Towards Quality Checkers for Web Site Designs

Melody Y. Ivory
EECS Department
UC Berkeley
Berkeley, CA 94720-1776
ivory@cs.berkeley.edu

Marti A. Hearst
SIMS
UC Berkeley
Berkeley, CA 94720-4600
hearst@sims.berkeley.edu

ABSTRACT

Building high-quality web sites is a challenging task, but many sites are necessarily created by non-professional designers. We are working towards creating design tools that incorporate a quality-checker, in analogy to grammar-checkers found in word processors, to help such designers. Towards this end, we have developed a system that computes a set of quantitative measures that can characterize the informational, navigational, and graphical aspects of the web site, and have shown that even a small set of such measures can be used to successfully distinguish highly-rated web sites from poorly-rated ones. We have also developed statistical profiles from these empirically-derived measures and incorporated these profiles into a rudimentary tool that provides some support for improving an implemented site. The next step is to develop an interactive web site evaluation tool that can help non-professional designers understand where and how their current design (an implemented site or early prototype) can be improved, based on its differences from sites designed for similar purposes.

Keywords
World Wide Web, Empirical Studies, Automated Usability Evaluation, Web Site Design

INTRODUCTION

What are the characteristics of high-quality web site designs? 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 operationalization of this advice. Furthermore, guidelines tend to conflict with one another [19], and the same advice is given independent of the type of web site being designed. Finally, guidelines require careful study and practice and may not be familiar to the occasional web designer.

Poor web site design is a serious problem; studies from industry suggest that poorly designed web sites lead to large losses in productivity and revenue [7, 8, 16, 21]. Thus, the question of how to improve the design of informational web sites is of critical importance. Although most prominent web sites are created by professional design firms, an enormous number of smaller sites are built by people, who, despite having little design experience or training, need to make information available online. As a consequence, the usability of web sites with local reach, such as non-profits and small businesses, is often substandard.

Our goal is the creation of an interactive tool to help steer occasional web site builders away from bad designs, and towards better ones; a kind of "quality checker" tool, similar in analogy 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. In a sense, we are mining existing web designs to create profiles of both bad and good design, to be applied to the design of new sites.

In this article, we present a set of quantitative measures that can characterize the informational, navigational, and graphical aspects of a web site. We summarize our earlier results that show that even a small set of such measures can be used to successfully distinguish highly-rated web designs from poorly-rated ones. We describe how we converted these measures into statistical profiles that characterize good web design for various types of sites. We also present a rudimentary design checking tool that provides some support for refining implemented sites. The next step is to leverage these results to develop an interactive web site evaluation tool that can help designers understand where and how their current design (an implemented site or an early prototype) can be improved, based on its differences from sites designed for similar purposes.

WEB PAGE AND SITE MEASURES

A web site interface is a complex mix of many elements (e.g., text, links, and graphics), formatting of these elements, and other aspects that affect the over-

Figure 1: Overview of web site design, derived from [Newman and Landay 2000].

Figure 2: Aspects associated with web site structure.

all interface quality. Consequently, web design entails a complex set of activities for addressing these diverse aspects [15]; Figure 1 depicts these activities. 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. Finally, experience design encompasses all three of the above, as well as properties that affect the overall user experience (e.g., download time, ads, popup windows, etc.) [23]. All of these design components entail some inquiry and analysis of intended users and tasks and affect the overall quality of the web interface.

Information, navigation, graphic, and experience design can be further refined into the aspects depicted in Figure 2. 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 2 addresses formatting of these building blocks, while the subsequent level addresses page-level (page formatting) aspects. The top two levels address the performance of pages and the architecture of sites (e.g., consistency of pages, breadth, and depth).

To build this chart, we conducted an extensive survey of web design literature, including texts written by recognized experts (e.g., [17, 25]) and published user studies in order to identify key features that impact the quality and usability of web interfaces. We then derived quantitative measures to assess many of the features discussed in the literature, including the amount of text, color usage, and site consistency. We have written a tool that can compute 157 page-level and site-level measures. We assessed the accuracy of this tool on a set of sample web pages and found high accuracy (84% on average) on 154 of the measures. The entire set of measures is summarized below.

