Chapter 6

Profiles of Highly-Rated Web Interfaces

6.1 Introduction

The focus of this chapter is the development of statistical models or profiles to support assessing Web site quality. Although statistical model development is a common methodology used for solving many problems from evaluating credit card applicants to personalizing information displayed to Web site visitors, this approach has not been previously used for evaluating Web sites. The use of extensive quantitative measures (described in Chapter 5) combined with Internet professionals' ratings for a large collection of Web sites makes it possible to apply model development techniques towards the problem of automated analysis of Web interfaces.

This chapter begins with a brief discussion of two prior studies that demonstrated the feasibility of developing statistical models to predict interface quality. Then, it describes the most recent study that shows that sophisticated statistical models can be developed to predict interface quality at both the page and site levels while taking into consideration the type of content on a site and the functional style of a page, for example. A shorter version of the study is scheduled for publication [Ivory and Hearst 2002].

6.2 Background: Prior Profile Development Work

Two prior studies by this author established that statistical models could be developed to predict interface quality from quantitative Web interface measures and corresponding expert ratings [Ivory *et al.* 2000; Ivory *et al.* 2001]. The first study reported a preliminary analysis of a collection of 428 Web pages [Ivory *et al.* 2000]. Each page corresponded to a site that had either been rated by Internet experts or had no rating. The expertise ratings were derived from a variety of sources, such as *PC Magazine's* Top 100, WiseCat's Top 100, and the final nominees for the 1999 Webby Awards; if a site was acknowledged by one of these sources, then it was considered to be rated. For each Web page, twelve quantitative measures having to do with page composition, layout, amount of information, and size (e.g., number of words, links, and colors) were computed.

Results showed that six measures – 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 groups (rated versus unrated) was able to classify pages into the two groups with 63% accuracy. The study also showed that the predictive accuracy could be improved by considering the functional type – home or other – in models.

The second study reported an analysis of 1,898 pages from 163 sites evaluated for the Webby Awards 2000 [The International Academy of Arts and Sciences 2000; Ivory *et al.* 2001]. At least three Internet professionals (referred to as expert judges) evaluated 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 [Sinha *et al.* 2001]. Pages were from sites in six topical categories – community, education, finance, health, living, and services – and represented several groups of sites, as rated by judges: good (top 33% of sites); "not good" (remaining 67% of sites), and poor (bottom 33% of sites). All of the quantitative measures examined in the first study were used, except for reading complexity. The reading complexity measure – the Gunning Fog Index [Gunning 1973] – was not used in this study because it was not computed for many small pages; the index requires at least a hundred words on a page for computation.

The analysis methodology of the first study was replicated to develop two linear discriminant classifier models: 1. distinguishing pages from good and not good sites; and 2. distinguishing pages from good and poor sites. For the first model, the predictive accuracy was 67% when content categories (e.g., community and education) were not taken into account and ranged between 70.7% and 77% when categories were assessed separately. The predictive accuracy of the second model ranged between 76% and 83%. Analysis of individual measures revealed that the word count could be used to characterize sub-groups of good pages. For example, good pages with a low word count (66 words on average as compared to 230 and 827 words for medium and large pages, respectively) had slightly more content, smaller page sizes, less graphics, and used more font variations than corresponding not-good pages.

6.3 Data Collection

The analysis in this chapter uses a large collection of pages and sites from the Webby Awards 2000 dataset [The International Academy of Arts and Sciences 2000] and is similar to the second study [Ivory *et al.* 2001]. This dataset as well as the analysis data are described below.

6.3.1 The Webby Awards 2000

The Webby Awards dataset is a unique untapped resource, as it appears to be the largest collection of Web sites rated along one set of criteria. For the 2000 awards, an initial pool of 2,909 sites were rated on overall quality as well as five specific criteria: content, structure & navigation, visual design, functionality, and interactivity. Additionally, the Web sites were assigned into 27 content categories, including news, personal, finance, services, sports, fashion, technology, arts, and weird. A panel of over 100 judges from The International Academy of Digital Arts & Sciences used a rigorous evaluation process to select winning sites. Webby Awards organizers state 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."

The judging takes place in three stages: review, nominating, and final; only the list of nominees for the final round are available to the public. Anyone can nominate any site to the review stage; nearly 3,000 Web sites were nominated in 2000. The analysis in this chapter focuses solely on sites evaluated during the review stage. Sinha *et al.* [2001] conducted an in depth analysis of judges' ratings for this stage and found that the content criterion was the best predictor of the overall score, while visual design was a weak predictor at best. However, all of the criteria are highly correlated and can be summarized with a single factor derived via principles component analysis [SPSS Inc. 1999]; this factor (referred to as the Webby factor) explained 91% of the variance in the criteria.

For the current study, sites were selected from six content categories – community, education, finance, health, living, and services – as described below.

Community. "Sites developed to facilitate and create community, connectedness and/or communication. These sites can target either a broad-based or niche audience."

Education. "Sites that are educational, promote education, or provide online curriculum. This could include educational content for children or adults, resources for educators, and 'distance learning' courses."

Finance. "Sites relating to financial services and/or information. These include online stock trading, financial news, or investor services."

Health. "Sites designed to provide information and resources to improve personal health. These may include medical news sites, health information, and online diagnosis. Health includes not only medicine but also includes alternative and mental health or fitness Web sites."

Living. "Sites which provide content about how to go about daily life or about elements that touch the personal side of life. Living includes gardening, home improvement, interior design, architecture, food, parenting, and similar subjects."

Services. "Sites that allow real world activities to be done online. These include sites that help people find jobs, houses, dates, or which otherwise facilitate offline activities from the keyboard."

These categories were selected because they were both information-centric (a primary goal is to convey information about some topic) and contained at least 100 sites. Although some sites in the financial and services categories had some functional aspects, such as looking up a stock chart or submitting a resume, most of the pages on these sites provided information. Three groups – good (top 33% of sites), average (middle 34% of sites), and poor (bottom 33% of sites) – were defined for analysis based on the overall score. Table 6.1 depicts the overall score used for defining the three classes of pages and sites. It is assumed that ratings not only apply to the site as a whole, but also to individual pages within the site.

6.3.2 Analysis Data

An early version of the Site Crawler Tool was used to download pages from 1,002 sites in the finance, education, community, health, services, and living categories. The crawler was

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

Table 6.1: Overall scores used to classify pages and sites as good (top 33%), average (middle 34%), or poor (bottom 33%). The rating scale is from one to ten. Average pages and sites fall in the range between the Top and Bottom cutoffs.

configured to crawl two levels from the home page on each site and to download fifteen level-one pages and 45 level-two pages (three pages linked from each level-one page). This number of pages was not retrievable on all of the sites. The early crawler version used for this data set did not follow redirects or download scripts, object files, or images used as form buttons. Hence, measures associated with these elements (e.g., script, object, and graphic bytes; Section 5.12) may be underestimated somewhat.

The Metrics Computation Tool was then used to compute page-level and site-level measures for downloaded pages that contained at least 30 words and were in English. The 30-word limit was used to eliminate blank or error message pages but still retain smaller content or form pages. The final data collection consists of 5,346 pages from 639 sites. The collection includes data for good, average, and poor pages and sites within each of the six content categories. Four of the good pages have missing Bobby measures; hence, some techniques, such as discriminant classification, exclude these pages, while other techniques, such as multiple linear regression and decision tree modeling, do not. Only 333 of the 639 sites have at least five downloaded pages as required for computing the variation measures (see Section 5.13.1); the site-level analysis will consider only this subset. The small number of sites with at least five pages can be attributed to several restrictions on the Site Crawler Tool, including a 12-minute crawling time limit on sites and the requirement that pages at subsequent crawling levels were not previously accessible.

All of the measures, except the two reading complexity measures, will be used for analyses throughout this chapter; reading complexity is excluded due to a high number of cases wherein reading complexity could not be computed[1]. However, individual measures used to derive reading complexity (e.g., fog big word count and fog sentence count; Section 5.4.11) are included in the analyses. All of the measures were carefully screened to remove outliers within the three classes of pages; the resulting data was normally distributed with equal variances.

6.4 Profile Development Methodology

The goal of profile development is to derive statistical models for classifying Web pages and sites into the good, average, and poor classes based on their quantitative measures. As discussed in Chapter 4, profile development encompasses statistical analysis of quantitative measures (page-level and site-level) and corresponding expert ratings. Prior studies included univariate and multivariate analyses [Ivory *et al.* 2000; Ivory *et al.* 2001]. Statistical techniques, such as correlation coefficients and t-tests for equality of means [Keppel and Zedeck 1989], revealed significant relationships among individual measures and expert ratings within each class of pages (i.e., rated vs. unrated and

[1]The Gunning Fog Index requires 100 words for computation; hence, it may not be possible to compute this index for pages with small amounts of text.

highly-rated vs. poorly-rated). Two multivariate techniques – multiple linear regression and linear discriminate analysis – illustrated significant relationships between measures as well as key measures for predicting the class of each page.

The prior analyses focused on describing key differences between the classes of pages; the analysis in this chapter expands on this work and also explores properties of highly-rated pages. The three-step analysis approach is described below.

1. Develop a model to classify pages into the three classes – good, average, and poor. Use linear discriminant classification and decision tree modeling as necessary.

2. For pages accurately classified by the model as good, identify groups of pages with common properties. Use K-means clustering [SPSS Inc. 1999] in this step.

3. Examine key relationships among measures in each cluster of good pages. Use descriptive statistics, ANOVAs, and correlation coefficients as necessary.

Comparisons are also made among good, average, and poor pages using techniques similar to step 3 above. The first and third steps are also followed for site-level analysis. This chapter replicates analyses performed in prior studies with several key differences: a larger sample of pages is used; site-level analysis is included; three classes of pages and sites are contrasted (good, average, and poor versus top and bottom); a larger number of quantitative measures are used; a page type is incorporated into the analysis; and machine learning techniques are used to develop models in some cases.

Profile development is approached in several phases in this chapter. First, profiles are developed across all of the pages irrespective of page types and content categories. Then, page types (home, link, content, form, and other; see Section 5.11) and content categories (community, education, finance, health, living, and services) are considered separately. Finally, profiles are developed across sites and within content categories across sites.

