Chapter 5

Web Interface Measures

5.1 Introduction

There is an abundance of design recommendations, recipes, and guidelines for building usable Web sites [Flanders and Willis 1998; Fleming 1998; Nielsen 1998c; Nielsen 1999b; Nielsen 2000; Rosenfeld and Morville 1998; Sano 1996; Schriver 1997; Shedroff 1999; Shneiderman 1997; Spool *et al.* 1999]. These guidelines address a broad range of Web site and page features, from the amount of content on a page to the breadth and depth of pages in the site. However, there is little consistency and overlap among them [Ratner *et al.* 1996] making it difficult to know which guidelines to adhere to. Furthermore, there is a wide gap between a heuristic such as "make the interface consistent" and the operationalization of this advice. Finally, most recommendations have not been empirically validated.

This chapter presents a set of 157 page-level and site-level measures based on an extensive survey of design recommendations from recognized experts and usability studies. The intent is to first quantify features discussed in the literature and to then determine their importance in producing highly-rated designs. Statistical models developed in Chapter 6 should facilitate the development of concrete, quantitative guidelines for improving Web interfaces; Chapter 10 demonstrates this for a subset of design guidelines.

This chapter begins with a view of Web interface structure and a summary of the 157 measures. The summary is followed by detailed discussions of all quantitative measures. Appendix C provides instructions for accessing an interactive appendix with visual depictions of all of the measures. Chapter 6 explores the use of these measures in developing profiles of quality Web interfaces.

5.2 Web Interface Structure

A Web interface is a mix of many elements (text, links, and graphics), formatting of these elements, and other aspects that affect the overall interface quality. Web interface design entails a complex set of activities for addressing these diverse aspects. To gain insight into Web design practices, Newman and Landay [2000] conducted an ethnographic study wherein they observed and interviewed eleven professional Web designers. One important finding was that most designers viewed Web interface design as being comprised of three components – information design, navigation design, and graphic design – as depicted in the Venn diagram in Figure 5.1. Information design focuses on determining an information structure (i.e., identifying and grouping content items) and developing category labels to reflect the information structure. Navigation design fo-

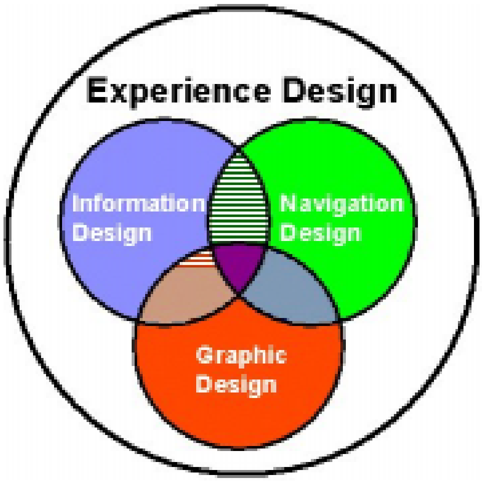

Figure 5.1: Overview of Web interface design. The Venn diagram is a modified version of the one in [Newman and Landay 2000]; it is reprinted with permission of the authors.

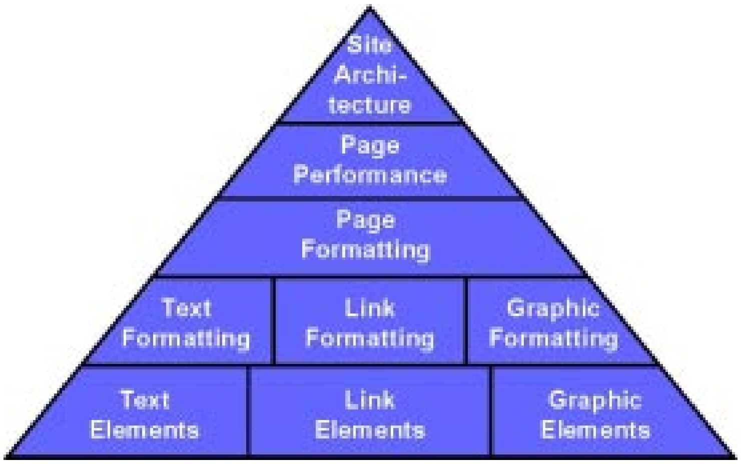

Figure 5.2: Aspects associated with Web interface structure.

cuses on developing navigation mechanisms (e.g., navigation bars and links) to facilitate interaction with the information structure. Finally, graphic design focuses on visual presentation and layout. All of these design components affect the overall quality of the Web interface. The Web design literature also discusses a larger, overarching aspect – experience design [Creative Good 1999; Shedroff 2001] – the outer circle of Figure 5.1. Experience design encompasses information, navigation, and graphic design. However, it also encompasses other aspects that affect the user experience, such as download time, the presence of graphical ads, popup windows, etc.

Information, navigation, graphic, and experience design can be further refined into the aspects depicted in Figure 5.2. The figure shows that text, link, and graphic elements are the building blocks of Web interfaces; all other aspects are based on these. The next level of Figure 5.2 addresses formatting of these building blocks, while the subsequent level addresses page-level formatting. The top two levels address the performance of pages and the architecture of sites, including the consistency, breadth, and depth of pages. The bottom three levels of Figure 5.2 are associated with information, navigation, and graphic design activities, while the top two levels – Page Performance and Site Architecture – are associated with experience design.

Metric	Description
Word Count	Total words on a page
Body Text %	Percentage of words that are body vs. display (i.e., headings) text
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 and images
Graphic %	Percentage of page bytes for images
Graphics Count	Total images on a page
Color Count	Total colors used
Font Count	Total font face, size, bolding, and italics combinations
Reading Complexity	Gunning Fog Index (ratios of words, sentences and words with more than 3 syllables)

Table 5.1: Web page metrics used in prior studies [Ivory *et al.* 2000; Ivory *et al.* 2001]. All measures, except Reading Complexity, were examined in both studies; Reading Complexity was not included in the second study.

Aspects presented in Figures 5.1 and 5.2 are used to organize discussions throughout this chapter.

5.3 Summary of Web Interface Measures

An extensive survey of Web design literature, including texts written by recognized experts (e.g., [Fleming 1998; Nielsen 2000; Sano 1996; Spool *et al.* 1999]) and published user studies (e.g., [Bernard and Mills 2000; Bernard *et al.* 2001; Boyarski *et al.* 1998; Larson and Czerwinski 1998]) was conducted to identify key features that impact the quality and usability of Web interfaces [Ivory *et al.* 2000]. HTML style guides were not consulted because they have been shown to be highly inconsistent [Ratner *et al.* 1996]. Sixty two features were identified from the literature, including: the amount of text on a page, fonts, colors, consistency of page layout in the site, use of frames, and others. As part of the analysis methodology, 157 quantitative measures were then developed to assess many of the 62 features.

Previous work by the author showed that twelve Web interface measures – Word Count, Body Text Percentage, Emphasized Body Text Percentage, Text Positioning Count, Text Cluster Count, Link Count, Page Size, Graphic Percentage, Graphics Count, Color Count, Font Count, and Reading Complexity – could be used to accurately distinguish pages from highly-rated sites [Ivory *et al.* 2000; Ivory *et al.* 2001]. Table 5.1 describes these measures. For this dissertation, 157 quantitative measures (including nine of the original twelve measures and variations of the other three) were developed to further assess aspects of the information, navigation, graphic, and experience design of Web interfaces. These measures provide some support for assessing 56 of the 62 features (90%) identified as impacting usability in the Web design literature. Measures developed in previous studies assessed less than 50% of these 62 features.

Guidelines provided in Section 3.4.1 for determining performance metrics were considered in developing the 157 measures. Specifically, a subset of measures with the following characteristics were implemented.

- **Low Variability:** measures are not a ratio of two or more variables; there is one exception to this rule – the average number of words in link text – which was developed to assess a feature reported in the literature.
- **Nonredundancy:** two measures do not convey essentially the same information.
- **Completeness:** measures reflect all aspects of the Web interface (i.e., information, navigation, graphic, and experience design).

A sample of fourteen Web pages with widely differing characteristics was used to validate the implemented measures. The actual value of each measure was manually computed and then used to determine the accuracy of computed results. For each page and each measure, the number of accurate hits and misses as well as the number of false positives and negatives were determined as described below.

Accurate Positive: an element is counted as it should have been counted (e.g., an actual word is counted by the tool as a word).

Accurate Negative: an element is not counted and it should not have been counted (e.g., an actual display word is not counted as a body word). This is relevant for discriminating measures that can label an object more than one way, such as deciding whether a word is a heading, body, or text link word.

False Positive: an element is counted and it should not have been counted (e.g., an actual body word is counted as a display word).

False Negative: an element is not counted and it should have been counted (e.g., an actual body word is not counted as a body word).

False positives and false negatives typically occur for discriminating measures (i.e., ones that entail deciding between two or more options). After counting the four types of hits and misses for a measure on a page, the following accuracies were then computed for the measure.

Hit Accuracy: the ratio of the number of accurate positives to the sum of the number of accurate positives and false negatives. False positives are not included in this computation, since they represent inaccurate hits.

Miss Accuracy: the ratio of the number of accurate negatives to the sum of the number of accurate negatives and false positives. False negatives are not included in this computation, since they represent inaccurate misses.

The average hit and miss accuracies for a measure is then the average hit and miss accuracies across all sample pages. Finally, the overall accuracy is computed over the average hit and miss accuracies. Metric tables in this chapter report the average hit, average miss, and overall accuracy for each measure across the 14-page sample. With a few exceptions, all of the measures are highly accurate (>84% overall accuracy across the sample). The least accurate measures – text positioning count (number of changes in text alignment from flush left) and text and link text cluster counts (areas highlighted with color, rules, lists, etc.) – require image processing as discussed later.

For measures developed in this chapter, the HTML Parser and Browser Emulator (see Chapter 4) was configured to simulate the Netscape Navigator 4.75 browser with default settings (fonts, link colors, menu bar, address bar, toolbar, and status line). The window size was fixed at 800 x 600 pixels.

The remainder of this chapter presents the 62 features derived from the literature survey, the measures developed to assess these features, and a short discussion of specific guidance from the literature when available. The discussion reflects the hierarchy presented in Figure 5.2. In each section, summary tables depict the measures, the Web interface aspects (information, navigation, graphic, and experience design) assessed by the measures, and their accuracy as described above. Two classes of measures were developed: discriminating (deciding between two or more options) and non-discriminating (a count). For discriminating measures, hit and miss accuracies are reported; they are not reported for non-discriminating measures. The overall accuracy is reported for both discriminating and non-discriminating measures.

5.4 Text Element Measures

Tables 5.2, 5.3, and 5.4 summarize 31 text element measures derived from the literature survey and discussed in this section. These measures provide insight about the following Web page features.

1. How much text is on the page?
2. What kind of text is on the page?
3. How good is the text on the page? Good in this context refers to whether or not the text contains an abundance of common words typically referred to as stop words. Good words are not stop words.
4. How complex is the text on the page? Complexity refers to the reading level required to understand the text as determined by the Gunning Fog Index [Gunning 1973].

5.4.1 Text Element Measures: Page Text

The text or visible (legible) words on a Web page has been discussed extensively in the literature [Flanders and Willis 1998; Landesman and Schroeder 2000; Nielsen 2000; Schriver 1997; Stein 1997] and is considered a major component of the information design. An analysis of experts' ratings of Web sites submitted for the Webby Awards 2000 revealed that content was by far the best predictor of ratings [Sinha *et al.* 2001]. The literature includes the following heuristics.

- Users prefer pages with more content as opposed to breaking content over multiple pages [Landesman and Schroeder 2000].
- Keep text short; use 50% less text than in print publications [Nielsen 2000].
- Break text up into smaller units on multiple pages [Flanders and Willis 1998; Nielsen 2000].

