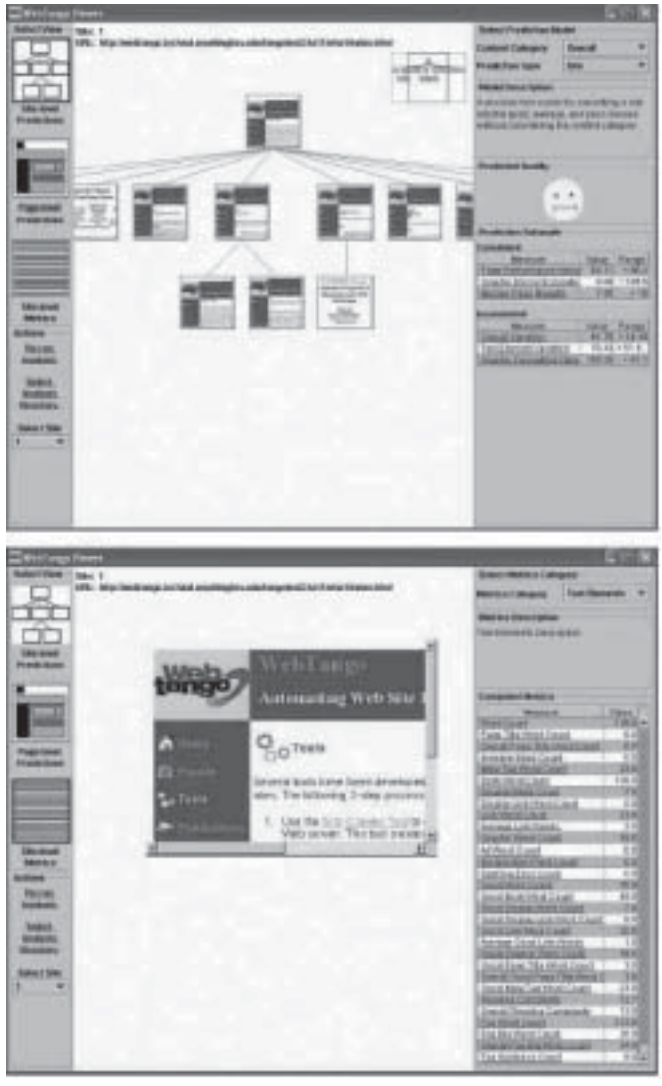

Figure 4: WebTango Viewer Tool. The top screen shot depicts site quality results. The bottom screen shot depicts page-level measures.

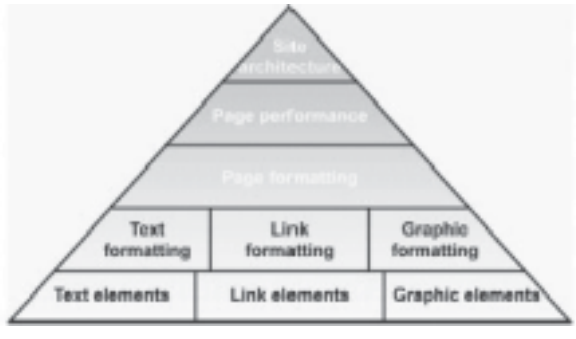


Figure 5: Web page and site structure. Text, link, and graphic elements are the building blocks of a web interface. Page- and site-level features use these elements to improve the user's experience.

We can further refine information, navigation, graphic, and experience design into the aspects depicted in Figure 5. The figure shows that text, link, and graphic elements are the building blocks of web interfaces; all other aspects are based on these. The next level of Figure 5 addresses formatting of these building blocks, while the subsequent level addresses page-level formatting. The top two levels address the performance of pages and the architecture of sites, including the consistency, breadth, and depth of pages. The bottom three levels of Figure 5 are associated with information, navigation, and graphic design activities, while the top two levels-Page performance and Site architecture-are associated with experience design.

### • 5.2 Web Interface Measures

We conducted an extensive survey of web design literature, including texts written by recognized experts (e.g., [Fleming, 1998; Nielsen, 2000; Sano, 1996; Spool *et al.,* 1999]) and published user studies (e.g., [Bernard and Mills, 2000; Bernard *et al.,* 2001; Boyarski *et al.,* 1998; Larson and Czerwinski, 1998]) to identify key features that impact the quality, usability, accessibility, performance, and so on of web interfaces [ivory *et al.,* 2000]. We did not consult HTML style guides, because researchers have shown them to be highly inconsistent [Ratner *et al.,* 1996]. We identified sixty-two features from the literature, including: the amount of text on a page, fonts, colors, consistency of page layout in the site, use of frames, and others. We then developed 157 quantitative measures to assess many of the 62 features.

Our early metric development work showed that we could use twelve web interface measures- Word Count, Body Text Percentage, Emphasized Body Text Percentage, Text Positioning Count, Text Cluster Count, Link Count, Page Size, Graphic Percentage, Graphics Count, Color Count, Font Count, and Reading Complexity-to accurately distinguish pages from highly rated interfaces [ivory *et al.,* 2000; Ivory *et al.,* 2001]. Table 1 describes the 157 quantitative measures (including nine of the original twelve measures and variations of the other three) that we de-veloped to assess aspects of the information, navigation, graphic, and experience design of web interfaces. These measures provide some support for assessing 56 of the 62 features (90 per-cent) identified as impacting usability and accessibility in the web design literature. Measures developed in previous studies assessed less than 50 percent of these 62 features.

In developing our web interface measures, we adhered to guidelines provided for determining performance metrics. Specifically, we implemented a subset of measures with the following characteristics.

- **Low variability:** measures are not a ratio of two or more variables; there is one exception to this rule-the average number of words in link text-which was developed to assess a feature reported in the literature.
- Nonredundancy: two measures do not convey essentially the same information.
- Completeness: measures reflect all aspects of the web interface (i.e., information, navi-gation, graphic, and experience design).

To validate the implemented measures, we used a sample of fourteen web pages t hat had widely differing characteristics. We manually computed the actual value of each measure and then compared these values to the automatically computed values. With a few exceptions, all the measures were highly accurate (greater than 84 percent accuracy across the sample). The least accurate measures-text positioning count (number of changes in text alignment from flush left) and text and link text cluster counts (areas highlighted with color, rules, lists, etc.)-require image processing to assess more accurately.

## 6 Development of Web Interface Profiles

The second and third steps in the WebTango process entail computing the page-level measures for a sample of rated pages and sites and then employing data mining approaches to develop models that can discriminate key features of high-quality interfaces; these steps correspond to the top part of Figure 2. We have completed four model-building efforts and, at the time of publication, are completing our fifth such effort. The studies have shown that it is possible to derive novel types of web design guidelines (statistical models) by dissecting good and poor designs. (Dissecting example designs is a common practice in developing design guidelines and patterns [Flanders and Willis, 1998; Heller and Rivers, 1996; Johnson, 2000; Norman, 1990; Smith and Mosier, 1986; van Duyne *et al.,* 2002].) Fur-

Table 1: Measures for assessing web design quality and usability. (Each category corresponds to a block in Figure 5.)

| *Category* | *Number of measures* | *Aspects measured* |
|---|---|---|
| Text elements | 31 | Amount of text, type, quality, and complexity. Includes visible and invisible text |
| Link elements | 6 | Number and type of links. |
| Graphic elements | 6 | Number and type of images. |
| Text formatting text | 24 | How body text is emphasized; whether some under-lined text is not in links; how text areas are highlighted; font styles and sizes; number of text colors; number of times text is repositioned. |
| Link formatting lined or | 7 | Colors used for links and whether there are text links that are not under-colored. |
| Graphic formatting | 7 | Minimum, maximum, and average image width and height; page area covered by images. |
| Page formatting | 27 | Color use, fonts, page size, use of interactive elements, page style control, and so on. Key measures include evaluating the quality of color combina tions (for text and panels) and predicting the functional type of a page. [a] |
| Page performance | 37 | Page download speed; page accessibility for people with disabilities; presence of HTML errors; and "scent" strength.[b] |
| Site architecture | 16 | Consistency of page elements, element formatting, page formatting and performance, and site size (number of pages or documents).[c] |

[a] The decision tree for predicting page type-home, link, content, form, or other-exhibited 84 percent accuracy for 1,770 pages.

[b] Our model predicts download speed with 86 percent accuracy. It considers the number and size of HTML, graphic, script, and object files and tables on the page. We use output from Bobby 3.2 (http://www.cast.org/bobby/) [WatchFire, 2002] runs to report accessibility errors. We report the total number of HTML errors determined by Weblint 1.02 [Bowers, 1996]. To assess scent quality, we report word overlap between the source and destination pages; the source link text and destination page; and the source and destination page titles.

[c] Consistency measures are based on coefficients of variation (standard deviation normalized by the mean) across measures for pages within the site. The site size measures only reflect the portion traversed by the crawler.

thermore, a preliminary study (see Section 8) suggests that web designers can use such guidelines to improve their sites. In this section, we summarize our four model-building efforts and then contrast models that we built during the last two efforts.

## 6.1 A Simple Prediction Model

The first study presented a preliminary analysis of a collection of over 400 informational web pages [Ivory *et al.,* 2000]. We labeled web pages as rated (that is, rated favorably by users or experts) and unrated (those that had not been so rated). For each web page, our early metrics computation script computed twelve quantitative measures having to do with page composition, layout, amount of information, and size (e.g., number of words, links, and colors). The analysis involved applying a linear discriminant classifier to the page types (rated and unrated) to assess if the measures could predict the pages' standings within these groups; the predictive accuracy was 63 percent and 6 measures-text cluster count, link count, page size, graphics count, color count, and reading complexity-were associated significantly with rated sites. Analysis of home pages revealed that they had measurably different characteristics than the other pages, suggesting the need for

context-sensitive models.

Exploring relationships among measures enabled us to hypothesize about key design aspects of rated pages (e.g., that they used a multi-level heading scheme, with a different color for each heading level, to facilitate scanning [Nielsen, 2000; Schriver, 1997; Spool *et al.,* 1999]); we validated our hypotheses by inspecting random page samples.

## 6.2 Context-Sensitive Prediction Models

The second study reported an analysis of 1,898 pages from sites evaluated for the Webby Awards 2000 [ivory *et al.,* 2001; The International Academy of Arts and Sciences, 2000]. For the first stage of the Webby Awards, anyone can submit a web site for review, thus sites vary widely from those that are well-designed to those that are poorly designed. At least three expert judges evaluated each submitted site on six criteria: content, structure and navigation, visual design, functionality, interactivity, and overall experience; the six criteria correlated highly. Web sites were also classified into 27 topical groups (content categories).