The profile development work revealed interesting correlations for measures within the good, average, and poor classes; however, it is not suggested that these correlations caused ratings. Causality can only be established with controlled usability studies; this will be the focus of future work. The study of pages and sites modified based on the developed profiles does suggest that some design aspects gleaned from the profiles are viewed favorably by users; Chapter 9 discusses this study.

6.5 Summary of Developed Profiles

Several statistical models were developed to classify pages and sites into the three classes (good, average, and poor) as depicted in Table 6.2. Another model was developed to map pages into one of the three clusters of good pages: small-page, large-page, and formatted-page. Each of these models encapsulates key predictor measures and relationships among measures for classifying pages and sites. All of the models are used by the Analysis Tool (see Chapters 4 and 8) and are summarized below. The remainder of this chapter discusses each model in detail.

6.5.1 Page-Level Models

Overall Page Quality (Section 6.6): 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 (see below). The model also reports the decision tree rule that generated the prediction.

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

Table 6.2: Page and site level classification accuracies. C&RT refers to the Classification and Regression Tree algorithm. LDA refers to the Linear Discriminant Analysis.

Closest Good Page Cluster (Section 6.7): 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 ten 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 (Section 6.8): 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 ten 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 good, average, and poor classes as determined from ANOVAs. A separate decision tree model predicts the functional type of a page based on page-level measures (see Section 5.11). The Analysis Tool also enables users to specify a page type for analysis.

Content Category Quality (Section 6.9): 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 ten measures that are consistent with the content category. Each model also reports the top ten measures that are inconsistent with the content category and acceptable metric values. In both cases, measures are ordered by their importance in distinguishing pages in the good, average, and poor classes as determined by ANOVAs. The Analysis Tool enables users to specify content categories for analysis.

An effort was made to develop good page clusters for each of the content categories and page types, but the number of good pages in each category was inadequate. K-means clustering typically converged onto two or three clusters; however, the cluster sizes were disproportionate with one or two clusters containing less than 30 pages, for instance.

Similarly, an effort was made to develop classification models that consider page type and content category combinations (i.e., predicting good community home pages or good finance link pages). Table 6.3 shows the distribution of pages into each content category and page type combination, and Table 6.4 shows the accuracy of discriminate classification models developed

Cont. Cat.	Page Type														
	Home			Link			Content			Form			Other		
	G	A	P	G	A	P	G	A	P	G	A	P	G	A	P
Comm.	67	33	30	190	108	51	212	155	101	51	53	19	22	16	15
Educ.	53	53	35	189	145	119	184	163	171	82	47	29	11	25	37
Finance	24	13	25	67	60	64	107	47	104	30	5	38	5	2	10
Health	20	33	26	61	145	90	67	218	153	21	46	21	3	15	9
Living	36	23	20	113	53	90	80	83	84	30	31	20	8	21	8
Services	15	32	23	60	70	65	71	86	108	20	35	19	5	19	21

Table 6.3: Number of good (G), average (A), and poor (P) pages used to develop models for each content category and page type combination. Four pages with missing Bobby measures were discarded by the discriminant classification algorithm.

for each combination. In some cases predictive accuracy is considerably lower, possibly due to inadequate data. It is also possible that page type mispredictions contribute to lower predictive accuracy. However, the results suggest that these models could be built with at least 60 pages for each combination.

6.5.2 Page-Level Models: Key Predictor Measures

Tables 6.5, 6.6, and 6.7 summarize page-level measures that were among the top ten measures in the models for classifying good, average, and poor pages (described above); ANOVAs were used to determine the top ten predictor measures for each model. The tables show that several measures – italicized body word count (text formatting), minimum font size (text formatting), minimum color use (page formatting), and Weblint errors (page performance) – were among the top ten predictors in over half of the models. The analysis showed that pages with many italicized body text words were consistent with poor pages. The analysis also showed that good pages use a smaller font size (< 9 pt), typically for copyright text, and an accent (color sparsely used; measured by the minimum color use). Finally, the analysis showed that good pages tended to contain more HTML coding (Weblint) errors, which correlated with page formatting measures, such as the table and interactive object counts.

Tables 6.5, 6.6, and 6.7 also show that the other predictor measures vary considerably across the page-level models. All but three models – other, health, and living page quality – have one or more text element measures as key predictors. Only a few models (overall, home, health, and living page quality) have link element measures as key predictors; the health page quality model has four link element measures as key predictors. The graphic element measures are used more so for the overall and page type quality models than for the content category quality models. The text formatting, graphic formatting, page formatting, and page performance measures are used fairly equally among the models. The link formatting measures are used more so for the content category quality models than for the overall and page type quality models.

The variation among key predictors in the models suggests that characteristics of pages vary depending upon the context – page type or content category. Consequently, design goals need to be clarified before applying the models towards assessing and improving a Web interface. Variation in key predictor measures plus the fact that it was possible to get accurate predictions from a small number of pages in the combined page type and content category models, suggest that an extensive set of page-level measures, such as the ones developed, is essential to model

Cont. Cat.	Page Type								
	Home			Link			Content		
	G	A	P	G	A	P	G	A	P
Comm.	79%	64%	83%	99%	100%	96%	97%	99%	92%
Educ.	94%	89%	94%	96%	90%	95%	94%	96%	92%
Finance	75%	85%	84%	100%	100%	100%	100%	100%	100%
Health	75%	64%	89%	98%	97%	98%	99%	99%	98%
Living	61%	70%	55%	93%	94%	92%	100%	98%	99%
Services	60%	69%	65%	100%	100%	100%	100%	100%	100%
Cont. Cat. Avg.	74%	73%	78%	98%	97%	97%	98%	98%	98%

Cont. Cat.	Page Type					
	Form			Other		
	G	A	P	G	A	P
Comm.	92%	93%	74%	73%	88%	80%
Educ.	83%	83%	90%	100%	88%	89%
Finance	100%	100%	95%	100%	100%	90%
Health	100%	98%	95%	100%	53%	100%
Living	100%	94%	85%	75%	95%	100%
Services	90%	100%	100%	100%	79%	95%
Cont. Cat. Avg.	94%	95%	90%	91%	84%	92%

Table 6.4: Classification accuracy for predicting good (G), average (A), and poor (P) pages for each content category and page type combination.

Measure	Overall Quality	Page Type Qual.					Content Cat. Qual.						Use Freq.
		H	L	C	F	O	C	E	F	H	L	S	
Text Element Measures													
Link Word Count		√	√										16.67%
Good Link Word Count	√		√										16.67%
Graphic Word Count					√			√					16.67%
Good Graphic Word Count								√	√				16.67% 16.67%
Spelling Error Count				√			√					√	25.00%
Link Element Measures													
Text Link Count	√												8.33%
Link Count		√								√			16.67%
Internal Link Count		√								√	√		25.00%
Redundant Link Count										√			8.33%
Link Graphic Count										√			8.33%
Graphic Element Measures													
Graphic Ad Count	√	√	√		√								33.33%
Animated Graphic Ad Count			√				√				√		25.00%
Text Formatting Measures													
Italicized Body Word Count	√	√		√	√	√		√		√			58.33%
Exclaimed Body Word Count		√											8.33%
Bolded Body Word Count						√							8.33%
Minimum Font Size	√		√	√	√		√	√			√		58.33%
Body Color Count					√								8.33%
Text Cluster Count			√										8.33%
Link Text Cluster Count												√	8.33%
Text Column Count			√							√		√	25.00%

Table 6.5: Key measures used for predictions in the page-level models; all of these measures were among the top 10 predictors for at least one model (Table 1 of 3). The page type quality models are for home (H), link (L), content (C), form (F), and other (O) pages. The content category models are for pages from community (C), education (E), finance (F), health (H), living (L), and services (S) sites. A √ indicates whether a measure was among the top 10 predictors for a model. The use frequency reflects the percentage of time a measure is among the top predictors across all of the models.

Measure	Overall Quality	Page Type Qual.					Content Cat. Qual.						Use Freq.
		H	L	C	F	O	C	E	F	H	L	S	
Link Formatting Measures													
Link Color Count											√		8.33%
Standard Link Color Count					√						√		16.67%
Non-Underlined Text Links							√		√			√	25.00%
Graphic Formatting Measures													
Minimum Graphic Width				√	√	√					√		33.33%
Minimum Graphic Height											√		8.33%
Page Formatting Measures													
Minimum Color Use	√		√	√		√	√	√				√	58.33%
Interactive Object Count					√			√					16.67%
Search Object Count									√		√	√	25.00%
Good Text Color Combinations									√				8.33%
Neutral Text Color Combinations						√							8.33%
Good Panel Color Combinations				√									8.33%
Bad Panel Color Combinations		√											8.33%
Vertical Scrolls				√									8.33%
Horizontal Scrolls				√									8.33%
Serif Font Count									√				8.33%
Undetermined Font Style Count							√						8.33%
Fixed Page Width Use								√					8.33%
Internal Stylesheet Use									√				8.33%
Self Containment											√		8.33%

Table 6.6: Key measures used for predictions in the page-level models; all of these measures were among the top 10 predictors for at least one model (Table 2 of 3). The page type quality models are for home (H), link (L), content (C), form (F), and other (O) pages. The content category models are for pages from community (C), education (E), finance (F), health (H), living (L), and services (S) sites. A √ indicates whether a measure was among the top 10 predictors for a model. The use frequency reflects the percentage of time a measure is among the top predictors across all of the models.

Measure	Overall Quality	Page Type Qual.					Content Cat. Qual.						Use Freq.
		H	L	C	F	O	C	E	F	H	L	S	
Page Performance Measures													
Graphic File Count												√	8.33%
HTML File Count									√				8.33%
Script File Count							√						8.33%
Script Bytes									√				8.33%
Object Count		√					√	√		√			33.33%
Table Count												√	8.33%
Bobby Approved												√	8.33%
Bobby Priority 1 Errors										√			8.33%
Bobby Priority 2 Errors	√	√						√					25.00%
Bobby Browser Errors				√	√			√		√			33.33%
Weblint Errors	√	√	√		√					√	√	√	58.33%
Visible Page Text Hits						√							8.33%
All Page Text Hits						√							8.33%
Visible Link Text Hits						√							8.33%
Visible Link Text Score						√							8.33%
All Link Text Hits						√							8.33%
Page Title Hits							√		√				16.67%
Page Title Score				√			√		√				25.00%

Table 6.7: Key measures used for predictions in the page-level models; all of these measures were among the top 10 predictors for at least one model (Table 3 of 3). The page type quality models are for home (H), link (L), content (C), form (F), and other (O) pages. The content category models are for pages from community (C), education (E), finance (F), health (H), living (L), and services (S) sites. A √ indicates whether a measure was among the top 10 predictors for a model. The use frequency reflects the percentage of time a measure is among the top predictors across all of the models.

development. Both prior metric studies used only a small subset of measures and a larger number of pages than the combined models, but the predictive accuracy was considerably less. Hence, it appears that the high accuracy of the developed models is largely attributable to the exhaustive set of measures.