Text Elements: 31 measures for assessing: the amount of text on a page; and the type, quality, and complexity of text on a page. The measures quantify both visible (e.g., all, link text, and heading words) and invisible text (e.g., meta tag keywords).

Link Elements: 6 measures for assessing the number and type of links (e.g., graphic and text links) on a page.

Graphic Elements: 6 measures for assessing the number and type of images (e.g., animated and link images) on a page.

Text Formatting: 24 measures for assessing: how body text (i.e., text that is not headings or links) is emphasized; whether there is underlined text that is not in text links on the page; font styles and sizes; the number of text colors; the number of times text is re-positioned on the page; and how text areas are highlighted.

Link Formatting: 3 measures for assessing whether there are text links that are not underlined and colors used for links.

Graphic Formatting: 7 measures for assessing the minimum, maximum, and average width and height of images as well as the amount of page area covered by them.

Page Formatting: 27 measures for assessing: color usage, fonts, page size, use of interactive elements, how the page style is controlled, and other page characteristics. Key measures include evaluating the quality of color combinations (for text and panels) and predicting the functional type of a page. For the latter measure, we developed a decision tree that exhibited 84% accuracy for predicting page types – home, link, content, form, or other – for 1,770 pages.

Page Performance: 37 measures for assessing: page download speed; whether the page is accessible to people with disabilities; whether there are HTML errors on the page; and whether there is strong "scent" to the page. We developed a comprehensive model for predicting download speed with 86% accuracy; the model considers the number and size of HTML, graphic, script, and object (e.g., applet) files along with the number of tables on the page. We use output from running Bobby 3.2 [3] for reporting accessibility errors. For assessing 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.

Site Architecture: 16 measures for assessing the consistency of page elements (i.e., text, link, and graphic elements), element formatting, page formatting and performance, as well as the size of the site (i.e., the number of pages or documents). The consistency measures are based on Coefficients of Variation (i.e., 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.

THE METHODOLOGY

Figure 3 shows the architecture of our system [9]. A special-purpose web site Crawler Tool can select pages at specific levels within the site; the depth of a page is determined based on whether the page was accessible from the previous level or not (i.e., a page at level two is not accessible from the home page but is accessible from a page that is connected directly to the home page). The Metrics Computation Tool computes 141 page-level and 16 site-level measures. The HTML Parser and Browser Emulator generates a detailed page model for use by the Site Crawler and Metrics Computation Tools.

The component for assessing the quality of web designs consists of two parts. The first determines key relationships and values for the measures described in the preceding section, analyzing a large number of sites that have been rated according to their quality and usability. The second, the Analysis Tool, uses the output of the metrics tool to show how a given design differs from highly-rated designs that serve a similar purpose. The current tool is in its infancy and only supports analysis of implemented sites; future work will focus on expanding the tool to support interactive web site evaluation at all phases of design.

Three prior studies established the validity of the first phase of this methodology. Based on study results, we believe that profiles developed from empirical data can potentially address limitations of existing assessment approaches, such as the lack of overlap in design guidelines and the lack of empirical validation [10, 11, 12].

The first study reported a preliminary analysis of a collection of 428 web pages [11]. Each page corresponded to a site that had either been highly rated by experts or had no rating. We derived the expertise ratings from a variety of sources, such as *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 having to do with page composition, layout, amount of information, and size (e.g., number of words, links, and colors). We wanted to assess if the measures can predict the pages' standings within the two groups and to determine characteristics of pages within each group.

Results showed that six metrics – text cluster count, link count, page size, graphics count, color count and reading complexity – were significantly associated with rated sites. Additionally, two strong pairwise correlations for rated sites, and five pairwise correlations for unrated sites were revealed. Predictions about how the pairwise correlations were manifested in the layout of the rated and unrated sites' pages were supported by inspection of randomly selected pages. A linear discriminant classifier applied to the page types (rated versus unrated) achieved a predictive accuracy of 63%.