We conducted a usability study of 57 evaluated sites to examine the relationship between Webby judges' scores and ratings assigned by participants (non experts) who used sites to complete tasks [ivory, 2001, Chapter 7]. Although the results suggested some relationship between expert and end user ratings, strong conclusions could not be drawn because the study was conducted at least six months after the judges reviewed the sites. Statistical analysis of judges' ratings revealed that content was the predominate factor in overall ratings and visual design was the least significant factor in most cases [Sinha *et al.,* 2001]. Furthermore, the analysis showed that assessment criteria varied in importance based on the content category, suggesting that judges considered the genres of sites.

For the second study, we obtained pages from sites in six categories-community, education, finance, health, living, and services-and computed the same quantitative measures examined in the first study, except for reading complexity. We grouped sites according to their overall score in the Webby standings as follows: *good* (top 33 percent of sites) versus either *not-good* (remaining 67 percent of sites) or *poor* (lowest 33 percent of sites). The analysis involved developing two statistical models to assess if the measures could predict the pages' standings within these groups. The first model used multiple linear regression to distinguish good from not-good sites; the predictive accuracy was 67 percent when content categories (e.g., community and education) were not considered, and even higher on average when the categories were assessed separately. The second model used discriminant classification analysis to compute statistics for good versus poor sites. The predictive accuracy of the second model ranged from 76-83 percent when categories were considered.

## 6.3 Elaborate Prediction Models

Our third study reported an analysis of 5,346 pages and 333 sites from the Webby Awards 2000 [ivory and Hearst, 2002a; Ivory and Hearst, 2002b]. The analysis used the 157 quantitative page-and site-level measures, the six content categories, and a page type classifier (for distinguishing among home pages, content pages, link pages, forms, and other pages). Using this extensive set of interface measures, we were able to develop more sophisticated statistical models for distinguishing pages and sites in the *good* (top 33 percent of sites), *average* (middle 34 percent of sites), and *poor* (bottom 33 percent of sites) groups. Specifically, we developed context-sensitive models to predict page and site standings based on the content category and on the functional type of pages; models to predict standings independent of the design context were also developed.

The accuracy of page-level models ranged from 93-96 percent, and the accuracy of site-level models ranged from 68-88 percent; the site-level accuracy was considerably less possibly due to inadequate data. We used K-means clustering to partition the web pages from good sites into three sub-groups (small-page, large-page, and formatted-page). These clusters had significantly different characteristics and provided more context for evaluating web designs.

Another limitation of the site-level models is that they do not take page-level quality into consideration. Thus, it is possible for a site to be classified as good even though all the pages in the site are classified as poor and vice versa. To remedy this situation, we compute the median predictions for pages in the site. We also report an aggregate site-quality prediction by applying some heuristics to the site-level and median page-level predictions. For instance, if the median page-level prediction is average and the site-level prediction is good, then the aggregate site-level quality is reported as average. These additional predictions need to be considered in determining the overall quality of a site.

We incorporated the models developed in this study into the current Analysis Tool prototype

## 6.4 Expanded Content Category Models

Our fourth study entailed the analysis of 4,833 pages and from 570 sites from the Webby Awards 2002 [Ivory, 2003c]. Similarly to the third study, we used the 157 quantitative page- and site-level measures, the six content categories, and the page type classifier. We also incorporated four additional informational content categories-activism, best practices, print & zines, and travel. To examine how well the approach can be applied to functional sites, we built models for commerce sites. The accuracy of page-level models ranged as follows: 90-95 percent (overall page quality), 93-96 percent (content category quality), and 80-85 percent (page type

Table 2: Profiles used to assess web page quality.

| *Profile* | *Model Type* | *Assessment* | *Output* |
|---|---|---|---|
| Overall page quality· | Decision tree | Classifies pages as good, average, or poor, of page type or content category. | Rule that generated the regardless prediction. |
| Closest good-page cluster | K-means clustering | Maps pages into small-, large-, or formatted-page clusters. | Distance between a page and the closestcluster's centroid. |
| | | | Top 10 measures consistent with the cluster.[a] |
| | | | Top 10 measures inconsistent with the cluster and accept able metric ranges.[a] |
| Page type quality | Discriminant | Classifies pages as good, average, classification | Top 10 measures consistent or poor, according to page type.with the page type.[a] |
| | | | Top 10 measures inconsistent with the page type and acceptable metric ranges.[a] |
| Content category quality | Discriminant | Classifies pages as good, average, classification | Top 10 measuresconsistent or poor, according to content category.with the content category.[a] |
| | | | Top 10 measures inconsistent with the content category and acceptable metric ranges." |

[a] Measures are ordered by their importance in distinguishing pages in the three clusters (or classes) as determined from analyses of variances (ANOVAs).

(Tables 2 and 3). The prototype reports predictions (i.e., whether an interface belongs to the good, average, or poor group based on its quantitative measures) along with justifications, such as decision tree rules and how interface measures are similar or different to/from measures for highly rated interfaces. The prototype enables designers to consider the context (e.g., the content type, functional type, page size, and overall site structure) in which pages and sites are designed during assessment. (Badre and Laskowski [2001] have shown context to be an important consideration in web site design.) The Viewer Tool enables designers to interactively explore quality assessments and quantitative measures.

Table 3: Profiles used predictions at the page

| *Profile* | *Model Type* | *Assessment* | *Output* |
|---|---|---|---|
| Overall page quality· | Decision tree | Classifies sites as good, average, or poor, regardless of content category. | • Rule that generated the prediction. |
| Median overall page quality | Statistical | Classifies sites as good, average, or poor, based on the median page quality (overall page quality model). | — |
| Aggregate overall page quality | Derived heuristics | Classifies sites as good, average, or poor, based on the median overall page quality and the overall site quality models. | — |
| Content category quality | Decision tree | Classifies sites as good, average, or poor, according to content category. | • Rule that generated the prediction. |
| Median content category quality | Statistical | Classifies sites as good, average, or poor, based on the median page quality (content category quality model). | |
| Aggregate content category quality | Derived heuristics | Classifies sites as good, average, or poor, based on the median content category quality and the site content category quality models. | — |

quality). These ranges are consistent with the ranges for the 2000 models; the page-type models actually improved in accuracy by five to ten percent. The dataset was not amenable to rebuilding the cluster models that we had built for the 2000 dataset; we believe this difference is attributable to the broader range of content categories represented. The accuracy of site-level models was 89-91 percent for the overall site quality and 82-92 percent for the content category models. We attribute this significant increase in prediction accuracy to the larger dataset that we used (720 versus 333 sites).

Future work will entail incorporating the original and the new prediction models into the Analysis Tool. We are currently developing models from the 2003 Webby Awards. The 2003 dataset contains measures for roughly three times as many pages and sites than those used in previous studies. In addition, it includes data for seven additional content categories- government & law, science, news, politics, spirituality, sports, and youth.

## 6.5 Evolution of the Web Interface Models

Given that we have used the same quantitative measures for the past three model-building efforts, it is possible to determine the degree to which the old models predict new ratings and vice versa. For example, the 2000 page-level models predicted the ratings of the 2002 pages with 85 percent consistency. The 2000 site-level models predicted the ratings of the 2002 sites with 78 percent consistency; we attribute this lower consistency to the fact that the 2000 site-level models are not as accurate as the page-level models.

When we examine model consistency in the opposite direction (i.e., how well do the 2002 models predict 2000 ratings), there is very little consistency among predictions (33-35 percent). One possible interpretation of this discrepancy is that the effective design practices (e.g., use of headings and link clustering) that designers used in 2000 are still being used. However, design practices appear to have evolved. While some practices are still the same, new findings and tools, as well as developed design expertise, has lead to new design practices. Hence, periodically rebuilding the models enables us to capture these evolving design practices. We will examine this hypothesis with the 2003 models and a future empirical study.

## 7 Application of the Web Interface Profiles

The fourth step in the WebTango process entails using the profiles to assess and improve the quality of other web designs; we elaborate on this task in this section. The intent of this section is three-fold: (1) to demonstrate how the models can be systematically applied to this problem, (2) to illustrate the type of design changes informed by the models and how they vary across models, and (3) to highlight the current limitations of the models. The example assessment closely follows the evaluation scenario depicted in Figure 1, which is the overarching goal of the WebTango Project. Currently, interpreting model predictions and determining appropriate design changes is a manual process, though the Viewer Tool provides some support. Future work will focus on automating recommendations for improving designs as well as implementing these recommendations. We will also incorporate mechanisms for identifying comparable high-quality designs.

We begin this section with a summary of the process for assessing web design quality. We then demonstrate use of the 2000 models for assessing and improving two example sites and show that the model output does inform design improvements.

### 7.1 Assessing Web Design Quality

Figures 1 and 3 depict how a web site designer might use the WebTango method and tools to evaluate a web design. Essentially, the designer submits a partially designed site (HTML representation) to the Analysis Tool, which uses the metrics tool to compute the quantitative measures. It then compares these measures to the profiles of highly rated designs; it can conduct this analysis based on the context-general content category, page size, and page type. The tool reports differences between the submitted design and similar well-designed sites (Tables 2 and 3). The designer can use the Viewer Tool to explore these results and to inform design improvements. He can repeat the assessment process as necessary.

- We describe two example assessments in the remainder of this section. The first example demonstrates the assessment of a single, simplistic page; the second example demonstrates the assessment of a small site with a more complex design. As background to the discussion, we mention briefly results from a small study wherein users evaluated the example pages and sites [ivory, 2001, Chapter 9]. Three students (two undergraduates and one graduate) used the profiles to refine (manually) four web sites, and the authors modified an additional site (described in the second assessment example). During the study, 13 participants made 15 page-level comparisons and four site-level ratings of the original and modified versions of the sites. Section 8 describes the study in detail.

### 7.2 Example Assessment # 1

Figures 6 and 7 depict the original and modified versions of an example page from our study (discussed in the subsequent section). The overall page quality model classifies the original page as poor, mainly because no font smaller than nine point was used and because images (not shown in the figure) at the bottom of the page are formatted in a way that makes the page longer than necessary. Good sites that contain nonessential information in the footer tend to signal this by placing this information in a smaller font size.