6.5.3 Site-Level Models

One limitation of the site-level models below is that they do not take page-level quality into consideration. Thus, it is possible for a site to be classified as good even though all of the pages in the site are classified as poor and vice versa. To remedy this situation, the median predictions for pages in the site are also reported by the Analysis Tool. These median page-level predictions need to be considered in determining the overall quality of a site.

Overall Site Quality (Section 6.10): a decision tree model for classifying a site into the good, average, and poor classes without considering the content category (see below). The model also reports the decision tree rule that generated the prediction.

Median Overall Page Quality (Section 6.10): predictions from the overall page quality model (described in Section 6.6) are used to derive the median overall page quality; the median overall page quality is then used to classify a site into the good, average, and poor classes. These predictions need to be considered in conjunction with predictions from the overall site quality model above. The accuracy of this model is the same as the accuracy of the overall page quality model; hence, no accuracy measure is reported in Table 6.2.

Content Category Quality (Section 6.11): 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.

Median Content Category Quality (Section 6.11): predictions from the page-level content category quality models (described in Section 6.9) are used to derive the median content category quality; the median content category quality is then used to classify a site into the good, average, and poor classes. These predictions need to be considered in conjunction with predictions from the site-level content category quality models above. The accuracy of these models are the same as the accuracy of the content category models for pages; hence, no accuracy measure is reported in Table 6.2.

To evaluate sites with the site-level models, the Site Crawler Tool (see Chapter 4) needs to be used to download pages from sites. The crawler should be configured to crawl three levels on each site and to download fifteen level-one pages and three level-two pages linked from each level-one page.

6.5.4 Site-Level Models: Key Predictor Measures

The maximum page depth was the only significant measure in site quality predictions, specifically for the overall site quality model; this is possibly due to inadequate data, the need for better site measures, or the need for model building methods. Table 6.2 shows that the site-level models are considerably less accurate than the page-level models. Future work will explore better measures and possibly model building methods to improve the site-level models.

Function	Squared Canonical Correlation	Wilks' Lambda	Chi-Square	Sig.	Classification Accuracy		
					Good	Average	Poor
1	0.44	0.394	4915.8	0.000	–	–	–
2	0.29	0.709	1810.5	0.000	–	–	–
Overall	–	–	–	–	76%	67%	74%

Table 6.8: Classification accuracy for predicting good, average, and poor pages using two linear discriminant functions.

6.6 Overall Page Quality

The goal of this section is to present an overall view of highly-rated pages that can be used in the future to assess pages without considering the six content categories studied. This analysis also does not consider page types. The data consists of 5,346 pages – 1,906 good pages (36%), 1,835 average pages (34%), and 1,605 poor pages (30%).

6.6.1 Overall Page Quality: Classification Model

Linear discriminant classification was used to develop a model for classifying all of the pages into the good, average, and poor classes. All of the measures, except external stylesheet use, met the criteria for inclusion in model development[2]. Table 6.8 summarizes key classification accuracy measures for this model. Since classification is performed for three groups, the algorithm derived 2 classification functions. The squared canonical correlation indicates the percentage of variance in the measures accounted for by each discriminant function. Wilks' Lambda indicates a complementary measure – the proportion of variance not explained by differences among groups. The corresponding Chi-Square for each Wilks' Lambda is computed for reporting significance; both discriminant functions have significant Wilks' Lambda. The model classifies pages with 72% accuracy overall.

Standardized coefficients for both discriminant functions illustrate key measures for classifying pages. Measures with standardized coefficients greater than one (in absolute value) are listed below.

Function 1: text element measures – meta tag, good meta tag, page title, and overall page title word counts; and page performance measures – page title and unique page title terms.

Function 2: text element measures – good page title, and overall good page title word counts, and fog sentence count; graphic element measure – graphic count; and page performance measures – visible page, visible link, all link, and all unique link text terms, page title and unique page title terms.

Classification accuracy was improved by developing a decision tree with the Classification and Regression Tree (C&RT) algorithm [Breiman *et al.* 1984]; 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% (96%, 94%, and 93% for good, average, and poor pages, respectively). The tree uses 71 of the measures; these measures represent all eight of the page-level metric categories – text element and formatting, link element and formatting, graphic element and formatting, and page formatting

[2]The page depth, reading complexity, and overall reading complexity measures were excluded from analyses in this section and others. Page type is also excluded except in sections examining this measure.

if ((Italicized Body Word Count is missing OR (Italicized Body Word Count $\leq$ 2.5)) AND (Minimum Font Size is missing OR (Minimum Font Size $\leq$ 9.5)) AND (Graphic Ad Count is not missing AND (Graphic Ad Count $>$ 2.5)))
Class = Good

This rule classifies pages as good pages if they have: two or fewer italicized body text words; use a font size of 9pt or less for some text; and more than two graphical ads.

---

if ((Italicized Body Word Count is missing OR (Italicized Body Word Count $\leq$ 2.5)) AND (Minimum Font Size is missing OR (Minimum Font Size $\leq$ 9.5)) AND (Graphic Ad Count is missing OR (Graphic Ad Count $\leq$ 2.5)) AND (Exclaimed Body Word Count is missing OR (Exclaimed Body Word Count $\leq$ 12.5)) AND (Exclaimed Body Word Count is not missing AND (Exclaimed Body Word Count $>$ 11.5)) AND (Bobby Priority 2 Errors is missing OR (Bobby Priority 2 Errors $\leq$ 5.5)) AND (Meta Tag Word Count is missing OR (Meta Tag Word Count $\leq$ 66)) AND (Emphasized Body Word Count is missing OR (Emphasized Body Word Count $\leq$ 174.5)) AND (Bad Panel Color Combinations is missing OR (Bad Panel Color Combinations $\leq$ 2.5)))
Class = Average

This rule classifies pages as average pages if they have: two or fewer italicized body text words; use a minimum font size of 9pt or less for some text; two or fewer graphical ads; twelve exclaimed body words (i.e., body text followed by exclamation points); five or fewer Bobby priority 2 errors; 66 or fewer meta tag words; 174 or fewer emphasized body words (i.e., body text that is colored, bolded, italicized, etc.); and less than two bad panel color combinations.

---

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

This rule classifies pages as poor pages if they have more than two italicized body text words.

Figure 6.1: Example decision tree rules for predicting page classes (Good, Average, and Poor).

and performance measures. Figure 6.1 depicts example rules for classifying pages into the good, average, and poor classes. Model predictions were retained for further analysis.

6.6.2 Overall Page Quality: Characteristics of Good, Average, and Poor Pages

Further analysis was conducted to determine significant differences between pages in the three classes. Specifically, one-way analyses of variance (ANOVAs) were computed to identify measures where the within-class variance is significantly different from the between-class variance. Correlation coefficients were also computed between pairs of predictor measures. The analysis only considered pages accurately classified by the decision tree – 1,822 good pages, 1,732 average pages, and 1,486 poor pages.

ANOVAs revealed that all but eight of the 71 measures – unique page title terms, graphic count, average font size, maximum graphic height, graphic link count, link graphic count, download time, and whether a page was Bobby approved – were significantly different between the three

classes of pages. Tables 6.9, 6.10, and 6.11 depict means and standard deviations for each measure. The contribution of each measure is reported by the F value; F values were sorted to determine a measure's rank. All of the F values are significant at the .05 level.

The top ten predictors are 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. Differences among good, average, and poor pages 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 among the classes; all differences were significant, except as noted below.

- Good pages surprisingly use minimum font sizes of nine points; 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.

- Average and poor pages have larger minimum color usages than good pages (five and six times vs. four times). Inspection of a random sample of good average, and poor pages suggest that this results from a tendency for good pages 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 these 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 interactive objects, tables, images, etc. This finding suggests that highly-rated pages tend not to conform to accessibility and good HTML coding standards. It is possible that in some cases good pages (and possibly average and poor pages) are unnecessarily penalized by these tools. As an example, Bobby requires alternative text to be provided for all images on a page. However, designers may frequently use blank images as spacers and may not provide alternative text for them, resulting in the page not being Bobby approved. On the other hand, if designers did provide alternative text for spacer images, this may actually impede blind users, since the text will be read by screen readers. Perhaps these tools need to consider the context in which page elements are being used.

- Good pages typically contain one graphical ad; poor pages are slightly more likely to contain graphical ads than average pages. An examination of ten sites suggests that ads on good sites are for well-known entities whereas ads on poor sites are for obscure entities. Kim and Fogg [1999] conducted a controlled study wherein 38 users rated Web pages (with and without graphical ads) on credibility ("high level of perceived trustworthiness and expertise") and found that pages with graphical ads were rated as more credible than those without graphical ads.

- Good pages contain about 27 text links, while average pages contain 22 and poor pages contain 19. Poor pages are also less likely to contain link text clusters (areas containing text links highlighted with color, lists, etc.), while good pages contain slightly more link clusters than average pages. There is a corresponding higher number of good link words (words in the link text that are not stop words or 'click') on good and average pages than on poor pages.