The second study reported an analysis of 1,898 pages from sites evaluated for the Webby Awards 2000 [12, 26]. For the Webbys, at least three expert judges evaluated each sites on six criteria: content, structure and navigation, visual design, functionality, interactivity, and overall experience; the six criteria were highly correlated and were summarized with one factor derived via principal components analysis [24]. Another useful aspect of the Webby Awards data is that web sites were classified into a number of topical groups.

For this study, we obtained pages from sites in six of these 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 score in the Webby standings as follows: "good" (top 33% of sites) versus either "not-good" (remaining 67% of sites) or "poor" (lowest 33% of sites). We wanted to assess if the measures can predict the pages' standings within these groups.

We developed two statistical models. The first used multiple linear regression to distinguish good from not-good sites; the predictive accuracy was 67% when content categories (e.g., community and education) were not taken

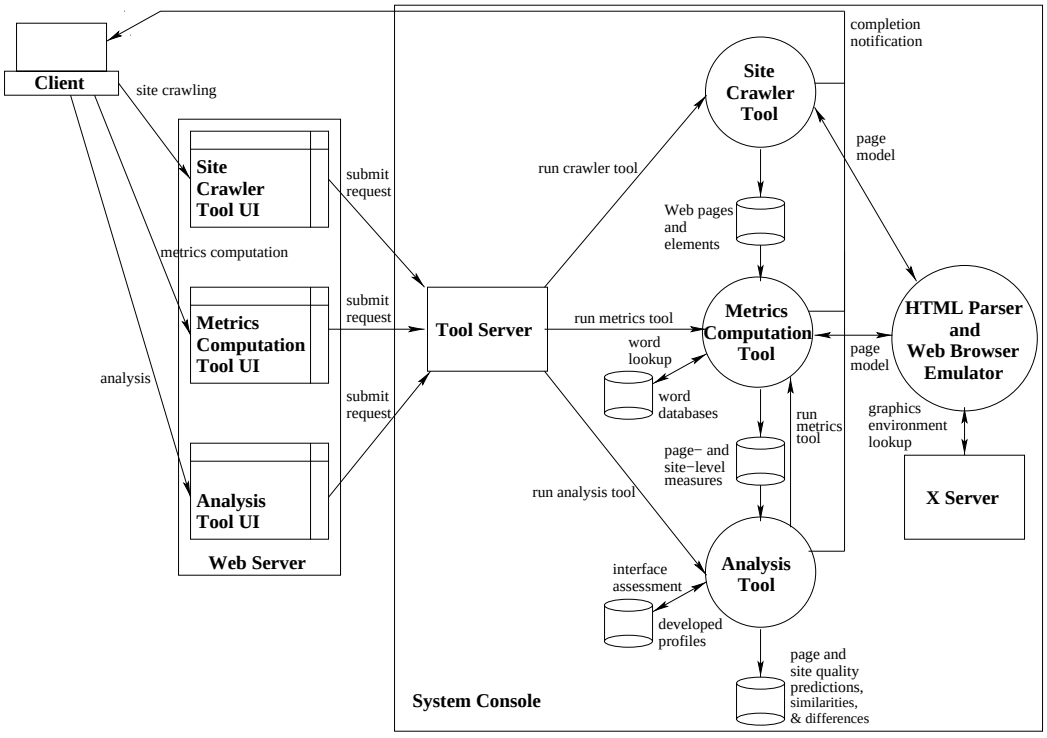


Figure 3: Architecture of the evaluation tools.

into account, and 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 predictive accuracy of the second model ranged from 76% to 83% when categories were taken into account.

The third study reported an analysis of page-level and site-level measures from 5,346 pages and 333 sites from the Webby Awards 2000 [10]. The analysis used all 157 of the measures discussed above, as well as the Webby content categories, and a page type classifier (for distinguishing among home pages, content pages, link pages, forms, and other pages). We developed more sophisticated profiles for distinguishing pages and sites in the good (top 33% of sites), average (middle 34% of sites), and poor (bottom 33% of sites) groups. The accuracy of page-level models ranged from 93%–96%, while the accuracy of site-level models ranged from 68%–88%; the site-level accuracy was considerably less possibly due to inadequate data. We also used K-means clustering to partition web pages from good sites into three subgroups (small-page, large-page, and formatted-page). These clusters have significantly different characteristics and provide more context for assessing web design quality.