The good-page cluster model provides insight about design quality. It reports that the page is 23.05 standard deviation units from the large-page cluster centroid. The model also reports several key deviations from the cluster, such as inadequate text and poor text positioning.

We modified the page based on the overall page quality and large-page cluster model. We improved text layout by introducing a second text column and reducing the top navigation area to one line. We also removed horizontal rules to reduce vertical scrolling as dictated by the large-page cluster model. Ten of the 13 study participants preferred the modified page to the original one after these conservative changes were made.

### 7.3 Example Assessment # 2

Figures 8-10 show three pages taken from a small (nine-page) site in the Yahoo Education/Health category. The site provides information about training programs offered to educators, parents, and children on numerous health issues, including leukemia and cerebral palsy. We selected the site because it was not in the training set or testing set, and also because on first glance, it appeared to have good features, such as clear and sharp images and a consistent page layout, but on further inspection it seemed to have problems. We focused on answering the following questions.

- Is this a high-quality site? Why or why not?
- Are these high-quality pages? Why or why not?
- What can be done to improve the quality of this site?

The first step was to download a representative set of pages from the site. For this particular site, only eight level-one pages were accessible, and no level-two pages were reachable, for a total of nine downloaded pages. Although there is a page containing links (Figure 9), the links are to pages that are external to the


Figure 6: Original version of the web page for the first assessment example.

site.

The next step was to use the Analysis Tool to compute site- and page-level measures and to apply the models to individual pages and to the site as a whole. Each model encapsulates relationships between key predictor measures and can be used to (1) generate quality predictions and (2) determine how pages and sites are consistent with or deviate from good pages and sites.

In the discussions below, when decision tree rules are used to generate predictions, the consequences are interpreted manually. When cluster models are applied, the score for each measure on an individual page is compared to that of the cluster centroid, and if the measure differs by more than one standard deviation unit from the centroid, the measure is reported as being inconsistent with the cluster. Cluster deviations are also interpreted manually. Although the Viewer Tool provides support for interpreting predictions, it was not developed when the assessment was done [Ivory, 2001; Ivory and Hearst, 2002b]. Detailed information about this example assessment is available in [ivory, 2001, Chapter 8].

### 7.3.1 Site-Level Assessment

The example site can be classified in both the health and education content categories, so we ran the site-level decision tree model initially without differentiating by content category. The site-level model predicted that the site is similar to poor sites overall; the median page quality prediction (i.e., median computed over the overall page quality model's predictions for the nine pages; poor) is consistent with the overall site quality model's prediction. The corresponding decision tree rule (top of Figure 11) reveals that the site has an unacceptable amount of variation (i.e., inconsistency) in link elements (31 percent), although variation for other site-level measures is acceptable. The combination of the link element variation and the lack of a comparable overall element variation violates patterns discovered on good sites.

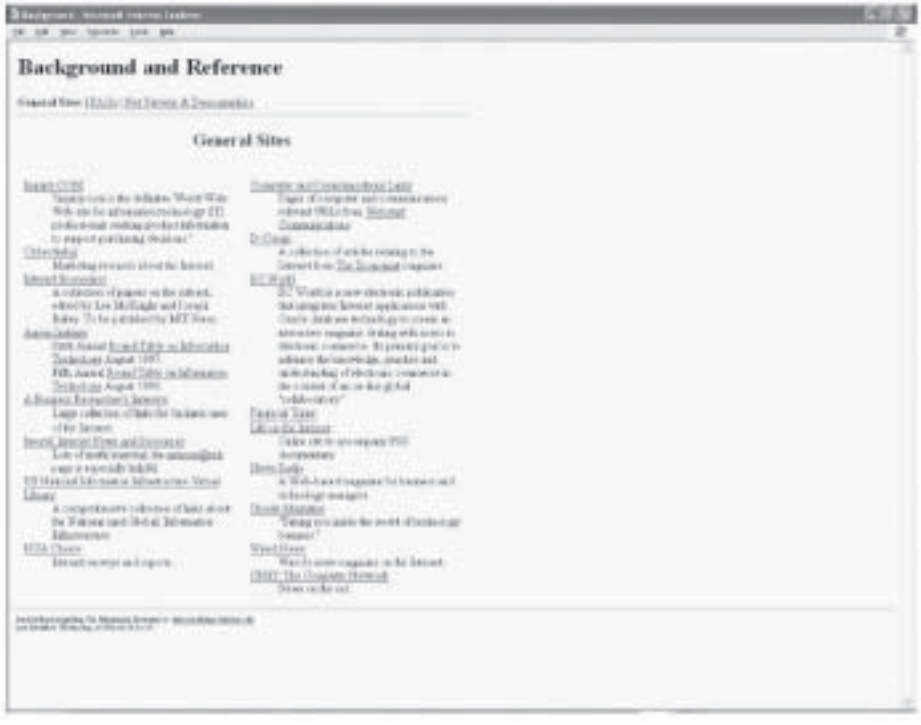


Figure 7: Modified version of the web page for the first assessment example. Students based their improve-ments on the overall page quality and closest good-page cluster models described in Table 2. (Some of the changes in the modified page are not visible.)

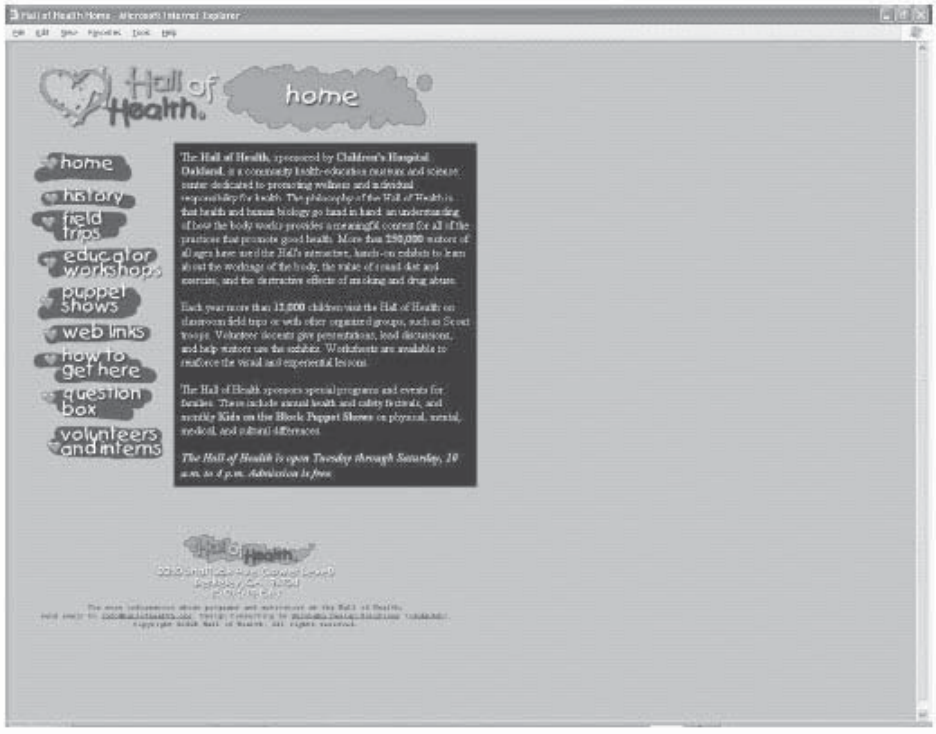


Figure 8: Home page taken from the example health education site (http: //www. hallof health. org/home . html; September 14, 2001).


Figure 9: Link page taken from the example health education site (http://www.hallofhealth.org/ weblinks.html; September 14, 2001).

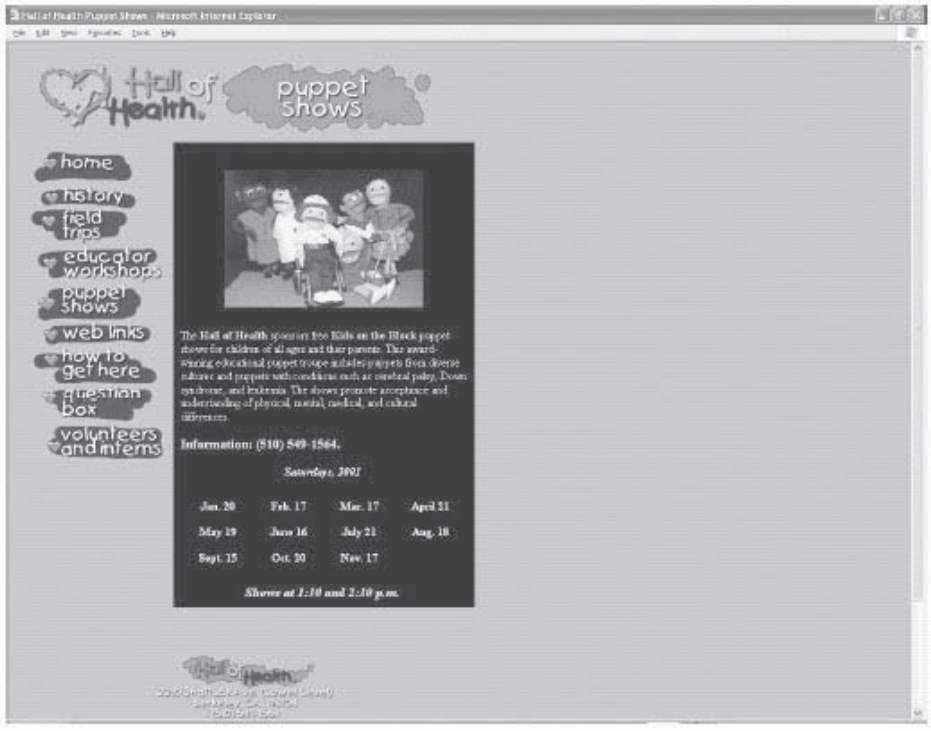


Figure 10: Content page taken from the example health education site (http://www.hallofhealth.org/puppetshows.html; September 14, 2001).

**Overall Site Quality**

if ((Page Performance Variation is missing OR (Page Performance Variation < 90.2)) AND (Overall Variation is not missing AND (Overall Variation < 14.49)) AND (Link Element Variation is missing OR (Link Element Variation > 29.195)) AND (Overall Element Varia-tion is missing OR (Overall Element Variation < 26.07))) Class = Poor

This rule classifies the site as poor because the pages have acceptable page performance, overall, and overall element variation, but they have more than 29.2 percent variation in link elements (30.68 percent).