As is often found in the literature, the first guideline contradicts the other two. Furthermore, there is no concrete guidance on how much text is enough or too much. Thus, a Word Count measure was developed to assess the amount of text on the page. The presence of invisible

		Aspects Assessed				**Accuracy**		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
How much text is on the page?								
Word Count	Total visible words	√				–	–	99.8%
Page Title Word Count	Number of words in the page's title (max of 64 chars)	√	√			–	–	100.0%
Overall Page Title Word Count	Number of words in the page's title (no char max)	√	√			–	–	100.0%
Invisible Word Count	Number of invisible words				√	–	–	100.0%
Meta Tag Word Count	Number of words in meta tags				√	–	–	100.0%
What kind of text is on the page?								
Body Word Count	Words that are body text (i.e., not headings or links)	√				99.9%	99.5%	99.8%
Display Word Count	Words that are display text (i.e., headings that are not links)	√				97.5%	100.0%	98.7%
Display Link Word Count	Words that are both link text and headings	√	√			75.5%	100.0%	87.7%
Link Word Count	Words that are link text and are not headings	√	√			99.7%	99.5%	99.6%
Average Link Words	Average number of words in link text		√			–	–	100.0%
Graphic Word Count	Number of words from <img> alt attributes	√				–	–	100.0%
Ad Word Count	Number of words possibly indicating ads ('advertisement' or 'sponsor')	√				100%	100%	100%
Exclamation Point Count	Number of exclamation points	√				–	–	100.0%
Spelling Error Count	Number of misspelled words				√	100.0%	99.9%	100.0%

Table 5.2: Summary of text element measures (Table 1 of 3). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

		Aspects Assessed				**Accuracy**		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
How good is the text on the page?								
Good Word Count	Total good visible words (i.e., not stop words)	√				100.0%	100.0%	100.0%
Good Body Word Count	Good body text words	√				100.0%	100.0%	100.0%
Good Display Word Count	Good display text words	√				100.0%	100.0%	100.0%
Good Display Link Word Count	Good combined display and link text words (i.e., not stop words or 'click')	√	√			100.0%	100.0%	100.0%
Good Link Word Count	Good link text words	√	√			100.0%	99.7%	99.8%
Average Good Link Words	Average number of good link text words		√			100.0%	100.0%	100.0%
Good Graphic Word Count	Number of good words from <img> alt attributes	√				100.0%	100.0%	100.0%
Good Page Title Word Count	Number of good page title words (max of 64 chars)	√	√			100.0%	100.0%	100.0%
Overall Good Page Title Word Count	Total number of good page title words (no char max)	√	√			100.0%	100.0%	100.0%
Good Meta Tag Word Count	Number of good meta tag words				√	100.0%	100.0%	100.0%

Table 5.3: Summary of text element measures (Table 2 of 3). Good in this context refers to the use of words that are not stop words or the word 'click' in the case of text links. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

Measure	**Description**	**Aspects Assessed**				**Accuracy**		
		ID	ND	GD	ED	Hit	Miss	Avg.
How complex is the text on the page?								
Reading Complexity	Gunning Fog Index computed over prose	√				–	–	97.6%
Overall Reading Complexity	Gunning Fog Index computed over all text	√				–	–	97.9%
Fog Word Count	Number of prose words (for Reading Complexity)	√				99.5%	100.0%	99.7%
Fog Big Word Count	Number of big prose words (for Reading Complexity)	√				94.9%	97.6%	96.2%
Overall Fog Big Word Count	Number of big words (for Overall Reading Complexity)	√				96.5%	98.5%	97.5%
Fog Sentence Count	Number of sentences (for Reading Complexity)	√				–	–	98.6%
Overall Fog Sentence Count	Number of sentences (for Overall Reading Complexity)	√				–	–	98.6%

Table 5.4: Summary of text element measures (Table 3 of 3). Complexity in this context refers to the reading level required to understand the text; this is determined by the Gunning Fog Index [Gunning 1973]. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

words (i.e., words formatted with the same foreground and background colors, thus making them illegible) is also measured, since such words may indicate spamming (techniques used to influence search engine ranking). These counts only include words in the static HTML and do not include words embedded in images, applets, and other objects. Image processing techniques are required to count words in these cases.

Word Count was examined in both of the prior metric studies [Ivory *et al.* 2000; Ivory *et al.* 2001] and was shown to be a key measure for distinguishing highly-rated pages in the second study. Furthermore, dividing the sample into three groups – low word count (average of 66 words), medium word count (average of 230 words), and high word count (average of 820 words) – revealed several key differences between highly-rated and poorly-rated Web pages.

5.4.2 Text Element Measures: Page Title

A page's title is also an important element of the information design [Berners-Lee 1995; Flanders and Willis 1998; Nielsen 2000]. Specific guidance includes the following.

- Use no more than 64 characters [Berners-Lee 1995].
- Use 2–6 words (40–60 characters) [Flanders and Willis 1998].
- Use different page titles for each page [Nielsen 2000].

To assess the number of words used in page titles, two measures – Page Title Word Count and Overall Page Title Word Count – were developed. The difference between these two measures is a restriction of 64 characters on the Page Title Word Count but not on the Overall Page Title Word Count. Prior experience with the Page Title Word Count revealed that extremely long page titles (e.g., >200 words) were sometimes used, possibly as a spamming technique. However, long titles are also used legitimately to make search engine results more readable; further analysis of good and poor Web pages needs to be conducted to know for certain that spamming is used. Nonetheless, only a portion of these words are actually visible in a Web browser; thus, the guideline from the HTML 2.0 specification (64 maximum characters) [Berners-Lee 1995] is used to determine the number of visible page title words. Deviations between these two measures may indicate spamming. To assess whether page titles varied between pages within a site, scent quality and page title consistency measures (discussed in Sections 5.12.7 and 5.13.1) were implemented.

5.4.3 Text Element Measures: Page Abstract

Although a page's abstract (i.e., meta tags used for search engines) is not a visible page feature, it does affect the experience design, especially when searching is used to locate information. Nielsen [2000] suggests that Web designers use meta tags with less than 150–200 characters on the page. To assess the use of meta tags, a Meta Tag Word Count was developed. It includes words used in both the description and keyword attributes of the meta tag.

5.4.4 Text Element Measures: Body Text

The interplay of display (i.e., headings) and body text affects users ability to scan the information on a page [Nielsen 1997; Schriver 1997]. Ideally, meaningful headings and sub-headings are used to help users locate specific information [Nielsen 1997]. Although no specific guidance was provided with respect to the amount of body vs. display text in a Web page, a Body Word Count and several Display Word Count measures were developed to assess the balance between these two features.

Detecting headings within a Web page is not a straight-forward activity. Prior experience revealed that only a small fraction of Web pages actually use the HTML header tags (<h1>, <h2>, <h3>, etc.) to format headings; the majority of pages use font and stylesheet tags. Hence, several heuristics were developed to detect headings; heuristics are based on the sample of fourteen Web pages used throughout metrics development. For example, if text is bolded, not embedded in other non-bolded text, and is not in a text link, then it is considered to be a heading. Another heuristic examines whether text is emphasized in some other manner (e.g., italicized or colored), is not embedded in other non-emphasized text, is formatted with a font size greater than twelve points, and is not in a text link to determine whether the text is a heading. The fixed font size of 12 points is not ideal, because it would be more appropriate to use font sizes used in the document as a baseline. The heuristics detected headings in the Web page sample with 98.7% overall. Similar heuristics were developed for detecting link headings with 87.7% overall accuracy; the hit accuracy was only 75.5%, while the miss accuracy was 100%. Using image processing techniques on an image of a Web page may result in more accurate detection of link headings.

The two prior metric studies used a Body Text Percentage measure (ratio of body text words to all words on the page). Both studies showed this measure to be important for distinguishing highly-rated pages. Despite its importance, this measure was abandoned because it directly violates the low variability (not a ratio of two or more measures) guidance for performance metrics. Analysis of the Word and Body Word Counts can provide the same insight as the Body Text Percentage measure.

5.4.5 Text Element Measures: Link Text

Words used in text links affect both the information and navigation design. Although the total number of link text words is measured by the Rating Game tool [Stein 1997], no specific guidance was found in the literature with respect to the appropriate amount of link text on a page. Two link text measures were developed – Link Word Count and Display Link Word Count. The latter measure reports the number of link text words that are also headings.

5.4.6 Text Element Measures: Link Text Length

The previous section discussed the total number of link text words, but this section focuses on the number of words in a text link. The literature states that the number of words used in each text link affects the information design and the navigation design in particular [Chi *et al.* 2000; Nielsen 2000; Sawyer and Schroeder 2000; Spool *et al.* 1999]. Specific guidance includes the following.

- Use 2–4 words in text links [Nielsen 2000].
- Use links with 7–12 "useful" words [Sawyer and Schroeder 2000].

These two guidelines are obviously contradictory. Hence, an Average Link Words measure was developed to assess the length of link text. This measure is the ratio of the total number of text link words (i.e., Link and Display Link Word Counts) to the number of text links. This ratio violates the low variability guidance for performance metrics; however, the measure was implemented because it allows for direct assessment of a feature with contradictory guidance in the literature. The second guideline includes a qualifier – "useful" words – this is assessed to some degree with a complementary measure, Average Good Link Words, which is discussed in Section 5.4.12.

5.4.7 Text Element Measures: Content Percentage

Nielsen [2000] suggests that the portion of the Web page devoted to content should represent 50–80% of the page. The content percentage can be computed by highlighting the portion of the page devoted to content and then determining what percentage of the page it represents. It is difficult to compute this percentage as prescribed without using image processing techniques. Instead, the following indirect measures were developed: the total number of words (Section 5.4.1); and the total number of words used for text links (Section 5.4.5).

5.4.8 Text Element Measures: Navigation Percentage

Similarly to the content percentage, Nielsen [2000] suggests that the portion of the Web page devoted to navigation should represent 20% of the page; this percentage should be higher for home and intermediate pages. The navigation percentage can be computed by highlighting the portion of the page devoted to navigation and then determining what percentage of the page it represents. It is difficult to compute this percentage as prescribed without using image processing techniques. Several indirect measures were developed, including: the total number of links of various kinds (e.g., text and graphic; Section 5.5.1); and the total number of words used for text links (Section 5.4.5).

5.4.9 Text Element Measures: Exclamation Points

The use of exclamation points in the page text has been discussed in the literature [Flanders and Willis 1998; Stein 1997]. The consensus is to minimize the use of exclamation points, since they are equivalent to blinking text [Flanders and Willis 1998]. The literature does not provide guidance on an acceptable number of exclamation points. Hence, an Exclamation Point Count was developed; this measure includes exclamation points in body, display, and link text.

5.4.10 Text Element Measures: Typographical Errors

As part of the Web Credibility Project at Stanford University [Kim and Fogg 1999; Fogg *et al.* 2000], a survey of over 1,400 Web users was conducted to identify aspects that impact the credibility of Web interfaces. Survey results showed that an amateurism factor, which includes the presence of spelling and other typographical errors, was negatively correlated with credibility [Fogg *et al.* 2000]. A follow-up controlled study further established that the presence of typographical errors decreased credibility [Kim and Fogg 1999]. The obvious guidance from this research is to avoid typographical errors. Flanders and Willis [1998] also suggest that Web designers avoid making typographical mistakes on pages.

The Metrics Computation Tool assesses one type of typographical error – spelling errors (English only). An extensive dictionary of English words, computer, Internet, medical terms, acronyms, and abbreviations, including the MRC psycholinguistic database [Coltheart 2001], is used for checking spelling errors. One limitation of the spelling error measure is that it ignores capitalized words, since they may be proper nouns typically not found in dictionaries. Hence, the number of spelling errors may be underreported. Another limitation is that it may count jargon words as spelling errors; thus, inflating the number of spelling errors. Other typographical errors, such as misplaced punctuation, are not assessed with quantitative measures, since more sophisticated computational linguistics is required.