We have incorporated these profiles into the Analysis Tool. To gain more insight about what they represent, we conducted a user study to examine the relationship between Webby judges' scores and ratings assigned by thirty participants who used sites to complete tasks. Analysis of objective and subjective data suggested some consistency between judges' ratings and usability ratings. However, concrete conclusions about profiles reflecting usability could not be drawn from the study due to the time difference between judges' scores and user ratings.

We have used the profiles to assess and refine 5 web sites and recently evaluated original and modified versions of these sites via a small study [9, 10]. For the study, thirteen participants completed 15 page-level comparisons (original vs. modified) and 4 site-level ratings (original and modified versions of two sites). Participants represented 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 pages modified based on the profiles over the original versions (58% to 43%), and participants rated modified sites higher than the original sites; differences were significant in both cases.

ASSESSING WEB DESIGN QUALITY

Figure 4 depicts a potential use scenario. A web designer would submit a partially designed site to the tool, which generates a number of quantitative measures. These would be compared to the profiles of highly-rated web

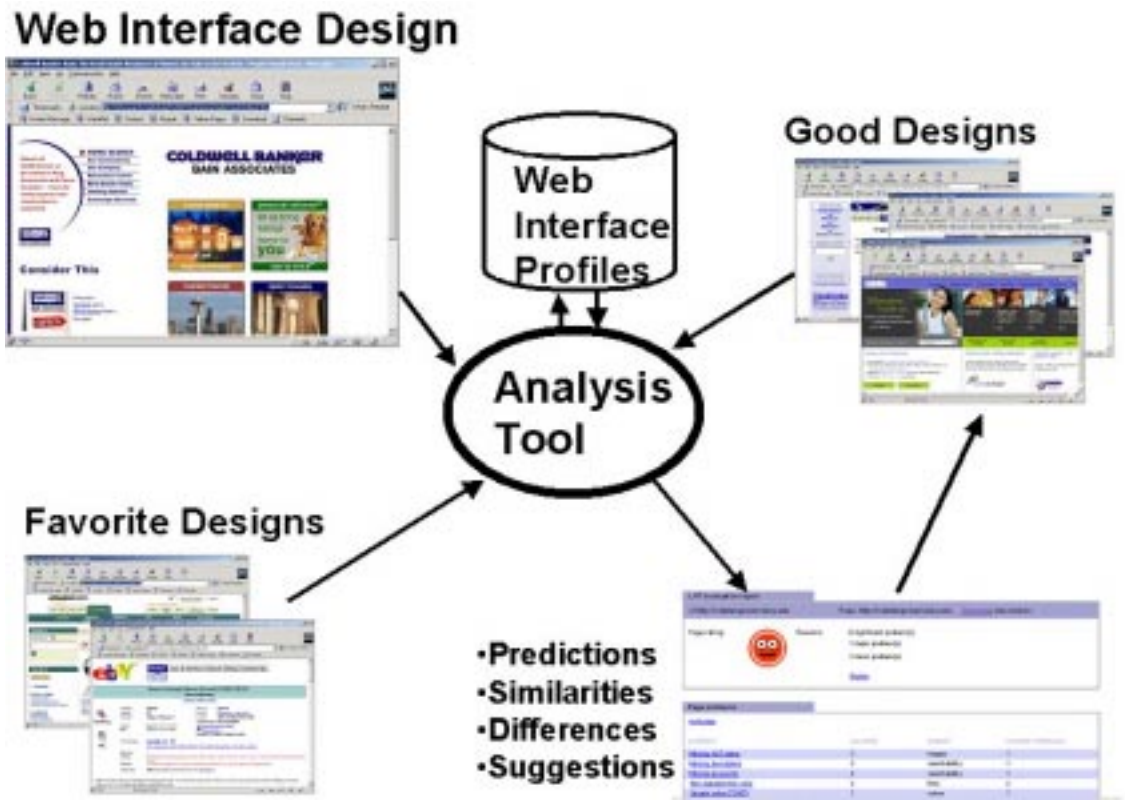


Figure 4: A use scenario for the web site design checking tool.

site designs within the same general category (health, living, finance, etc.), size, and page type (home page, content page, form, etc.). The tool would report on how the design differs from similar well-designed sites and offer links to those designs, along with specific suggestions for improvements. The designer could use these results to choose between alternative designs as well as to inform design improvements. This assessment process could be repeated as necessary.