**Health Site Quality**

if ((Graphic Element Variation is not missing AND (Graphic Element Variation < 32.695)) AND (Text Element Variation is missing OR (Text Element Variation > 47.45)) AND (Text Element Variation is missing OR (Text Element Variation < 92.25))) Class = Poor

This rule classifies the site as poor because the pages have acceptable graphic element variation, but they have between 47.45 percent and 92.25 percent variation in text elements (53.18 percent).

**Education Site Quality**

if ((Median Page Breadth is missing OR (Median Page Breadth < 11.25)) AND (Page Title Variation is missing OR (Page Title Variation < 196.7)) AND (Page Formatting Variation is missing OR (Page Formatting Variation < 27.785)) AND (Page Title Variation is missing OR (Page Title Variation < 132.495)) AND (Graphic Formatting Variation is not missing AND (Graphic Formatting Variation < 16.165))) Class = Good

This rule classifies the site as good due to an acceptable combination of measures: the median page breadth (8) is less than twelve; and pages in the site have very little similarity in page titles (37.5 percent), page formatting variation (0 percent), and graphic formatting variation (3.19 percent).

Figure 11: Decision tree rules reported for the example health education site. The rules were reported by the overall (top), health (middle), and education (bottom) site quality models.

The major source of link element variation is the text link count. Eight out of nine pages have from two to four text links; the remaining page has 27 text links, and acts as a links page (see Figure 9). The decision tree rule suggests that a link element variation level below 29 percent is typical on good sites. One interpretation of this finding is that good sites strive to keep the navigation structure consistent among pages and may even distribute links over multiple pages to maintain this consistency. Hence, the rule may indicate the need to similarly redistribute the links on this page.

We also assessed site quality according to the two applicable content categories-health and education. The decision tree for health sites predicted that the site is a poor health site (middle of Figure 11). In this case the problem is inadequate text element variation. Most of the pages on the site contain paragraphs of text without headings and use only one font face (serif); this may actually make it harder for users to scan the page to find the information they are looking for [Nielsen, 2000; Spool *et al.,* 1999]. The median health page quality prediction (poor) is consistent with the health site prediction.

The decision tree for education sites made a prediction contrary to that for sites overall and health sites; it found this site to be consistent with good education sites (bottom of Figure 11). Good health and good education sites are similar with respect to graphic formatting variation, but are quite different on the other measures, which is the cause for this disparity. However, as will be discussed below, the median education page quality is poor.

### 7.3.2 Page-Level Assessment

The decision tree model for predicting page quality reports that all nine of the pages are con-sistent with poor pages. The home page (Figure 8) contains seventeen italicized words in the body text; the model considers pages with more than two italicized words in the body text to be poor pages (see rule at the top of Figure 12). Schriver [1997] suggests that italicized text should be avoided because it is harder to read on computer screens than in printed documents

We developed a color measure, Minimum Color Count, to track the number of times each color is used on a page and to report the minimum number of times a color is used; this measure detects the use of an accent or sparsely-used color. The model classifies the content page (Figure 10) as poor mainly because the minimum number of times a color is used is sixteen and all of the text, including the copyright text at the bottom of the page, is formatted with a font size greater than 9pt (bottom of Figure 12). Good pages tend to have an accent color that they use sparingly, whereas poor pages seem to overuse accent colors. Good pages also tend to use a smaller font size for copyright or footer text, unlike poor pages. Additionally, the example content page contains 34 colored body text words, which is twice the average number found on good pages; in the extreme case, a large number of colored words could result in the uncolored words standing out more so than the colored words. The same prediction and decision tree rule is reported for the link page.

**Home Page**

if ((Italicized Body Word Count is not missing AND (Italicized Body Word Count > 2.5))) Class = Poor

This rule classifies the home page as poor because it contains more than two italicized words (17) in the body text.

**Link and Content Page**

if ((Italicized Body Word Count is missing OR (Italicized Body Word Count < 2.5)) AND (Minimum Font Size is not missing AND (Minimum Font Size > 9.5)) AND (Minimum Graphic Height is missing OR (Minimum Graphic Height < 36)) AND (Minimum Color Use is not missing AND (Minimum Color Use > 15.5))) Class = Poor

This rule classifies both the link and content pages as poor because they contain an ac-ceptable number of italicized words in the body text and contain at least one image with a height less than 37 pixels, but all of the text is formatted with a font greater than 9pt and all of the colors are used more than fifteen times. Good pages tend to use a font smaller than 9pt typically for copyright text, and they use an accent color.

Figure 12: Decision tree rules reported for the three example pages. These rules were reported by the overall page quality model.

To gain more insight about ways to improve page quality, we used the tool to map each page into one of the three clusters of good pages-small-page, large-page, or formatted-page. All pages mapped to the small-page cluster and were far from the cluster centroid (median distance of 10.9 standard deviation units); the page from the dataset that is closest to the center of this cluster has a distance of 4.0 standard deviation units. Pages in the example site deviate on key measures that distinguish pages in this cluster, including the graphic ad, text link, link text cluster, interactive object, and link word counts. Table 4 summarizes, for the sample content page, the ten key measures (i.e., measures that play a major role in distinguishing pages in this cluster) that deviate from the cluster centroid; deviations are similar for other pages in the site. Most of these deviations, including two of the top ten measures (text link count and good link word count), can be attributed to the fact that the site provides predominately graphical links instead of textual links for navigation. Table 4 also shows deviation on the page height (vertical scrolls), the use of words formatted with sans serif fonts (sans serif word count), and the overall use of fonts (font count-combinations of a font face, size, bolding, and italics).

We also evaluated the quality of these pages using the more context-sensitive page quality models for health and education pages (as opposed to the overall model). The health model predicted that all but two pages were poor health pages, which mirrors the results of the site-level model. However, the education model predicted that all pages were poor education pages, which contradicts the corresponding site-level prediction. In both cases, prediction rationales were similar to the issued mentioned above. The contrast between site- and page-level predictions demonstrates the need to incorporate page-level predictions into the site-level prediction, as previously discussed.

Finally, we evaluated the quality of these pages using the models for each page type-home, link, content, form, and other. The page type decision tree made accurate predictions for six of the nine pages, but inaccurately predicted that three pages were consistent with link pages; visual inspection suggested that these pages were actually content pages. Mispredictions were mainly due to an improper balance of link, body, and display text stemming from an overuse of image links. After correcting the page type predictions, the models classified all nine pages as poor pages. The page type quality models reported several deviations that were also reported by other models, including the minimum font size, minimum color use, sans serif word count,and text link count.

Table 4: Top ten measures that deviate from the small-page cluster for the example content page.[a]

| *Measure* | *Value* | *Cluster Range* |
|---|---|---|
| Vertical Scrolls | 2.0 | (0.56-2.00) |
| Text Column Count | 5.0 | (0.62-4.36) |
| All Page Text Terms | 129.0 | (138.59-353.24) |
| Link Count | 12.0 | (12.40-41.24) |
| Text Link Count | 2.0 | (4.97-27.98) |
| Good Link Word Count | 3.0 | (7.43-49.67) |
| Bobby Browser Errors | 6.0 | (7.54-14.99) |
| Font Count | 6.0 | (3.64-5.80) |
| Sans Serif Word Count | 0.0 | (13.91-253.57) |
| Display Word Count | 33.0 | (1.13-18.67) |

[a] The measures are presented in their order of importance, as determined by analyses of variances (ANOVAs). Each range reflects one standard deviation unit around the metric value at the cluster centroid. The page's measures are 8.33 standard deviation units from the cluster centroid.

### 7.3.3 Summary of Assessment Findings

The models provide some direct insight for resolving design issues associated with some of the measures. For example, decision tree rules reported by the overall page quality model indicate inconsistent measures with a ">" in the threshold (e.g., italicized body word count > 2.5); they also indicate consistent measures with a "<" in the threshold. The Viewer Tool parses decision trees rules and displays consistent and inconsistent measures. The designer could explore ways to reduce measure values below the thresholds, such as removing italics or text coloring, changing font sizes, breaking text into multiple columns, etc. The same guidance holds for the other decision tree models.

Similarly to the decision tree rules, the cluster and discriminant classification models provide ranges for acceptable metric values; they also report the top ten measures that deviate from the underlying models. Some of the model deviations are straightforward to correct, provided the designer understands the model output and relevant measures. Other model deviations are not as straightforward to correct, such as introducing additional links and content or reducing the reading complexity. Future work on automating design

changes should make it easier to interpret and use the models to improve designs.

Based on the analysis, we derived a list of possible ways to improve the site. The changes below are ordered based on their potential impact (how much they mitigate measures that were reported frequently as being inconsistent). The recommendations only apply to the results generated during the initial application of the models; subsequent model applications revealed further changes that are not discussed here. No recommendations are made to address the accessibility and Weblint errors, since the roles of these measures in improving design quality are unclear. Specific changes made as well as the results of the changes are discussed in the next section.

1. Increase the number of text links and corresponding link text (text link, link word, and good link word counts). This will simultaneously increase the total number of links andinternal links (link and internal link count) and decrease link element variation.
2. Use a smaller font size for some text, such as the footer text (minimum font size).
3. Decrease color overuse for page text and introduce an accent color (minimum color use).
4. Minimize or eliminate the use of italicized words in body text (italicized body word count).
5. Minimize text positioning (changes from flush left and columns where text starts; text positioning and column counts).
6. Minimize font combinations (font face, size, bolding, and italics combinations; font count).
7. Reduce the sizes of images (average graphic width, minimum graphic height, and graphic pixels).
8. Improve the page layout to reduce vertical scrolling (vertical scrolls).
9. Use tables with explicit widths to control the page layout (fixed page width use).
10. Vary the text elements and the formatting of text elements on the page (element variation,good body and display word counts, sans serif word count).
11. Reduce the number of colors used for body text (body color count).

### 7.3.4 Improving the Site

Although the example site is somewhat aesthetically pleasing and highly consistent across pages within the site, the individual pages and the site as a whole are classified as being of poor quality. We modified the pages to incorporate a subset of the recommendations discussed above.

