Statistical Profiles of Highly-Rated Web Sites

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

We are creating an interactive tool to help non-professional web site builders create high quality designs. We have previously reported that quantitative measures of web page structure can predict whether a site will be highly or poorly rated by experts, with accuracies ranging from 67–80%. In this paper we extend that work in several ways. First, we compute a much larger set of measures (157 versus 11), over a much larger collection of pages (5300 vs. 1900), achieving much higher overall accuracy (94% on average) when contrasting good, average, and poor pages. Second, we introduce new classes of measures that can make assessments at the site level and according to page type (home page, content page, etc.). Finally, we create statistical profiles of good sites, and apply them to an existing design, showing how that design can be changed to better match high-quality designs.

Keywords

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

INTRODUCTION

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.

There are books filled with web design guidelines, but there is a wide gap between a heuristic such as “make the interface consistent” and the operationalization of this advice. Furthermore, guidelines can conflict with one another with little advice about what to do in these cases [12]. And finally, guidelines that require careful study and practice may not be familiar to the occasional web designer.

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 spell 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 pages to create profiles of both bad and good design, to be applied to the design of new sites.

In earlier work we introduced a methodology whereby we compute a number of measures of web page structure and use these measures to predict scores assigned to the sites by expert judges [7, 8]. In particular, the most recent preceding work computed 11 measures and built two models using linear discriminant analysis. One compared the top-rated 33% of the sites against the remaining 67%, and the other compared the top 33% against the bottom 33%. The results of that study, applied to 1,898 pages across 163 sites, ranged in accuracy from 67–80%. That study also found that classification accuracy improved when the pages were subdivided into topical categories and that good pages could be clustered meaningfully according to the number of words on pages.

In this paper we extend that work in a number of ways. We modified the tool to compute an order of magnitude more measures, including some that measure page performance, and some that compute consistency of page measures at the site level. We also applied the analysis to more than twice as many pages and three times as many sites, and used machine learning algorithms to improve the predictions. We find a significant improvement in accuracy, measuring finer distinctions.

We are also concerned that different types of pages have different characteristics; for example, home pages seem to differ in structure from content pages, which in turn differ from pages that consist mainly of web forms or links to other sites. In order to develop different models for each of these types of pages, we created a classifier that can automatically distinguish among them. Finally, we show an example of how the results of such analyses can be used to make suggestions about how to change the site to better conform with highly-rated sites.

The next sections describe related work, give an overview of the 157 page-level and site-level measures, describe the

CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.
Copyright 2001 ACM 1-58113-453-3/02/0004…$5.00.

statistical models and the accuracy of their predictions using the new measures, and show the example site analysis. Many details about the measures, statistical models, and example site analysis have been omitted; a more in-depth discussion can be found in [6] .

RELATED WORK

Most methods for evaluating web site quality assess static HTML according to a number of pre-determined guidelines, such as whether all graphics contain ALT attributes (e.g., [4]). Other techniques compare quantitative web page measures – such as the number of links or graphics – to thresholds [18]. 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.

Simulation has also been used for web site evaluation. For example, WebCriteria's Site Profile [19] 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 [5] 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 content 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 [12]; 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.

WEB PAGE AND SITE MEASURES

Web design can be characterized according to information, navigation, graphic, and experience design [10, 13]. We conducted an extensive survey of web design literature, including texts written by recognized experts (e.g., [11, 15]) in order to identify key features that affect these design aspects, and thus the overall quality of a web site. We organize 157 of these features into the general classes summarized below; the number of features in each class is in parenthesis.

Text Elements: (31) 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) the number and type of links (e.g., graphic and text links) on a page.

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

Text Formatting: (24) 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) whether there are text links that are not underlined and colors used for links.

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

Page Formatting: (27) color usage, fonts, 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).

Page Performance: (37) page size, page download speed; accessibility of the page for people with disabilities; whether there are HTML errors on the page; and whether there is strong "scent" to the page. We developed a model for predicting download speed that has 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 [4] and Weblint 1.02 [1] for reporting accessibility and HTML errors, respectively. For accessing 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) the consistency of page elements (i.e., text, link, and graphic elements), element formatting, page formatting and performance as well as the depth, breadth, and size of the site (i.e., the number of pages or documents). The site architecture measures only reflect the portion traversed by the crawler (i.e., the total number of pages crawled as well as the crawling breadth and depth).