The current version of the Analysis Tool supports many aspects of this scenario. Specifically, the tool enables the designer to iteratively assess the quality of an implemented site based on the following profiles.

Overall Page Quality: a decision tree model for classifying a page into the good, average, and poor classes without considering the functional type of a page or the content category. The model also reports the decision tree rule that generated the prediction.

Closest Good Page Cluster: a K-means clustering model for mapping a page into one of the three good page clusters – small-page, large-page, and formatted-page. The model reports the distance between a page and the closest cluster's centroid and the top 10 measures that are consistent with this cluster. The model also reports the top 10 measures that are inconsistent with the cluster as well as acceptable metric ranges. In both cases, measures are ordered by their importance in distinguishing pages in the three clusters as determined from ANOVAs.

Page Type Quality: discriminant classification models for classifying a page into the good, average, and poor classes when considering the functional type of a page – home, link, content, form, and other. The model reports the top 10 measures that are consistent with the page type. The model also reports the top 10 measures that are inconsistent with the page type and acceptable metric values. In both cases, measures are ordered by their importance in distinguishing pages in the three classes as determined from ANOVAs.

Content Category Quality (Page-Level): discriminant classification models for classifying a page into the good, average, and poor classes when considering the content category of the site – community, education, finance, health, living, and services. Each model reports the top 10 measures that are consistent with the content category. Each model also reports the top 10 measures that are inconsistent with the content category and acceptable metric values. In both cases, measures are or-

dered by their importance in distinguishing pages in the three classes as determined by ANOVAs.

Overall Site Quality: a decision tree model for classifying a site into the good, average, and poor classes without considering the content category. The model also reports the decision tree rule that generated the prediction.

Content Category Quality (Site-Level): decision tree models for classifying a site into the good, average, and poor classes when considering the content category of the site – community, education, finance, health, living, and services. Each model reports the decision tree rule that generated the prediction.

These profiles make it possible to take into consideration the context in which pages and sites are designed. As an example, Figure 5 depicts the original and modified version of an example page from the previously discussed study. The overall page quality model classifies the original page as poor, mainly because it used font sizes greater than 9pt for all text (good pages were found to use a smaller font size for copyright and footer text) and because the heights of images at the bottom of the page were too large. The good page cluster model provides more insight about design quality and reported that the page was 23.05 standard deviation units from the large-page cluster centroid. The model also reports a number of key deviations from the cluster, such as the amount and positioning of text being inadequate.

Changes were made to the modified version of the page based on the overall page quality and good page cluster models. For example, the amount of text was distributed over multiple pages, which resulted in the page being consistent with the small-page cluster as opposed to the large-page cluster. The layout of text was improved (a second text column was introduced and the top navigation area was reduced to one line) and horizontal rules were removed to reduce vertical scrolling as dictated by the small-page cluster model. Ten of the thirteen participants in the user study preferred the modified page over the original version after these conservative changes were made.

Currently, the designer has to manually interpret model output in order to identify appropriate changes; the tool also does not provide links to good designs in order to help inform design improvements. For the study of original and modified designs (discussed above), 3 of the 5 study sites were modified by undergraduate and graduates students without prior Web design experience; thus, demonstrating that it is possible for others to interpret model output and modify designs accordingly.

Future work will focus on automating recommendations for improvement as well as implementing these recommendations. Future work also entails identifying and presenting good designs for sites designed for similar purposes; this should help to inform design improvements. Another limitation of the current tool is that it only supports refinement of an implemented site; future work will focus on expanding this approach to support the early stages of Web design.