- To improve the color and text link counts and simultaneously reduce the link count variation, a link text cluster (i.e., an area of text links shaded with a different background color to make it stand out) was added as a footer at the bottom of each page; the textual links in the cluster mirror the content of the graphical links. It was not necessary to split the link page into multiple pages, because adding the footer decreased the link element variation from 31 to 7 percent.
- To improve text formatting and the text element variation score: headings were used to break up paragraphs; additional font variations were used-Arial font (sans serif) for body text and Trebuchet (serif) for headings; and the font size of the copyright text was reduced to 9pt. The color of headings was also changed to gold for consistency with the models. We implemented all changes via an internal style sheet; the addition of the style sheet also improved the self-containment scores (i.e., degree to which all page elements are rendered solely via the HTML and image files).
- To improve the emphasized (i.e., bolded, colored, italicized, etc.) body text scores, italics and colors within body text were converted to bold, uncolored body text on all pages. Colored, non-italicized body text was also converted to uncolored body text.
- To improve the minimum color usage scores, a color accent was added to the vertical barsbetween the text links in the footer of each page. A browser-

Figure 13: Modified home page for the example health education site. Gold headings were added, sans serif fonts were used for body text, colored and italicized body text was removed, and a fixed page width of 640 pixels was used. A footer navigation bar was added to the bottom of the page, an accent color was added to the footer navigation bar, footer elements were reorganized to reduce vertical scrolling, and the font size of footer text was reduced; none of these changes are visible in the screen shot. See Figure 15 for the footer navigation bar.

Figure 14: Modified link page for the example health education site. Gold headings were added, sans serif fonts were used for body text, colored and italicized body text was removed, the sizes of images were reduced, and a fixed page width of 640 pixels was used. A footer navigation bar was added to the bottom of the page, an accent color was added to the footer navigation bar, footer elements were reorganized to reduce vertical scrolling, and the font size of footer text was reduced; none of these changes are visible in the screen shot. See Figure 15 for the footer navigation bar.

Figure 15: Modified content page for the example health education site. Gold headings were added, sans serif fonts were used for body text, colored and italicized body text was removed, the sizes of images were reduced, and a fixed page width of 640 pixels was used. A footer navigation bar was added to the bottom of the page, an accent color was added to the footer navigation bar, footer elements were reorganized to reduce vertical scrolling, and the font size of footer text was reduced; not all of these changes are visible in the screen shot.

safe color was selected asdictated by a subsequent prediction by the overall page quality model.

- To reduce vertical scrolling, the logo and copyright notice at the bottom of the pages wereplaced adjacent to each other in one table row. The sizes of images and borders around them were also reduced to improve space utilization. Furthermore, text was wrapped to the left of the images versus images not being inlined with text.
- To further improve the page layout, fixed widths (640 pixels) were used for the main layout table.

Figures 13-15 depict the revised pages corresponding to the pages in Figures 8-10; many of the changes are not visible, because they appear at the bottom of the pages. Furthermore, we implemented only a subset of the potential changes.

After making these changes, the models classified all pages correctly by functional type, and they rated them as good pages overall as well as good health pages. Figure 16 depicts the complex decision tree rule that classified all the pages as good overall. The median distance to the small-page cluster was 4.7 as compared to 10.9 standard deviation units for the original pages. Eight pages were rated as average pages based on their functional type; one was rated as poor. In addition, five of the nine pages were rated as average education pages; the four remaining pages were rated as poor. These differences in predictions demonstrate the potential difficulty of satisfying all the models simultaneously. Hence, a clear design objective needs to be chosen prior to making any changes, since the models could reveal a different set of changes to make.

The site was still classified as a poor site overall, but for a different reason-too much text element variation. The original site had very little variation in text elements (body and display text in particular); adding headings to pages increased the text element variation (75.5 percent) above the acceptable threshold of 51.8 percent. Ensuring that all pages contain similar amounts of display text is probably the simplest way to resolve this issue. Some pages, such as the example link page, have long headings, while other pages have relatively short headings. The site was also classified as a poor health site and a good education site, consistent with classifications before the modifications; the same decision tree rules were reported (see Figure 11). The median overall page, education page, and health page quality predictions contradicted the site-level models.

## 8 Evaluation of the Web Interface Profiles

The final step in the WebTango process entails validating the web interface profiles. This section presents findings from an empirical study of pages and sites modified based on the 2000 models [Ivory, 2001; Ivory and Hearst, 2002a; Ivory and Hearst, 2002b]. The study examines (1) whether it is possible for others to use the profiles to modify designs and (2) whether the resulting designs are of a higher quality than the original ones. A

detailed discussion of the study can be found in [Ivory, 2001, Chapter 9].

## 8.1 Study Design

We conducted a study to determine whether changes made based on two profiles-the overall page quality and the good cluster models (see Table 2)-improve design quality (i.e., usability, accessibility, performance, and so on). We randomly selected five study sites from various Yahoo categories, such as finance and education; study sites included the two discussed in Section 7. Two undergraduate students and a graduate student modified three of the study sites, while the tool developer modified the other two sites. The students had little or no training in web design and had very little experience with building web sites. Furthermore, they did not have prior experience with the Analysis Tool, the quantitative measures, nor the profiles. They made straightforward changes based directly on the decision tree rules and cluster model results. Thirteen participants (people who had and did not have web design experience) completed a within-subjects experiment wherein they performed two types of tasks. The first task-page-level analysis-required participants to explore the original and modified versions of a web page and to select the design that they felt exhibited the highest quality; there were a total of fifteen comparisons for pages from three sites. The second task-site-level analysis-required participants to explore a collection of pages from a web site and to rate the quality of the site on a 5-point scale. Participants rated original and modified versions of two sites; there were a total of four site ratings. Given that only a subset of pages were modified for each site, it was not feasible to have participants attempt to complete information-seeking tasks during this study.

## 8.2 Study Results

The page-level analysis focused on testing the hypoth-

if ((Minimum Font Size is missing OR (Minimum Font Size < 9.5)) AND (Graphic Ad Count is missing OR (Graphic Ad Count < 2.5)) AND (Exclaimed Body Word Count is missing OR (Exclaimed Body Word Count < 11.5)) AND (Minimum Graphic Height is missing OR (Minimum Graphic Height < 38.5)) AND (Vertical Scrolls is missing OR (Vertical Scrolls < 3.5)) AND (Bad Panel Color Combinations is missing OR (Bad Panel Color Combinations < 2.5)) AND (Object Count is missing OR (Object Count < 4.5)) AND (Good Meta Tag Word Count is missing OR (Good Meta Tag Word Count < 42.5)) AND (Minimum Color Use is missing OR (Minimum Color Use < 12.5)) AND (Horizontal Scrolls is missing OR (Horizontal Scrolls < 0.5)) AND (Weblint Errors is missing OR (Weblint Errors < 54.5)) AND (Colored Body Word Count is missing OR (Colored Body Word Count > 0.5)) AND (Emphasized Body Word Count is missing OR (Emphasized Body Word Count < 183)) AND (Bolded Body Word Count is missing OR (Bolded Body Word Count < 43.5)) AND (Script Bytes is not missing AND (Script Bytes < 173.5)) AND (Text Positioning Count is missing OR (Text Positioning Count < 9)) AND (Serif Word Count is missing OR (Serif Word Count < 325.5)) AND (Italicized Body Word Count is missing OR (Italicized Body Word Count < 1.5)) AND (Graphic Count is not missing AND (Graphic Count < 15.5)) AND (Minimum Graphic Width is missing OR (Minimum Graphic Width < 97.5)) AND (Bobby Browser Errors is missing OR (Bobby Browser Errors > 6.5))) Class = Good

This rule classifies a page as a good page because it: uses a smaller font size for some text; has fewer than sixteen images and no graphical ads; uses at least one image with a height smaller than 39 pixels as well as at least one image with a width smaller than 98 pixels; has fewer than 183 total emphasized (i.e., italicized, bolded, colored, etc.) body words, but has fewer than 11.5 exclaimed body words (i.e., body words followed by exclamation points), fewer than 44 bolded body words, fewer than two italicized body words, and at least 1 colored body word; requires fewer than four vertical scrolls and no horizontal scrolls; starts text in nine or fewer vertical positions; uses fewer than 2.5 bad panel color combinations and uses an accent color; uses no scripts, applets, or other objects; uses fewer than 43 good meta tag words and has fewer than 325 words formatted with serif fonts; and has fewer than 55 Weblint errors and more than six Bobby browser errors.

Figure 16: Decision tree rule reported for all the modified example pages, This rule was reported by the overall page quality model.

esis that pages modified based on the overall page quality and the good cluster models are of a higher quality than the original pages. The remodeling did not entail changing the content on pages, except for redistributing content over multiple pages, adding headings, etc. as dictated by the models. Thus, most of the differences between the original and modified pages were minimal and focused on the page layout and text formatting.

The results showed that modified pages were preferred 57.4 percent of the time, while the original pages were preferred 42.6 percent of the time. The Chi-Square test [Easton and Mc-Coll, 1997] revealed this difference to be significant ($\%^2 = 4.3$, asymptotic significance of .038). Participants preferred the modified pages more so than the original ones in ten of the fifteen

29 comparisons. Their comments about why they preferred the modified pages supported the changes made based on the profiles. In particular, participants felt that the modified pages were easier to read, required less scrolling, were cleaner, used better color schemes, made better use of whitespace, used headings, eliminated italics, and used better fonts.

Comments also revealed that in a few cases, mainly for four pages on the same site, partic-ipants responded negatively to changes made based on the profiles. For example, in modifying pages to minimize vertical scrolling, one of the web designers accidentally introduced horizontal scrolling. Participants preferred the original version of the pages, because the width of text was restricted to fit within the browser window (i.e., the page did not require horizontal scrolling). If we excluded responses for pages on this site, then the results for the two remaining sites show that modified pages were preferred 66.9 percent of the time, while the original pages were preferred 33.1 percent of the time; this difference was highly significant (% = 14.9, asymptotic significance of .000).

The site-level analysis focused on testing the hypothesis that sites with pages modified based on the overall page quality and the good page cluster models are of a higher quality than the original sites. Similarly to the page-level analysis, students used the profiles to modify individual pages in the site. Students relied on the median overall page quality (see Table 3) for site level assessments; as the quality of the individual pages improved, so did the median overall page quality. Students did not use the overall site quality model, due to the discrepancy between page- and site-level predictions that existed at the time of the study.

Participants rated the quality of the original sites as 3.0 on average *(a* = 1.36); however, they rated the quality of modified sites as 3.5 on average *(a* = 1.03). A paired samples t-test [Easton and McColl, 1997] revealed that this difference was significant *(p* = .025); this means that each participant tended to rate the modified version higher than the original version. Similarly to the page-level analysis, participant comments provided support for many of the changes made based on the profiles.

