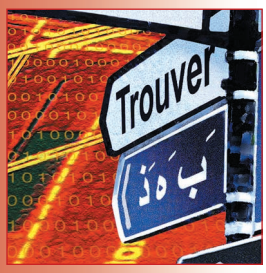

Improving Web Site Design

Using quantitative measures of the informational, navigational, and graphical aspects of a Web site, a quality checker aims to help nonprofessional designers improve their sites.

Melody Y. Ivory and Marti A. Hearst
University of California, Berkeley

Poorly designed Web sites can lead to lost productivity and revenue. The question of how to improve the design of informational Web sites is thus of critical importance. Although most prominent Web sites are created by professional design firms, many smaller sites are built by people with little design experience or training. As a consequence, Web sites with local reach, such as those belonging to nonprofits and small businesses, often have substandard usability.

What makes a high-quality Web site design? Although there are books filled with Web design guidelines, there is a wide gap between a heuristic such as "make the interface consistent" and the implementation of this advice. Furthermore, guidelines tend to conflict, and they offer the same advice for all types of Web sites, regardless of their purpose. Finally, guidelines require careful study and practice and might not be familiar to the occasional Web designer.

As part of the WebTango project, we explore automated approaches for helping designers improve their sites. Our goal is to create an interactive tool that helps steer occasional Web site builders away from bad designs and toward better ones – a "quality checker" tool, analogous to a grammar checker in a word processor. What distinguishes our work from most others is that this tool is based on empirically derived measures computed over thousands of Web pages. We converted these measures, which characterize the informational, navigational, and graphical aspects of a Web site, into profiles for a variety of site types. Our rudimentary design-checking tool uses these profiles to assess Web site designs; future versions will also suggest design improvements.

Many of the software tools described in this article are available online at webtango.berkeley.edu.

Web Page and Site Measures

A Web site interface is a complex mix of text, links, graphic elements, formatting, and other aspects that affect the site's overall quality. Consequently, Web site design entails a broad set of activities for addressing these diverse aspects.[1]

- *Information design* focuses on identi-

 1089-7801/02/$17.00 ©2002 IEEE

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

Category	No. 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	24	How body text is emphasized; whether some underlined text is not in text links; how text areas are highlighted; font styles and sizes; number of text colors; number of times text is repositioned.
Link formatting	3	Colors used for links and whether there are text links that are not underlined or 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 combinations (for text and panels) and predicting the functional type of a page.[1]
Page performance	37	Page download speed; page accessibility for people with disabilities; presence of HTML errors; and "scent" strength.[2]
Site architecture	16	Consistency of page elements, element formatting, page formatting and performance, and site size (number of pages or documents).[3]

1. The decision tree for predicting page type—home, link, content, form, or other—exhibited 84 percent accuracy for 1,770 pages.
2. 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 (www.cast.org/bobby/) runs to report accessibility errors. 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.
3. 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.

fying and grouping content items and developing category labels to reflect the site's information structure.

- *Navigation design* focuses on developing mechanisms (such as navigation bars and links) to facilitate interaction with the information structure.
- *Graphic design* focuses on visual presentation.
- *Experience design* encompasses all three of these categories, as well as properties that affect the overall user experience (download time, ads, popup windows, and so on).[2]

All of these design components entail some inquiry and analysis into the tasks that users are likely to undertake.

Information, navigation, graphic, and experience design can be further refined into the aspects depicted in Figure 1. The bottom levels correspond to information, navigation, and graphic design (for example, text elements and formatting reflect the information design); the top levels correspond to experience design. The figure shows that text, link, and graphic elements are the building blocks of Web interfaces. Aspects on the next level address the formatting of these building blocks, and the subsequent level addresses page formatting. The top two levels address page performance and site architecture (page consistency, breadth, depth, and so on).

