Empirically Validated Web Page Design Metrics

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

Rashmi R. Sinha
SIMS/Psychology Department
UC Berkeley
Berkeley, CA 94720-5050
sinha@sims.berkeley.edu

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

ABSTRACT

A quantitative analysis of a large collection of expert-rated web sites reveals that page-level metrics can accurately predict if a site will be highly rated. The analysis also provides empirical evidence that important metrics, including page composition, page formatting, and overall page characteristics, differ among web site categories such as education, community, living, and finance. These results provide an empirical foundation for web site design guidelines and also suggest which metrics can be most important for evaluation via user studies.

Keywords

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

INTRODUCTION

There is currently much debate about what constitutes good web site design [19, 21]. Many detailed usability guidelines have been developed for both general user interfaces and for web page design [6, 16]. However, designers have historically experienced difficulties following design guidelines [2, 7, 15, 24]. Guidelines are often stated at such a high level that it is unclear how to operationalize them. A typical example can be found in Fleming's book [10] which suggests ten principles of successful navigation design including: be easily learned, remain consistent, provide feedback, provide clear visual messages, and support users' goals and behaviors. Fleming also suggests differentiating design among sites intended for community, learning, information, shopping, identity, and entertainment. Although these goals align well with common sense, they are not justified with empirical evidence and are mute on actual implementation.

Other web-based guidelines are more straightforward to implement. For example, Jakob Nielsen's alertbox column [18] of May 1996 (updated in 1999) claims that the top ten mistakes of web site design include using frames, long pages, non-standard link colors, and overly long download times. These are based on anecdotal observational evidence. Another column (March 15, 1997) provides guidelines on how to write for the web, asserting that since users scan web pages rather than read them, web page design should aid scannability by using headlines, using colored text for emphasis, and using 50% less text (less than what is not stated) since it is more difficult to read on the screen than on paper. Although reasonable, guidelines like these are not usually supported with empirical evidence.

Furthermore, there is no general agreement about which web design guidelines are correct. A recent survey of 21 web guidelines found little consistency among them [21]. We suspect this might result from the fact that there is a lack of empirical validation for such guidelines.

Surprisingly, no studies have derived web design guidelines directly from web sites that have been assessed by human judges. In this paper we report the results of empirical analyses of the page-level elements on a large collection of expert-reviewed web sites. These metrics concern page composition (e.g., word count, link count, graphic count), page formatting (e.g., emphasized text, text positioning, and text clusters), and overall page characteristics (e.g., page size and download speed). The results of this analysis allows us to predict with 65% accuracy if a web page will be assigned a very high or a very low rating by human judges. Even more interestingly, if we constrain predictions to be among pages within categories such as education, community, living, and finance, the prediction accuracy increases to 80% on average.

The remainder of this paper describes related work, our methodology, including the judged web dataset, the metrics, and the data collection process; the results of the study in detail, and finally our conclusions.

RELATED WORK

Most quantitative methods for evaluating web sites focus on statistical analysis of usage patterns in server logs [5, 8, 11, 12, 26, 27]. 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 esti-

SIGCHI'01, March 31-April 4, 2001, Seattle, WA, USA.

mates 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 [4, 22]. Other techniques compare quantitative web page measures – such as the number of links or graphics – to thresholds [25, 27, 28]. However, concrete thresholds for a wider class of quantitative web page measures still remain to be established; our work is a first step towards this end.

The Design Advisor [9] 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 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 [29] 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 [3] 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.

Zhu and Gauch [30] gathered web site quality ratings criteria from a set of expert sites, including Internet Scout, Lycos Top 5%, Argus Clearinghouse, and the Internet Public Library. For each site they computed web page currency, availability, authority, popularity, cohesiveness, and information-to-noise ratio. This last metric is the only one related to the kind of metrics discussed below, and is computed as the number of bytes taken up by words divided by the total number of bytes in the page; in essense a word percentage measure. The authors assessed these metrics in terms of how well they aided in various information retrieval tasks, finding that weighted combinations of metrics improved search over text content alone. They did not relate these metrics to web site usability or attempt to predict the judges ratings outside the context of search.