## 8.3 Study Implications

The study demonstrated that it was possible for people, other than the tool developer, to interpret and apply the models. However, it also demonstrated the need to ensure that errors are not introduced during this process. Both the page- and site-level results were promising, in that they showed that participants responded favorably to the changes that were made to the designs, based on the models, even though they were conservative for this first study. It is possible that less conservative changes would have resulted in larger differences in page preferences and site ratings. We will re-examine this question with future studies after we have implemented automated recommendations and possibly modifications in some manner.

# 9 Analysis of Web Design Guidelines

Another application of the quantitative measures and profiles is to revisit web design recom-mendations, specifically for recommendations that are contradictory, vague, or not empirically validated. Ideally, quantifying effective design patterns and revealing concrete thresholds (i.e., numerical ranges for measures) will provide the extra guidance that novice or occasional web designers need to build better sites. We have contrasted thresholds derived from our 2000 and 2002 models for several aspects of web interfaces, including the amount of text, font styles and sizes, and colors. The statistical models revealed quantitative thresholds that validate and, in some cases, invalidate advice in the literature. The design patterns extend beyond individual measures to show, for instance, that the amount of text formatting is proportional to the amount of text on a page (i.e., a change in one aspect may necessitate a change in another). We also 30 found that the guidance varies slightly depending on the design context (e.g., page style) and that fundamental design patterns have not changed radically over recent years.

Detailed discussions of our analysis of web design guidelines can be found in [Ivory, 2001, Chapter 10] and [Ivory, 2003b]. Our examinations demonstrate that the methodology makes it possible to derive web design

guidelines directly from empirical data.

## 10 Future WebTango Research Directions

Research that we have conducted on the WebTango Project represents an important first step toward enabling non-professional or occassional designers to iteratively improve the quality of their web site designs. The methodology and tools are still in their infancy and provide support only for refining an implemented site; thus, there are many ways in which they can be improved. We summarize several future research plans.

### 10.1 Advancing Research on Web Design and Evaluation

We will conduct research to facilitate advances in web interface design and evaluation that has broad implications for other automated web site evaluation researchers and for web practitioners. We are designing studies to identify an empirically validated web site design process, which we plan to support with a design tool. We have conducted a preliminary study and comparison of several existing automated evaluation tools [Ivory *et al.,* 2003; Ivory and Chevalier, 2002; Ivory, 2003a] and will continue examining the efficacy of such tools, including our own. We will also work to develop a corpus of usability-tested sites, as a way to further research on web site design and the validation of automated evaluation methodologies.

### 10.2 Improving the WebTango Measures

We plan to expand the set of quantitative measures to assess other design aspects, such as the reuse of web interface elements across pages in a site. A major limitation of the current set of measures is that they do not assess content quality. Future work will explore using text analysis techniques to possibly derive other measures of content quality. Another limitation is that the measures do not adequately gauge accessibility for the disabled; we found that the good web pages tended not to be accessible, as determined by the Bobby tool. Future work will examine other ways to measure accessibility, for instance computing the nesting level of tables. (Although tables may help sighted users scan pages, they may impede blind users.)

We are currently exploring the use of image processing techniques to classify design compo-nents, to improve the accuracy of existing measures, to enable the development of new ones, and to enable support for non-HTML pages and early design representations. Supporting early de-sign representations also requires adjustments to the profiles, such as ignoring certain measures during analysis.

### 10.3 Developing a Robust Evaluation Tool

All the WebTango tools need to be reimplemented as part of a robust, open source browser, such as Mozilla; this redevelopment will enable support for framesets, scripts, applets, and other objects, as well as real-time analysis. Real-time analysis is crucial for developing an interactive evaluation tool to support iterative design and evaluation.

Other key components of the interactive evaluation tool include: (1) recommending design improvements based on model predictions, (2) applying recommendations so users can preview the changes, and (3) showing comparable designs for exploration. Some model deviations are easy to correct, such as removing or changing text formatting, using good color combinations, 31 and resizing images; it is possible to automatically modify the HTML to incorporate these types of changes. Other changes, such as reducing vertical scrolling, adding text columns, improving readability, and adding links are not as straightforward. We consider the automated correction of design deviations to be crucial for improving the current state of the web. Our empirical studies of novice and professional designers demonstrate that designers find it extremely difficult to implement design guidelines, even with automated evaluation tools [ivory and Chevalier, 2002; Chevalier and Ivory, 2003a; Chevalier and Ivory, 2003b; Ivory *et al.,* 2003]; thus, automating design changes should help tremendously.

More work needs to be done to better understand the profiles before interactive evaluation can be supported. For example, factor analysis or multidimensional scaling techniques could be used to reduce the number of measures and to gain more insight about relationships among measures. This would enable recommendations and modifications to be based on combinations of measures versus individual measures.

## 11 Conclusions

Non-professional and occasional web site designers need additional support in creating high-quality sites. Automated evaluation and other approaches have been developed to address this need. We described the various approaches to automated web site evaluation and then elaborated on one methodology-the WebTango method. The WebTango method is consistent with measurement approaches used in the performance evaluation domain and guideline review approaches used in

the usability evaluation domain. Unlike other web assessment techniques, this approach uses quantitative not capture users' subjective preferences. For example, one study has shown that perceived download speed is more important than actual download speed [Scanlon and Schroeder, 2000]. Although we can measure actual download speed, it may not be possible to assess perceived speed. Nonetheless, the methodology can be viewed as a reverse engineering of design decisions that were presumably informed by user input.

## 12 Acknowledgments

This research was supported by a Hellman Faculty Fund Award, a Microsoft Research Grant, a Gates Millennium Fellowship, a GAANN fellowship, and a Lucent Cooperative Research Fellowship Program grant. We thank everyone that has contributed to the WebTango Project, including: Marti Hearst, Rashmi Sinha, Roderick Megraw, Alissa Harrison, Young-Mi Shin, 32 Shiquing Yu, Mary Deaton, Nicole Elger, Tina Marie, Wenchun Wang, Deep Debroy, Toni Wadjiji, Chantrelle Nielsen, Wai-ling Ho-Ching, Stephen Demby, David Lai, and Aline Chevalier. We thank Maya Draisin and Tiffany Shlain at the International Academy of Digital Arts and Sciences for making the Webby Awards data available. We thank Tom Phelps for his assistance with the Metrics Computation Tool.

## References

Adaptive Technology Research Center (2002). A-Prompt project. Available at http://aprompt.snow.utoronto.ca/

Ahlberg and Shneiderman (1994). Visual informa-tion seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the Conference on Human Factors in Computing Systems,*, Boston, MA, 313-317. New York: ACM Press.

AnyBrowser.com (2002) Any Browser.com: Your Source for Browser Com-patibility Verificaiton, Available at http://www.anybrowser.com/ScreenSizeTest. html.

Autowebmaker (2003). Automated site creation - Autowebmaker. Available at http://www.autowebmaker.com/

Bacheldor (1999). Push for performance. *Information Week,* September 20:18-20.

Badre, Aalbert and Laskowski, Sahron (2001).. The cultural context of web genres: Context vs. style. In *Proceedings of the 7th Conference on Human Fac-tors & the Web,* Madison, WI, Available at http://www.optavia.com/hfweb/7thconferenceproceedings.zip/Laskowski.pdf

Beirekdar, Abdo, Vanderdonckt, Jean, Noirhomme-Fraiture, Monique (2002). KWARESMI - knowledge-based web automated evaluation with reconfigurable guidelines optimization. In *PreProceedings of the 9th International Workshop on the De-sign, Specification, and Verification of Interactive Systems,* 362-376, Rostock, Ger-many, June 12-14 2002. Available at http://www.isys.ucl.ac.be/bchi/publications/ 2002/Beirekdar-DSVIS2002.pdf

Bernard, Michael and Mills, Melissa (2000). So, what size and type of font should I use on my website? *Usability News,* Available at http://wsupsy. psy.twsu.edu/surl/usabilitynews/2S/font.htm

Bernard, Michaael, Liao, Chia Hui Mills, Melissa(2001). The effects of font type and size on the legibility and reading time of online text by older adults. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 2, 175-176, Seattle, WA.

Blackboard, Inc. Welcome to blackboard. Available at http://www. blackboard.com/

Blackmon, Marilyn Hughes., Poison, Peter G, Kitajima, Muneo, Lewis, Clayton(2002). Cognitive walkthrough for the web. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 4 of *CHI Letters,* 463-470, Minneapo-lis, MN.

Blackmon, Marilyn Hughes, Kitajima, Muneo, Poison, Peter G (2003). Re-pairing usability problems identified by the cognitive walkthrough for the web. In *Proceedings of the Conference on Human Factors in Computing Systems,* pages 497-504, Fort Lauderdale, FL.BMC Software(2002). SiteAngel. Available at http://www.bmc.com/ siteangel/

Bowers, Neil (1996). Weblint: quality assurance for the World Wide Web. In *Proceedings of the Fifth International World Wide Web Conference,* Paris, France. Amsterdam, The Netherlands: Elsevier Science Publishers. Available at http://www5conf.inria.fr/ fich_html/papers/P34/Overview.html

Boyarski, Dan, Neuwirth, Christine, Forlizzi, Jodi, Regli, Susan Harkness.(1998) A study of fonts designed for screen display. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 1, 87-94. New York: ACM Press.

Brajnik, Giorgio (2000). Automatic web usability evaluation: Where is the limit? In *Proceedings of the 6th Conference on Human Factors & the Web,* Austin, TX.

Biichner, Alex G, Mulvenna, Maurice D(1998) Discov-

ering inter-net marketing intelligence through online analytical web usage mining. *SIGMOD Record,* 27(4):54-61.

Byrne, Michael D, John, Bonnie E, Wehrle, Neil S, Crow, Daaavid C (1999). The tangled web we wove: A taxonomy of WWW use. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 1,544-551, Pittsburg, PA. New York: ACM Press.

Card, Stuart K., Pirolli, Peter, Van Der Wege, Mija, Morrison, Julie B,. Reeder, Robert W, Schraedley, Paamela, K, Boshart, Jenea(2001). Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability. In *Proceedings Conference on Human Factors in Computing Systems,*498-505.

Chevalier, Aline, Ivory, Melody Y(2003a). Can novice designers apply usability criteria and recommendations to make web sites easier to use? In *Proceedings of the 10th International Conference on Human-Computer Interaction,* Crete, Greece, June 22-27 2003. In press.