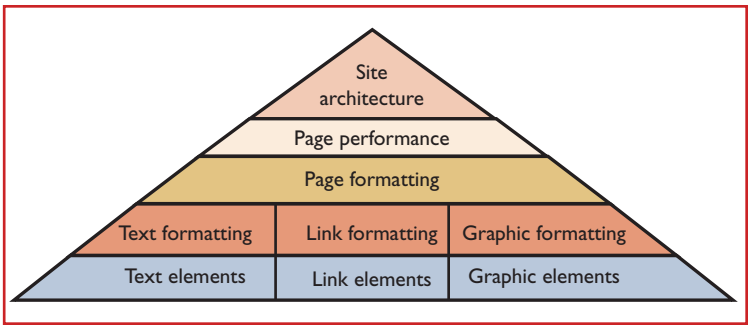

Figure 1. Web site structure. Text, link, and graphic elements are the building blocks of a Web interface. Page- and site-level features use these elements to improve the user's experience.

To build this chart, we surveyed Web design literature[3-5] and published user studies[5] to identify key features that impact Web interface quality and usability. We derived quantitative measures to assess features such as text amount, color, and site consistency, which are discussed in the literature. We then developed a tool that can compute 157 page- and site-level measures. We assessed the tool's accuracy in computing measures for a set of sample Web pages and