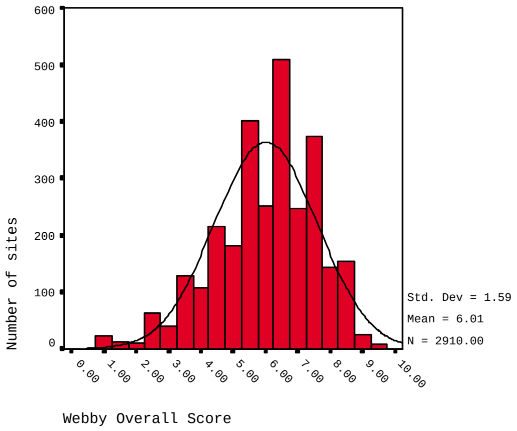

Figure 1: Histogram of the the overall scores assigned to the sites considered for the 2000 Webby Awards. The x axis is the overall score and the y axis is the number of sites assigned this score.

The most closely related work is our earlier study [13] in which we reported a preliminary analysis of a collection of 428 web pages. Each page corresponded to a site that had either been highly rated by experts, or had no rating. The expertise ratings were derived from a variety of sources, such as *PC Magazine* Top 100, WiseCat's Top 100, and the final nominees for the Webby Awards. For each web page, we captured 12 quantitative measures having to do with page composition, layout, amount of information, and size (e.g., number of words, links, and colors). We found that 6 metrics – text cluster count, link count, page size, graphics count, color count and reading complexity – were significantly associated with rated sites. Additionally, we found 2 strong pairwise correlations for rated sites, and 5 pairwise correlations for unrated sites. Our 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 work reported in this paper expands on that preliminary analysis in several ways. First, rather than comparing highly rated sites to unrated sites, we are comparing sites that have been rated on a single scale, and according to several measures, by one set of judges. Second, the sites within this dataset have been classified into topics (such as financial, educational, community), thus allowing us to see if preferred values for metrics vary according to type of category. Finally, informed by the results of our preliminary study, we have improved our metrics and analyze a larger number of web pages. This work further validates our preliminary analysis.

METHODOLOGY

This study computes quantitative web page attributes (e.g., number of fonts, images, and words) from web pages evaluated for the 2000 Webby Awards [20]. The Webby organizers place web sites into 27 categories, including news, personal, finance, services, sports, fashion, and travel. A panel of over 100 judges from The International Academy of Digital Arts & Sciences use a rigorous evaluation process to select winning sites.[1] Webby organizers describe the judge selection criteria as follows: “Site Reviewers are Internet professionals who work with and on the Internet. They have clearly demonstrable familiarity with the category in which they review and have been individually required to produce evidence of such expertise. The site reviewers are given different sites in their category for review and they are all prohibited from reviewing any site with which they have any personal or professional affiliation. The Academy regularly inspects the work of each reviewer for fairness and accuracy.”

Judges rate web sites based on six criteria: content, structure & navigation, visual design, functionality, interactivity, and overall experience. Figure 1 shows the distribution of the overall criterion across all of the judged sites. We suspected that the six criteria were highly correlated, suggesting that there was one factor underlying them all. To test this hypothesis, we used a principles component analysis to examine the underlying factor structure. The first factor accounted for 91% of the variance in the six criteria. In the experiments reported below, we used both the overall Webby score and the extracted factor for doing discriminant classification.

For our study, we selected sites from six topical categories – financial, educational, community, health, service, and living – because these categories contained at least 100 information-centric sites (in which the primary goal is to convey information about some topic). We used the overall score to define two groups of sites for analysis: good (top 33% of sites), and not-good (remaining 67% of sites). Specifically, we wanted to determine if there are significant differences between the groups – both overall and within each category. Furthermore, we wanted to construct models for predicting group membership. These models would enable us to establish concrete thresholds for each metric, evaluate them with user studies, and eventually provide guidance for design improvement. We also used the composite rating to group sites into two categories: top 33% of sites, and bottom 33% of sites. The cutoffs for both sets, based on the overall criterion (ranging from 1 to 10) are:

	Overall	Community	Education	Finance
Top	6.97	6.58	6.47	6.6
Bottom	5.47	5.66	5.38	5.8
	Health	Living	Services	
Top	7.9	6.66	7.54	
Bottom	6.4	5.66	5.9	

The following section introduces the metrics and describes the data collected for this analysis.