Measure	Mean			Std. Dev.			F val.	Rank
	Good	Avg.	Poor	Good	Avg.	Poor		
Text Element Measures								
Good Link Word Count	47.5	36.2	31.8	42.2	34.5	31.0	83.8	10
Good Meta Tag Word Count	15.3	10.9	17.0	23.5	17.2	24.5	33.8	23
Meta Tag Word Count	20.8	14.3	21.6	32.1	22.8	31.3	31.6	24
Good Word Count	204.0	175.4	185.5	157.9	142.9	144.2	16.9	32
Word Count	378.0	326.2	345.5	298.3	271.1	273.3	15.4	37
Good Page Title Word Count	3.1	3.3	3.4	1.7	1.7	1.9	10.9	44
Exclamation Point Count	1.2	1.0	1.0	1.5	1.4	1.4	9.8	45
Display Word Count	17.1	14.8	16.4	17.5	16.0	17.0	8.2	52
Fog Big Word Count	33.7	32.2	36.4	34.3	34.6	37.6	5.7	59
Body Word Count	264.5	243.3	265.1	232.6	224.9	238.5	4.9	60
Good Body Word Count	127.3	118.7	129.2	114.9	113.4	118.3	3.9	62
Link Element Measures								
Text Link Count	27.4	21.5	18.6	23.0	19.5	16.9	84.2	9
Link Count	41.2	34.1	31.8	28.8	23.9	21.7	63.5	14
Redundant Link Count	7.7	6.6	6.7	7.7	6.4	6.8	13.7	39
Graphic Element Measures								
Graphic Ad Count	1.2	0.7	0.7	1.4	0.9	0.8	103.7	5
Redundant Graphic Count	9.4	7.9	8.8	12.4	10.2	11.0	8.1	54
Text Formatting Measures								
Minimum Font Size	9.0	9.3	9.3	0.2	0.7	0.7	247.4	1
Italicized Body Word Count	0.5	0.5	1.1	0.9	0.9	1.7	139.3	3
Link Text Count	1.2	1.0	0.6	1.4	1.4	0.8	91.3	6
Text Column Count	4.1	3.4	2.8	3.7	3.0	2.4	76.4	11
Exclaimed Body Word Count	4.1	3.0	2.3	6.4	4.8	3.7	52.1	18
Text Cluster Count	2.2	1.9	1.5	2.3	2.3	1.5	41.2	21
Capitalized Body Word Count	1.7	1.2	1.8	2.5	1.5	2.5	37.7	22
Text Positioning Count	3.5	2.8	3.3	3.3	2.8	3.2	20.9	28
Sans Serif Word Count	227.4	185.7	214.4	239.7	203.4	228.1	15.9	34
Display Color Count	1.5	1.3	1.4	0.9	1.0	0.9	15.5	35
Emphasized Body Word Count	52.8	51.5	61.4	60.8	56.3	69.2	11.8	43

Table 6.9: Means and standard deviations for good, average, and poor pages (Table 1 of 3). All measures are significantly different (.05 level) and sorted within each category by their contribution to predictions (Rank column); the rank reflects the size of the F value.

Measure	Mean			Std. Dev.			F val.	Rank
	Good	Avg.	Poor	Good	Avg.	Poor		
Text Formatting Measures								
Bolded Body Word Count	11.4	12.8	14.0	16.1	17.4	20.2	9.1	48
Colored Body Word Count	17.6	20.2	20.6	24.4	26.4	27.4	6.7	55
Serif Word Count	92.4	83.0	83.2	133.2	116.5	124.9	3.2	63
Link Formatting Measures								
Standard Link Color Count	1.1	1.4	1.5	1.2	1.2	1.3	58.8	16
Graphic Formatting Measures								
Minimum Graphic Width	22.9	31.9	20.4	32.5	43.8	28.5	47.1	20
Minimum Graphic Height	10.1	8.7	9.0	13.7	11.1	12.6	6.2	57
Maximum Graphic Width	436.9	456.3	439.5	190.1	175.2	176.0	5.8	58
Page Formatting Measures								
Minimum Color Use	3.5	5.3	6.6	3.1	5.1	6.7	156.2	2
Interactive Object Count	3.2	2.2	1.7	3.7	2.9	2.5	91.1	7
Bad Panel Color Combinations	1.0	0.7	0.6	1.3	0.8	0.8	70.6	12
Vertical Scrolls	2.0	1.6	1.6	1.4	1.0	1.0	60.2	15
Horizontal Scrolls	0.1	0.1	0.0	0.3	0.3	0.2	28.7	25
Color Count	8.0	7.6	7.5	2.7	2.4	2.3	23.0	26
Font Count	5.6	5.3	5.1	2.2	2.5	2.4	22.7	27
Good Text Color Combinations	3.4	3.1	3.1	2.2	1.8	1.7	19.2	29
Page Pixels	802K	731K	747K	475K	439K	445K	11.8	42
Good Panel Color Combinations	0.4	0.5	0.6	0.7	0.7	0.8	8.9	49
Browser-Safe Color Count	5.0	4.9	4.9	1.6	1.6	1.5	4.7	61

Table 6.10: Means and standard deviations for good, average, and poor pages (Table 2 of 3). All measures are significantly different (.05 level) and sorted within each category by their contribution to predictions (Rank column); the rank reflects the size of the F value.

Measure	Mean			Std. Dev.			F val.	Rank
	Good	Avg.	Poor	Good	Avg.	Poor		
Page Performance Measures								
Weblint Errors	34.5	26.3	19.1	36.6	27.8	18.4	116.1	4
Bobby Priority 2 Errors	4.0	3.6	3.5	1.3	1.1	1.0	87.3	8
Bobby Browser Errors	13.9	11.5	11.8	7.5	6.3	5.8	66.3	13
Object Count	1.9	1.3	1.4	2.1	1.4	1.4	55.1	17
Script File Count	0.5	0.2	0.2	1.2	0.6	0.6	52.0	19
All Page Text Terms	305.2	265.7	303.3	214.7	191.3	232.5	18.9	30
Visible Page Text Terms	258.7	221.8	254.9	203.4	175.4	212.4	18.3	31
Visible Page Text Hits	31.0	28.5	26.4	24.1	22.9	21.9	16.5	33
Table Count	8.1	7.4	6.9	6.2	5.7	5.7	15.4	36
Script Bytes	1.2K	1.1K	924.0	1.4K	1.5K	1.1K	14.3	38
HTML Bytes	15.4K	14.3K	13.8K	9.5K	9.8K	8.2K	13.0	40
Bobby Priority 1 Errors	1.3	1.2	1.3	1.0	0.8	0.8	13.0	41
All Unique Link Text Terms	133.5	119.4	125.8	101.9	91.2	92.8	9.7	46
All Page Text Score	141.1	126.3	138.2	106.3	101.5	114.2	9.3	47
Visible Page Text Score	91.5	81.0	85.8	77.4	70.8	77.6	8.7	50
Visible Unique Link Text Terms	120.4	108.5	111.4	96.2	87.9	87.6	8.2	51
All Page Text Hits	41.7	37.8	39.2	29.8	28.3	30.2	8.1	53
Page Title Terms	3.8	3.9	4.1	2.4	2.1	2.4	6.6	56

Table 6.11: Means and standard deviations for good, average, and poor pages (Table 3 of 3). All measures are significantly different (.05 level) and sorted within each category by their contribution to predictions (Rank column); the rank reflects the size of the F value.

- Good pages appear to be more interactive than pages in the other classes; they contain about three interactive objects (e.g., search button or pulldown menu). Average and poor pages contain about two interactive objects.

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 as described below.

- Good pages appear to use colors in various ways. 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 color 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).

- Average pages appear to use bad panel color combinations to format link text clusters. Furthermore, the number of good links words is correlated with the number of redundant links suggesting that good link words may appear for example in navigation bars and footers but not necessarily in the text. The average and minimum font sizes are correlated suggesting little variance in text sizes on average pages.

- Poor pages appear to use color to a lesser degree than good and average pages; however, when colors are used, they are typically overused as discussed previously. Color count is correlated with the number of interactive objects suggesting that color is used to highlight these objects. The number of good text color combinations is also correlated with the number of interactive objects, which are typically formatted with a white background and black text. This suggests that poor pages tend to use multiple formatting techniques at once. There is a correlation between redundant graphic count and graphic link count, which suggests that image links are presented multiple times. This is in contrast to good pages that repeat links in multiple forms.

Figures 6.2, 6.3, and 6.4 depict example good, average, and poor pages, respectively. These pages demonstrate many of the discussed properties.

6.7 Good Page Clusters

The decision tree model presented in Section 6.6 accurately classified 1,822 (96%) of the 1,906 good pages from both the training and test sets; these pages were retained for cluster analysis. K-means clustering [SPSS Inc. 1999] was used to identify sub-groups of similar good pages. This method requires all measures to be on the same scale; hence, measures were transformed into Z scores. Each Z score is a standard deviation unit that indicates the relative position of each value within the distribution (i.e., $Z_i = \frac{x_i - \bar{x}}{\sigma}$, where x_i is the original value, $\bar{x}$ is the mean, and σ is the standard deviation).

K-means clustering converged onto three clusters of good pages. The first cluster consists of 450 pages (24.5%), the second cluster consists of 364 pages (20%), and the final cluster consists of 1,008 pages (55.3%). Tables 6.12, 6.13, and 6.14 contrast means and standard deviations for the three clusters and depict the rank of each measure based on ANOVA results(i.e., F values). All

Figure 6.2: Good page exhibiting several key properties of this class: links repeated in multiple forms (text and images), graphical ads, interactive objects, multi-level colored headings, navigation bars, and variations in font sizes, including a smaller size for the footer.

Figure 6.3: Average page exhibiting several key properties of the class: bad panel and text color combinations for link text clusters, redundant links, and similar average and minimum font sizes (10pt and 9pt).