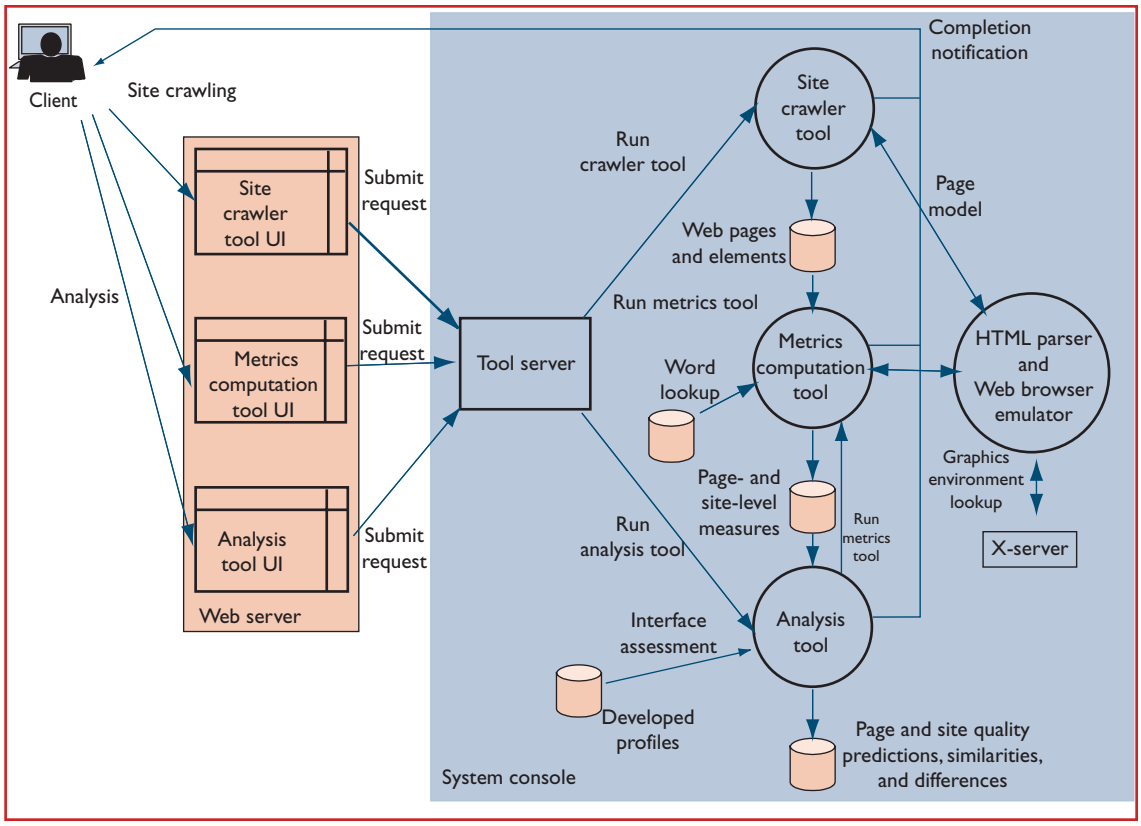

Figure 2. WebTango architecture. The crawler tool selects pages to be measured by the metrics computation tool. The analysis tool uses this information to evaluate a submitted Web site design.

found high accuracy (84 percent on average) on 154 of the measures. Table 1 (previous page) summarizes the entire set of measures.

System Architecture

Figure 2 shows the WebTango architecture.[5] The designer runs the Web site *crawler tool* to download a sample of pages for analysis. The designer specifies a starting page, typically the homepage, and the tool randomly selects pages at successive levels from the starting page. It determines page depth based on whether the page is accessible from the previous level (for example, a page at level two is inaccessible from the starting page but is accessible from a page directly connected to the starting page). The crawler attempts to select only informational pages — that is, not advertisements, Flash pages, login pages, and so on.

The designer then runs the *analysis tool* on the sample to get quality assessments. The analysis tool interacts with the *metrics computation tool*, which calculates the 141 page-level and 16 site-level measures described in Table 1 for those pages. The designer can iteratively run the analysis tool on the sample without rerunning the crawler.

The HTML *parser and browser emulator* generates a detailed page model. The crawler tool uses this model to determine pages to crawl at each level. The model also contains information about each page element, including size, position, and formatting, which the metrics computation tool uses to calculate page-level measures.

The analysis tool uses the metrics computation tool output to show how a given design differs from highly rated designs with a similar purpose. It uses several statistical models derived from an analysis of more than 300 sites that were rated according to their quality and usability. These models encapsulate key relationships and values for the measures described in Table 1. The current tool supports only the analysis of implemented sites; future work will focus on expanding the tool to support interactive evaluation at all design phases.

Predicting Page and Site Ratings

We performed three studies to test the validity of the model-building phase of our methodology. Results showed that profiles developed from empirical data can potentially address limitations in existing assessment approaches, such as inconsistencies in design guidelines and the absence of validation mechanisms.[6-8]

Related Work in Evaluating Web Designs

Automated support for evaluating Web designs is an underexplored research area. Still, several tools have been developed toward this end. We summarize several classes of these tools below.

Quantitative Analysis Tools

Most quantitative methods for evaluating Web sites focus on statistical analysis of usage patterns in server logs.[1,2] Traffic-based analysis (for example, pages-per-visitor or visitors-per-page) and time-based analysis (such as click paths and page-view durations) provide data that the evaluator must interpret to identify usability problems. Because Web server logs give incomplete traces of user behavior, and because network latencies can skew timing estimates, this analysis is largely inconclusive.

Other techniques compare quantitative Web page measures — such as the number of links or graphics — to thresholds.[3] Concrete thresholds for a wider class of quantitative Web page and site measures still remain to be established, however; our methodology works toward this end.

Simulation Tools

Simulation has also been used for Web site evaluation. For example, WebCriteria's Site Profile (www.webcriteria.com) attempts to mimic a user's information-seeking behavior within an implemented site model. This tool uses an idealized user model that follows an explicit, prespecified navigation path through the site and estimates several metrics, such as page load and optimal navigation times.

Chi, Pirolli, and Pitkow have developed a simulation approach for generating a site's navigation paths based on content similarity, server log data, and linking structure.[1] The simulation models hypothetical users traversing the site from specified start pages, using information *scent* (common keywords between the user's goal and content on linked pages) to make navigation decisions. Neither of these approaches accounts for the impact of various Web page attributes, such as text amount or link layout.