Web Page Metrics

From a list of 42 web page attributes associated with effective design and usability [13], we developed an automated tool to compute the 11 metrics that we focus on in this study (see Table 1). (This subset was chosen primarily because it was the easiest to compute; we are in the process of extending the tool to compute a wider range of metrics.) The tool functions similarly to the Netscape Navigator browser in processing web pages and cascading stylesheets; it has limited support for inline frames, but does not support framesets, applets, scripts or other embedded objects. We analyzed the accuracy of the computed metrics using a set of 5 pages with widely different features, such as use of stylesheets, style tags, and forms. Overall, the metrics are about 85% accurate with text cluster and text positioning counts range from 38% to 74% accuracy.

Data Collection

We used the metrics tool to collect data for 1,898 pages from the six Webby Awards categories. These pages are from 163 sites and from 3 different levels in the site – the home page, pages directly accessible from the home page (level 1), and pages accessible from level 1 but not directly accessible from the home page (level 2). We attempted to capture 15 level 1 pages and 45 level 2 pages from each site. Because not every website has many pages at each level, our collection consists of an average of 11 pages per site.

RESULTS

We employed several statistical techniques, including linear regression, and linear discriminant analysis, and t-test for equality of means, to examine differences between the good and not-good groups. The following sections discuss the findings in detail.

Distinguishing Good Pages

We used Linear Discriminant analysis to discriminate good from not-good pages, and top from bottom pages. This technique is suitable for cases where the predicted variable is dichotomous in nature. We built two predictive models for identifying good webpages using linear discriminant analysis:

- Model 1: A simple, conservative model that distinguishes “good” (top 33%) from “not good” (bottom 67%) websites, using the overall Webby criterion as the predictor.

[1] Webby Awards judging has three rounds. The data used in this study are derived from the first round of judging; only the list of nominees for the last round is available to the public. Throughout this paper, we assume a score assigned to a site applies uniformly to all the pages within that site.

Metric	Description
Word Count	Total words on a page
Body Text %	Percentage of words that are body vs. display text (i.e., headers)
Emphasized Body Text %	Portion of body text that is emphasized (e.g., bold, capitalized or near !'s)
Text Positioning Count	Changes in text position from flush left
Text Cluster Count	Text areas highlighted with color, bordered regions, rules or lists
Link Count	Total links on a page
Page Size	Total bytes for the page as well as elements graphics and stylesheets
Graphic %	Percentage of page bytes that are for graphics
Graphics Count	Total graphics on a page (not including graphics specified in scripts, applets and objects)
Color Count	Total colors employed
Font Count	Total fonts employed (i.e., face + size + bold + italic)

Table 1: Web page metrics computed for this study.

- Model 2: A more complex model that uses the Webby factor and distinguishes "top" (top 33%) from "bottom" (bottom 33%) pages.

Tables 2 and 3 summarize the accuracy of the predictions for both models for the entire sample as well as within each category. We report the Wilks Lambda along with the associated Chi-square for each of the models; all of the discriminant functions have significant Wilks Lambda. The squared canonical correlation indicates the percentage of variance in the metrics accounted for by the discriminant function. The final and most important test for the model is the classification accuracy.

For Model 1, the overall accuracy is 67% (50.4% and 78.4% for good and not-good pages, respectively) if categories are not taken into account (see Table 2). Classification accuracy is higher on average when categories are assessed separately (70.7% for good pages and 77% for not-good pages). Our earlier results [13] achieved 63% overall accuracy but had a smaller sample size, did not have separation into category types, and had to distinguish between rated sites versus non-rated sites, meaning that good sites may have been included among the non-rated sites.

Interestingly, the average percentage of variance explained (33%) within categories is more than double the variance explained across the dataset. The health category has the highest percentage of variance explained and also has the highest classification accuracy of 89% (80.9% and 94.6% for good and not-good pages, respectively). The accuracy for this model is indicative of the predictive power of this approach. In the future we plan to use more metrics and a larger dataset in our analysis.

The model with the smallest percentage of variance explained (20% for the living category) is also the model with the lowest classification accuracy of 55% (47.4% and 62.3% for good and not-good pages, respectively). We partially attribute this lower accuracy to a smaller sample size; there are only 118 pages in this category.

The results for Model 2 are shown in Table 3. The average category accuracy increases to 73.8% for predicting the top pages and 86.6% for predicting the bottom pages. (This prediction does not comment on intermediate pages, however.) The higher accuracy is caused both by the relatively larger differences between top and bottom pages (as opposed to top versus the rest) and by the use of the Webby factor.