Chevalier, Aline, Ivory, Melody Y(2003b). Web site designs: Influences of designer's experience and design constraints. *International Journal of Human-Computer Studies,* 58(l):57-87, 2003.

Chi, Ed H. Pirolli, Peter, Pitkow, James(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, 1*61-168, The Hague, The Netherlands, April 2000. New York: ACM Press.

Chi, Ed H. Pirolli, Peter, Pitkow, Chen, Kim Pitkow,James(2001). Using information scent to model user information needs and actions on the web. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 1, 490-497, Seattle, WA, March 2001. New York: ACM Press.

Chi, Ed H., Adam Rosien, and Jeffrey Heer(2002). Lumberjack: intelligent discovery and analysis of web user traffic composition. In *WEBKDD,* Edmonton, Canada, July 23 2002.

Chi, Ed H, Rosien, Aadam, Supattanasiri, Gesara, Williams, Amanda, Royer, Christi-aan, Chow, Celia, Robles, Erica Brinda Dalai, Chen, Julie Cousins, Steve.(2003). The bloodhound project: Automating discovery of web usability issues using the infoscent™ simulator. In *Proceedings of the Conference on Human Factors in Computing Systems,* 505-512. New York: ACM Press, April 5-10 2003.

Chi, Ed H (2002). Improving web usability through visualization. *IEEE Internet Comput-ing,* 6(2):64-71.

Comber, Tim (1995). Building usable web pages: An HCI perspective. In Roger Debre-ceny and Allan Ellis, editors, *Proceedings of the First Australian World Wide Web Confer-ence,* 119-124, Ballina, Australia, April 1995. Ballina, Australia: Norsearch. Available at http://www.scu.edu.au/sponsored/ausweb/ausweb95/papers/hypertext/comber/

Computer Science and Telecommu-nications Board (1997). *More Than Screen Deep: Toward Every-Citizen Interfaces to the National Information Infrastructure.* National Academy Press, Washington, D.C.

Cooley,R, Srivastava, J, Mobasher, B(1997). Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).*\

Coyne, Kara Pernice, Nielsen, Jakob(2001). Beyond alt text: Making the web easy to use for users with disabilities. Nielsen Norman Group Report.

Creative Good(1997). White paper one: Building a great customer expe-rience to develop brand, increase loyalty and grow revenues. Available at http://www.creativegood.com/creativegood-whitepaper.pdf

Cugini, John, Scholtz, Jean(1999). VISVIP: 3D visualization of paths through web sites. In *Proceedings of the International Workshop on Web-Based Information Visualization,* 259-263, Florence, Italy, September 1999. Institute of Electrical and Electronics Engineers.

Detweiler, Mark C. Omanson, Richard C(1996). Ameritech web page user interface standards and design guidelines. Ameritech Corporation, Chicago, IL, Available at http://www.ameritech.com/corporate/testtown/library/standard/ web_guidelines/index.html.

Drott, M. Carl(1998). Using web server logs to improve site design. In *Proceedings of the 16th International Conference on Systems Documentation,* 43-50, Quebec, Canada, September 1998. New York: ACM Press. Available http://drott. cis.drexel. edu/SIGDOC98/Logpaper2.html

Dyreson,Curtis E. (1997). Using an incomplete data cube as a summary data sieve. *Data Engineering Bulletin,* 20(l):19-Easiwebmaker(2003). Easiwebmaker - automated site creation and management tools. Available at http://www.easiwebmaker.ie

Easton, Valerie J. McColl, John, H(1997) . Statistics glossary vl.l. Available at http://www.stats.gla.ac.uk/steps/

glossary/index.html

Electronic Software Publishing Corporation(2002). LinkScan, 2002. Available at http://www.elsop.com/ linkscan/

Enviz, Inc(2001). Enviz Insight. Available at http://www. enviz. com/solutions/ insight.html, October 2001

Etgen, Michael, Cantor, Judy(2001). What does getting WET (Web Event-logging Tool) mean for web usability. In *Proceedings of the 5th Conference on Human Factors & the Web,* Gaithersburg, Maryland, June 1999. Available at http://www.nist. gov/itl/div894/vvrg/hfweb/ proceedings/etgen-cantor/index.html

Etzion, Oren (1996). The world-wide web: Quagmire or gold mine? *Communications of the ACM,* 39(ll):65-68.

Exodus(2002). Monitoring & management services. Available at http: //www. exodus. net/solutions/management/performance_monitoring.html, 2002

Faraday, Peter(2000). Visually critiquing web pages. In *Proceedings of the 6th Con-ference on Human Factors & the Web,* Austin, TX.

Farkas, David K. Farkas, Jean, B (2000). Guidelines for designing web navigation. *Technical Communication Online,* 47(3), August 2000. Available at http://www. techcomm-online.org/issues/v47n3/pdf/0410.pdf

Flanders,Vincent Willis, Michael(1998). *Web Pages That Suck: Learn Good Design by Looking at Bad Design.* San Francisco, CA: SYBEX.

Fleming, Jennifer(1998). *Web Navigation: Designing the User Experience.* O'Reilly & Associates, Sebastopol, CA.

Forrester Research(1999). Why most web sites fail. Available at http: //www.forrester.com/Research/ ReportExcerpt/O,1082,1285,00.html

Freshwater Software(2002). SiteSeer. Available at http:/ /www. freshwater.com/SiteSeer.htm

Fuller, Rodney, de Graaff, Johannes J(1996). Measuring user moti-vation from server log files. In *Proceedings of the 2nd Conference on Human Factors & the Web,* Redmond, WA, October 1996. Available at http:// www.microsoft.com/usability/ webconf/fuller/fuller.htm

Heer, Jeffrey, Chi, Ed H(2002). Separating the swarm: categorization meth-ods for user sessions on the web. In *Proceedings of the Conference on Human Factors in Computing Systems,* 243-250. New York: ACM Press.

Helfrich, Brian., Landay, James A(1999). QUIP: quantitative user interface profiling. Unpublished manuscript, 1999. Available at http://www.nano-sim.org/ quip

Heller, Hagan., Rivers, David(1996). Design lessons from the best of the world wide web. In *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems,* volume 2 of *Tutorial 12,* 350-351.

Hochheiser, Harry., Shneiderman, Ben(2001).Using interac-tive visualizations of WWW log data to characterize access patterns and inform site design. *Journal of the American Society for Information Science and Technology,* 52(4):331-343.

Hong, Jason., .Heer, Jeffrey., Waterson, Sarah.,. Landay, James A(2001). We-bquilt: A proxy-based approach to remote web usability testing. *ACM Transactions on Information Systems,* 19:263-285.

Ivory, Melody Y., Chevalier, Aline(2002). A study of automated web site evaluation tools. Technical Report 02-10-01, University of Wasington, Department of Computer Science and Engineering, 2002. Available at ftp:/ /ftp.cs.washington.edu/tr/ 2002/10/UW-CSE-02-10-01.pdf

Ivory, Melody Y., and Hearst,Marti A(2001). State of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys,* 33(4) :470-516.

Ivory , Melody Y., Hearst, Marti A (2002a). Improving web site design. *IEEE Internet Computing,* 6(2):56-63.

Ivory , Melody Y., Hearst, Marti A (2002b). Statistical profiles of highly-rated web site interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 4 of *CHI Letters,* pages 367-374, Minneapolis, MN.

Ivory , Melody Y., Sinha, Rasmi, R., Hearst, Marti A (2002b). . Hearst. Preliminary find-ings on quantitative measures for distinguishing highly rated information-centric web pages. In *Proceedings of the 6th Conference on Human Factors & the Web,* Austin, TX, June 2000. Available at http: //www. tri. sbc. com/hf web/ivory/ paper. html.

Ivory,Melody Y., Sinha, Rashmi R., Hearst, Marti A(2001). Empirically val-idated web page design metrics. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 1, pages 53-60, Seattle, WA.

Ivory, Melody Y., Mankoff, Jennifer., Le, Audrey(2003). Using automated tools to improve web site usage by users with diverse abilities. *IT&Society,* 1(3).

Ivory, Melody Y(2001). *An Empirical Foundation for Automated Web Interface Eval-uation.* PhD thesis, University of California, Berkeley, Computer Science Divi-

sion. Available at http://www.ischool.Washington.edu/myivory/thesis/index.html

Ivory, Melody Y(2003a). Automated web site evaluation. In *Human-Computer Interac-tion Series,* volume 3. Dordrecht, The Netherlands: Kluwer Academic Publishers. In press.

Ivory, Melody Y.(2003b). Characteristics of web site designs: Reality vs. recommenda-tions. In *Proceedings of the 10th International Conference on Human-Computer Interaction,* Crete, Greece, June 22-27 2003. In press.

Ivory, Melody Y(2003c). Using design knowledge to teach web designers. Available at http://webtango.ischool.Washington.edu/talks/designknowtalk.pdf, 2003

Jackson-Sanborn,Emily., Odess-Harnish, Kerri., Warren, Nicki(2002). Website accessibility: a study of ADA compliance. Technical Reports TR-2001-05, University of North Carolina - Chapel Hill, School of Information and Library Science, 2002. Available at http://ils.unc.edu/ils/research/reports/accessibility.pdf

Jain, Raj. *The Art of Computer Systems Performance Analysis.* New York: Wiley-Interscience.

Johnson, Jeff(2000). *GUI Bloopers Don'ts and Do's for Software Developers and Web Designers.* San Francisco, CA: Morgan Kaufmann Publishers.

Kitajima, M., Blackmon, M.H., Poison, P.G(2000). A comprehension-based model of web navigation and its application to web usability analysis. In S. McDonald, Y. Waern, and G. Cockton, editors, *People and Computers XIV - Usability or Else! (Pro-ceedings ofHCI2000),*357-373.

Kosala, Raymond, Blockeel, Hendrik(2000). Web mining research: a survey. *A CM SIGKDD Explorations Newsletter,* 2(1):1-15, 2000.

Landauer, Thomas K., Dumais, Susan T(1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,* 104:211-240.

Larson, Kevin., Czerwinski, Mary(1998). Web page design: Impli-cations of memory, structure and scent for information retrieval. In *Proceedings of the Con-ference on Human Factors in Computing Systems,* volume 1, 25-32, Los Angeles, CA, April 1998. New York: ACM Press.