5.4.11 Text Element Measures: Readability

The literature survey revealed numerous discussions of the readability or required reading level of text [Berners-Lee 1995; Flanders and Willis 1998; Gunning 1973; Nielsen 2000; Schriver 1997; Spool *et al.* 1999]. Spool *et al.* [1999] determined that the Gunning Fog Index (GFI) [Gunning 1973] was the only readability measure correlated with Web interfaces. To compute this index, a passage containing at least 100 words needs to be selected from the page. Then, the number of words (Fog Word Count), the number of words with more than two syllables (Fog Big Word Count), and the number of sentences (Fog Sentence Count) needs to be computed. Equation 5.1 demonstrates how to compute the Gunning Fog Index from these measures.

$$GFI \;=\; \left(\frac{Fog\ Word\ Count}{Fog\ Sentence\ Count} + \frac{Fog\ Big\ Word\ Count}{Fog\ Word\ Count} * 100.0\right) * 0.4 \tag{5.1}$$

This measure was originally developed for printed documents; however, Spool *et al.* [1999] discovered that this measure correlated with the scannability of Web pages. Specific guidance for the GFI in both the print and Web domains is provided below. As can be expected, guidance is contradictory in the two domains.

- A lower Gunning Fog Index of 7–8 is ideal for printed documents [Gunning 1973]; only a 7th or 8th grade education is required to read them.
- A higher Gunning Fog Index (∼15.3) is ideal for Web pages; pages that are reported to require a college education to read them facilitate scanning [Spool *et al.* 1999].

Two reading complexity measures were developed – the Reading Complexity computed over prose text (sentences) and the Overall Reading Complexity computed over all of the text, including links and bulleted lists. Prior experience with computing the Gunning Fog Index over all of the text suggested that this measure was not truly a measure of reading complexity, rather it measured the degree to which text is broken up to facilitate page scanning [Ivory *et al.* 2000; Spool *et al.* 1999]. Hence, the reading complexity measure computed over just the prose text was developed to be more consistent with the intended use of the Gunning Fog Index.

All of the measures used in computing the reading complexity measures, such as the number of sentences and big words, are reported by the Metrics Computation Tool. The number of big words is determined by first looking up words in the MRC psycholinguistic database [Coltheart 2001], which contains the number of syllables for over 100,000 words. If the word is not in the database, then the algorithm used by the UNIX style program (counting consonant vowel pairs) [Cherry and Vesterman 1981] is used.

A previous study of the reading complexity computed over all of the page text showed this measure to be important for distinguishing unrated (from sites that have not been identified by reputable sources as exhibiting high quality) and rated (from sites that have been identified by reputable sources as exhibiting high quality) pages [Ivory *et al.* 2000]; this study is discussed briefly in Chapter 6. The reading complexity of rated pages was consistent with the second guideline presented above. The study also showed that reading complexity values above 15.8 were not associated with rated pages; thus, there appears to be a threshold above which higher reading complexity (i.e., reported as more difficult to read) may impede scanning.

5.4.12 Text Element Measures: Information Quality

The literature extensively discusses ways to improve the quality of information (e.g., content appropriateness, relevance, language, tone, and freshness) [Flanders and Willis 1998; Nielsen 2000; Rosenfeld and Morville 1998; Schriver 1997; Spool *et al.* 1999]. Specific guidance includes the following.

- Support efficient, easy first-time use (i.e., logical grouping of content and organization) [Rosenfeld and Morville 1998].
- Update content often [Flanders and Willis 1998; Nielsen 2000].

It is extremely difficult to assess the quality of information without user input. Hence, the implemented measures provide only limited coverage of the broad spectrum of aspects that influence information quality. Specifically, the Metrics Computation Tool computes the number of good words, link words, display words, body words, page title words, meta tag words, and so on. A word is determined to be good if it is not a stop word; an extensive list of 525 common English words is used for this assessment. For determining good link words, the word 'click' is also considered as a stop word. None of the quantitative measures assess the guidelines depicted above.

5.5 Link Element Measures

Table 5.5 summarizes six link element measures derived from the literature survey and discussed in this section. (These measures are different than the link text measures presented in the previous section.) These measures provide insight about the following Web page features.

1. How many links are on the page?
2. What kind of links are on the page?

5.5.1 Link Element Measures: Links

Links are an essential element of the navigation design and are discussed extensively in the literature [Flanders and Willis 1998; Furnas 1997; Larson and Czerwinski 1998; Rosenfeld and Morville 1998; Sano 1996; Schriver 1997; Spool *et al.* 1999; Zaphiris and Mtei 1997]. Several usability studies have been conducted to provide the following guidance about the breadth (i.e., how many links are presented on a page), depth (i.e., how many levels must be traversed to find information), and other aspects of the navigation structure.

- Use moderate levels of breadth with minimal depth (e.g., two levels) in the information architecture [Larson and Czerwinski 1998].
- Minimize depth [Zaphiris and Mtei 1997].
- A large number of links impedes navigation [Spool *et al.* 1999].
- Avoid broken links [Flanders and Willis 1998; Nielsen 2000; Spool *et al.* 1999].

		Aspects Assessed				Accuracy		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
How many links are on the page?								
Link Count	Total number of links		√			–	–	99.8%
What kind of links are on the page?								
Text Link Count	Number of text links	√	√			99.7%	100.0%	99.9%
Link Graphic Count	Number of image links		√	√		100.0%	100.0%	100.0%
Page Link Count	Number of links to other sections (i.e., anchors) within the page		√			100.0%	100.0%	100.0%
Internal Link Count	Number of links that point to destination pages within the site		√			100.0%	100.0%	100.0%
Redundant Link Count	Number of links that point to the same destination page as other links on the page		√			100.0%	100.0%	100.0%

Table 5.5: Summary of link element measures. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

A Link Count was developed to assess the total number of links (graphical and text) on a page. This measure captures the breadth of links on a page, but it does not assess the depth of links. It was examined in both of the prior metric studies and found to be important for distinguishing highly-rated pages. The maximum page depth measure (discussed in Section 5.13.2) provides some insight into navigation structure. For instance, the deepest level that the crawler could traverse on the site (i.e., not find pages that hadn't been seen before) may reflect the depth of the information architecture.

Broken links have been mentioned repeatedly in the literature as a usability problem. Given that there are numerous tools for detecting broken links, such as Weblint [Bowers 1996], no measure was developed.

5.5.2 Link Element Measures: Text Links

Text links are considered to be the most important type of links in the literature [Flanders and Willis 1998; Sawyer and Schroeder 2000; Scanlon and Schroeder 2000c; Spool *et al.* 1999]. The consensus in the literature is that text rather than image links should be used [Flanders and Willis 1998; Spool *et al.* 1999]. A Text Link Count was developed to measure the number of text links on a Web page.

5.5.3 Link Element Measures: Link Graphics

The literature extensively discusses the use of graphical links [Flanders and Willis 1998; Sano 1996; Sawyer and Schroeder 2000; Scanlon and Schroeder 2000c; Spool *et al.* 1999]. Specific guidance includes the following.

- Avoid using graphical text links; they are typically ignored [Sawyer and Schroeder 2000; Spool *et al.* 1999] or may impede navigation [Scanlon and Schroeder 2000c; Spool *et al.* 1999].
- Use corresponding text links [Flanders and Willis 1998; Sano 1996].

A Link Graphic Count was developed to assess the number of links that are images. A similar measure was computed for the number of graphics that are also links and is discussed in Section 5.6.2. Although the Link Graphic Count measures the use of graphical links, there is no direct measure of whether equivalent text links is provided. However, the Redundant Link Count (discussed in Section 5.5.8) may provide some indirect insight about the use of equivalent text links.

5.5.4 Link Element Measures: Within-Page Links

The use of within-page links or links that point to other areas in the same page has been found to be problematic [Nielsen 2000; Sawyer and Schroeder 2000; Spool *et al.* 1999]. The consensus is that these types of links should be avoided, since they may be confusing [Nielsen 2000; Sawyer and Schroeder 2000; Spool *et al.* 1999]. A Page Link Count was developed to report the presence of within-page links.

5.5.5 Link Element Measures: External Links

The use of external links or links that point to other sites has also been reported as potentially problematic [Nielsen 2000; Spool *et al.* 1999]. One rationale is that users may not be aware that they left the original site [Spool *et al.* 1999]. Nielsen [1997] presents the contrary view that external links increase credibility based on a study of nineteen users reading Web pages.

Nielsen [2000] suggests that Web designers use a different link color scheme to signify external links and to better inform users that they are leaving the site.

An Internal Link Count was developed to measure the use of internal as opposed to external links. This measure considers links that point to the same domain (e.g., www1.cnet.com and www2.cnet.com) as being internal. The difference between the Link and Internal Link Counts lies in the number of external links. No measure was developed to determine if different color schemes are used for internal and external links.

5.5.6 Link Element Measures: Embedded Links

Embedding links within text on the page has been discussed in the literature [Rosenfeld and Morville 1998; Spool *et al.* 1999]. The consensus is that Web designers should avoid surrounding links with text, since they are difficult to scan [Rosenfeld and Morville 1998; Spool *et al.* 1999]. No measure was developed to detect the presence of embedded links, since image processing is required for accurate detection.

5.5.7 Link Element Measures: Wrapped Links

Spool *et al.* [1999] suggests that wrapped links or links spanning multiple lines should be avoided. During usability studies, users interpreted each line of a wrapped link as being a separate link versus all lines comprising a single link. No measure was developed to detect the presence of wrapped links, since it requires an accurate simulation of browser rendering. Use of a browser that can more accurately represent the layout of page elements, such as Mozilla or Opera, would facilitate detection of wrapped links.

5.5.8 Link Element Measures: Redundant Links

The surveyed literature espouses the value of redundant or multiple links to the same content [Sawyer and Schroeder 2000; Spool *et al.* 1999; Spool *et al.* 2000], while one study of different types of links (e.g., neighborhood, parent, index, etc.) found redundant links to confuse users within cyber shopping malls [Kim and Yoo 2000]. Specific guidance includes the following.

- Use multiple links to the same content with appropriate scent in each area [Spool *et al.* 2000].
- Use different forms for repeated links (e.g., text, graphical text, or image) [Sawyer and Schroeder 2000].
- Redundant links may cause confusion [Kim and Yoo 2000].

A Redundant Link Count was developed to assess the use of repeated links. No measure was implemented to detect whether different forms of links are used or whether different link text is used to label links that point to the same location.

5.5.9 Link Element Measures: Navigation Quality

Much has been said about many aspects of the navigation structure, such as the clarity of links, the use of scent (hints about the contents on a linked page), the relevance of links, and the use of effective navigation schemes [Chi *et al.* 2000; Fleming 1998; Furnas 1997; Larson and Czerwinski 1998; Miller and Remington 2000; Nielsen 2000; Rosenfeld and Morville 1998; Sawyer *et al.* 2000; Spool *et al.* 1999; Spool *et al.* 2000]. Specific guidance on these aspects includes the following.

- Use clear headings with related links (i.e., link clustering) to enhance scent [Spool *et al.* 2000].
- Expose multiple levels of the information architecture (i.e., link clustering with headings) [Sawyer *et al.* 2000].
- Effective navigation requires small pages (views), few clicks between pages, and strong scent [Furnas 1997].
- Weak scent (i.e., ambiguous link text) impedes navigation [Chi *et al.* 2000; Miller and Remington 2000; Spool *et al.* 1999; Spool *et al.* 2000].
- Similar link text across links impedes navigation [Spool *et al.* 1999].
- Do not use a shell strategy (i.e., fixed navigation bar content); use navigation bars at the top and bottom of pages vs. down the sides [Spool *et al.* 1999].
- Avoid using 'Click Here' for link text [Nielsen 2000].
- Support multiple modes of finding information (i.e., directed searching and browsing) [Fleming 1998; Rosenfeld and Morville 1998].
- The navigation scheme should be easy to learn, consistent, efficient, and relevant to the type of site (i.e., entertainment and information sites use different navigation schemes) [Fleming 1998].
- Use breadcrumbs (i.e., displaying a navigation trail) vs. long navigation bars [Nielsen 2000].