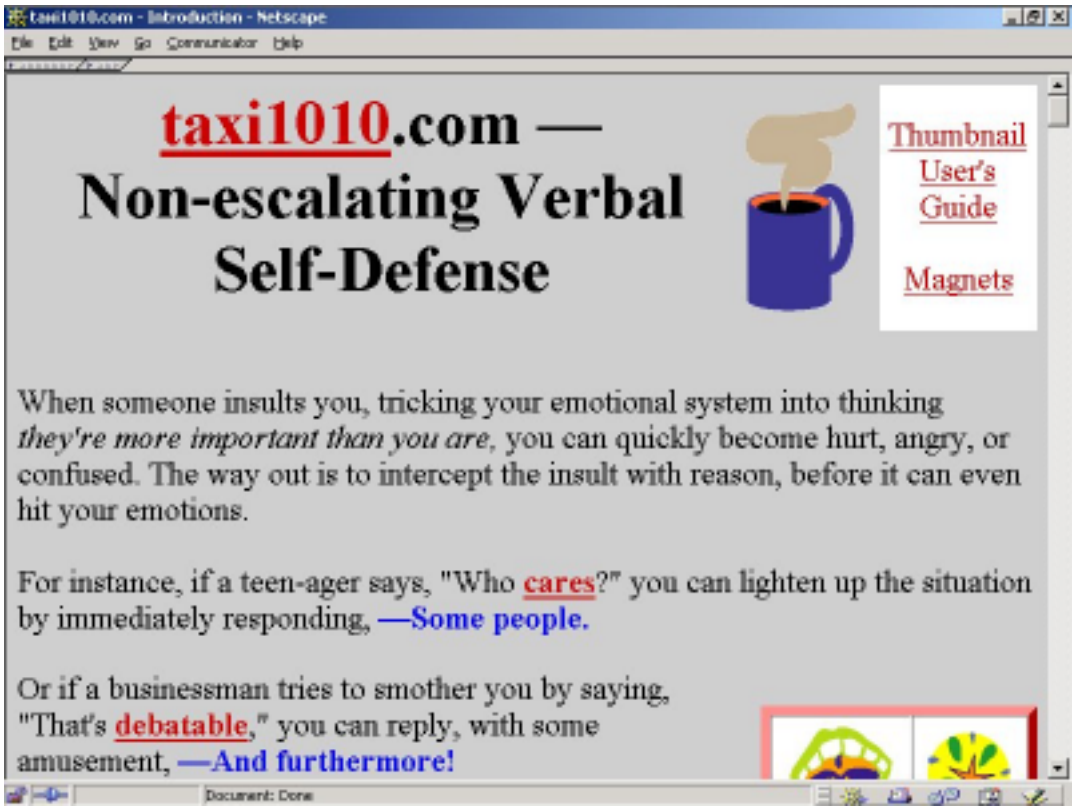

Figure 6.4: Poor page exhibiting several key properties of the class: italicized body text, repeated image links (images at the bottom right also appear in an area that is not visible), and minimal link text clustering.

of the decision tree measures were significantly different among clusters, except for the maximum graphic height.

Nine of the top ten measures are associated with the amount of text on a page, including the word count, good word count, HTML bytes, and vertical scrolls. Hence, the second and third clusters could be characterized as representing large and small pages. The large-page cluster is consistent with a group identified in a prior study by the author [Ivory *et al.* 2001], while the small-page cluster is consistent with two groups identified in the same study – low and medium word count pages. The remaining top ten measure – table count – distinguishes pages in the first cluster as ones that are highly formatted. Pages in the formatted-page cluster contain on average 120 more words than pages in the small-page cluster.

Pages closest to the centroid of each cluster are depicted in Figures 6.5, 6.6, and 6.7. All of the pages exhibit most of the properties previously discussed for good pages. Pages in the small-page and large-page clusters are similar in many ways, except for the amount of text on a page. ANOVAs contrasting these two clusters revealed that they are similar on 13 measures – the number of meta tag and good meta tag words, script bytes, minimum color use, good page title words, use of standard link colors, horizontal scrolls, graphical links, download time, and Bobby approval and priority 2 errors.

Pages in the formatted-page cluster are quite distinct from pages in the small-page and large-page clusters. They use more text positioning and columns, tables, and text color and panel color combinations. They also contain more graphics and redundant graphics, graphical ads, and have smaller minimum image widths and heights; correlations suggest that many of these graphics are possibly for organizing pages. Pages in this cluster also contain more interactive objects and colors. The example good page presented in Figure 6.2 belongs to this cluster.

6.8 Page Type Quality

The goal of this section is to show differences when the page type – home, link, content, form, or other – is included in the analysis. This analysis does not consider content categories. Table 6.15 summarizes the analyzed data.

Linear discriminant analysis was used to distinguish good, average, and poor pages within each page type, yielding an overall accuracy of 82%. Table 6.16 summarizes the accuracy of each page type model. These models are 7–15% less accurate than the overall page quality model. Recall that page types are predicted by a decision tree model with 84% overall accuracy (see Section 5.11). It is possible that mispredicted pages within each page type category may have negatively impacted model development.

ANOVAs revealed that the top ten predictor variables varied across page types, although several predictors from the overall page model were often among the top ten, including the interactive object count, minimum font size, italicized body word count, minimum color use, graphic ad count, and Bobby and Weblint errors. Key predictors in each page type model are discussed below. An effort was made to develop good page clusters for each of the page types, but the number of good pages for each type was inadequate.

Home Pages. The top ten measures for classifying good, average, and poor home pages include the following: graphic ad count (graphic element); object count, Weblint errors, and Bobby priority 2 errors (page performance); bad panel color combinations (page formatting); link and internal link counts (link element); italicized and exclaimed body word counts (text formatting); and link word count (text element). Good home pages contain considerably more links (internal links in particular) and a corresponding higher number of link words

Measure	Mean			Std. Dev.			F val.	Rank
	FP	LP	SP	FP	LP	SP		
Text Element Measures								
Word Count	360.6	849.2	215.7	194.7	221.2	140.2	1726.5	1
Good Word Count	203.1	449.3	115.9	110.5	112.8	74.5	1690.6	2
Body Word Count	200.3	621.1	164.4	138.8	196.1	132.4	1317.7	5
Good Body Word Count	92.4	303.5	79.2	70.6	94.5	65.9	1314.0	6
Fog Big Word Count	23.7	79.1	21.7	20.5	36.4	23.0	730.7	11
Good Link Word Count	80.0	62.3	27.8	41.9	47.2	26.5	364.9	22
Display Word Count	23.3	29.4	9.8	17.4	20.3	12.1	251.8	32
Exclamation Point Count	1.8	1.5	0.8	1.6	1.7	1.2	96.0	52
Good Meta Tag Word Count	23.7	12.0	12.7	25.3	20.6	22.7	38.6	60
Meta Tag Word Count	32.0	16.7	17.3	34.5	28.4	31.2	36.0	61
Good Page Title Word Count	2.8	3.3	3.2	1.6	1.7	1.7	16.8	66
Link Element Measures								
Link Count	67.6	49.5	26.3	25.6	31.2	17.4	525.6	18
Text Link Count	46.7	35.2	16.0	22.3	25.4	13.7	427.6	21
Redundant Link Count	12.6	8.4	5.2	8.6	8.3	5.7	171.2	39
Link Graphic Count	17.3	10.3	9.8	8.9	8.4	8.2	135.3	46
Graphic Element Measures								
Graphic Count	45.9	20.9	17.5	20.3	17.9	14.5	477.2	19
Redundant Graphic Count	21.7	7.5	4.7	13.7	10.7	7.9	452.6	20
Graphic Ad Count	2.2	0.9	0.8	1.5	1.3	1.2	230.1	33
Graphic Link Count	14.7	7.5	7.5	8.2	7.3	7.1	164.7	42
Text Formatting Measures								
Text Column Count	8.2	3.8	2.5	3.7	3.3	2.3	598.5	14
Link Text	2.4	1.2	0.7	1.4	1.4	1.0	311.3	26
Cluster Count								
Sans Serif Word Count	284.4	422.7	131.5	197.0	343.4	140.1	283.0	27
Text Cluster Count	3.7	3.0	1.2	2.4	2.6	1.5	278.1	29
Display Color Count	2.2	1.7	1.1	0.9	0.8	0.8	261.9	30
Text Positioning Count	5.8	3.7	2.3	3.5	3.4	2.5	214.1	34
Emphasized Body Word Count	56.1	102.0	33.6	51.3	73.1	48.2	202.8	36
Capitalized Body Word Count	1.5	3.5	1.2	2.3	2.9	2.1	136.6	44

Table 6.12: Means and standard deviations for the 3 clusters of good pages – formatted-page (FP), large-page (LP), and small-page (SP) (Table 1 of 3). All measures are significantly different (.05 level) and sorted within each category by their contribution (Rank column); the rank reflects the size of the F value.

Measure	Mean			Std. Dev.			F val.	Rank
	FP	LP	SP	FP	LP	SP		
Text Formatting Measures								
Bolded Body Word Count	12.6	20.9	7.3	14.9	19.6	13.5	108.9	50
Italicized Body Word Count	0.5	1.0	0.3	0.9	1.0	0.7	98.6	51
Serif Word Count	66.5	167.8	76.7	113.6	178.0	111.0	72.9	56
Colored Body Word Count	26.1	21.6	12.3	24.8	28.2	21.2	55.4	58
Average Font Size	10.2	11.0	10.8	1.0	1.2	1.2	52.1	59
Exclaimed Body Word Count	5.5	5.4	3.1	7.0	7.3	5.6	29.9	64
Minimum Font Size	8.9	9.0	9.0	0.3	0.2	0.2	7.3	70
Link Formatting Measures								
Standard Link Color Count	0.9	1.2	1.2	1.0	1.2	1.2	9.9	68
Graphic Formatting Measures								
Minimum Graphic Width	3.1	24.9	30.9	7.0	33.5	35.2	136.0	45
Minimum Graphic Height	2.3	10.8	13.3	4.1	14.1	15.0	116.8	48
Maximum Graphic Width	493.2	429.5	414.4	130.9	191.2	206.4	30.6	63
Page Formatting Measures								
Vertical Scrolls	2.1	3.8	1.2	1.2	1.3	0.9	748.8	9
Page Pixels	888K	1.4M	557K	393K	449K	284K	736.2	10
Color Count	11.1	7.8	6.8	2.6	2.2	1.7	687.7	12
Good Text Color Combinations	5.7	3.4	2.4	2.1	2.1	1.5	533.9	17
Bad Panel Color Combinations	2.1	0.9	0.5	1.3	1.2	1.0	312.3	25
Font Count	6.9	6.8	4.7	2.0	2.2	1.7	281.5	28
Good Panel Color Combinations	1.0	0.4	0.2	0.8	0.7	0.6	197.7	37
Interactive Object Count	5.7	2.6	2.2	3.8	3.3	3.3	169.5	41
Browser-Safe Color Count	6.0	5.0	4.7	1.8	1.4	1.3	129.6	47
Minimum Color Use	2.5	3.9	3.8	2.3	3.2	3.3	32.4	62
Horizontal Scrolls	0.0	0.1	0.1	0.2	0.3	0.3	10.3	67

Table 6.13: Means and standard deviations for the 3 clusters of good pages – formatted-page (FP), large-page (LP), and small-page (SP) (Table 2 of 3). All measures are significantly different (.05 level) and sorted within each category by their contribution (Rank column); the rank reflects the size of the F value.