In related work [23] analyzing the Webby Award criteria in detail, we found that the content criterion was the best predictor of the overall score, while visual design was a weak predictor at best. Here we see that the metrics are able to better predict the Webby factor than the overall score. We think this happens because the overall criterion is an abstract judgement of site quality, while the Webby factor (consisting of contributions from content, structure & navigation, visual design, functionality, interactivity, and as well as overall ratings) reflects aspects of the specific criteria which are more easily captured by the metrics.

The Role of Individual Metrics

To gain insight about predictor metrics in these categories, we also employed multiple linear regression analysis to predict the overall Webby scores. We used a backward elimination method wherein all of the metrics are entered into the equation initially, and then one by one, the least predictive metric is eliminated. This process is repeated until the Adjusted R Square shows a significant reduction with the elimination of a predictor. Table 4 shows the details of the analysis. The adjusted R^2 for all of the regression analyses was significant at the .01 level, meaning that the metrics explained about 10% of the variance in the overall score for the whole dataset. This indicates that a linear combination of our metrics could significantly predict the overall score.

We used standardized Beta coefficients from regression equations to determine the significance of the metrics in predicting good vs. not-good pages. Table 5 illustrates which of the metrics make significant contributions to predictions as well as the nature of their contributions (positive or negative). Significant metrics across the dataset are fairly consistent with

Category	Squared Canonical Correlation	Wilks Lambda	Chi-square	Sig.	Sample Size	Classification Accuracy	
						Good	Not-Good
Overall	0.13	0.87	268.15	0.000	1898	50.40%	78.40%
Community	0.23	0.77	111.87	0.000	430	67.70%	64.80%
Education	0.26	0.74	111.52	0.000	373	69.00%	90.20%
Finance	0.37	0.63	85.61	0.000	190	63.20%	88.00%
Health	0.60	0.4	104.8	0.000	121	94.60%	80.90%
Living	0.20	0.8	24.07	0.012	118	47.40%	62.30%
Services	0.34	0.66	100.6	0.000	311	82.50%	75.80%
Cat. Avg.						70.70%	77.00%

Table 2: Classification accuracy for predicting good and not-good pages. The overall accuracy ignores category labels. Discriminant analysis rejects some data items.

Category	Squared Canonical Correlation	Wilks Lambda	Chi-square	Sig.	Sample Size	Classification Accuracy	
						Top	Bottom
Overall	0.14	0.86	192.85	0.000	1286	66.8%	63.4%
Community	0.60	0.40	275.78	0.000	305	83.2%	91.9%
Education	0.28	0.72	118.93	0.000	368	75.7%	73.2%
Finance	0.47	0.53	85.74	0.000	142	76.5%	93.4%
Health	0.65	0.35	165.19	0.000	165	93.0%	87.3%
Living	0.22	0.79	24.46	0.010	106	42.3%	75.9 %
Services	0.36	0.64	90.51	0.000	208	85.7%	74.8%
Cat. Avg.						76.07%	82.75%

Table 3: Classification accuracy for predicting the top 33% versus the bottom 33% according to the Webby factor. The overall accuracy ignores category labels.

profiles discussed in the next section; most of the metrics found to be individually significant play a major role in the overall quality of pages.

Profiles of Good Pages

Word count was significantly correlated with 9 other metrics (all but emphasized body text percentage), so we used it to subdivide the pages into three groups, depending on their size: low (avg. word count = 66.38), medium (avg. word count = 229.87) and high (avg. word count = 827.15). Partitioning pages based on the word count metric created interesting profiles of good versus not-good pages. In addition, the regression score and discriminant analysis classification accuracy increases somewhat when the dataset is divided in this manner; Model 1 is most accurate for pages that fall into the medium-size group.

To develop profiles of pages based on overall ratings, we compared the means and standard deviations of all metrics for good and not-good pages with low, medium, and high word counts (see Table 6). We employed t-tests for equality of means to determine their significance and also report 2-tailed significance values. Different metrics were significant among the different size groups, with the exception of graphic percentage, which is significant across all groups. The data suggests that good pages have relatively fewer graphics; this is consistent with our previously discussed finding that visual design was a weak predictor of overall rating [23].