Many of these suggestions are difficult to measure in an automated manner. However, some guidelines, such as the number of good link words (excludes the word 'click'; Section 5.4.12), support for searching (Section 5.10.7), use of link clustering (Section 5.7.7), and the consistency of link elements and formatting (Section 5.13.1), are measured. Several site architecture measures, such as the maximum page depth and breadth (as determined by the crawling depth and breadth; Section 5.13.2), may also provide some insight about the navigation structure.

5.6 Graphic Element Measures

Table 5.6 summarizes six graphic element measures developed and discussed in this section. These measures assess the following features of Web pages.

1. How many graphics are on the page?
2. What kind of graphics are on the page?

5.6.1 Graphic Element Measures: Graphics

Images are a key element of the graphic design, and image use is discussed extensively in the literature [Ambühler and Lindenmeyer 1999; Flanders and Willis 1998; Nielsen 2000; Sano 1996; Schriver 1997; Spool *et al.* 1999; Stein 1997]. Scanlon and Schroeder [2000c] identified and provided guidance for the following four categories of graphics.

- **Content Graphics** - provide content (i.e., see vs. read); users typically do not complain about the download speed of content graphics.

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
How many graphics are on the page?								
Graphic Count	Total number of images			√		–	–	100.0%
What kind of graphics are on the page?								
Redundant Graphic Count	Number of images that point to the same image files as other images			√		100.0%	100.0%	100.0%
Graphic Link Count	Number of images that are links		√	√		100.0%	100.0%	100.0%
Animated Graphic Count	Number of animated images			√		100.0%	100.0%	100.0%
Graphic Ad Count	Number of images that possibly indicate ads			√		97.9%	96.1%	97.0%
Animated Graphic Ad Count	Number of animated images that possibly indicate ads			√		97.2%	99.9%	98.6%

Table 5.6: Summary of graphic element measures. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

- **Navigation Graphics** - help users navigate; these should typically be avoided.
- **Organizer Graphics** - bullets, rules, etc. that direct users' attention on the page; these are typically ignored by users.
- **Ornamental Graphics** - logos and other images that ornament the page; these images are typically the least effective and most costly in terms of download speed.

In some cases, one graphic may serve multiple roles. Furthermore, the role of a graphic may not be apparent from looking at it. Regardless of the role of graphics, the consensus in the literature is that the number of images needs to be minimized to improve download speed. Nielsen [2000] also suggests that text rendered as images (one type of content graphic) should be eliminated except for image captions. A Graphic Count measure was developed to quantify the use of images on a Web page. Text rendered as images is not assessed, since it would require image processing. Another limitation of this measure is that it includes images that may not be visible on the page, such as spacer images; image processing techniques are required to accurately report the number of visible images on the page. The use of navigation graphics is assessed with the Graphic Links measure discussed in the following section. Although the use of organizer and ornamental graphics is not directly measured, the developed Redundant Graphic Count (repeated use of image files) may provide an indirect measurement of these types of images.

The Graphic Count measure was examined in both prior metric studies. The first study determined that rated pages contained more images than pages that were not rated; inspection of a random sample of pages revealed that this higher number of graphics was attributable to organizer graphics. The second study found that in all cases, highly-rated pages contained fewer images. (Results from the second study are more definitive because they are based on analysis of expert ratings rather than rated and unrated sites). Even though rated pages contained more images in the first study, graphic count was not a key predictor of rated or highly-rated pages in either study.

5.6.2 Graphic Element Measures: Graphical Links

The use of images as links is also discussed in the literature [Flanders and Willis 1998; Sawyer and Schroeder 2000; Scanlon and Schroeder 2000c; Spool *et al.* 1999]. Specific guidance about graphical links is below. This guidance was also reported for the Link Graphic Count discussed in Section 5.5.3. The Graphic Link Count was developed to quantify the use of images as links. This measure varies from the Link Graphic Count in Section 5.5.3 when image maps are used on pages, since every area of an image map is counted as a separate image link.

- Avoid using graphical text links; they are typically ignored [Sawyer and Schroeder 2000; Spool *et al.* 1999] or may impede navigation [Scanlon and Schroeder 2000c; Spool *et al.* 1999].
- Use corresponding text links [Flanders and Willis 1998; Sano 1996].

5.6.3 Graphic Element Measures: Graphical Ads

Several literature sources discuss the presence of graphical ads on Web pages [Kim and Fogg 1999; Klee and Schroeder 2000; Nielsen 2000]. Specific guidance includes the following.

- Ads affect the user experience; integrate ads with content [Klee and Schroeder 2000].
- Usability dictates that ads should be eliminated [Nielsen 2000].

- Ads increase credibility [Kim and Fogg 1999].

The guidelines are obviously contradictory. Kim and Fogg [1999] describe credibility as a "high level of perceived trustworthiness and expertise," which appears to be related to usability, although this link has yet to be established. A controlled study wherein 38 users rated Web pages (with and without graphical ads) on credibility showed that pages with graphical ads were rated as more credible than those without graphical ads. A Graphic Ad Count was developed to gain more insight about the use of ads on Web pages. This measure uses a number of heuristics (e.g., whether the image is a link to a known advertising agency or contains words indicative of advertisement in the URL) to detect the presence of ads with 97% overall accuracy.

5.6.4 Graphic Element Measures: Animation

The use of animated images and scrolling text is often debated in the literature [Flanders and Willis 1998; Nielsen 2000; Spool *et al.* 1999]. Specific guidance on animation includes the following.

- Minimize animated graphics [Flanders and Willis 1998].
- Avoid using animation unless it is appropriate (e.g., showing transitions over time) [Nielsen 2000].
- Animation is irritating to users; it impedes scanning [Spool *et al.* 1999].

None of these guidelines provide concrete guidance about how much animation is too much. Hence, an Animated Graphic Count was developed to quantify the use of animated images. Similarly, an Animated Graphic Ad Count was developed to quantify the use of animated graphical ads. Animated images are counted as ads if they are used as links to pages from well-known advertising agencies (DoubleClick, HitBox, Adforce, etc.), contain the word ad or advertising in the URL, or are to pages on external sites. No measure was developed to detect the use of scrolling text, since scrolling text is typically implemented using scripts.

5.6.5 Graphic Formatting Measures: Graphic Quality

The quality of images used on the page, including their appropriateness and optimization (i.e., bytes and resolution), is discussed in the literature [Flanders and Willis 1998; Nielsen 2000; Sano 1996; Schriver 1997; Spool *et al.* 1999]. The only concrete guidance provided is for Web designers to use smaller images with fewer colors [Flanders and Willis 1998; Sano 1996]. Several measures were developed to assess the sizes of images as well as the screen area covered by images (Section 5.9). Image processing is not used; thus, the use of colors within images is not assessed.

5.7 Text Formatting Measures

Tables 5.7, 5.8, and 5.9 summarize 24 text formatting measures developed and discussed in this section. These measures assess the following aspects of Web interfaces.

1. How is body text emphasized?
2. Is there underlined text that is not in text links?

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
How is body text emphasized?								
Emphasized Body Word Count	Body text words that are emphasized (e.g., bolded or capitalized)	√		√		98.6%	98.1%	98.4%
Bolded Body Word Count	Body text words that are bolded	√		√		100.0%	99.9%	99.9%
Capitalized Body Word Count	Body text words that are capitalized	√		√		99.4%	99.7%	99.6%
Colored Body Word Count	Body text words that are a color other than the default text color	√		√		96.8%	95.1%	96.0%
Exclaimed Body Word Count	Body text words that are near exclamation points	√				100.0%	100.0%	100.0%
Italicized Body Word Count	Body text words that are italicized	√		√		100.0%	99.5%	99.8%
Is there underlined text (not in text links)?								
Underlined Word Count	Number of words that are underlined but are not text links	√	√	√		100.0%	100.0%	100.0%

Table 5.7: Summary of text formatting measures (Table 1 of 3). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

3. What font styles are used?
4. What font sizes are used?
5. How many text colors are used?
6. How many times is text re-positioned?
7. How are text areas highlighted?

5.7.1 Text Formatting Measures: Text Emphasis

Several literature sources provide guidance about emphasized text (e.g., bolded, italicized, and capitalized text) on the page [Flanders and Willis 1998; Nielsen 2000; Schriver 1997]. Specific guidance includes the following.

- Avoid mixing text attributes (e.g., color, bolding, and size) [Flanders and Willis 1998].

Measure	**Description**	**Aspects Assessed**				**Accuracy**		
		ID	ND	GD	ED	Hit	Miss	Avg.
What font styles are used?								
Serif Word Count	Number of words formatted with serif font faces	✓		✓		99.9%	100.0%	99.9%
Sans Serif Word Count	Number of words formatted with sans serif font faces	✓		✓		99.9%	91.7%	95.8%
Undetermined Font Style Word Count	Number of words formatted with undetermined font faces	✓		✓		100.0%	100.0%	100.0%
Font Style	Whether text is predominately sans serif, serif, or undetermined font styles	✓		✓		–	–	100.0%
What font sizes are used?								
Minimum Font Size	Smallest font size (in points) used for text	✓		✓		–	–	100.0%
Maximum Font Size	Largest font size used for text	✓		✓		–	–	100.0%
Average Font Size	Predominate font size used for text	✓		✓		–	–	100.0%
How many text colors are used?								
Body Color Count	Number of colors used for body text	✓		✓		100.0%	95.6%	97.8%
Display Color Count	Number of colors used for display text	✓		✓		100.0%	100.0%	100.0%

Table 5.8: Summary of text formatting measures (Table 2 of 3). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a $\surd$. Hit and miss accuracies are only reported for discriminating measures.

		Aspects Assessed				**Accuracy**		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
How many times is text re-positioned?								
Text Positioning Count	Number of text areas that change position from flush left			√		–	–	87.7%
Text Column Count	Number of x positions (i.e., columns) where text starts			√		–	–	95.0%
How are text areas highlighted?								
Text Cluster Count	Number of text areas that are highlighted in some manner	√		√		–	–	68.3%
Link Text Cluster Count	Number of link text areas that are highlighted in some manner	√	√	√		–	–	77.8%
Border Cluster Count	Number of text and link text areas that are highlighted with bordered regions	√		√		77.8%	100.0%	88.9%
Color Cluster Count	Number of text and link text areas that are highlighted with colored regions	√		√		93.6%	84.7%	89.2%
List Cluster Count	Number of text and link text areas that are highlighted with lists	√		√		91.7%	100.0%	95.8%
Rule Cluster Count	Number of text and link text areas that are highlighted with horizontal or vertical rules	√		√		60.8%	98.8%	79.8%

Table 5.9: Summary of text formatting measures (Table 3 of 3). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

- Minimize blinking text [Flanders and Willis 1998; Nielsen 2000].
- Avoid italicizing and underlining text [Schriver 1997].
- Avoid using all caps for text [Nielsen 2000].

No measure was developed to assess text emphasis across the entire page, since emphasis is used for both display and body text. However, measures were developed to assess body text emphasis as discussed in the next section.

5.7.2 Text Formatting Measures: Body Text Emphasis

The guidance provided in the previous section also applies to emphasized body text; thus, it is repeated below.

- Avoid mixing text attributes (e.g., color, bolding, and size) [Flanders and Willis 1998].
- Minimize blinking text [Flanders and Willis 1998; Nielsen 2000].
- Avoid italicizing and underlining text [Schriver 1997].
- Avoid using all caps for text [Nielsen 2000].

Numerous measures were developed to quantify body text emphasis, including an Emphasized Body Word Count (total number of body words that are bolded, italicized, or emphasized in some other way) and Bolded, Italicized, and Colored Body Word Counts. An Underlined Word Count was also developed to detect the presence of words that are underlined but not in text links, since it is possible for such words to be mistaken for links. The metrics do not directly measure the mixing of text emphasis, although this may be apparent from the individual measures.

An Emphasized Body Text Percentage measure was examined during both prior metric studies. This measure did not make a significant contribution for distinguishing rated pages in the first study, but it did make a significant contribution for distinguishing highly-rated pages in the second one. This measure was abandoned because it violates the low variability guidance for performance metrics.