Guideline Review Tools

Some approaches, such as Bobby (www.cast.org/bobby/), assess static HTML according to a number of predetermined guidelines (whether all graphics contain ALT attributes, for example). A similar analysis technique, the Design Advisor,[4] uses heuristics about the attentional effects of various elements, such as motion, size, images, and color, to determine and superimpose a scanning path on a Web page. The heuristics are based on empirical results from eye-tracking studies of multimedia presentations. They have not been validated for Web pages, however.

Brajnik surveyed 11 automated Web site analysis methods, including static analysis tools and Site Profile.[5] He found that these tools address only a few usability features, such as download time, presence of alternative text for images, and HTML and link validity. Existing tools do not address other usability aspects, such as consistency and information organization. Ratner, Grose, and Forsythe have also shown that HTML guidelines themselves show little consistency[6]; hence, tools based on these guidelines might be suspect. Another major limitation of existing tools is that they are not based on empirical data.

Similar guideline review approaches evaluate the quality of graphical interfaces. For example, Parush et al. developed and validated a tool for computing the complexity of dialog boxes imple-

continued on p. 60

Developing a Simple Prediction Model

Our first study reported a preliminary analysis of 428 Web pages.[7] Each page corresponded to a site that was either rated highly by experts or was unrated. We derived the expertise ratings from a variety of sources, including *PC Magazine*'s Top 100, WiseCat's Top 100, and the final nominees for the Webby Awards. For each Web page, we computed 12 quantitative measures related to page composition, layout, amount of information, and size (such as the number of words, links, and colors). We wanted to assess whether the measures could predict page standings within the two groups, and to determine characteristics of pages within each group.

We found that six measures — text cluster count, link count, page size, graphics count, color count, and reading complexity — were significantly different for pages in the two groups. For example, text clustering was used to a larger degree in rated pages than in unrated pages. Such clustering facilitates *scanning* — quickly skimming text to find needed information.[3] Additionally, results revealed two strong pairwise correlations for pages from rated sites, and five pairwise correlations for pages from unrated sites. The rated pages had correlations between link and text cluster counts as well as between font and color counts, which suggested that clustering was used to organize links into groups and that color was used mainly for display text. Similar correlations between measures on unrated pages revealed several design patterns, including the use of color to highlight body and display text, use of multiple colors for text links, and use of image links as opposed to text links. An inspection of randomly selected pages supported our predictions about how the layout of the rated and unrated sites' pages manifested the pairwise correlations.

We used a linear discriminant classification method to investigate relationships among measures and to predict whether pages should be classified as rated or unrated. The linear discriminant classifier

Related Work in Evaluating Web Designs (cont.)

continued from p. 59

mented with Microsoft Visual Basic.[7] The tool considers changes in screen element size, element alignment and grouping, and screen space utilization in its calculations. AIDE (semi-Automated Interface Designer and Evaluator), a more advanced tool, helps designers assess and compare different design options using quantitative task-sensitive and task-independent metrics, including efficiency (distance of cursor movement), vertical and horizontal alignment of elements, horizontal and vertical balance, and designer-specified constraints (such as element positioning).[8] An optimization algorithm automatically generates initial UI layouts.

Sherlock focuses on task-independent consistency checking (for example, same widget placement and labels) within the UI or across multiple UIs. It evaluates visual properties of dialog boxes, terminology (for example, it identifies confusing terms and checks spelling), as well as button sizes and labels.[9] Other automated critique tools, such as KRI/AG tool (knowledge-based review of user interface)[10] and IDA (user interface design assistance),[11] perform rule-based interface critiques.