Category	Adj. R Square	Std. Err.	F value	Sig.	Sample Size
Overall	0.10	1.92	20.40	.000	1898
Community	0.36	1.76	22.52	.000	430
Education	0.16	1.53	10.34	.000	536
Finance	0.24	1.90	7.78	.000	234
Health	0.56	0.79	27.98	.000	234
Living	0.11	1.62	2.70	.000	153
Services	0.27	1.82	11.15	.000	311

Table 4: Linear regression results for predicting overall rating for good and not-good pages. The F value and corresponding significance level shows the linear combination of the metrics to be related to the overall rating.

Returning to Table 5, we see that in most cases, the positive or negative contribution of a metric aligns with differences in the means of good vs. bad pages depicted in Table 6, with the exception that page size and link count in the medium word count category appear to have opposite contribution than expected, since in general good pages are smaller and have more links on average than not-good pages. Looking in detail at Tables 5 and 6, we can create profiles of the good pages that fall within low, medium, and high word counts:

Low Word Count. Good pages have slightly more content, smaller page sizes, less graphics and employ more font variations than not-good pages. The smaller page sizes and graphics count suggests faster download times for these

	Word Count			**Category**						
Metric	Low	Med.	High	**Com.**	**Edu.**	**Fin.**	**Hlth.**	**Lvng**	**Serv.**	**Freq.**
Word Count			⇓	⇓	⇑		⇓		⇓	4
Body Text %	⇑			⇑	⇓	⇑				3
Emp. Body Text %		⇓	⇓			⇓	⇑		⇓	3
Text Pos. Count	⇓		⇓	⇓			⇓			2
Text Clus. Count	⇑						⇓			1
Link Count		⇓					⇑		⇑	2
Page Size	⇑	⇑	⇑	⇑			⇑	⇑	⇓	4
Graphic %	⇓	⇓	⇓	⇓	⇓	⇑	⇓			4
Graphics Count	⇓	⇓	⇓	⇓			⇓			2
Color Count		⇑	⇑	⇑		⇑	⇑			3
Font Count	⇑			⇑		⇓	⇓		⇑	4

Table 5: Significant beta coefficients for all metrics in terms of whether they make a positive (⇑), negative (⇓), or no contribution in predicting good pages. The frequency column summarizes the number of times a metric made significant contributions within the categories.

pages (this was corroborated by a download time metric, not discussed in detail here). Correlations between font count and body text suggest that good pages vary fonts used between header and body text.

Medium Word Count. Good pages emphasize less of the body text; if too much text is emphasized, the unintended effect occurs of making the unemphasized text stands out more than emphasized text. Based on text positioning and text cluster count, medium-sized good pages appear to organize text into clusters (e.g., lists and shaded table areas). The negative correlations between body text and color count suggests that good medium-sized pages use colors to distinguish headers.

High Word Count. Large good pages exhibit a number of differences from not-good pages. Although both groups have comparable word counts, good pages have less body text, suggesting pages have more headers and text links than not-good pages (we verified this with hand-inspection of some pages). As mentioned above, headers are thought to improve scannability, while generous numbers of links can facilitate information seeking provided they are meaningful and clearly marked.

DISCUSSION

It is quite remarkable that the simple, superficial metrics used in this study are capable of predicting expert's judgements with some degree of accuracy. It may be the case that these computer-accessible metrics reflect at some level the more complex psychological principles by which the Webby judges rate the sites. It turns out that similar results have been found in related problems. For example, one informal study of computer science grant proposals found that superficial features such as font size, inclusion of a summary section and section numbering distinguishes between proposals that are funded and those that are not [1]. As another example, programs can assign grades to student essays using only superficial metrics (such as average word length, essay length, number of commas, number of prepositions) and achieve correlations with teachers' scores that are close to those of between-teacher correlations [14].

There are two logical explanations for this effect; the first is that there is a causal relationship between these metrics and deeper aspects of information architecture. The second possibility is that high quality in superficial attributes is generally accompanied by high quality in all aspects of a work. In other words, those who do a good job do a good job overall. It may be the case that those site developers who have high-quality content are willing to pay for professional designers to develop the other aspects of their sites. Nevertheless, the fact that these metrics can predict a difference between good and not-good sites indicates that there are better and worse ways to arrange the superficial aspects of web pages. By reporting these, we hope to help web site developers who cannot afford to hire professional design firms.

There is some question as to whether or not the Webby Awards judgements are good indicators of web site *usability*, or whether they assess other measures of quality. We have conducted a task-based user study on a small subset of the web sites within our sample, using the WAMMI Usability Questionnaire [17]. We plan to report the results of this study in future work.