Measure	Mean			Std. Dev.			F val.	Rank
	FP	LP	SP	FP	LP	SP		
Page Performance Measures								
Visible Unique Link Text Terms	99.2	270.6	75.6	75.6	66.6	47.2	1529.7	3
All Unique Link Text Terms	113.5	287.9	86.7	82.9	75.3	51.9	1335.3	4
HTML Bytes	25.1K	19.9K	9.4K	7.9K	8.3K	5.1K	986.9	7
Table Count	15.3	8.0	4.9	5.5	5.7	3.5	821.9	8
Visible Page Text Terms	189.3	520.6	195.0	152.1	158.2	156.4	646.6	13
All Page Text Terms	234.6	569.9	241.1	177.1	176.1	162.8	570.0	15
Weblint Errors	73.0	34.1	17.5	37.1	34.5	20.8	569.1	16
Bobby Browser Errors	20.6	13.0	11.2	6.6	7.0	5.9	357.1	23
Visible Page Text Score	126.6	239.8	111.8	106.7	105.2	83.0	259.8	31
Visible Page Text Hits	33.3	49.8	23.2	28.5	24.8	16.7	207.0	35
Bobby Priority 1 Errors	2.0	1.3	1.1	1.0	1.0	0.8	173.6	38
Object Count	3.3	1.6	1.3	2.2	2.2	1.7	170.5	40
All Page Text Hits	43.3	62.8	33.4	34.7	29.7	22.8	160.6	43
All Page Text Score	141.1	126.3	138.2	106.3	101.5	114.2	9.3	47
Bobby Priority 2 Errors	4.7	3.8	3.7	1.2	1.3	1.1	114.1	49
Page Title Terms	2.8	4.8	4.0	2.3	2.4	2.3	85.8	53
Script File Count	1.1	0.5	0.2	1.8	1.2	0.7	84.6	54
Unique Page Title Terms	2.8	4.7	3.9	2.3	2.3	2.2	83.2	55
Script Bytes	1.8K	1.0K	940.2	1.6K	1.4K	1.2K	68.1	57
Bobby Approved	0.1	0.2	0.2	0.3	0.4	0.4	29.7	65
Download Time	14.5	16.2	16.4	9.6	8.2	8.2	8.1	69

Table 6.14: Means and standard deviations for the 3 clusters of good pages – formatted-page (FP), large-page (LP), and small-page (SP) (Table 3 of 3). All measures are significantly different (.05 level) and sorted within each category by their contribution (Rank column); the rank reflects the size of the F value.

Page Type	Good	Average	Poor	Total
Home	213	187	159	559
Link	680	581	479	1740
Content	721	752	721	2194
Form	234	217	146	597
Other	54	98	100	252
Total	1902	1835	1605	5342

Table 6.15: Number of pages used to develop the page type quality models. Four pages with missing Bobby measures were discarded by the discriminant classification algorithm.

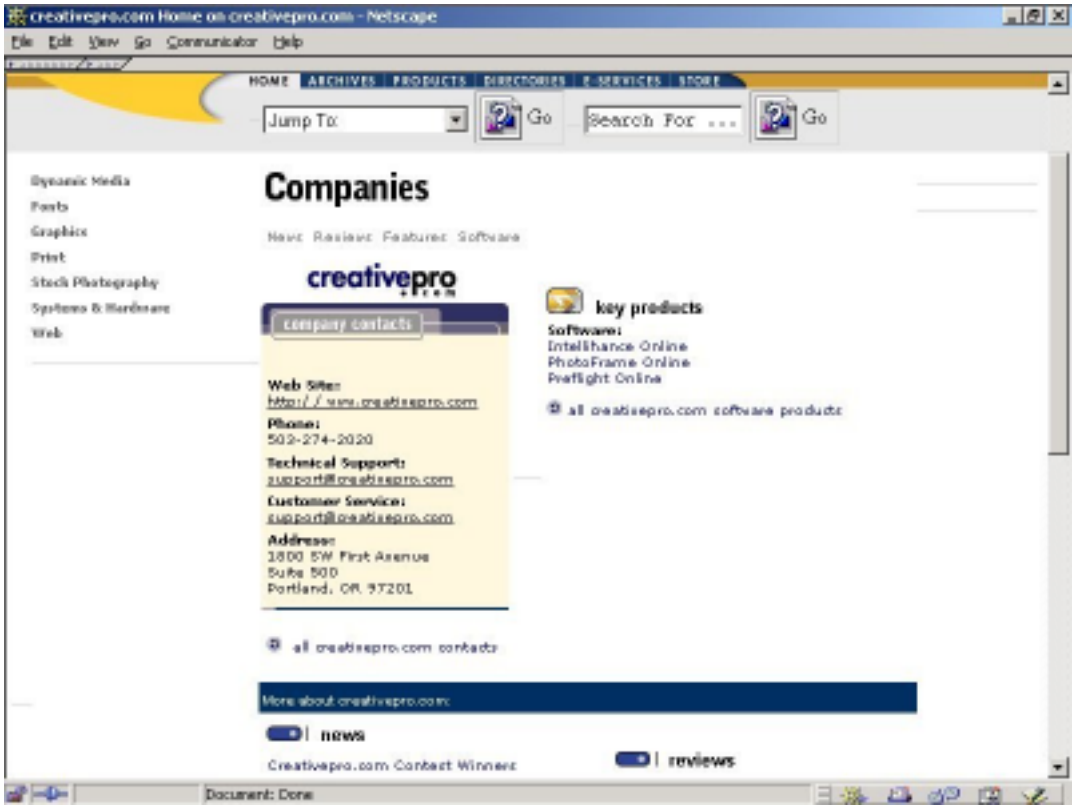

Figure 6.5: Page closest to the centroid of the formatted-page cluster (distance = 5.33 total standard deviation units of difference across all measures). Missing images are caused by a deficiency in the early version of the Site Crawler Tool.

Page Type	Sample Size	Classification Accuracy		
		Good	Average	Poor
Home	559	83.6%	80.2%	84.9%
Link	1740	80.1%	71.1%	78.3%
Content	2194	79.9%	74.2%	79.6%
Form	597	81.6%	77.0%	87.0%
Other	252	88.9%	76.5%	83.0%
Page Type Average		82.8%	75.8%	82.6%

Table 6.16: Classification accuracy for predicting good, average, and poor pages within page types.

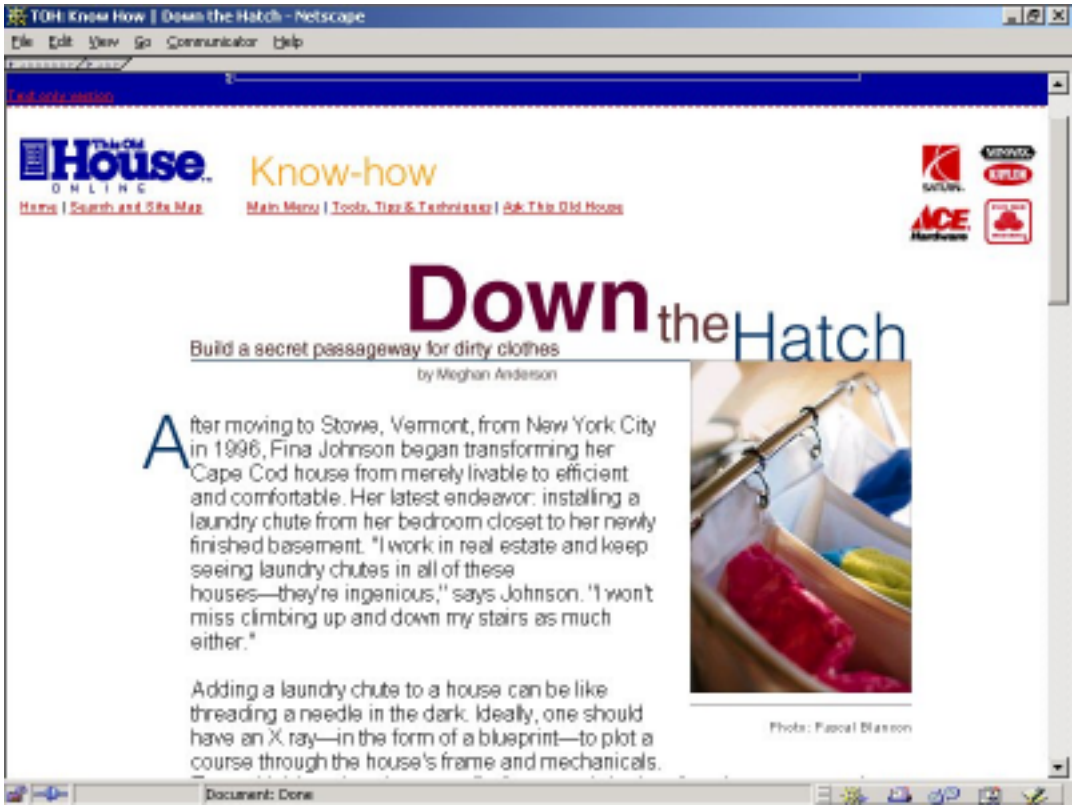

Figure 6.6: Page closest to the centroid of the large-page cluster (distance = 5.96 total standard deviation units of difference across all measures). Missing images are caused by a deficiency in the early version of the Site Crawler Tool.