References

1. E.H. Chi, P. Pirolli, and J. Pitkow, "The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, New York, Apr. 2000, pp. 161-168.
2. M.C. Drott, "Using Web Server Logs to Improve Site Design," *Proc. 16th Int'l Conf. Systems Documentation*, ACM Press, New York, Sept. 1998, pp. 43-50.
3. Y.L. Theng and G. Marsden, "Authoring Tools: Towards Continuous Usability Testing of Web Documents," *Proc. 1st Int'l Workshop Hypermedia Development*, June 1998; also available at www.eng.uts.edu.au/~dbl/HypDev/ht98w/YinLeng/HT98_YinLeng.html.
4. P. Faraday, "Visually Critiquing Web Pages," *Proc. 6th Conf. Human Factors and the Web*, June 2000.
5. G. Brajnik, "Automatic Web Usability Evaluation: Where Need to be Done?" *Proc. 6th Conf. Human Factors and the Web*, June 2000.
6. J. Ratner, E.M. Grose, and C. Forsythe, "Characterization and Assessment of HTML Style Guides," *Proc. Conf. Human Factors in Computing Systems*, vol. 2, ACM Press, New York, Apr. 1996, pp. 115-116.
7. A. Parush, R. Nadir, and A. Shtub, "Evaluating the Layout of Graphical User Interface Screens: Validation of a Numerical, Computerized Model," *Int'l J. Human Computer Interaction*, vol. 10, no. 4, 1998, pp. 343-360.
8. A. Sears, "AIDE: A Step Toward Metric-Based Interface Development Tools," *Proc. 8th ACM Symp. User Interface Software and Technology*, ACM Press, New York, 1995, pp. 101-110.
9. R. Mahajan and B. Shneiderman, "Visual and Textual Consistency Checking Tools for Graphical User Interfaces," tech. report CS-TR-3639, Univ. of Maryland, College Park, May 1996; also available at www.isr.umd.edu/TechReports/ISR/1996/TR_96-46/TR_96-46.phtml.
10. J. Lowgren and T. Nordqvist, "Knowledge-Based Evaluation as Design Support for Graphical User Interfaces," *Proc. Conf. Human Factors in Computing Systems*, ACM Press, New York, May 1992, pp. 181-188.
11. H. Reiterer, "A User Interface Design Assistant Approach," *Proc. IFIP 13th World Computer Congress*, vol. 2, K. Brunnstein and E. Raubold, eds., Elsevier Science Publishers, Amsterdam, Aug. 1994, pp. 180-187.

achieved a predictive accuracy of 63 percent, showing that the quantitative measures could characterize some differences between the two groups.

Developing Context-Sensitive Prediction Models

In our second study, we analyzed 1,898 pages from sites evaluated for the Webby Awards 2000.[8] For the Webbys, at least three expert judges evaluated each site on six criteria: content, structure and navigation, visual design, functionality, interactivity, and overall experience. The six criteria were highly correlated, which enabled us to use principal components analysis to summarize the criteria as one number or factor.[9] Another useful aspect of the Webby Awards data is that it classifies Web sites into topical groups.

For this study, we obtained pages from sites in six content categories — community, education, finance, health, living, and services — and computed the same quantitative measures examined in the first study, except for reading complexity. For the analysis, we grouped the sites according to their overall Webby scores as good (top 33 percent of sites), not-good (remaining 67 percent of sites), or poor (lowest 33 percent of sites).

To assess whether the measures could predict page standings within these groups, we developed two statistical models. The first used multiple linear regression to distinguish good from not-good sites. Predictive accuracy proved to be 67 percent when content categories were not taken into account and were even higher on average when categories were assessed separately. The second model used discriminant classification analysis to compute statistics for good versus poor sites. The second model's predictive accuracy ranged from 76 to 83 percent when categories were taken into account.

Creating Profiles

The third study analyzed page- and site-level measures from 5,346 pages and 333 sites from the Webby Awards 2000.[6] The analysis used all 157 measures discussed in Table 1, as well as the Webby content categories and a page type classifier (for distinguishing among homepages, content pages, link pages, forms, and other pages). We developed more sophisticated profiles for dis-

Table 2. Profiles used to assess Web site quality.