We have written a specialized crawling tool to download pages. The crawler is configured to access pages from different levels of the site, where level zero is the home page, level one refers to pages one link away from the home page, level two refers to pages one link away from the level one pages, and so on. The standard settings are to download the home page, up to 15 level-one pages and 45 level-two pages (3 from each of the level-one pages).

We also built a software tool to compute the 157 measures for downloaded pages. To assess its accuracy, we manually computed the values for a set of example web pages and

Assessent Type	Analysis Method	Classification Accuracy		
		Good	Average	Poor
Page Level (5346 pages)				
Overall	C&RT	96%	94%	93%
Content	LDA	92%	91%	94%
Page Type	LDA	84%	78%	84%
Site Level (333 sites)				
Overall	C&RT	88%	83%	68%
Content	C&RT	71%	79%	64%

Table 1: Page and site level classification accuracies. C&RT refers to the Classification and Regression Tree algorithm. LDA refers to Linear Discriminant Analysis.

compared the results of the tool against these values. The accuracy was high (84% on average) on 154 of the measures. The three measures with lower accuracy are 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.).

COMPUTING STATISTICAL PROFILES

This analysis develops profiles of highly-rated Web pages and sites by contrasting quantitative measures from sites evaluated for the 2000 Webby Awards [17]. A panel of over 100 judges from The International Academy of Digital Arts & Sciences used a rigorous evaluation process to select winning sites. Judges rated sites based on six criteria: content, structure & navigation, visual design, functionality, interactivity, and overall experience. (For more information, see [14].)

We defined three classes of sites for analysis – good (top 33% of sites), average (middle 34% of sites), and poor (bottom 33% of sites) – based on the overall score. It is assumed that ratings not only apply to the site as a whole, but also to individual pages within the site.

We selected sites from six topical categories – community, education, finance, health, living, and services – because each of these categories contained at least 100 information-centric sites (in which the primary goal is to convey information about some topic). The data collection consists of 5,346 pages from 639 of these sites.

We also developed a classifier for labeling a page type as one of: home page, content page, link page, form, or other. We did this by labeling 1,770 pages by hand and training a decision tree classifier on 70% of these pages, using the 141 page-level measures as input to the classifier. Its accuracy on the remaining 30% of pages is 75%, and the overall accuracy is 84%.

PAGE-LEVEL ANALYSIS

Analysis Across Pages

We used the Classification and Regression Tree (C&RT) algorithm [3] to develop a model for classifying the pages into the good, average, and poor classes; this method generates binary trees and uses pruning to minimize overfitting. The data consists of 5,346 pages – 1,906 good pages (36%), 1,835 average pages (34%), and 1,605 poor pages (30%); 70% of the data was used for training and 30% for the test sample. The resulting tree contains 144 rules and has an overall accuracy of 94%. (Table 1 summarizes the accuracies for each of the three classes of pages.) 71 of the 141 page-level measures are significant according to the C&RT algorithm; these measures represent all 8 of the page-level metric categories.

We used one-way analyses of variance (ANOVAs) in order to identify measures where the within-class variance was significantly different from the between-class variance. We also computed correlation coefficients between pairs of predictor measures. The analysis only considered pages accurately classified by the decision tree. Some of the differences among good, average, and poor pages based on the top ten predictors (minimum font size, minimum color use, italicized body word count, Weblint errors, graphic ad count, link text cluster count, interactive object count, Bobby priority 2 errors, text link count, and good link word count) are described below. ANOVAs were also computed between pairs of classes (i.e., good vs. average, good vs. poor, and average vs. poor) to gain more insight about similarities and differences between classes; all of the differences were significant, except as noted below.