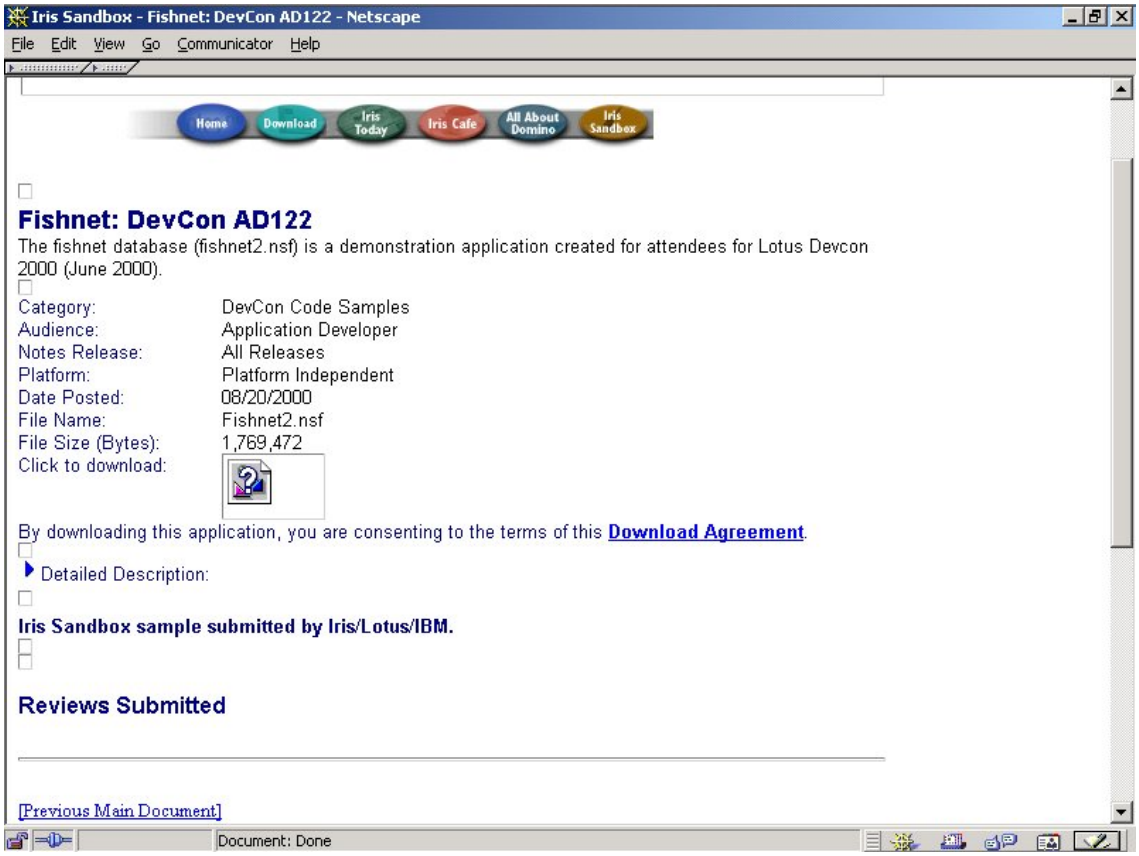

Figure 6.7: Page closest to the centroid of the small-page cluster (distance = 3.98 total standard deviation units of difference across all measures). Missing images are caused by a deficiency in the early version of the Site Crawler Tool.

than average and poor home pages. As was found for good pages overall, good home pages contain two graphical ads on average, while average and poor home pages contain an average of one graphical ad, respectively. Finally, good home pages use exclamation points to emphasize body text as opposed to italics; the converse is true for poor home pages.

Link Pages. The top ten measures for classifying link pages include: minimum font size, text cluster count, and text column count (text formatting); minimum color use and interactive object count (page formatting); Weblint errors (page performance); graphic and animated graphic ad counts (graphic element); and link and good link word counts (text element). Good link pages contain considerably more link and good link words than average and poor link pages. They also contain more text clusters, which suggests that text headings are used to organize groups of links. Good link pages also contain five text columns versus four and three for average and poor link pages, which suggests that links are organized into multiple columns. The last two claims were verified by examining a random sample of good link pages. Finally, good link pages are more likely to use an accent color unlike average and poor link pages.

Content Pages. The top predictor measures for classifying content pages include: minimum font size and italicized body word count (text formatting); minimum color use, good panel color combinations, and horizontal and vertical scroll counts (page formatting); spelling error count (text element); page title score and Bobby browser errors (page performance); and minimum graphic width (graphic formatting). Good content pages require an average of 2.4 vertical scrolls to read, while average and poor content pages require an average of 1.9 scrolls. This is because good content pages contain considerably more words than pages in the other categories; the difference is significant. Good content pages also require slightly more horizontal scrolls (.14 versus .1); however, horizontal scrolling is typically not required for content pages in any of the classes.

Average and poor content pages are more likely to contain good panel color combinations than good content pages suggesting that good content pages minimize colored areas on the page. Good content pages are also less likely to use page titles that are similar to source page titles suggesting the use of unique page titles among pages. Finally, good content pages appear to contain two spelling errors versus average and poor content pages that contain one spelling error. Inspection of a random sample of ten pages in each class revealed that most spelling errors (according to the Metrics Computation Tool) on good content pages are due to the use of jargon and abbreviations, such as cyberspace, messaging, groupware, busdev, and imusic. On the other hand, spelling errors on average and poor content pages tended to be true errors.

Form Pages. The top ten measures for classifying good, average, and poor form pages include the following: interactive object count (page formatting); minimum font size, body color count, and italicized body word count (text formatting); standard link color count (link formatting); graphic ad count (graphic element); Bobby browser and Weblint errors (page performance); minimum graphic width (graphic formatting); and graphic word count (text element). Many of the differences among form pages in the three classes mirror differences found with the overall page quality model. In addition, good form pages contain an average of eight interactive objects, while average and poor form pages contain an average of six interactive objects, respectively. Good form pages also use fewer than two body text colors unlike average and poor form pages that use more than two body text colors.

Content Category	Good	Average	Poor	Total
Community	542	365	216	1123
Education	518	433	391	1342
Finance	232	127	241	600
Health	172	457	299	928
Living	267	211	222	700
Services	171	242	236	649
Total	1902	1835	1605	5342

Table 6.17: Number of pages used to develop the content category quality models. Four pages with missing Bobby measures were discarded by the discriminant classification algorithm.

Content Category	Sample Size	Classification Accuracy		
		Good	Average	Poor
Community	1123	91.5%	87.9%	85.6%
Education	1342	87.8%	83.6%	85.7%
Finance	600	98.3%	98.4%	96.7%
Health	928	89.0%	93.4%	94.3%
Living	700	90.6%	89.1%	90.5%
Services	649	95.3%	95.0%	95.3%
Content Category Average		92.1%	91.2%	91.4%

Table 6.18: Classification accuracy for predicting good, average, and poor pages within content categories.

Other Pages. Recall that this page type broadly represents all remaining graphical (e.g., splash pages, image maps, and Flash) and non-graphical (e.g., blank, under construction, error, applets, text-based forms, and redirect) pages; thus, the pages have widely different features. The top ten measures for classifying other pages include the following: minimum color use and neutral text color combinations (page formatting); all link text hits, visible link text hits and score, all page text hits, visible page text hits (page performance); minimum graphic width (graphic formatting); and italicized and bolded body word counts (text formatting). Many of the top predictor measures are associated with the quality of scent between the source page's text and link text and the destination page's text. Good other pages have fewer common words with source link and page text than average pages but more common words than poor pages on all of the scent measures. Good other pages are more likely to contain neutral text color combinations than average and poor other pages. However, the three classes contain an equal number of good text color combinations, which suggests that good other pages use multiple text color combinations. Finally, good other pages contain about three bolded body words, while average and poor other pages contain six and twelve, respectively.

6.9 Content Category Quality (Pages)

The goal of this section is to show differences when the content category – community, education, finance, health, living, or services – is included in the analysis. The page type is not considered. Table 6.17 summarizes the analyzed data for each content category.

Linear discriminant analysis was used to distinguish good, average, and poor pages within each category. The classification models have an overall accuracy of 91%; Table 6.18 summarizes the accuracy of the each model. The average accuracy for the content category models is about

7–15% higher than the models developed for page types in Section 6.8. ANOVAs computed over pages accurately classified by each model revealed that the top ten predictor variables varied across content categories, although several predictors from the overall page model as well as the page type models were often among the top ten, including the minimum font size, italicized body word count, minimum color use, and Bobby and Weblint errors. Key predictors in each content category model are discussed below.

Community Pages. The top ten predictor measures for classifying good, average, and poor community pages include the following: minimum font size and undetermined font count (text formatting); spelling error count (text element); script file count, object count, and page title hits and score (page performance); minimum color use (page formatting); non-underlined text links (link formatting); and animated graphic ad count. Average and poor community pages were more likely to use fonts that are not recognized as serif or sans serif fonts (i.e., fonts that are not in the extensive lookup tables) than good pages. Good community pages are more likely to use scripts and one animated graphic on pages. Good community pages are more likely to contain text links without visible underlines than poor community pages but less so than average community pages. Finally, good community pages tend to use more distinct page titles between pages than the average and poor pages.

Education Pages. Key measures for classifying education pages include: Bobby priority 2 and browser errors and object bytes (page performance); minimum font size and italicized body word count (text formatting); minimum color use, fixed page width use, and interactive object count (page formatting); graphic and good graphic word counts (text element). Good education pages are less likely to use objects, such as applets, than poor education pages; there is no significant difference between good and average education pages. Good education pages are more likely to use a fixed page width (typically controlled by tables) than average and poor education pages. Good education pages contain about three interactive objects, while average and poor pages contain two and one, respectively. Similarly to other good pages, good education pages were slightly more likely to contain one graphical ad; however, this measure did not play a major role in classifying pages.

Finance Pages. The top ten predictor measures for classifying finance pages include the following: HTML file count, script bytes, and page title hits and score (page performance); serif font count, search object count, good text color combinations, and internal stylesheet use (page formatting); good graphic word count (text element); and non-underlined text links (link formatting). There are considerable differences in formatting for good finance pages compared to average and poor finance pages. Good finance pages consist of multiple HTML files due to the use of external stylesheets; they are also more likely to use internal stylesheets. Good finance pages contain an average of 1.1K bytes (σ = 1.2K) for scripts, while average and poor finance pages contain 1.9K (σ = 1.8K) and 883 bytes (σ = 812), respectively. Good finance pages use more distinct page titles between pages. They use more good text color combinations and search objects than average pages but fewer than poor pages; there was no difference in the number of bad and neutral text color combinations. Finally, good finance pages are more likely to contain text links without visible underlines and less likely to use serif fonts.