Profile	Model	Assessment	Output
Page-level			
Overall page quality	Decision tree model	Classifies pages as good, average, or poor, regardless of page type or content category.	■ decision tree rule that generated the prediction.
Closest good-page cluster	*K*-means clustering model	Maps pages into small-, large-, or formatted-page clusters.	■ distance between a page and the closest cluster's centroid. ■ top 10 measures consistent with this cluster. ■ top 10 measures inconsistent with the cluster and acceptable metric ranges.[1]
Page type quality	Discriminant classification model	Classifies pages as good, average, or poor, according to page type.	■ top 10 measures consistent with the page type. ■ top 10 measures inconsistent with the page type and acceptable metric values[1]
Content category quality	Discriminant classification model	Classifies pages as good, average, or poor, according to content category.	■ top 10 measures consistent with the content category ■ top 10 measures inconsistent with the content category and acceptable metric values[1]
Site-level			
Overall site quality	Decision tree model	Classifies sites as good, average, or poor, regardless of content category.	■ decision tree rule that generated the prediction.
Content category quality	Decision tree model	Classifies sites as good, average, or poor, according to content category.	■ decision tree rule that generated the prediction.

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

tinguishing pages and sites in the good, average (middle 34 percent of sites), and poor groups. The accuracy of page-level models ranged from 93 percent to 96 percent, while the accuracy of site-level models ranged from 68 to 88 percent (possibly due to inadequate data). Although we can find correlations between values for measures and expert ratings, we cannot yet claim that the measures caused the sites to be highly rated. The sites might have received high ratings for reasons other than what the measures assess, such as content quality.

We also used *K*-means clustering to partition Web pages from good sites into small-, large-, and formatted-page subgroups. (Pages in the latter subgroup contain, on average, 120 more words than pages in the small-page subgroup and use more text positioning, columns, and tables, as well as text and panel color combinations.) These clusters have significantly different characteristics and provide more context for assessing Web design quality.

We incorporated these profiles into the analysis tool. To gain more insight into what the profiles represent, we conducted a user study to examine the relationship between Webby judges' scores and the ratings assigned by 30 participants who used the sites to complete tasks. Our analysis of the objective and subjective data suggested some consistency between judges' ratings and usability ratings. We could not draw concrete conclusions about profiles reflecting usability, however, because the study was conducted at least six months after sites were reviewed by Webby judges; hence, sites might have undergone major changes in the interim.

Applying Models to Web Site Design

We used the profiles to assess and refine (by hand) five Web sites. We then conducted a small study to evaluate these sites.[5,6] Only minor and conservative changes were made to the sites. For the study, 13 participants completed 15 page-level comparisons and four site-level ratings of the original and modified versions. Participants repre-

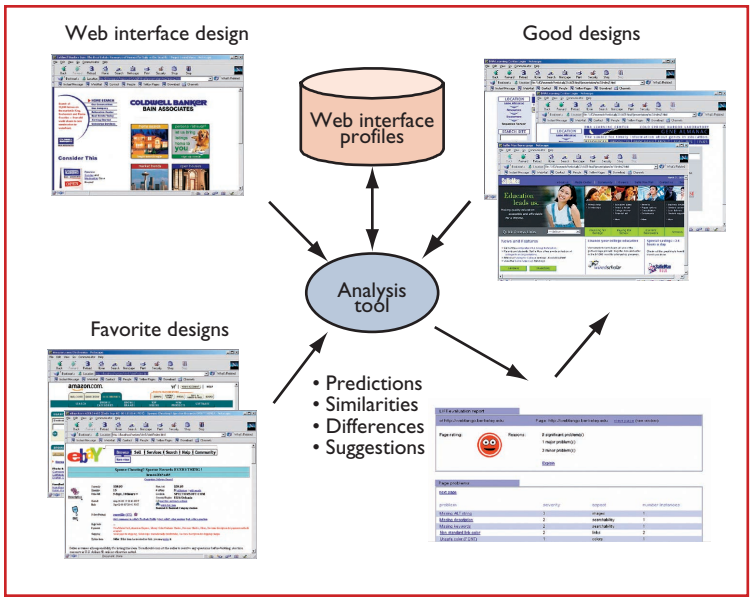

Figure 3. Sample results from a submitted Web site design. The analysis tool compares features of the submitted design to features of highly rated sites. In future it will suggest improvements as well as links to those "good" designs.