5.7.3 Text Formatting Measures: Font Styles

Font style (e.g., serif or sans serif) has been discussed extensively in the literature [Bernard and Mills 2000; Bernard *et al.* 2001; Boyarski *et al.* 1998; Nielsen 2000; Schriver 1997]. Several usability studies have been conducted wherein serif and sans serif fonts were compared with respect to reading speed and user preference [Bernard and Mills 2000; Bernard *et al.* 2001]. Boyarski *et al.* [1998] have also compared fonts designed for computer screens to those designed for print. Specific guidance on font styles includes the following.

- Use serif fonts for faster reading by older adults [Bernard *et al.* 2001].
- Sans serif fonts have a slight advantage over serif fonts and are more preferred [Bernard and Mills 2000; Schriver 1997].
- Use fonts designed for computer screens (e.g., Verdana and Georgia) rather than fonts designed for print (e.g., Times New Roman) [Boyarski *et al.* 1998].

- Use only a few sizes from one or two typeface families; use one serif and one sans serif font for contrast [Schriver 1997].

- Use sans serif fonts for smaller text and serif fonts for larger text [Nielsen 2000].

Measures were developed to determine the predominant font style used for text. Specifically, the number of words formatted with serif, sans serif, and undetermined font styles are computed. Tables of serif and sans serif font names were compiled from the literature and Web sites that classify fonts (e.g., www.buyfonts.com). The HTML Parser and Browser Emulator assumes that the first valid font name specified is the one used by the browser. The font style of a word is then determined by looking up the font name in the tables. A Font Style measure is also computed as the maximum over the font style word counts; it identifies the predominate font style (sans serif, serif, or undetermined font styles).

The measures do not capture whether fonts designed for computer screens as opposed to those designed for print are used, whether a few font sizes from multiple typeface families are used, or whether sans serif fonts are used for small text and serif fonts are used for larger text. However, these measure in conjunction with the font point size measures (discussed below) provide some indirect insight about the font face and size combinations used.

5.7.4 Text Formatting Measures: Font Point Sizes

Font sizes (e.g., 9 pt and 14 pt) used in Web pages has been discussed extensively in the literature [Bernard *et al.* 2001; Flanders and Willis 1998; Nielsen 2000; Schriver 1997]. Specific guidance on font sizes includes the following.

- Use 14 pt fonts for older adults [Bernard *et al.* 2001].

- Use font sizes greater than 9 pt [Flanders and Willis 1998; Schriver 1997].

- Use 10 to 11 pt (or higher) for body text and 14 pt (or higher) for display text; use larger point sizes for serif faces [Schriver 1997].

Measures were developed to assess the maximum, minimum, and average point sizes of text on Web pages. The average point size was determined by tracking point sizes used for each word on the page and then dividing by the total number of words (i.e., Word Count). No measures were developed to track fonts sizes used for body as opposed to display text or for sans serif as opposed to serif fonts.

5.7.5 Text Formatting Measures: Text Colors

Flanders and Willis [1998] encourages Web designers to minimize the number of text colors. In addition to the Colored Body Word Count (discussed in Section 5.7.2), Body and Display Text Color Counts were developed; these measures report the number of unique colors used for body and display text. The measures do not assess if different colors are used for body and display text.

One limitation of the color measures is that they may count colors that are not discriminable or perceptually close. For example, it is possible that users may consider text colored with two different shades of blue as being the same color; color counts would be inflated in this case. Future work will consider the perceptual closeness of colors in metrics computation.

5.7.6 Text Formatting Measures: Text Positioning

Changing the alignment of text has been discussed in the literature [Flanders and Willis 1998; Nielsen 2000; Schriver 1997], and specific guidance includes the following.

- Avoid centering large blocks of text or links; left-justified text is preferred [Flanders and Willis 1998].
- Use left-justified, ragged-right margins for text; do not center items in vertical lists [Schriver 1997].

A Text Positioning Count was developed to quantify the number of times that text areas change position from flush left. This measure was examined in both prior metric studies and determined to be important for distinguishing rated and highly-rated pages. A similar measure, Text Column Count, was developed to determine the number of unique positions or columns where text starts on the page. Ideally, this measure will provide some insight about the complexity of the page layout, since tables are typically used in a nested manner to control text positioning beyond using the alignment tags. No measure was developed to assess the alignment of items in a list or the amount of text that is aligned.

5.7.7 Text Formatting Measures: Text Clustering

Use of text clusters or highlighting text areas in some manner (e.g., enclosing text in bordered or colored regions) is encouraged as a way to emphasize important text on the page [Landesman and Schroeder 2000; Nielsen 2000; Sawyer and Schroeder 2000; Schriver 1997]. Text clustering is not to be confused with text emphasis; in the latter case, only the text is emphasized (e.g., bolded or italicized) as opposed to the entire area surrounding the text in the first case. Specific guidance on text clustering includes the following.

- Use text clustering in small amounts and clusters should not contain much continuous text [Schriver 1997].
- Delineate links with bullets, spaces, etc. when they are in a group [Landesman and Schroeder 2000; Sawyer and Schroeder 2000]. This type of formatting is considered to be a form of text clustering.

Several measures of text and link text clustering were developed, including Text and Link Text Cluster Counts. If more than 50% of the text in a cluster is link text, then the cluster is considered a link text cluster and vice versa for text clusters. Use of specific clustering techniques, such as colored, list, or rule clusters, is also quantified. For colored text clusters, proximity to other colored clusters is considered. For example, if a colored cluster is adjacent to another colored cluster with a perceptually-close color, then only one text cluster is counted. Perceptual color closeness is determined by first translating the RGB (red, green, and blue) colors [Foley *et al.* 1990] into CIE Luv (luminancy and chrominancy) colors [Bourgin 1994]. The CIE Luv color space is device-independent and uniform (i.e., all colors are equidistant from each other). If the Euclidean distance of two Luv colors is less than 5, then the colors are considered to be perceptually close [Jackson *et al.* 1994].

Detecting some forms of text clustering, such as manual rule or list clusters (see Sections 5.7.8 and 5.7.9 below), is not as accurate as all of the other measures developed; image processing techniques are required to improve the accuracy of these measures.

The Text Cluster Count (encompassing text and link text clusters) was studied in both prior metric studies. Although it was shown in both cases that rated and highly-rated pages used text clustering more so than the other pages, this measure did not play an important role in classifying these pages.

5.7.8 Text Formatting Measures: Lists

A list is one form of text clustering as discussed above. Schriver [1997] suggests that Web designers minimize the number of lists on a Web page. A List Cluster Count was developed to quantify the number of lists on the page. One limitation of this measure is that it can only capture lists created with HTML list tags (e.g., <ul> or <ol>). Hence, it was only 79% accurate overall with only a 60% hit accuracy for the Web page sample. Image processing techniques are required to detect manually-created lists (e.g., rows of items separated with <br> tags) and to consequently improve accuracy.

5.7.9 Text Formatting Measures: Rules

Horizontal and vertical rules are other forms of text clustering. Specific guidance on the use of rules includes the following.

- Minimize the number of rules [Schriver 1997].
- Horizontal rules that are the full width of the page may be interpreted as the end of a page and possibly discourage scrolling [Spool *et al.* 1999].

A Rule Cluster Count was developed to quantify the number of rules on the page. One limitation of this measure is that it can only capture rules created with the HTML rule tag (<hr>). Image processing techniques are required to detect manually-created rules (e.g., an image containing a thin line). No measure was developed to detect whether vertical rules span the full width of pages.

5.7.10 Text Formatting Measures: Text in Clusters

As discussed above, the literature also discusses the amount of text contained in text clusters [Furnas 1997; Nielsen 2000; Schriver 1997]. Schriver [1997] suggests that Web designers minimize the amount of continuous text in clusters [Schriver 1997]. No measure was developed to assess the amount of text in clusters, since the current cluster measures are not highly accurate.

5.8 Link Formatting Measures

Table 5.10 summarizes three link formatting measures developed and discussed in this section. These measures assess the following Web interface features.

1. Are there text links that are not underlined?
2. What colors are used for links?

		Aspects Assessed				Accuracy		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
Are there text links that are not underlined?								
Non-Underlined Text Links	Whether there are text links without visible underlines	√	√	√		–	–	91.7%
What colors are used for links?								
Link Color Count	Number of colors used for text links		√	√		100.0%	99.2%	99.6%
Standard Link Color Count	Number of default browser colors used for text links		√	√		100.0%	100.0%	100.0%

Table 5.10: Summary of link formatting measures. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

5.8.1 Link Formatting Measures: Non-Underlined Links

It is possible for text links without underlines to be overlooked, since users are accustomed to the converse. Sawyer and Schroeder [2000] suggest that Web designers avoid using non-underlined text links, because they may be confused with text or considered a placeholder. A Non-Underlined Text Links measure was developed to detect the presence of text links without visible underlines.

5.8.2 Link Formatting Measures: Link Colors

The surveyed literature provides some guidance on colors used for text links [Nielsen 2000; Sawyer and Schroeder 2000; Spool *et al.* 1999], including the following.

- Use distinct link and visited link colors [Nielsen 2000; Sawyer and Schroeder 2000].
- Use link and visited link colors that are similar to default browser colors (shades of blue, red, and purple) [Nielsen 2000].
- Use default browser colors for links [Spool *et al.* 1999].

A Link Color Count was developed to assess the number of colors used for text links. No measures were developed to detect whether separate colors are used for unvisited and visited links, nor whether colors are similar to the default browser colors; such measures will be developed in future work. A Standard Link Color Count was developed to measure the number of default browser colors used (e.g., blue used for unvisited links).

One limitation of the color measures is that they may count colors that are not discriminable or perceptually close. For example, it is possible that users may consider link text colored with two different shades of blue as being the same color; color counts would be inflated in this case. Future work will consider the perceptual closeness of colors in metrics computation.

5.9 Graphic Formatting Measures

Table 5.11 summarizes the seven graphic formatting measures developed for assessing the following features of Web interfaces.

Measure	**Description**	**Aspects Assessed**				**Accuracy**		
		ID	ND	GD	ED	Hit	Miss	Avg.
How tall are graphics?								
Minimum Graphic Height	Minimum image height			√		–	–	100.0%
Maximum Graphic Height	Maximum image height			√		–	–	100.0%
Average Graphic Height	Average image height			√		–	–	100.0%
How wide are graphics?								
Minimum Graphic Width	Minimum image width			√		–	–	100.0%
Maximum Graphic Width	Maximum image width			√		–	–	100.0%
Average Graphic Width	Average image width			√		–	–	100.0%
How much page area is covered by graphics?								
Graphic Pixels	Total page area covered by images			√		–	–	100.0%

Table 5.11: Summary of graphic formatting measures. All of the measures are expressed in pixels. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

1. How tall are graphics?

2. How wide are graphics?

3. How much page area is covered by graphics?

Several literature sources discuss the sizes of images (the number of pixels in an image) [Flanders and Willis 1998; Nielsen 2000; Schriver 1997]. Flanders and Willis [1998] suggest that Web designers avoid using large graphics, although they do not quantify what is considered a large graphic. Hence, several measures were developed to assess the sizes of images: the minimum, maximum, and average height of images; the minimum, maximum, and average width of images; and the total pixel area covered by images. The latter measure in conjunction with a measure of the pixel area required for the page (discussed in Section 5.10.8) provides insight about what portion of the page is covered by images.

5.10 Page Formatting Measures

Tables 5.12, 5.13, and 5.14 summarize the 27 page formatting measures developed and discussed in this section. These measures assess the following aspects of Web interfaces.

1. How are colors used across the page?
2. What fonts are used across the page?
3. How big is the page? Big in this context refers to the width and height of pages.
4. Are there interactive elements on the page?
5. How is the page's style controlled?
6. Information about other page characteristics (e.g., the page's functional type and self-containment). Section 5.11 discusses functional types for pages, including home, link, and form pages. Self-containment in this context refers to the degree to which the page is rendered with HTML code and images, as opposed to being rendered with stylesheets, applets, scripts, etc.