RELATED WORK

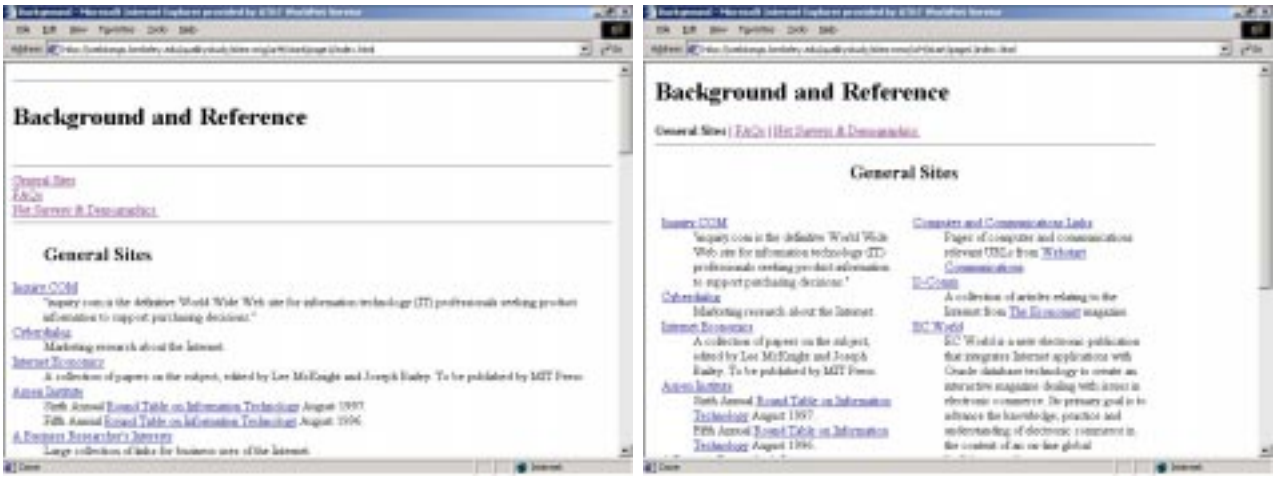
Most quantitative methods for evaluating web sites focus on statistical analysis of usage patterns in server logs (e.g., [4, 5]). Traffic-based analysis (e.g., pages-per-visitor or visitors-per-page) and time-based analysis (e.g., click paths and page-view durations) provide data that the evaluator must interpret in order to identify usability problems. This analysis is largely inconclusive since web server logs provide incomplete traces of user behavior, and because timing estimates may be skewed by network latencies.

Other approaches assess static HTML according to a number of pre-determined guidelines, such as whether all graphics contain ALT attributes (e.g., [3]). Other techniques compare quantitative web page measures – such as the number of links or graphics – to thresholds [27]. However, concrete thresholds for a wider class of quantitative web page and site measures still remain to be established; the methodology presented in this paper is working towards this end.

A similar analysis technique, the Design Advisor [6], 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 author developed the heuristics based on empirical results from eye tracking studies of multimedia presentations. However, the heuristics have not been validated for web pages.

Simulation has also been used for web site evaluation. For example, WebCriteria's Site Profile[1] attempts to mimic a user's information-seeking behavior within a model of an implemented site. This tool uses an idealized user model that follows an explicit, pre-specified navigation path through the site and estimates several metrics, such as page load and optimal navigation times. As another example, Chi, Pirolli, and Pitkow [4] have developed a simulation approach for generating navigation paths for a site based on content similarity among pages, server log data, and linking structure. The simulation models hypothetical users traversing the site from specified start pages, making use of information scent (i.e., common keywords between the user's goal and con-

[1] http://www.webcriteria.com



Original Example Page Modified Example Page

Figure 5: The original and modified version of an example page from a study site. Some of the changes in the modified page are not visible.

tent on linked pages) to make navigation decisions. Neither of these approaches account for the impact of various web page attributes, such as the amount of text or layout of links.