CONCLUSIONS AND FUTURE WORK

The Webby Awards dataset is possibly the largest human-rated corpus of web sites available. Any site that is submitted is initially examined by three judges on six criteria. As such it is a statistically rigorous collection. However, since the criteria for judging are so broad, it is unclear or unknown what the specific web page components are that judges actually use for their assessments. As such it is not possible for those who would like to look to these expert-rated sites to learn

	Low Word Count			Medium Word Count			High Word Count		
	Mean & (Std. Dev.)			Mean & (Std. Dev.)			Mean & (Std. Dev.)		
Metric	G	NG	Sig.	G	NG	Sig.	G	NG	Sig.
Word Count	74.6 (45.7)	62.7 (38.0)	**0.002**	231.77 (59.49)	228.1 (59.5)	0.430	803.0 (629.8)	844.7 (675.2)	0.426
Body Text %	62.4 (45.7)	60.0 (29.7)	0.337	68.18 (21.91)	68.5 (23.6)	0.793	73.5 (25.3)	80.3 (20.8)	**0.000**
Emp. Body Text %	12.1 (20.0)	14.5 (23.3)	0.180	9.28 (13.86)	18.1 (22.5)	**0.000**	11.2 (19.8)	17.1 (22.3)	**0.001**
Text Pos. Count	1.4 (4.7)	1.6 (2.6)	0.605	4.59 (9.18)	3.5 (6.9)	0.096	6.1 (16.7)	7.3 (19.1)	0.403
Text Clus. Count	1.3 (4.6)	1.1 (2.3)	0.477	5.17 (10.07)	3.4 (6.0)	**0.007**	7.8 (14.9)	7.8 (17.6)	0.973
Link Count	74.6 (14.4)	16.0 (14.1)	0.202	35.68 (22.54)	36.2 (24.0)	0.764	61.1 (50.3)	51.7 (48.4)	**0.019**
Page Size	23041.2 (35362.2)	32617.6 (43350.7)	**0.004**	53429.98 (76395.69)	46753.0 (36435.8)	0.163	77877.7 (104845.7)	50905.0 (39019.0)	**0.000**
Graphic %	28.8 (36.5)	48.9 (36.7)	**0.000**	40.64 (35.18)	56.0 (27.6)	**0.000**	37.8 (36.0)	45.4 (26.5)	**0.004**
Graphics Count	11.4 (14.5)	15.0 (15.8)	**0.005**	24.88 (23.07)	26.2 (20.7)	0.451	25.3 (26.2)	25.8 (27.0)	0.835
Color Count	6.1 (1.9)	5.9 (1.8)	0.224	7.47 (2.42)	7.1 (2.4)	**0.045**	8.1 (3.0)	7.2 (2.4)	**0.000**
Font Count	3.7 (1.9)	3.2 (2.1)	**0.001**	5.42 (1.95)	5.3 (2.4)	0.320	6.7 (3.0)	6.7 (3.1)	0.999

Table 6: Means and standard deviations (in parenthesis) for the good (G) and not-good (NG) groups based on the low, medium, and high word count categories. The table also contains t-test results (2-tailed significance) for each profile; bold text denotes significant differences (i.e., $p < 0.05$).

how to improve their own designs to derive value from these results. We hope that the type of analysis that we present here opens the way towards a new, bottom-up methodology for creating empirically justified, reproducible interface design recommendations, heuristics, and guidelines.

We are developing a prototype analysis tool that will enable designers to compare their pages to profiles of good pages in each subject category. However, the lack of agreement over guidelines suggests there is no one path to good design; good web page design might be due to a combination of a number of metrics. For example, it is possible that some good pages use many text clusters, many links, and many colors. Another good design profile might make use of less text, proportionally fewer colors, and more graphics. Both might be equally valid paths to the same end: good web page design. Thus we do not plan to simply present a rating, nor do we plan on stating that a given metric exceeds a cutoff point. Rather, we plan to develop a set of profiles of good designs for each category, and show how the designer's pages differ from the various profiles.

It is important to keep in mind that metrics of the type explored here are only one piece of the web site design puzzle; this work is part of a larger project whose goals are to develop techniques to empirically investigate all aspects of web site design, and to develop tools to help designers assess and improve the quality of their web sites.