The HTML Parser and Browser Emulator (see Chapter 4) was configured to simulate the Netscape Navigator 4.75 browser with default window settings (menu bar, address bar, toolbar, and status line). Currently, most monitors are 800 x 600 pixels [DreamInk 2000]; thus, the window size was fixed at 800 x 600 pixels for computing page formatting measures.

5.10.1 Page Formatting Measures: Colors

Colors used on computer screens, in Web interfaces in particular, is discussed extensively in the literature [Ambühler and Lindenmeyer 1999; Flanders and Willis 1998; Kaiser 1998; Murch 1985; Sano 1996; Schriver 1997; Stein 1997]. Specific guidance includes the following.

- Use no more than 6 discriminable colors [Murch 1985].
- Use browser-safe colors [Kaiser 1998].
- Use 256 (i.e., 8 bit) color palettes [Sano 1996].

In addition to the color measures developed to assess text and link formatting (Sections 5.7.5 and 5.8.2), three measures were developed to assess color use at the page level. A Color Count measures the total number of unique colors used for text, links, backgrounds, table areas, etc.; this measure made significant contributions in distinguishing Web pages in both prior metric studies. For each color used, the number of times it is used on the page is also tracked. For example, for every word on the page, the corresponding color use count is updated. The Minimum Color Use measure reports the minimum number of times a color is used. The average and maximum number of times a color is used is not reported, since these measures would be proportional to the amount of text on the page. This measure detects the use of an accent or sparsely-used color.

Use of "good" colors is also assessed through several measures. The Web design literatures encourages the use of browser-safe colors or colors whose RGB values are all evenly divisible by 51 [Kaiser 1998]. The Browser-Safe Color Count reports the number of such colors. Other measures to assess the quality of color combinations were developed and are discussed below.

One limitation of the color measures is that they may count colors that are not discriminable or perceptually close. For example, it is possible that users may consider two shades of blue with distinct RGB values as being the same; color counts would be inflated in this case. Future work will consider the perceptual closeness of colors in metrics computation.

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
How are colors used across the page?								
Color Count	Number of colors used			√		–	–	99.5%
Minimum Color Use	Minimum number of times a color is used			√		–	–	100.0%
Browser-Safe Color Count	Number of browser-safe colors used			√		100.0%	100.0%	100.0%
Good Text Color Combination	Number of good text or thin line color combinations			√		100.0%	100.0%	100.0%
Neutral Text Color Combinations	Number of neutral text or thin line color combinations			√		100.0%	100.0%	100.0%
Bad Text Color Combinations	Number of bad text or thin line color combinations			√		100.0%	100.0%	100.0%
Good Panel Color Combinations	Number of good thick line or panel color combinations			√		100.0%	100.0%	100.0%
Neutral Panel Color Combinations	Number of neutral thick line or panel color combinations			√		100.0%	100.0%	100.0%
Bad Panel Color Combinations	Number of bad thick line or panel color combinations			√		100.0%	100.0%	100.0%

Table 5.12: Summary of page formatting measures (Table 1 of 3). The quality of text and panel color combinations is assessed based on research in [Murch 1985]. The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

Measure	**Description**	**Aspects Assessed**				**Accuracy**		
		ID	ND	GD	ED	Hit	Miss	Avg.
What fonts are used across the page?								
Font Count	Number of fonts used			√		–	–	95.7%
Serif Font Count	Number of serif font faces used			√		–	–	100.0%
Sans Serif Font Count	Number of sans serif font faces used			√		–	–	100.0%
Undetermined Font Style Count	Number of undetermined font faces used			√		–	–	100.0%
How big is the page?								
Page Height	Height of page in pixels (600 pixel screen height)			√		–	–	88.2%
Page Width	Width of page in pixels (800 pixel screen width)			√		–	–	95.7%
Page Pixels	Total screen area required to render the page			√		–	–	84.0%
Vertical Scrolls	Number of vertical scrolls required to view the entire page				√	–	–	85.9%
Horizontal Scrolls	Number of horizontal scrolls required to view the entire page				√	–	–	100.0%
Are there interactive elements on the page?								
Interactive Element Count	Number of text fields, buttons, and other form objects				√	–	–	100.0%
Search Element Count	Number of forms for performing a search				√	–	–	91.7%

Table 5.13: Summary of page formatting measures (Table 2 of 3). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

		Aspects Assessed				Accuracy		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
How is the page's style controlled?								
External Stylesheet Use	Whether an external stylesheet file is used to format the page				√	–	–	100.0%
Internal Stylesheet Use	Whether an internal stylesheet is used within the <head> tag to format the page				√	–	–	100.0%
Fixed Page Width Use	Whether tables are used to create a specific page width			√		–	–	100.0%
Other page characteristics (e.g., page's functional type and self-containment)								
Page Depth	Level of the page within the site (determined by (crawling order)		√			–	–	–
Page Type	The page's functional type				√	–	–	84.0%
Self Containment	The degree to which all page elements are rendered solely via the HTML and image files				√	–	–	100.0%
Spamming Use	Whether the page uses invisible text or long page titles possibly indicating spamming				√	–	–	100.0%

Table 5.14: Summary of page formatting measures (Table 3 of 3). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

5.10.2 Page Formatting Measures: Color Combinations

The quality of color combinations used on computer screens, in Web pages in particular, has also been discussed in the literature [Flanders and Willis 1998; Murch 1985; Nielsen 2000]. Specific guidance includes the following.

- Use color combinations determined to be good (i.e., high contrast) via research studies [Murch 1985].
- Avoid using black backgrounds [Flanders and Willis 1998].
- Use high contrast between background and text [Flanders and Willis 1998; Nielsen 2000].

Murch [1985] conducted a study with sixteen users wherein color combinations on computer screens were examined; combinations of white, black, red, green, blue, cyan, magenta, and yellow were used. From this study, the author developed tables summarizing foreground and background color combinations that were preferred or rejected by at least 25% of the participants. Color combinations were examined for text and thin lines (i.e., less than three pixels in width or height) as well as for thick lines and panels (i.e., large shaded areas, such as a colored navigation bar).

Based on this information, six measures were developed to report the number of good, neutral, and bad color combinations for text (text and thin lines) and panels (panels and thick lines). Color combinations are only reported as being bad or good if 40% (seven or more) of the participants rejected or preferred the color combinations in Murch's study; this is a more stringent criteria than the 25% cutoff used by Murch. The previously-discussed CIE Luv color space is used as a starting point for mapping RGB values from a Web page into one of the eight study colors. A CIE Luv color is then mapped into the closest Munsell color [Foley *et al.* 1990]; this final mapping identifies a corresponding hue (i.e., one of the eight study colors). Over 700 Munsell color swatches are used for this mapping; the mapping algorithm and Munsell color swatch information was provided courtesy of Chris Healey [Healey 1996].

The Metrics Computation Tool analyzes all foreground and background color combinations used in a Web page, provided these color combinations are embedded within the HTML code. As previously discussed, the hues identified for foreground and background colors are used to determine if the color combination is good, bad, or neutral (i.e., not rejected or preferred by at least 40% of study participants). If two colors map into the same hue, then the distance between the colors is considered to determine whether it is a bad (i.e., colors are too close) or neutral (i.e., colors are distinct enough) color combination. Experimenting with various color distances revealed that a Euclidean distance of 25 was adequate for determining if two colors are distinct enough; this distance is the square of the previously-discussed perceptual closeness distance [Jackson *et al.* 1994].

Cultural information is not taken into consideration when determining the quality of color combinations. Whether the page background is black is also not reported. Color combinations used in images, applets, etc. are not assessed, since this assessment requires image processing techniques.

One limitation of the color combination measures is that they may count color combinations that are not discriminable or perceptually close. For example, it is possible that users may consider two areas with white foreground text and backgrounds that are shades of blue with distinct RGB values as being the same; color combination counts would be inflated in this case. Future work will consider the perceptual closeness of colors in metrics computation. Another limitation is that the color palette is not evaluated to determine if a good selection of colors is used; this will be explored in future work.

5.10.3 Page Formatting Measures: Fonts

A font is a combination of four features: a font face, a font size, whether text is bolded, and whether text is italicized [Schriver 1997]. Several sources discuss fonts [Nielsen 2000; Schriver 1997; Stein 1997], and Nielsen [2000] suggests that Web designers use no more than two fonts and possibly one for special text. A Font Count was developed to report the total number of unique fonts used on a page; this measure played a significant role in distinguishing highly-rated pages in the second metrics study. The number of each type of font face – serif, sans serif, and undetermined font styles – is also reported. The measures do not report whether a font or font face is used for special text nor do they distinguish fonts used for body and display text.

The font measures in this section are distinct from the ones in Section 5.7.3. The latter ones assess the number of words formatted with serif and sans serif font styles; thus, proving some insight about how fonts are used.

5.10.4 Page Formatting Measures: Line Length

The width of text lines on the page is discussed in the literature [Flanders and Willis 1998; Schriver 1997], and specific guidance includes the following.

- Keep line lengths to 40–60 characters [Schriver 1997].
- Keep text between 9 to 15 words per line [Flanders and Willis 1998].

No measure was developed to assess text line lengths, since lengths vary considerably when multiple columns are used. Image processing techniques are required to accurately determine line lengths.

5.10.5 Page Formatting Measures: Leading

Spacing between consecutive text lines on a page is more of a concern for print documents than Web documents, since Web pages typically use consistent spacing as dictated by browsers. Stylesheet parameters can also control leading. Schriver [1997] suggests that leading be 120% of the font face's point size and even larger between paragraphs. Given that leading is mainly controlled by the browser, no measure was developed to assess this feature.

5.10.6 Page Formatting Measures: Framesets

Use of framesets is an often debated topic in Web design literature, since they typically confuse users [Flanders and Willis 1998; Nielsen 2000; Stein 1997]. Specific guidance includes the following.

- Avoid using framesets [Nielsen 2000].
- Use tables instead of framesets [Flanders and Willis 1998].

Although simple to measure, no measure was developed to report the use of framesets, since the tool was not designed to support them. However, the tool does support inline frames, since text equivalents are typically provided by designers; inline frames are only supported by the Internet Explorer browser. All metric studies, including the study discussed in the next chapter, excluded sites that used framesets.

5.10.7 Page Formatting Measures: Interactive Elements

Use of buttons, text boxes, pull-down menus, and other interactive elements has been discussed extensively in the literature [Flanders and Willis 1998; Rosenfeld and Morville 1998; Sawyer and Schroeder 2000; Scanlon and Schroeder 2000b; Spool *et al.* 1999]. Specific guidance on interactive elements includes the following.

- Avoid using mouseovers and pull-down menus for navigation [Rosenfeld and Morville 1998; Sawyer and Schroeder 2000].
- Support search; users use search half of the time [Scanlon and Schroeder 2000b].
- Make the scope and results of searching clear [Spool *et al.* 1999].
- Do not use form buttons as links (i.e., overuse buttons) [Flanders and Willis 1998].

An Interactive Element Count was developed to report the total number of form elements used (all buttons, select menus, checkboxes, etc.). A Search Element Count was also developed to determine whether searching is supported on the page. Several heuristics were developed to detect search forms with 92% overall accuracy. For example, if the form or some element in the form contains the term "search," then it is considered a search form. One limitation of the measure is that it does not detect whether the search is for the site itself or for the Web at large.

No measures were developed to detect the use of form elements for navigation or whether the scope and results of searching are clear.

5.10.8 Page Formatting Measures: Screen Size

Much guidance is provided in the literature on the width and height of Web interfaces [Flanders and Willis 1998; Nielsen 2000; Sano 1996; Sawyer *et al.* 2000], including the following.

- Limit horizontal width to 572 pixels or less [Sano 1996].
- Restrict page width and height to 595 x 295 pixels [Flanders and Willis 1998].
- Restrict page width to less than 600 pixels [Nielsen 2000].
- Longer pages are better; use 800 x 600 pixels; and avoid horizontal scrolling [Sawyer *et al.* 2000].