Brajnik [2] surveyed 11 automated web site analysis methods, including the previously mentioned static analysis tools and WebCriteria's Site Profile. The survey revealed that these tools address only a sparse set of usability features, such as download time, presence of alternative text for images, and validation of HTML and links. Other usability aspects, such as consistency and information organization are unaddressed by existing tools. Ratner, Grose, and Forsythe have also shown that HTML guidelines themselves show little consistency [19]; hence, tools developed based on these guidelines may be suspect. Another major limitation of existing tools is that they are not based on empirical data.

Similar automated analysis and critique approaches have been developed for evaluating the quality of graphical interfaces. For example, Parush *et al.* [18] developed and validated a tool for computing the complexity of dialog boxes implemented with Microsoft Visual Basic. The tool considers changes in the size of screen elements, the alignment and grouping of elements, as well as the utilization of screen space in its calculations. AIDE (semi-Automated Interface Designer and Evaluator) [22] is a more advanced tool that helps designers assess and compare different design options using quantitative task-sensitive and task-independent metrics, including efficiency (i.e., distance of cursor movement), vertical and horizontal alignment of elements, horizontal and vertical balance, and designer-specified constraints (e.g., position of elements). AIDE also uses an optimization algorithm to automatically generate initial UI layouts.

Sherlock [14] focuses on task-independent consistency checking (e.g., same widget placement and labels) within the UI or across multiple UIs; it evaluates visual properties of dialog boxes, terminology (e.g., identify confusing terms and check spelling), as well as button sizes and labels. Other automated critique tools, such as KRI/AG tool (Knowledge-based Review of user Interface) [13] and IDA (user Interface Design Assistance) [20], perform rule-based critiques of interfaces.

DISCUSSION

Although it is possible to find correlations between values for measures and expert ratings, we cannot yet claim that the values found for the measures are what cause the sites to be highly rated. It is possible that the highly-rated sites received those ratings for reasons other than what is assessed with the measures, such as the quality of the content of the site.

Experiments with the tool have shown that it helps the designer to refine certain design aspects, such as the amount of text on the page, text formatting, color combinations, font usage, and other page layout considerations. The study of modified sites provides preliminary evidence that the profiles can provide insight on how to take good content that is poorly presented and improve its presentation, thus improving users' experience in accessing that content. And, because it is possible to empirically find commonalities among the presentation elements of the highly-rated sites, this provides strong evidence that the presentational aspects of highly-rated sites that differ from those of poorly-rated sites are in fact important for good design.

This design checking approach is not intended to be used

as a substitute for user responses, but rather as a complement. This approach and other automated tools cannot help the designer to assess whether the site meets users' needs, whether the site meets company objectives, and other design and usability aspects; these aspects can only be assessed via user input. Furthermore, it is not the case that the issues identified by automated tools are true usability issues. Several studies, such as the one conducted by Bailey *et al.* [1], have contrasted expert reviews and usability testing and found little overlap in findings between the two methods.

CONCLUSIONS

We are developing an interactive tool to enable non-professional web site builders to contrast their sites to similar highly-rated ones. Our approach entails computing over 150 quantitative measures to assess page-level and site-level aspects of a site's information, navigation, graphic, and experience design. Three empirical studies have demonstrated our ability to characterize quality sites with high accuracy. The studies suggest that the methodology can be viewed as a reverse engineering of design decisions that went into producing high quality designs, and ideally users informed these decisions. This approach allows us to provide concrete guidance on similarities and differences to highly-rated sites and ultimately will allow us to offer suggestions for improvements. We will also use empirical findings to further clarify prescriptive guidelines that pervade web design literature.

We have made many of the software tools available online at http://webtango.berkeley.edu/

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 Webby Awards 2000 data available; and Tom Phelps for his assistance with the extended Metrics Computation Tool.

REFERENCES

1. Robert W. Bailey, Robert W. Allan, and P. Raiello. Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the Human Factors Society 36th Annual Meeting*, volume 1 of *COMPUTER SYSTEMS: Usability and Rapid Prototyping*, pages 409–413, 1992.

2. Giorgio Brajnik. Automatic web usability evaluation: Where is the limit? In *Proceedings of the 6th Conference on Human Factors & the Web*, Austin, TX, June 2000. Available at http://www.tri.sbc.com/hfweb/brajnik/hfweb-brajnik.html.