sented three groups: professional designers (4), nonprofessional designers who had built Web sites (3), and people who had no experience building Web sites (6). The results showed that participants preferred modified pages (57 percent) to the original versions (43 percent), and participants rated modified sites as 3.5 out of 5.0 and original sites as 3.0, a significant difference.

Assessing Web Design Quality

Figure 3 shows how a Web site designer might use the WebTango system once it is completed. The designer submits a partially designed site to the analysis tool, which generates several quantitative measures. The tool compares these measures to the profiles of highly rated designs in the same general content category, size, and page type. The tool reports differences between the submitted design and similar well-designed sites and offers links to those sites, along with specific suggestions for improvement. The designer uses these results to inform design improvements. Designers can repeat the assessment process as necessary.

The current version of the analysis tool lets the designer iteratively assess an implemented site's quality based on the profiles described in Table 2. These profiles let us consider the context in which pages and sites are designed.

Figure 4 depicts the original and modified versions of an example page from our study. 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.

The good-page cluster model provides insight about design quality, and 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 small 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 small-page cluster model. Ten of the 13 study participants preferred the modified page to the original after these conservative changes were made.

Conclusions and Future Work

The study of modified sites provides preliminary evidence that the profiles can provide insight for improving content presentation, thus improving the user experience in accessing that content. Moreover, three of the five sites in our study of original and modified designs were modified by undergraduate and graduate students with little or no prior Web design experience, which demonstrates that nonprofessionals can interpret model output and modify designs accordingly. Finally, the fact that we empirically located commonalities among presentation elements from highly rated sites underscores the importance of identifying those elements for good design.

This design-checking approach is not intended to replace usability testing, but rather to complement it. Automated tools cannot help designers assess certain usability aspects, such as whether a site meets user needs or company objectives, which can only be assessed via user input. Furthermore, automated tools may not identify true usability issues. Several studies have contrasted expert reviews and usability testing and found little overlap between the two.[10] However, the tool can be used to address potential design issues before conducting usability testing. Furthermore, tool results may be helpful in identifying aspects to focus on during testing, such as text readability or whether page layouts facilitate information search.

Future work will focus on automating and implementing recommendations and identifying good designs for similar types of sites. The current tool only supports refinement of an implemented site; future work will focus on supporting the early stages of Web design.

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 Rashmi Sinha and Deep Debroy for ongoing participation in this work; Maya Draisin and Tiffany Shlain at the International Academy of Digital Arts and Sciences for making the data from the Webby Awards 2000 available; and Tom Phelps for his assistance with the extended metrics computation tool.