Differences in page width and height recommendations reflect changes in the dominant screen sizes over the years. Currently, most monitors are 800 x 600 pixels [DreamInk 2000]. Thus, measures of the width and height of a page were developed using an 800 x 600 screen size. The available width for pages is determined based on the Netscape Navigator 4.75 browser; small areas are occupied on the left and right of the browser window for a scroll bar and border. A Page Pixels measure was developed to report the total number of pixels covered by the page; this measure is the page width multiplied by the page height.

5.10.9 Page Formatting Measures: Screen Coverage

The total screen area covered (i.e., non whitespace) is also discussed in the literature [Sawyer *et al.* 2000; Schriver 1997; Spool *et al.* 1999]. The consensus is that Web designers minimize whitespace on the page [Sawyer *et al.* 2000; Spool *et al.* 1999]. No measure was developed to determine the total screen area covered, since this computation may require image processing techniques. For example, blank images are often used extensively as spacers; counting areas covered by spacers could inflate screen coverage.

5.10.10 Page Formatting Measures: Text Density

The screen area covered by text is discussed in the literature [Sawyer *et al.* 2000; Schriver 1997; Spool *et al.* 1999]; specific guidance includes the following.

- Text should cover no more than 25–30% of the screen area [Schriver 1997].
- Greater text density facilitates page scanning [Sawyer *et al.* 2000; Spool *et al.* 1999].

These guidelines obviously contradict each other. However, no measure was developed to determine the text density, since this computation may require image processing techniques to compute accurately.

5.10.11 Page Formatting Measures: Scrolling

Vertical and horizontal scrolling is discussed extensively in the literature [Flanders and Willis 1998; Nielsen 2000; Spool *et al.* 1999]; specific guidance includes the following.

- Minimize scrolling [Spool *et al.* 1999].
- Minimize vertical scrolling to 2 screens [Flanders and Willis 1998].
- Users should not be required to scroll [Nielsen 2000].

The number of vertical and horizontal scrolls required to view the entire page is measured. An 800 x 600 screen size is used for these computations, since most monitors are at least 800 x 600 pixels [DreamInk 2000]. The available width for pages is determined based on the Netscape Navigator 4.75 browser; small areas are occupied on the left and right of the browser window for a scroll bar and border. Similarly, the available height is adjusted to account for the Netscape default menu bar, toolbar, address bar, and status bar.

5.10.12 Page Formatting Measures: Stylesheets

Use of stylesheets to control page layout has been discussed in the literature [Flanders and Willis 1998; Nielsen 2000], and specific recommendations include the following.

- Use stylesheets to enforce consistency [Flanders and Willis 1998; Nielsen 2000].
- Use external rather than embedded stylesheets [Nielsen 2000].

Two measures – External and Internal Stylesheet Use – were developed to detect the use of external stylesheet files and internal stylesheets within the <head> tag of Web pages, respectively. Both of these measures are associated with experience design in Table 5.14, since the affects of stylesheets are more overarching than just information, navigation, and graphic design. For instance, use of external and internal stylesheets can potentially improve consistency across pages and reduce download time. Use of embedded style tags (i.e., style attributes within HTML tags or within the body of Web pages) is not detected, since this is not considered as a consistent use of stylesheets to control page layout. A related measure – Fixed Page Width – detects whether tables are used to force a fixed page width or not.

5.10.13 Page Formatting Measures: Layout Quality

All of the measures discussed in this section provide some insight about some aspects of layout quality. Other high-level aspects discussed in the literature include: aesthetics, alignment, balance, and the presence of distractions (e.g., popup windows or spawning browser windows) [Flanders and Willis 1998; Nielsen 2000; Sano 1996; Scanlon and Schroeder 2000a; Schriver 1997]. High-level features such as aesthetics are difficult to measure in an automated manner, since this is largely subjective and requires input from users. Other aspects, such as alignment and balance, require image processing techniques to assess; hence, no measures were implemented to assess them. Although it is suggested that Web designers minimize distractions (e.g., spawning browser windows) [Scanlon and Schroeder 2000a], the presence of distractions is not detected, since these elements are typically implemented with scripting.

Other features associated with layout quality, such as page depth, self-containment, and spamming use are measured. The page depth is determined by the crawling order within a site, and does not necessarily reflect the actual depth of the page in the site. Self-containment reflects the degree to which all elements required to render the page are contained in the HTML page itself and image files, as opposed to being rendered by external stylesheet files, scripts, applets, and other objects. For example, if a page does not use external stylesheets, scripts, applets, etc., then it is considered to have high self-containment. If long page titles (more than 64 characters) or a series of invisible words are in the page, then spamming use is reported for the page.

5.11 Page Function

Another aspect of page formatting is the primary function of Web pages (i.e., whether pages are primarily for content or links), which needs to be considered during design [Flanders and Willis 1998; Sano 1996; Stein 1997]. The first metric study showed that considering page function (home vs. other pages) leads to more accurate predictions. Although several studies have focused on automatically predicting the genre of a site (e.g., commercial or academic) [Bauer and Scharl 2000; Hoffman *et al.* 1995], few studies have been conducted to determine ways for automatically predicting the function of a page [Karlgren 2000; Pirolli *et al.* 1996; Stein 1997]. Stein [1997] considers the ratio of text link words to all of the words on a page. If this ratio is high, then the page is considered to be primarily for links; no guidance is given for what constitutes a high ratio.

Karlgren [2000] surveyed over 600 Web users to determine a set of eleven genres (e.g., home pages, searchable indices, journalistic, reports, FAQs, and others) and used 40 linguistic (e.g., number of adverbs, characters, long words, present participles, etc.) and Web (e.g., number of images and links) measures for predicting page type. Karlgren developed a decision tree to generate predictions; however, the author did not report the accuracy of the tree nor provide model details. This work was incorporated into a search interface that clusters search results by genre.

Pirolli *et al.* [1996] present a set of functional categories for Web pages, including home (organizational and personal), index, source index (sub-site home pages), reference, destination (sink pages such as acronym, copyright, and bibliographic references), and content. Pages can belong to multiple categories in their classification scheme. The authors developed a classification algorithm based on features that a Web page exhibits (e.g., page size, number of inbound and outbound links, depth of children, and similarity to children). For example, reference pages had a relatively larger page size and fewer inbound and outbound links. Page classification was used in conjunction with text similarity, hyperlink structure, and usage data to predict the interests (i.e., plausible information-seeking goals) of Web site users.

The work in [Karlgren 2000] and [Pirolli *et al.* 1996] was used as a starting point for developing an approach for predicting page types. Examination of the eleven genres in [Karlgren 2000] revealed that most qualified the type of text on the page (e.g., journalistic, discussions, lists, and FAQs), while the remaining genres were for pages with different functionality (e.g., links, forms, and home pages). Similarly, the categories in [Pirolli *et al.* 1996] contain multiple text, link, and home page types. Hence, these categories were collapsed into the following five page types.

Home: main entry pages to a site that typically provide a broad overview of site contents.

Link: pages that mainly provide one or more lists of links. Links may be annotated with text or grouped with headings (e.g., yahoo directory, redirect page, or sitemap). This functional type includes category pages (entry pages to sub-sites or major content areas).

Content: pages that mainly provide text. This functional type includes reference (e.g., a glossary, FAQ, search and site tips, and acronyms) and legal (e.g., disclaimers, privacy statements, terms, policies, and copyright notices) pages.

Form: pages that are primarily HTML forms.

Other: 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.

With the assistance of an undergraduate student, Deep Debroy, a sample of 1,770 Web pages were classified into the five categories above; each page was assigned to only one category. The pages came from 6 types of sites (Community, Education, Finance, etc.) with low to high expert ratings (see Chapter 6); there were at least 223 pages for each functional type. Multiple Linear Regression analysis [Keppel and Zedeck 1989] was used to identify a subset of the measures for model building. The Classification and Regression Tree (C&RT) [Breiman *et al.* 1984] method with 70% training and 30% test samples was used on this subset of measures. The resulting tree contains 70 rules and has an accuracy of 87% and 75% for the training and test samples, respectively. Cross validation [Schaffer 1993] was also used to further assess whether the model is generalizable to new data; the tree has a 5-fold cross validation accuracy of 75%. Figures 5.3 and 5.4 present example rules for predicting each page type.

5.12 Page Performance Measures

Tables 5.15–5.19 summarize 37 page performance measures developed to answer the following questions.

1. How fast is the page rendered?

if ((All Page Text Score is not missing AND (All Page Text Score $\leq$ 10.5)) AND (HTML Bytes is missing OR (HTML Bytes $>$ 2735)) AND (Link Count is missing OR (Link Count $>$ 8.5)) AND (Meta Tag Word Count is missing OR (Meta Tag Word Count $>$ 18.5)))

PageType = Home

This rule classifies pages as home pages if they have: minimal content similarity between source and destination page text (good visible and invisible words; this is typically because home pages are the first pages crawled and are therefore not compared to a source page); an HTML page size of more than 2.7K; 8.5 or more links; and 18.5 or more meta tag words.

if ((All Page Text Score is missing OR (All Page Text Score $>$ 10.5)) AND (Good Body Word Count is not missing AND (Good Body Word Count $\leq$ 86.5)) AND (Link Word Count is missing OR (Link Word Count $\leq$ 73)) AND (Good Body Word Count is not missing AND (Good Body Word Count $\leq$ 22.5)) AND (Link Count is not missing AND (Link Count $>$ 21.5)) AND (Script Bytes is missing OR (Script Bytes $\leq$ 2678.5)) AND (Interactive Object Count is missing OR (Interactive Object Count $\leq$ 1.5)))

PageType = Link

This rule classifies pages as link pages if they have: some content similarity between source and destination page text (good visible and invisible words); between 22.5 and 86.5 good body words; 73 or fewer link words; 1.5 or fewer interactive objects; more than 21.5 links; and less than 2.7K of script bytes (possibly for form validation).

if ((All Page Text Score is missing OR (All Page Text Score $>$ 10.5)) AND (Link Word Count is not missing AND (Link Word Count $>$ 77.5)) AND (Good Body Word Count is not missing AND (Good Body Word Count $>$ 278)) AND (Exclaimed Body Word Count is not missing AND (Exclaimed Body Word Count $\leq$ 0.5)))

PageType = Content

This rule classifies pages as content pages if they have: some content similarity between source and destination page text (good visible and invisible words); more than 77.5 link words; more than 278 good body words; and virtually no body words that are near exclamation points.

Figure 5.3: Example decision tree rules for predicting page types (Home, Link, and Content pages).

if ((All Page Text Score is not missing AND (All Page Text Score $\leq$ 10.5)) AND (HTML Bytes is not missing AND (HTML Bytes $\leq$ 2735)) AND (Search Object Count is not missing AND (Search Object Count $>$ 0.5)))

PageType = Form

This rule classifies pages as form pages if they have: minimal content similarity between source and destination page text; 2.7K or fewer HTML bytes; and search support.

if ((All Page Text Score is not missing AND (All Page Text Score $\leq$ 10.5)) AND (HTML Bytes is not missing AND (HTML Bytes $\leq$ 2735)) AND (Search Object Count is missing OR (Search Object Count $\leq$ 0.5)))

PageType = Other

This rule classifies pages as other pages if they have: minimal content similarity between source and destination page text; 2.7K or fewer HTML bytes; and virtually no search support.

Figure 5.4: Example decision tree rules for predicting page types (Form and Other pages).

2. Is the page accessible to people with disabilities?
3. Are there HTML errors on the page?
4. Is there strong “scent” to the page? Scent in this context refers to whether the page text provides hints about the contents on a linked page without having to navigate to the page.