3. CAST. Bobby. http://www.cast.org/bobby/, 2000.

4. Ed H. Chi, Peter Pirolli, and James Pitkow. 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*, pages 161–168, The Hague, The Netherlands, April 2000. New York, NY: ACM Press.

5. M. Carl Drott. Using web server logs to improve site design. In *Proceedings of the 16th International Conference on Systems Documentation*, pages 43–50, Quebec, Canada, September 1998. New York, NY: ACM Press.

6. Peter Faraday. Visually critiquing web pages. In *Proceedings of the 6th Conference on Human Factors & the Web*, Austin, TX, June 2000. Available at http://www.tri.sbc.com/hfweb/faraday/faraday.htm.

7. Susan Greenwood. E-commerce report: Good web site design can lead to healthy sales. In *The New York Times*, August 30 1999.

8. Mark Hurst. Holiday '99 e-commerce. http://www.creativegood.com, Sept 1999.

9. Melody Y. Ivory. *An Empirical Foundation for Automated Web Interface Evaluation.* PhD thesis, University of California, Berkeley, Computer Science Division, 2001. In preparation.

10. Melody Y. Ivory and Marti A. Hearst. Statistical profiles of highly-rated web site interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*, volume 1, Minneapolis, MN, April 2002. To appear.

11. Melody Y. Ivory, Rashmi R. Sinha, and Marti A. Hearst. Preliminary findings 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/hfweb/ivory/paper.html.

12. Melody Y. Ivory, Rashmi R. Sinha, and Marti A. Hearst. Empirically validated web page design metrics. In *Proceedings of the Conference on Human Factors in Computing Systems*, volume 1, pages 53–60, Seattle, WA, March 2001. New York, NY: ACM Press.

13. Jonas Lowgren and Tommy Nordqvist. Knowledge-based evaluation as design support for graphical user interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 181–188, Monterey, CA, May 1992. New York, NY: ACM Press.

14. Rohit Mahajan and Ben Shneiderman. Visual & textual
consistency checking tools for graphical user interfaces. Technical Report CS-TR-3639, University of Maryland, College Park, May 1996. Available at http://www.isr.umd.edu/TechReports/ISR/1996/TR_96-46/TR_96-46.phtml.

15. Mark W. Newman and James A. Landay. Sitemaps, storyboards, and specifications: A sketch of web site design practice. In *Proceedings of Designing Interactive Systems: DIS 2000*, Automatic Support in Design and Use, pages 263–274, August 2000.

16. Jakob Nielsen. The alertbox: Current issues in web usability. http://www.useit.com/alertbox.

17. Jakob Nielsen. *Designing Web Usability: The Practice of Simplicity*. Indianapolis, IN: New Riders Publishing, 2000.

18. Avraham Parush, Ronen Nadir, and Avraham Shtub. Evaluating the layout of graphical user interface screens: Validation of a numerical, computerized model. *International Journal of Human Computer Interaction*, 10(4):343–360, 1998.

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

20. H. Reiterer. A user interface design assistant approach. In Klaus Brunnstein and Eckart Raubold, editors, *Proceedings of the IFIP 13th World Computer Congress*, volume 2, pages 180–187, Hamburg, Germany, August 1994. Amsterdam, The Netherlands: Elsevier Science Publishers.

21. Vividence Research. Tangled web 2001. http://www.vividence.com, June 2001.

22. Andrew Sears. AIDE: A step toward metric-based interface development tools. In *Proceedings of the 8th ACM Symposium on User Interface Software and Technology*, pages 101–110, Pittsburg, PA, November 1995. New York, NY: ACM Press.

23. Nathan Shedroff. *Experience Design 1*. Indianapolis, IN: New Riders Publishing, 2001.

24. Rashmi Sinha, Marti Hearst, and Melody Ivory. 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.

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

26. The International Academy of Arts and Sciences. The webby awards 2000 judging criteria. Available at http://www.webbyawards.com/judging/criteria.html, 2000.

27. Yin Leng Theng and Gil Marsden. 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.