- Good pages use minimum font sizes of 9 points or less; however, the standard deviation is smaller than those for the other two classes, indicating less variance. Inspection of a random sample of good pages revealed that this minimum font size is often used for footer text, such as copyright notices. There is no significant difference between the minimum font sizes employed on average and poor pages.
- The minimum color use metric reports the minimum number of times a color is used on a page. Average and poor pages have larger minimum color usages than good pages, which suggests that colors are possibly overused. Good pages tend to have at least one sparsely used accent color.
- Good and average pages rarely contain italicized words within body text; there is no significant difference between the two classes. Poor pages contain one italicized body word on average.
- Good pages contain the most Bobby priority 2 and Weblint errors (average of 35 and 19, respectively), while poor pages contain the fewest errors. There were correlations between these errors and the number of of interactive objects, tables, images, etc. This finding suggests that highly-rated pages tend not to conform to accessibility standards.
- Good pages typically contain one graphical ad; poor pages are slightly more likely to contain graphical ads than average pages. An examination of 10 sites suggests that ads on good sites are for well-known entities (companies with recognizable brands like Saturn and American Express) whereas ads on poor sites are for obscure entities. This result makes more sense in the light of the fact that a controlled study in which 38 users rated Web pages – with

and without graphical ads – on credibility ("high level of perceived trustworthiness and expertise") found that pages with graphical ads were rated as more credible than those without graphical ads [9].

- Good pages contain significantly more links than average pages, which in turn contain more links than poor pages. Poor pages are also less likely to contain link text clusters (areas of text links highlighted with color or lists such as a nagivation bar), while good pages contain slightly more link text clusters than average pages. There is a corresponding higher number of content words on links on good and average pages than on poor pages.
- Good pages appear to be more interactive than pages in the other classes; they contain 3 interactive objects (e.g., search button, text box, or pulldown menu). Average and poor pages contain 2 interactive objects on average.

Exploring large correlations (i.e., $r \geq .5$ in absolute value) between pairs of measures within each sample provided more insight about differences among the classes. For example, on good pages, correlation between the color and display color counts suggests that these pages use a multi-level heading scheme wherein headings at each level are different colors. There is also a correlation between good text and good panel color combinations suggesting these pages use colored areas and colored text simultaneously (e.g., in navigation bars). Good pages also use tables to control the formatting of text links and images. Correlations between redundant link and graphic link counts coupled with a medium-strength correlation between redundant link and text link counts suggest that links are presented multiple times in different forms (e.g., as an image in a navigation bar and as text in a footer).

Characterizing Sub-groups of Good Pages

The model above reflects design features that are common across all good pages, but there are obviously many ways to create good pages. We used K-means clustering [16] to identify 3 sub-groups of good pages. ANOVAs revealed the key differences among the clusters; nine of the top ten measures are associated with the amount of text on a page, including the word count, HTML bytes, and vertical scrolls. The large-page cluster (364 pages) and the small-page cluster (1008 pages) can be characterized as consisting of high and low word count (this is consistent with groups identified in a prior study [8]). The other top ten measure – table count – distinguishes pages in the formatted-page cluster (450 pages). These pages contain on average 120 more words than pages in the small-page cluster and use more text positioning and columns, tables, as well as text and panel color combinations. The three cluster models provide more insight about design practices than the overall model, since it is possible to determine on a measure-by-measure basis how pages are similar to or deviate from cluster centroids.

Analysis Within Content Categories

We used linear discriminant analysis to derive equations for distinguishing good, average, and poor pages within each content category – community, education, finance, health, living, and services, with an overall accuracy of 91%. ANOVAs computed over pages accurately classified by each model revealed that the top 10 predictor variables varied across content categories. For example, the health page model uses four link element measures (internal, redundant, graphic, and total link counts) for classifying pages, while the living page model uses four link and graphic formatting measures (link and standard link color counts, and minimum graphic width and height). Similarly to the cluster models, the content category models enable a measure-by-measure assessment of similarities and differences between a page and the underlying model.

Analysis Within Page Types

We used linear discriminant analysis to derive equations for distinguishing good, average, and poor pages within each page type, yielding an overall accuracy of 82%; the average accuracy for page type models is about 7–15% less than the models developed for content categories possibly due to mispredicted page types. Similarly to the content category analysis, ANOVAs revealed that the top 10 predictor variables varied across page type categories. For example, the content page model uses four page formatting measures (minimum color use, good panel color combinations, and vertical and horizontal scrolls) for classifying pages, while the other page model uses five page performance measures for assessing the similarity of content between source page and link text and destination page text. The models enable a measure-by-measure assessment of similarities and differences between a page and the underlying model.