ACKNOWLEDGMENTS

This research was supported by a Hellman Faculty Fund Award, a Gates Millennium Fellowship, a GAANN fellowship, and a Lucent Cooperative Research Fellowship Program grant. We thank Maya Draisin and Tiffany Shlain at the International Academy of Digital Arts & Sciences for making the WebbyAwards 2000 data available; Nigel Claridge and Jurek Kirakowski of WAMMI for agreeing to analyze our user study data; and Tom Phelps for his assistance with the extended metrics tool.

REFERENCES

1. Daniel Berleant. Does typography affect proposal assessment? *Communications of the ACM*, 43(8):24–25, 2000.

2. Jose A. Borges, Israel Morales, and Nestor J. Rodriguez. Guidelines for designing usable world wide web pages. In *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, volume 2, pages 277–278, 1996.

3. Giorgio Brajnik. Automatic web usability evaluation: Where is the limit? In *Proceedings of the 6th Conference on Human Factors and the Web*, 2000.

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 infor-

mation scent, usage, and usability of a web site. In *Proceedings of ACM CHI 00 Conference on Conference on Human Factors in Computing Systems*, 2000.

6. Tim Comber. Building usable web pages: An HCI perspective. In Roger Debreceny and Allan Ellis, editors, *Proceedings of the First Australian World Wide Web Conference AusWeb'95*, pages 119–124. Norsearch, Ballina, 1995.

7. Flavio de Souza and Nigel Bevan. The use of guidelines in menu interface design: Evaluation of a draft standard. In *Proceedings of IFIP INTERACT'90: Human-Computer Interaction*, Detailed Design: Menus, pages 435–440, 1990.

8. M. Carl Drott. Using web server logs to improve site design. In *ACM 16th International Conference on Systems Documentation*, Getting Feedback on your Web Site, pages 43–50, 1998.

9. Peter Faraday. Visually critiquing web pages. In *Proceedings of the 6th Conference on Human Factors and the Web*, 2000.

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

11. Rodney Fuller and Johannes J. de Graaff. Measuring user motivation from server log files. In *Proceedings of the Human Factors and the Web 2 Conference*, October 1996.

12. Harry Hochheiser and Ben Shneiderman. Understanding patterns of user visits to web sites: Interactive starfield visualizations of WWW log data. In *Proceedings of ASIS '99*, 1999.

13. Melody Y. Ivory, Rashmi 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 and the Web*, 2000.

14. Karen Kukich. Beyond automated essay scoring. *IEEE Intelligent Systems*, 15(5):22–27, September/October 2000.

15. Jonas Lowgren and Tommy Nordqvist. Knowledge-based evaluation as design support for graphical user interfaces. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*, pages 181–188, 1992.

16. Patrick J. Lynch and Sarah Horton. *Web Style Guide: Basic Design Principles for Creating Web Sites*. Yale University Press, 1999.

17. Nomos Management. WAMMI web usability qusetionnaire. http://www.nomos.se/wammi, 1999.

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

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

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

21. Julie Ratner, Eric M. Grose, and Chris Forsythe. Characterization and assessment of HTML style guides. In *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, volume 2, pages 115–116, 1996.

22. Jean Scholtz and Sharon Laskowski. Developing usability tools and techniques for designing and testing web sites. In *Proceedings of the 4th Conference on Human Factors & the Web*, 1998.

23. Rashmi Sinha, Melody Y. Ivory, and Marti A. Hearst. Content or graphics? an empirical analysis of criteria for award-winning websites. Submitted for publication, 2001.

24. Sidney L. Smith. Standards versus guidelines for designing user interface software. *Behaviour and Information Technology*, 5(1):47–61, 1986.

25. Lincoln D. Stein. The rating game. http://stein.cshl.org/ lstein/ rater/, 1997.

26. Terry Sullivan. Reading reader reaction: A proposal for inferential analysis of web server log files. In *Proceedings of the Human Factors and the Web 3 Conference*, Practices & Reflections, June 1997.

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*, 1998.

28. Harold Thimbleby. Gentler: A tool for systematic web authoring. *International Journal of Human-Computer Studies*, 47(1):139–168, 1997.

29. Web Criteria. Max, and the objective measurement of web sites. http://www.webcriteria.com, 1999.

30. Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd Annual International ACM/SIGIR Conference*, pages 288–295, Athens, Greece, 2000.