References

1. M.W. Newman and J.A. Landay, "Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice," *Proc. Designing Interactive Systems: DIS 2000, Automatic Support in Design and Use*, Aug. 2000, ACM Press, New York, pp. 263-274.
2. N. Shedro, *Experience Design 1*, New Riders Publishing, Indianapolis, Ind., 2001.
3. J. Nielsen, *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing, Indianapolis, Ind., 2000.
4. J.M. Spool et al., *Web Site Usability: A Designer's Guide*, Morgan Kaufmann, San Francisco, 1999.
5. M.Y. Ivory, *An Empirical Foundation for Automated Web Interface Evaluation*, doctoral dissertation, Univ. of California, Berkeley, Computer Science Dept., 2001.
6. M.Y. Ivory and M.A. Hearst, "Statistical Profiles of Highly Rated Web Site Interfaces," *Proc. Conf. Human Factors in Computing Systems*, vol. 1, ACM Press, New York, Apr. 2002, to appear.
7. M.Y. Ivory, R.R. Sinha, and M.A. Hearst, "Preliminary Findings on Quantitative Measures for Distinguishing Highly Rated Information-Centric Web Pages," *Proc. 6th Conf. Human Factors and the Web*, June 2000.
8. M.Y. Ivory, R.R. Sinha, and M.A. Hearst, "Empirically Validated Web Page Design Metrics, *Proc. Conf. Human Factors in Computing Systems*, vol. 1, ACM Press, New York, Mar. 2001, pp. 53-60.
9. R. Sinha, M. Hearst, and M. Ivory, "Content or Graphics? An Empirical Analysis of Criteria for Award-Winning Websites," *Proc. 7th Conf. Human Factors and the Web*, June 2001; also available at www.optavia.com/hfweb/7thconferenceproceedings.zip/Sinha.pdf.
10. R.W. Bailey, R.W. Allan, and P. Raiello, "Usability Testing vs. Heuristic Evaluation: A Head-to-Head Comparison," *Proc. Human Factors Soc. 36th Ann. Meeting*, Human Factors Soc., Santa Monica, Calif., 1992, pp. 409-413.

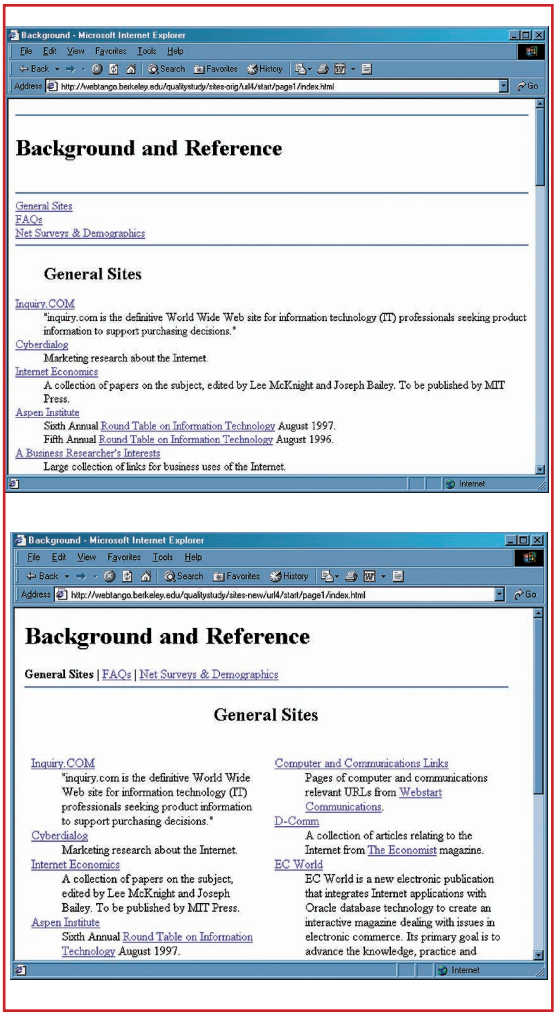

Figure 4. Original and modified versions of a Web page. Students based their improvements on the overall page quality and closest good-page cluster models described in Table 2. (Some of the changes in the modified page are not visible.)

Melody Ivory is a postdoctoral researcher in the School of Information Management and Systems at the University of California, Berkeley. Her research interests include user interfaces, automated Web interface evaluation and text analysis, and data mining. She received an MS and a PhD in computer science from UC Berkeley.

Marti Hearst is an assistant professor in the School of Information Management and Systems at UC Berkeley. Her research interests include user interfaces and visualization for information retrieval, empirical computational linguistics, and text data mining. She received BA, MS, and PhD degrees in computer science from UC Berkeley.

Readers can contact the authors at {ivory, hearst}@sims.berkeley.edu.