Discussion

Page-level results in this study are somewhat similar to results in our earlier empirical studies [7, 8]. In particular, good pages were found to contain more words and links, use colored headings, and use more fonts and text clustering. However, none of the previous 11 measures were among the top ten predictors in any of the models, although the word count measure was important for distinguishing clusters of good pages. Nonetheless, the accuracy of predictions afforded by the new measures improved from 70–80% to over 90% in most cases.

SITE-LEVEL ANALYSIS

Site-level analysis explores two types of site architecture measures – the consistency of pages in the site and the site structure. The consistency of pages across the site is computed using Coefficients of Variation (i.e., standard deviation normalized by the mean). We determine the average variation for all measures within each general class of measures, as well as overall element variation (the 3 element classes), overall formatting variation (the 4 formatting classes), and overall variation (all classes except site architecture), as well as word overlap for page titles between pairs of pages. The site variation measures require at least 5 pages on a site for reliable results; hence, we only have such measures for 333 of the

Measure	Good	Average	Poor
Maximum Depth	1.75	1.81	1.94
Median Breadth	7.34	7.21	7.05
Maximum Breadth	9.14	8.95	8.80

Table 2: Site structure: averages across good, average, and poor sites.

639 sites. These are subdivided into 121 good sites (36%), 118 average sites (35%), and 94 poor sites (29%).

The second kind of site architecture measure is that of site structure, based on how deeply and broadly the crawler could traverse the site given the crawler configuration.

Analysis Across Sites

To assign sites into the good, average, and poor classes, we used the C&RT algorithm trained on 70% of the data. The resulting tree contains 50 rules and has an overall accuracy of 81% (see Table 1 for more details). The accuracy of site predictions is lower than that of the page-level models possibly because of a smaller training set; it is also possible that the site-level measures or prediction method need to be improved.

ANOVAs for correctly classified sites revealed that the sites only differed significantly on the maximum depth measure. Table 2 shows that the median and maximum breadths crawled on the good sites are slightly higher than for average and poor sites, although not significantly different. This suggests that the information architectures of good and average sites emphasize breadth over depth.

The lack of significant differences on all but one measure suggests that relationships among measures is very important for classifying sites, more so than with page classification. Examining large, unique correlations between measures on accurately-classified sites revealed interesting differences, such as:

- Correlations between text element and text formatting variation on good sites suggest that text formatting is altered as the amount of text increases on pages. Good sites also have slightly more variation on both of these measures than average and poor sites.
- There were 13 unique correlations between measures on poor sites. Most of the correlations suggest that formatting variation (text, link, graphic, and page) play a major role in the overall and page performance variation measures as opposed to the element variation measures. Poor pages tend to have less formatting variation than average and poor sites, but they have slightly more variation in page performance and element variation.

Analysis Within Content Categories

We also used the C&RT algorithm to develop models for classifying the 333 sites into the good, average, and poor classes within the 6 content categories. Table 1 summarizes the classification accuracy of the models; accuracy for predicting poor pages is lower in most cases possibly due to having fewer sites. ANOVAs for correctly classified sites did not reveal significant differences in measures. Future work will entail developing a larger sample size, especially of poor sites, in order to improve predictions. Our analysis suggests that a minimum of 35 sites per class and content category is needed to improve accuracy.

EXAMPLE WEB SITE ASSESSMENT

This section describes the application of the profiles developed above to the assessment and improvement of an example web site. The intent of this section is: (1) to demonstrate how the models can be systematically applied to this problem, and (2) to highlight current limitations of the models. The section also summarizes results from a small study of the site designs.

Figure 1 shows three pages taken from a small (9 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 is to download a representative set of pages from the site. For this particular site, only 8 level one pages were accessible, and no level two pages were reachable, for a total of 9 downloaded pages. Although there is a page containing links (middle of Figure 1), the links are to pages external to the site.

The next step is to use the analysis tool to compute site-level 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 (i) generate quality predictions and (ii) determine how pages and sites are consistent with or deviate from good ones. Currently, interpreting model predictions to determine appropriate design changes is a manual process.

Site-level Assessment

The example site can be classified in both the health and education content categories, so we initially ran the site-level decision tree model without differentiating by content category. The model predicted that the site was similar to poor sites overall. The corresponding decision tree rule revealed that the site had a 31% variation in the link element measure, although variation for other site-level measures was low. The combination of the link element variation and the lack of a comparable element variation violated patterns discovered on good sites.