Health Pages. The top ten predictor measures for classifying good, average, and poor health pages include the following: Bobby priority 1 and browser errors, Weblint errors, and object count (page performance); italicized body word count (text formatting); link, link graphic,

internal, and redundant link counts (link element); and text column count (page formatting). There are several key differences in links on the good, average, and poor health pages. Specifically, good health pages contain an average of 48 links, while average and poor pages contain 33 and 28, respectively. Furthermore, good health pages contain more image, internal, and redundant links than average and poor health pages. Good health page also start text in about five different places on the page, while average and poor pages start text in three and two different places, respectively. This suggests that multiple columns are used to layout links on the page, which was confirmed by a higher link text cluster count for good health pages. Good health pages are also more likely to use scripts.

Living Pages. The top ten measures for classifying living pages include: minimum font size (text formatting); self containment and search object count (page formatting); Weblint errors (page performance); link and standard link color counts (link formatting); minimum graphic width and height (graphic formatting); animated graphic ad count (graphic element); and internal link count (link element). Good and average living pages use an average of three colors for links, while poor living pages use four. However, good living pages are less likely to use standard (browser default) link colors than average and poor living pages. They typically contain one animated graphical ad, one search form, and considerably more internal links than pages in the other classes. Good living pages are slightly less self-contained (i.e., the page can be rendered solely with the HTML code and associated images as opposed to requiring stylesheets, scripts, etc.) than average pages, but more so than poor pages; the self-containment measure on good living pages is largely due to the use of scripts.

Services Pages. The top ten predictor measures for classifying services pages include the following: minimum color use, text column count, and search object count (page formatting); Weblint errors, Bobby approved, table count, and graphic file count (page performance); spelling error count (text element); link text cluster count (text formatting); and non-underlined text links (link formatting). Good services pages are more likely to be Bobby approved than average and poor pages; this is the only case where this result was found. These pages start text in more places and use more link text clusters than average and poor pages. Good services pages are also more likely to contain links without visible underlines and use fewer graphic files than average and poor services pages.

6.10 Overall Site Quality

The goal of this section is to present an overall view of highly-rated sites that does not consider content categories. Most site-level measures require data from at least five pages for computation. Thus, the analyzed data only consists of 333 sites – 121 good sites, 118 average sites, and 94 poor sites.

To assign sites into the good, average, and poor classes, the C&RT trained on 70% of the data was used. The resulting tree contains 50 rules and has an overall accuracy of 81% (see Table 6.2 for more details). The accuracy of site predictions is lower than that of the other page-level models most likely because of a smaller training set; it is also possible that the site-level measures or prediction method need to be improved. Figure 6.8 depicts example decision tree rules for classifying sites.

ANOVAs for correctly classified sites revealed that the sites only differed significantly on the maximum depth measure. Table 6.19 shows that the median and maximum breadths crawled on the good sites are slightly higher than for average and poor sites, although not significantly

if ((Page Performance Variation is missing OR (Page Performance Variation $\leq$ 90.2)) AND (Overall Variation is not missing AND (Overall Variation $\leq$ 14.49)) AND (Link Element Variation is not missing AND (Link Element Variation $\leq$ 29.195)) AND (Link Element Variation is missing OR (Link Element Variation $>$ 20.98)))

Class = Good

This rule classifies sites as good sites if they have: 90.2% or less variation in page performance; 14.49% or less variation across all measures; and a link element variation between 20.98% and 29.2%.

---

if ((Page Performance Variation is missing OR (Page Performance Variation $\leq$ 90.2)) AND (Overall Variation is missing OR (Overall Variation $>$ 14.49)) AND (Graphic Element Variation is missing OR (Graphic Element Variation $\leq$ 185.5)) AND (Text Element Variation is missing OR (Text Element Variation $>$ 51.845)) AND (Graphic Formatting Variation is not missing AND (Graphic Formatting Variation $>$ 81.1)) AND (Median Page Breadth is missing OR (Median Page Breadth $\leq$ 10)))

Class = Average

This rule classifies sites as average sites if they have: 90.2% or less variation in page performance; variation across all measures greater than 14.49%; graphical element variation of 185.5% or less; graphic formatting variation greater than 81.1%; text element variation greater than 51.85%; and a medium breadth of ten pages or fewer at each level.

---

if ((Page Performance Variation is not missing AND (Page Performance Variation $>$ 90.2)))

Class = Poor

This rule classifies sites as poor sites if they have variations in page performance greater than 90.2%.

Figure 6.8: Example decision tree rules for predicting site classes (Good, Average, and Poor).

Measure	Mean			Std. Dev.		
	Good	Average	Poor	Good	Average	Poor
Maximum Depth	1.75	1.81	1.94	0.43	0.40	0.24
Median Breadth	7.34	7.21	7.05	4.85	3.99	4.11
Maximum Breadth	9.14	8.95	8.80	4.85	3.99	4.11

Table 6.19: Means and standard deviations for site architecture measures. These measures possibly provide some insight about the information architecture on good, average, and poor sites.

different. This suggests that the information architectures of good and average sites emphasize breadth over depth [Larson and Czerwinski 1998; Zaphiris and Mtei 1997].

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

- 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. Text formatting variation is also correlated with the maximum and median breadth at each level and the number of pages crawled on the site, which provides further support that text formatting varies among pages in good sites.

- Average sites only had one unique correlation – between graphic element and overall element variation. The overall element variation considers the amount of variation across pages on text, link, and graphic element measures, including the number of good display text words, text links, and animated images. The graphic element variation measure considers a subset of measures examined for the overall element variation measure, namely the graphic element measures. The large correlation between these two measures suggests that the overall element variation predominantly reflects variation in graphic elements as opposed to the variation in text and link elements.

- There were thirteen unique correlations between measures on poor sites. Most of the correlations suggest that variations in formatting (text, link, graphic, and page formatting variation) play a major role in the overall variation and page performance variation measures as opposed to variation in elements (text, link, and graphic element variation). Poor pages tend to have less formatting variation than average and good sites, but they have slightly more variation in page performance and element variation.

One of the limitations of the overall site quality model is that it does not consider the quality of pages in its predictions, because it is based on a completely different set of measures. Consequently, it is possible for the overall site quality model to predict that a site is consistent with good sites, but the overall page quality model predicts that all of the pages in the site are consistent with poor pages. Recall that the assumption throughout this chapter is that Webby judges' ratings apply to the site as a whole as well as to all of the pages within the site. Hence, it is not possible to incorporate page quality into the site quality model at this time. To remedy this situation, the median computed over predictions for individual pages in the site is reported

Content Category	Good	Average	Poor	Total
Community	34	22	15	71
Education	29	30	24	83
Finance	15	6	14	35
Health	12	27	17	56
Living	20	17	10	47
Services	11	16	14	41
Total	121	118	94	333

Table 6.20: Number of sites used to develop the content category quality models.

Content Category	Sample Size	Classification Accuracy		
		Good	Average	Poor
Community	71	88.2%	59.1%	66.7%
Education	83	79.3%	80.0%	66.7%
Finance	35	73.3%	83.3%	71.4%
Health	56	50.0%	81.5%	82.4%
Living	47	90.0%	76.5%	60.0%
Services	41	81.8%	93.8%	35.7%
Content Category Average		77.1%	79.0%	63.8%

Table 6.21: Classification accuracy for predicting good, average, and poor sites within content categories.

by the Analysis Tool; the predictions are generated by the overall page quality model discussed in Section 6.6. The median page-level prediction can be considered in conjunction with the site quality prediction in assessing the quality of a site.

6.11 Content Category Quality (Sites)

The goal of this section is to present an overall view of highly-rated sites that considers content categories; Table 6.20 summarizes the analyzed data. The C&RT was used to develop models for classifying the 333 sites into the good, average, and poor classes within the six content categories. Table 6.21 summarizes the classification accuracy of the decision tree models; accuracy for predicting poor sites is lower in most cases due to 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. The analysis suggests that a minimum of 35 sites per class and content category is needed to improve accuracy.

Similarly to the overall site quality model, the Analysis Tool reports the median computed over predictions for individual pages in the site as a way to incorporate page quality into assessing site quality within each content category. The predictions are generated by the content category models for pages; these models are discussed in Section 6.9.

6.12 Summary

This chapter presented several models for predicting expert ratings of pages and sites and hopefully for assessing Web interface quality as well. Page-level models were developed for the six content categories and five page types in addition to a model that classifies pages across

content categories and page types. Page-level models were also developed for content category and page type combinations, although these models were not discussed in detail; this model building effort demonstrated that it was possible to develop highly-accurate models, provided there were at least 60 pages for a content category and page type combination. Similarly, site-level models were developed across content categories as well as within content categories. Due to a smaller sample of site-level measures, the site-level models were not as accurate as page-level models.

Several key correlations were highlighted by the page-level models, including the use of an accent color on good pages, the use of fonts smaller than 9 pt for copyright and footer text on good pages, and the use of italicized body text on poor pages. The page type models showed that the measures found to be important for predictions were relevant to the functional style of pages. For example, it was found that good form pages use more interactive objects than average and poor form pages. Similarly, it was found that good link pages use more links than average and poor link pages. Overall, the key predictor measures varied across the models suggesting that an exhaustive set of page-level measures, such as the ones developed, is necessary for accurate predictions. The analysis also suggests that a broader set of site-level measures needs to be developed to improve predictions. The maximum crawling depth was the only key predictor for good sites; this measure in conjunction with the breadth measures suggest that good sites emphasize breadth over depth, which has been suggested in the literature.

Although some characteristics of pages and sites were presented for each model, more work needs to be done to better understand the design decisions encapsulated in the developed profiles. This is especially important for future work on supporting automated critique of Web interfaces. Furthermore, the efficacy of the developed profiles needs to be established via user feedback; this will be addressed in the remaining chapters. Specifically, Chapter 7 suggests that there is a relationship between expert ratings and usability ratings, Chapter 8 demonstrates that the profiles can be used to assess and improve the quality of Web sites, Chapter 9 shows that users prefer pages and sites modified based on the profiles over the original ones, and Chapter 10 demonstrates that the profiles can be used to examine established Web design guidelines.