Levine., Rick (1996). Guide to web style. Sun Microsytems, 1996. Available at http: //www.sun.com/styleguide/

Lin, James Landay, James A(2002). Damask: A tool for early-stage design and prototyping of multi-device user interfaces. In *Proceedings of The 8th International Conference on Distributed Multimedia Systems,* pages 573-580, San Francisco, CA.

Lin, James., Newman, Mark., Hong, Jason I., Landay, Jams A(1999). DENIM: Finding a tighter fit between tools and practice for web site design. Submitted for publication.

Lyardet, Fernando., Rossi, Gustavo and Schwabe, Daniel(1999). Discovering and using design patterns in the WWW. *Multimedia Tools and Applications,* 8(3):293-308.

Lycos, Inc(2003). Tripod. Available at http://www.tripod.lycos.com/

Lynch, Patrick J., Horton, Sarah (1999). *Web Style Guide: Basic Design Principles for Creating Web Sites.* Princeton, NJ: Yale University Press, 1999. Available at http://info.med.yale.edu/cairn/manual

Lynch, Gene., Palmiter, Susan., Tilt, Chirs(1999). The Max model: A standard web site user model. In *Proceedings of the 5th Conference on Human Factors & the Web,* Gaithersburg, Maryland.

Miller, Craig S., Remington, Roger W(2000) . A computational model of web navigation: Exploring interactions between hierarchical depth and link ambiguity. In *Proceedings of the 6th Conference on Human Factors & the Web,* Austin, TX, June 2000. Available at http: // www. tri. sbc. com/hf web/miller/article. html

Miller Craig S., Remington, Roger W.(2002). Effects of structure and lable ambiguity on information navigation. In *Proceedings of the Conference on Human Factors in Computing Systems,* Extended Abstracts, pages 630-631, Minneapolis, MN.

National Cancer Institute(2001). Research-based web design & us-ability guidelines. Available at http://usability.gov/guidelines/

NetIQ(2002). Webtrends reporting center. Available at http://www.netiq.com/ products/wrc/default.asp

Newman, Mark W., Landay, James A(2000). . Sitemaps, storyboards, and specifications: A sketch of web site design practice. In *Proceedings of Designing Interactive Systems: DIS 2000,* pages 263-274, New York.

Nielsen, Jakob Web(1998).Web Usability: Why and how. *Users First!,* September 14, 1998. Available at http://www.zdnet.com/devhead/stories/articles/

0,4413,2137433, 00.html

Nielsen, Jakob(1999). User interface directions for the Web. *Communications of the ACM,* 42(I):65-72, January 1999.

Nielsen, Jakob(2000). *Designing Web Usability: The Practice of Simplicity.* Indianapo-lis, IN: New Riders Publishing.

Norman, Donald A(1990). *The Design of Everyday Things.* Doubleday, New York.

Paciello, Michael G., Paciello, Mike(2000). *Web Accessibility for People With Disabilities.* CMP Books, Gilroy, CA.

Paganelli, Laila., Paterno, Fabio(2002). Intelligent analysis of user interactions with web applications. In *Proceedings of the 7th International Conference on Intelligent User Interfaces,* pages 111-118. New York: ACM Press.

Pennsylvania's Initiative on Assistive Technology(2001). Pennsylvania's Initiative on Assistive Technology. Wave 2.0. Available at http://www.wave .webaim. org: 8081/ wave/index. jsp

Ratner, Julie., Grose, Eric M, Forsythe, Cyris(1996). Characterization and assessment of HTML style guides. In *Proceedings of the Conference on Human Factors in Computing Systems,* volume 2, 115-116, Vancouver, Canada, April 1996. New York: ACM Press.

Rosenfeld, Louis Morville, Peter(1998). *Information Architecture for the World Wide Web.* O'Reilly & Associates, Sebastopol, CA.

Sano, Darrell(1996). *Designing Large-Scale Web Sites: A Visual Design Methodology.* Wiley Computer Publishing, John Wiley & Sons, Inc., New York.

Sauer, C. H., K. M. Chandy, K.M(1981). *Computer Systems Performance Modeling.* Englewood Cliffs, NJ: Prentice Hall.

Scanlon, Tara., Schroeder., Will(2000). Report 1: What people do with web sites. In *Designing Information-Rich Web Sites.* Bradford, MA: User Interface Engineering.

Scholtz, Jean., Laskowski, Sharon(1998). Developing usability tools and techniques for designing and testing web sites. In *Proceedings of the 4th Conference on Human Factors & the Web,* Basking Ridge, NJ, June 1998. Available at http://www. research.att.com/conf/ hfweb/proceedings/scholtz/index.html.

Schriver, Karen A(1997). *Dynamics in Document Design.* Wiley Computer Publish-ing, John Wiley & Sons, Inc., New York.

Schwartz, Matthew., (2000). Web site makeover. *Computerworld,* January 31, 2000. Available at http:// www.computerworld.com/home/print.nsf/all/ 000126e3e2

Shedroff, Nathan(1999). Recipe for a successful web site. Available at http://www. nathan.com/thoughts/recipe

Shedroff, Nathan. *Experience Design 1.* Indianapolis, IN: New Riders Publish-ing.

Shneiderman, Ben (1997). Designing information-abundant web sites: Issues and recommendations. *International Journal of Human-Computer Studies,* 47(I):5-29.

Sinha, Rashmi., Hearst, Marti., Ivory.. elody, Y(2001). Content or graphics? an empirical analysis of criteria for award-winning websites. In *Proceedings of the 7th Conference on Human Factors & the Web,* Madison, WI, June 2001. Available at http://www. optavia. com/ hfweb/7thconferenceproceedings.zip/Sinha.pdf

Smith, Sidney L., Mosier, Jane N(1986). Guidelines for designing user interface software. Technical Report ESD-TR-86-278, The MITRE Corporation, Bedford, MA 01730.

Spiliopoulou. Myra., Pohle, Carsten., Faulstich, Lukas(1999). Improv-ing the effectiveness of a web site with web usage mining. In *WEBKDD,* San Diego, CA, 1999. Available at http://www.acm.org/sigs/sigkdd/proceedings/webkdd99/papers/ paperl8-myra.ps

Spiliopoulou, Myra(2000). Web usage mining for web site evaluation. *Communi-cations of the ACM,* 43(8):127-134.

Spool, Jared M., Scanlon, Tara., Schroeder, Will., Snyder, Carolyn., DeAngelo, Terri(1999). *Web Site Usability: A Designer's Guide.* Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Stein, Lincoln D(1997). The rating game. Available at http://stein.cshl.org/ "Istein/rater/

Sullivan, Terry(1997). Reading reader reaction: A proposal for inferential analy-sis of web server log files. In *Proceedings of the 3rd Conference on Human Factors & the Web,* Boulder, CO, June 1997. Available at http:/ /www.research.att.com/conf/hfweb/ conferences/ denverS.zip

Thatcher, Jim., Waddell, Cynthia., Henry, Shawn., Swierenga, Sarah., Urban, Mark., Burks, Michael,Regan, Bob., Bohman, Paul (2002). *Constructing Accessible Web Sites.* Glasshaus, England.

The International Academy of Arts and Sciences ( 2000) The International Academy of Arts and Sciences. The Webby Awards 2000 judging criteria. Available at http:/ /www. webbyawards.com/judging/criteria.html

Theng, Yin Leng., Marsden, Gil (1998). Authoring tools: Towards continuous usability testing of web documents. In *Proceedings of the 1st International Workshop on Hypermedia Development,* Pittsburg, PA, June 1998. Available at http: //www.eng.uts.edu.au/~dbl/HypDev/ht98w/YinLeng/HT98_YinLeng.html

Tiedtke,Thomas., Martin, Christian., Gerth, Norbert(2002). AWUSA - a tool for automated website usability analysis. In *PreProceedings of the 9th International Workshop on the Design, Specification, and Verification of Interactive Systems,* Rostock, Germany, June 12-14 2002. Available at http://www.uni-paderborn.de/cs/ag-szwillus/ lehre/ss02/seminar/semband/MarcoWeissenborn/WebUsabilityTools/AWUS1605.pdf

Turns, Jennifer., Wagner, Tracey(2002). Listening to the learners: A case study in health information website design. In *Proceedings of the Annual Conference of the Society of Technical Communication.*

UsableNet (2000) UsableNet. LIFT online. Available at http://www.usablenet.com

UsableNet (2002) UsableNet. LIFT - Nielsen Norman Group Edition. Available at http: //www.usablenet.com/products_services/lfd_nng/lfd_nng.html, 2002

van Duyne, Douglas K.., Landay, James A., Hong, Jason, I(2002). *The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience.* Boston: Addison-Wesley.

Vividence Corporation (2002) Vividence Corporation. Customer experience management: The Vividence approach and methodology. Available at http: //www.vividence. com/resources/ public/what+we+do/methodology/method.pdf

WatchFire(2002). Bobby worldwide. Available at http://bobby, watchf ire. com/ bobby/html/en/index.jsp

Web Criteria(1999). Web Criteria. Max, and the objective measurement of web sites, 1999. [WebCT, Inc., ] WebCT, Inc. WebCT.com. Available at http://www.webct.com

Williams, Thomas R(2000). Guidelines for designing and evaluating the display of information on the web. *Technical Communication Online,* 47(3), August 2000. Available at http://www.techcomm-online.org/issues/v47n3/pdf/0411.pdf

Wilson, TIm(1999). The cost of downtime. *InternetWeek.com,* July 30, 1999. Available at http://www.internetwk.com/sitereliability/sitelead.htm

World Wide Web Consortium (1999). Web content accessi-bility guidelines 1.0. Geneva, Switzerland, 1999. Available at http://www.w3.org/TR/ WAI-WEBCONTENT

World Wide Web Consortium (2001).W3C HTML validation service. Available at http://validator.w3.org/, 2001

World Wide Web Consortium (2002) World Wide Web Consortium. Evaluation, repair, and transformation tools for web content accessibility. Available at http://www.w3.org/WAI/ ER/existingtools.html.

Yahoo! Inc.,(2003). Yahoo! GeoCities. Available at http://geocities.yahoo, com/home/, 2003

Zaiane, Osmar R., Xin, Man., Han, Jiawei(1998). Discovering web access pat-terns and trends by applying OLAP and data mining technology on web logs. In : *Advances in Digital Libraries,* 19-29.

Zona Research, Inc.(1999). The economic impacts of unacceptable web site download speeds.