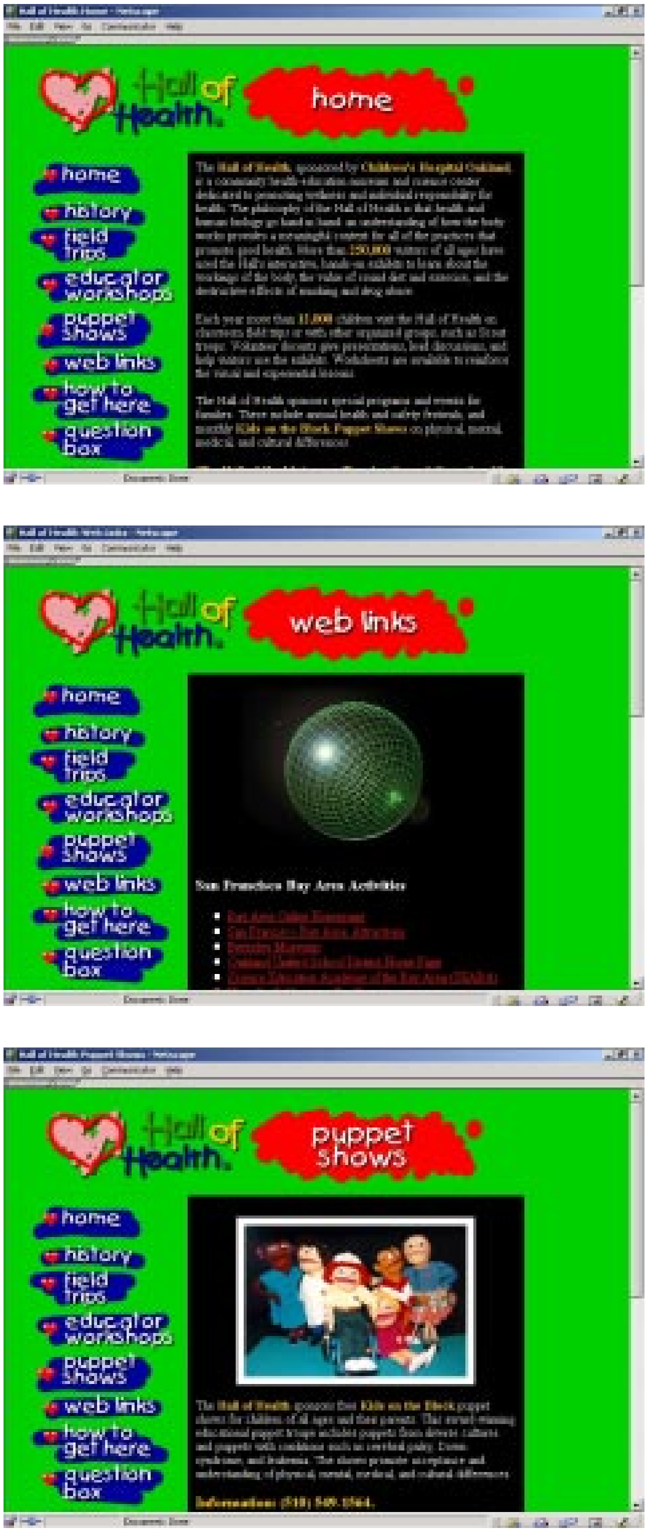


Figure 1: Home (top), link (middle), and content (bottom) pages taken from the example health education site.

The major source of link element variation was the text link count. Eight out of nine pages had from 2 – 4 text links; the remaining page had 27 text links, and acts as a links page (see middle of Figure 1). The decision tree rule suggests that a link element variation level below 29% is typical on good sites.

We also assessed site quality according to the two applicable content categories. The decision tree for health sites predicted that this was a poor health site. In this case the problem was inadequate text element variation. Most of the pages on the site contain paragraphs of text without headings and use only one font face (serif).

The decision tree for site-level quality for education sites makes a prediction contrary to that for sites overall and health sites; it found this site to be consistent with good education sites. 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 seen below, the models predict this to be a poor education site at the page level.

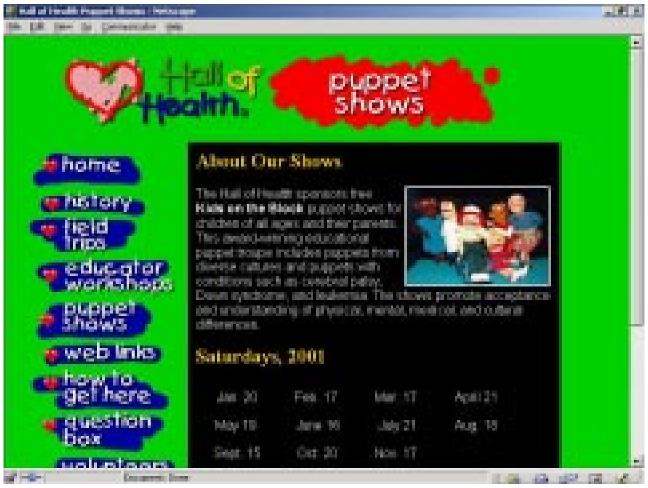
Page-level Assessment

The decision tree model for predicting page quality reports that all 9 of the pages were consistent with poor pages. The home page (top of Figure 1) contains 17 italicized body words; pages with more than 2.5 italicized body words are considered poor pages in the model.

The content page (bottom of Figure 1) is classified as poor mainly because the minimum number of times a color is used is 16. Good pages tend to have an accent color that they use sparingly, whereas poor pages seem to overuse accent colors. Additionally, the example content page contains 34 colored body text words, which is twice the average number found on good pages.

To gain more insight about ways to improve page quality, we mapped each page into one of the 3 clusters of good pages – small-page, large-page, and formatted-page. All of the pages mapped into the small-page cluster and are far from the cluster centroid with a median distance of 10.9 standard deviation units. Pages in the example site deviated on key measures that distinguish pages in this cluster, including the graphic ad, text link, link text cluster, interactive object, and content link word counts. Most of these deviations can be attributed to the fact that the site provides predominately graphical links versus text links for navigation.

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). All but two of the pages were predicted to be poor health pages, which mirrors the results of the site-level model. However, all of the pages were also predicted to be poor education pages, contrasting with the site-level model. In both cases, predictions were based on the features mentioned above.



About Our Shows

The Hall of Health sponsors free **Kids on the Block** puppet shows for children of all ages and their parents. This award-winning educational puppet troupe includes puppets from diverse cultures and puppets with conditions such as cerebral palsy, Down syndrome, and leukemia. The shows promote acceptance and understanding of physical, mental, medical, and cultural differences.

Saturdays, 2001

Jan. 20	Feb. 17	Mar. 17	April 21
May 19	June 16	July 21	Aug. 18
Sept. 15	Oct. 20	Nov. 17	

Figure 2: Revised content page for the example health education site. Many of the changes are not visible, including a set of text links at the bottom of the page that mirrors the graphical links, removal of colored and italicized body text words, and addition of an accent color.

The contrast between site-level and page-level predictions demonstrate the need to incorporate page-level predictions into the site-level prediction. For example, a site can only be considered a good site if the site-level measures are consistent with good sites AND most of the pages are consistent with good pages. At the site level, the example site was highly consistent on page formatting, graphic formatting, and page titles; however, the page quality predictions show that several design aspects, such as text formatting and link elements, need to be improved. If the site-level model for education sites incorporated page-level measures, then this site would be considered a poor education site.

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 6 of the 9 pages, but inaccurately predicted that 3 pages were consistent with link pages; visual inspection suggested that these pages were actually content pages. The mispredictions were mainly due to an improper balance between links and text and the lack of text links. This is a common misclassification problem in the page type model. After correcting the page type predictions, all 9 of the pages were classified as poor pages.

Improving the Site

Although the example site is somewhat aesthetically pleasing and highly consistent, the individual pages and the site as a whole are mostly consistent with poor pages and sites. We used the observations generated by the analysis discussed above to revise the pages manually. A subset of these changes are described below; Figure 2 depicts the modified version of the content page (bottom of Figure 1).

To improve the color count and reduce the link count variation, we added a link text cluster (an area shaded with a different background color to make it stand out) that mirrors the content of the graphical links, as a footer at the bottom of each page. To improve the text element and text formatting variation score, we added headings to break up paragraphs and added font variations – arial font (sans serif) for body text and trebuchet (serif) for headings, and reduced the font size of the copyright text to 9pt. To improve the emphasized (colored, bolded, italicized, etc.) body text score, we converted italics and colors within body text to bold, uncolored body text. To improve the minimum color usage scores, we added a color accent to the vertical bars between the text links in the footer of each page. To reduce vertical scrolling, images were resized and footer elements were moved to reduce page heights.

After making these changes, all of the pages were classified correctly by functional type, and they were rated as good pages overall as well as good health pages. The median distance to the small-page cluster was 4.7 standard deviation units as compared to 10.9 standard deviation units for the original pages. In addition, 8 of the 9 pages were rated as average pages based on their functional types. Lastly, 5 of the 9 pages were rated as average education pages, and the other 4 were rated as poor. 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%) above the acceptable threshold of 51.8%. The site was also classified as a poor health site and a good education site, consistent with classifications before the remodeling.

Evaluation and Discussion

We have recently completed a small study in which 13 participants completed page-level comparisons (original vs. modified) and four 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 process described above was replicated by undergraduate students, a graduate student, and the authors for four additional example sites. (This exercise demonstrates that it is possible for others to interpret model output and modify designs accordingly.) The results showed that participants preferred pages modified based on the Web interface profiles over the original versions (58% to 43%), and participants rated modified sites (including the example site) higher than the original sites; differences were significant in both cases.

CONCLUSIONS

We have computed over 150 quantitative measures to assess page-level and site-level aspects of a site's information, navigation, and graphic design. Three empirical studies have demonstrated our ability to categorize sites according to quality ratings (as evaluated by Webby Awards judges) with high accuracy. From these results we have constructed profiles of web site design that reflect a pages' content type, functional type, and size, as well as overall site structure. These profiles can address limitations of using static design guidelines, by providing suggestions for improvements that reflect the context and particulars of a given site design.

The next step is to develop an interactive tool that helps non-expert designers apply the results of the recommendations. Such a tool may be able to simultaneously educate novice designers about these subtle design aspects and aide them in producing quality designs. We intend to investigate the efficacy of such an analysis tool in a future study.

This approach is not without drawbacks. The analysis tool cannot make recommendations about how to improve the content of the site, nor about the clarity and appropriateness of text. It also cannot make recommendations about subtle aesthetic design decisions. Furthermore, one may naturally question what these profiles represent – highly usable, aesthetically-pleasing, or perhaps merely popular sites. A small study showed that users preferred pages and sites modified based on the profiles over the original versions; however, future studies are needed to better understand the design practices encapsulated in the models. Nonetheless, the methodology can be viewed as a reverse engineering of design decisions that went into producing high quality designs.

More details about the profiles, study, and tools are available at http://webtango.berkeley.edu/.

ACKNOWLEDGEMENTS

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; Tom Phelps for his assistance with the extended Metrics Computation Tool; and the anonymous CHI reviewers for their thoughtful comments.

REFERENCES

1. Neil Bowers. Weblint: quality assurance for the World Wide Web. In *Proceedings of the Fifth International World Wide Web Conference*, Paris, France, May 1996. Amsterdam, The Netherlands: Elsevier Science Publishers.
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.
3. Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.
4. CAST. Bobby. http://www.cast.org/bobby/, 2000.
5. 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.
6. Melody Y. Ivory. *An Empirical Foundation for Automated Web Interface Evaluation*. PhD thesis, University of California, Berkeley, Computer Science Division, 2001. In preparation.
7. 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.
8. 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.
9. N. Kim and B. J. Fogg. World wide web credibility: What effects do advertisements and typos have on the perceived credibility of web page information? Unpublished thesis, Stanford University, 1999.
10. 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.
11. Jakob Nielsen. *Designing Web Usability: The Practice of Simplicity*. Indianapolis, IN: New Riders Publishing, 2000.
12. 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.
13. Nathan Shedroff. *Experience Design 1*. Indianapolis, IN: New Riders Publishing, 2001.
14. 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.
15. 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.
16. SPSS Inc. *SPSS Base 10.0 Applications Guide*. Chicago, IL: SPSS Inc., 1999.
17. The International Academy of Arts and Sciences. The webby awards 2000 judging criteria. http://www.webbyawards.com/judging/criteria.html, 2000.
18. 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.
19. Web Criteria. Max, and the objective measurement of web sites. http://www.webcriteria.com, 1999.