5.12.1 Page Performance Measures: Page Bytes

The total bytes for the Web page (HTML code, stylesheets, objects, and images) plays an important role in how fast the page loads in the browser [Flanders and Willis 1998; Nielsen 2000]. Nielsen [2000] recommends that Web designers keep page bytes below 34K for fast loading. Most HTML authoring tools, such as Microsoft FrontPage and Macromedia Dreamweaver, provide download speed estimates; these estimates typically only consider the total bytes for pages. Experiments using only the total page bytes to estimate download speed revealed that such estimates only account for roughly 50% of download speed. Hence, a more accurate model of download speed was developed; this model is discussed in Section 5.12.4.

A Page Bytes measure was developed and shown to make significant contributions in distinguishing Web pages in both prior metric studies. The percentage of page bytes attributable to graphics was also shown to play a significant role in distinguishing highly-rated pages in the second study. These measures are not used in the download speed computation, since different types of bytes (e.g., graphic bytes or script bytes) influence download speed differently. The page bytes measure does not capture bytes that may be embedded in objects, stylesheets, scripts, etc. For example, if a stylesheet imports a secondary stylesheet, page bytes for the first stylesheet are counted but not for the second one.

The page bytes and graphic percentage measures were replaced with an HTML bytes measure – total bytes for HTML tags, excluding tags for scripts, layers, and other objects – and the number of HTML files. Similar measures were developed for graphics, scripts, and objects as discussed in the sections below.

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
How fast is the page rendered?								
Table Count	Number of HTML tables used to render the page				√	–	–	100.0%
HTML File Count	Number of HTML files, including stylesheet files				√	100.0%	100.0%	100.0%
HTML Bytes	Total bytes for HTML tags, text, and stylesheet tags				√	100.0%	100.0%	100.0%
Graphic File Count	Number of image files				√	100.0%	100.0%	100.0%
Graphic Bytes	Total bytes for image files				√	100.0%	100.0%	100.0%
Script File Count	Number of script files				√	100.0%	100.0%	100.0%
Script Bytes	Total bytes for scripts (embedded in script tags and in script files)				√	100.0%	100.0%	100.0%
Object File Count	Number of object files (e.g., for applets, layers, sound, etc.)				√	100.0%	100.0%	100.0%
Object Bytes	Total bytes for object tags and object files				√	100.0%	100.0%	100.0%
Object Count	Number of scripts, applets, objects, etc.				√	–	–	100.0%
Download Time	Time for a page to fully load over a 56.6K modem (41.2K connection speed)				√	–	–	86.0%

Table 5.15: Summary of page performance measures (Table 1 of 5). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

		Aspects Assessed				**Accuracy**		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
Is the page accessible to people with disabilities?								
Bobby Approved	Whether the page was approved by Bobby as being accessible to people with disabilities				√	–	–	100.0%
Bobby Priority 1 Errors	Number of Bobby priority 1 errors reported				√	–	–	100.0%
Bobby Priority 2 Errors	Number of Bobby priority 2 errors reported				√	–	–	100.0%
Bobby Priority 3 Errors	Number of Bobby priority 3 errors reported				√	–	–	100.0%
Bobby Browser Errors	Number of Bobby browser compatibility errors reported				√	–	–	100.0%
Are there HTML errors in the page?								
Weblint Errors	Number of HTML syntax errors reported by Weblint				√	–	–	100.0%
Is there strong “scent” to the page?								
Visible Page Text Terms	Maximum good visible words on source & destination pages	√	√			–	–	100.0%
Visible Unique Page Text Terms	Maximum good visible, unique words on source & destination pages	√	√			100.0%	100.0%	100.0%
Visible Page Text Hits	Common visible words on source & destination pages	√	√			100.0%	100.0%	100.0%

Table 5.16: Summary of page performance measures (Table 2 of 5). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

		Aspects Assessed				**Accuracy**		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
Is there strong “scent” to the page?								
Visible Page Text Score	Score for common visible words on source & destination pages	√	√			100.0%	100.0%	100.0%
All Page Text Terms	Maximum good visible and invisible words on source & destination pages	√	√			–	–	100.0%
All Unique Page Text Terms	Maximum good visible and invisible, unique words on source & destination pages	√	√			100.0%	100.0%	100.0%
All Page Text Hits	Common visible and invisible words on source & destination pages	√	√			100.0%	100.0%	100.0%
All Page Text Score	Score for common visible and invisible words on source & destination pages	√	√			100.0%	100.0%	100.0%
Visible Link Text Terms	Maximum good visible words in the link text & destination page	√	√			–	–	100.0%
Visible Unique Link Text Terms	Maximum good visible, unique words in the link text & destination page	√	√			100.0%	100.0%	100.0%
Visible Link Text Hits	Common visible words in the link text & destination page	√	√			100.0%	100.0%	100.0%

Table 5.17: Summary of page performance measures (Table 3 of 5). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
Is there strong "scent" to the page?								
Visible Link Text Score	Score for common visible words in the link text & destination page	√	√			100.0%	100.0%	100.0%
All Link Text Terms	Maximum good visible and invisible words in the link text & destination page	√	√			–	–	100.0%
All Unique Link Text Terms	Maximum good visible and invisible, unique words in the link text & destination page	√	√			100.0%	100.0%	100.0%
All Link Text Hits	Common visible and invisible words in the link text & destination page	√	√			100.0%	100.0%	100.0%
All Link Text Score	Score for common visible and invisible words in the link text & destination page	√	√			100.0%	100.0%	100.0%
Page Title Terms	Maximum good page title words on source & destination pages	√	√			–	–	100.0%
Unique Page Title Terms	Maximum good, unique page title words on source & destination pages	√	√			100.0%	100.0%	100.0%

Table 5.18: Summary of page performance measures (Table 4 of 5). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
Is there strong "scent" to the page?								
Page Title Hits	Common page title words on source & destination pages	$\surd$	$\surd$			100.0%	100.0%	100.0%
Page Title Score	Score for common page title words on source & destination pages	$\surd$	$\surd$			100.0%	100.0%	100.0%

Table 5.19: Summary of page performance measures (Table 5 of 5). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a $\surd$. Hit and miss accuracies are only reported for discriminating measures.

5.12.2 Page Performance Measures: Graphic Bytes

The literature discusses the total bytes for images on the page [Flanders and Willis 1998; Nielsen 2000]. Specifically, Flanders and Willis [1998] suggest that Web designers keep graphic bytes to less than 35K and to optimize images. This recommendation obviously contradicts the recommendation from Nielsen to keep the Web page and all elements under 34K [Nielsen 2000]. A graphic bytes measure was developed to report this information. The number of graphic files is also reported.

5.12.3 Page Performance Measures: Objects

Use of applets, controls, scripts, marquees, plug-ins, etc. is discussed extensively in the literature [Ambühler and Lindenmeyer 1999; Flanders and Willis 1998; Nielsen 2000; Rosenfeld and Morville 1998; Spool *et al.* 1999; Stein 1997]; specific guidance includes the following.

- Avoid gratuitous use of technology [Nielsen 2000; Rosenfeld and Morville 1998].
- Minimize the use of video [Nielsen 2000].
- Avoid using sound files [Flanders and Willis 1998].

Separate measures were developed to track the number of script and layer files as well as the total bytes for these elements, including bytes for the HTML tags within the Web page. Similarly, the number of object files and the total bytes for objects is measured. These measures are used in conjunction with previously-discussed measures to estimate the download speed (see discussion below). The number of bytes and files do not capture bytes or files that may be embedded in objects and scripts. For example, if an object loads images, then the bytes for the loaded images are not counted although bytes for the object itself are.

The total number of objects – scripts (both scripts within the HTML page and in an external script file), layers, applets, etc. – is also reported.

5.12.4 Page Performance Measures: Download Speed

The time for a page to fully load is considered a critical issue for Web interfaces [Flanders and Willis 1998; Nielsen 2000; Scanlon and Schroeder 2000a; Spool *et al.* 1999; Zona Research 1999]. Specific guidance includes the following.

- Download speed should be no more than 10 seconds [Nielsen 2000].
- Home pages greater than 40K result in significant bailouts [Zona Research 1999].
- Perceived download speed matters more than actual download speed [Scanlon and Schroeder 2000a].

Many HTML authoring tools provide download speed estimates to assist Web designers with optimizing pages. Experience with these estimates on the sample revealed that such estimates only account for about 50% of the actual download speed; there is also a major discrepancy for download speed estimates provided by the Bobby tool [Clark and Dardailler 1999; Cooper 1999]. These discrepancies are mostly due to the fact that these tools use a simplistic model, typically based on the total bytes for pages and in some cases, such as Bobby, incorporate a latency measure proportional to the total number of files.

A model of download speed was developed based on the actual download speeds for the sample. A 56.6K modem was used to measure actual download speeds, since this is currently the most prevalent modem [DreamInk 2000]. The measurement activity revealed that it is rarely possible to achieve a 56.6K connection speed with a 56.6K modem; this is due to various technological limitations of the analog modem [Bash 1997] and possibly poor telephone connections. For 50 connection sessions to three different Internet service providers at various times of the day, the average and median connection speed were 41.2K with 42.6K as a close second. Hence, a connection speed of 41.2K was used to measure download speed for the sample; this value is also used for estimating download speed. All fourteen of the Web pages were downloaded twelve times each. Several warm-up downloads were performed and browser caching was disabled during this measurement activity.

Several measures believed to impact download speed were developed: Graphic Bytes and Graphic File Count (Section 5.12.2), HTML Bytes and HTML File Count (Section 5.12.1), Script Bytes and Script File Count (Section 5.12.3), Object Bytes and Object File Count (Section 5.12.3), Object Count (Section 5.12.3), and Table Count (the number of <table> tags in the Web page). These measures and the actual download speeds were used for Multiple Linear Regression analysis [Keppel and Zedeck 1989] to determine an equation for predicting download speed. A backward elimination method was used wherein all of the measures were entered into the equation initially, and then one by one, the least predictive measure was eliminated. This process was repeated until the Adjusted R Square (predictive accuracy) showed a significant increase with the elimination of a predictor. All of the measures were retained by this method, except for the Object Count.

Equation 5.2 depicts the model developed to predict the download speed for a Web page. The Adjusted R Square for Equation 5.2 is .86 indicating that the measures explained about 86% of the variance in predictions. The F value is 123.06 and significant at the $p < .01$ level; this indicates that the linear combination of the measures significantly predicts the download speed. Table 5.20 shows the relative contribution of each measure to download speed predictions; all measure contributions are significant at the $p < .01$ level.

$$DownloadSpeed = -0.181 + 0.0003085 * GraphicBytes + \tag{5.2}$$

$$\begin{aligned}
&0.0002492 * HTMLBytes + \\
&0.0005748 * ScriptBytes + \\
&0.003211 * ObjectBytes + \\
&10.140 * HTMLFileCount + \\
&-0.385 * GraphicFileCount + \\
&-1.378 * ScriptFileCount + \\
&-6.243 * ObjectFileCount + \\
&-0.463 * TableCount
\end{aligned}$$

5.12.5 Page Performance Measures: Accessibility

Whether a Web interface is accessible to people with disabilities is discussed in the literature [Clark and Dardailler 1999; Cooper 1999; Nielsen 2000; Web Accessibility Initiative 1999]. The consensus is that Web designers adhere to accessibility principles to create sites that serve a broad user community. In other words, accessible interfaces are of higher quality than non-accessible ones, although this claim has not been empirically examined.

Results of running the Bobby tool (version 3.2) [Clark and Dardailler 1999; Cooper 1999] is reported for each Web page; default options are used. Specifically, whether the page is Bobby approved, the number of priority 1, 2, and 3 errors, as well as the number of browser compatibility errors is reported.

5.12.6 Page Performance Measures: HTML Errors

Whether Web interfaces contain HTML or browser-incompatibility errors is also discussed in the literature [Bowers 1996; Kim and Fogg 1999; Fogg *et al.* 2000]. One survey showed that HTML errors decrease credibility [Kim and Fogg 1999; Fogg *et al.* 2000]. Hence, the results of running the Weblint tool (version 1.02) [Bowers 1996], specifically the total number of Weblint errors is reported. Weblint detects numerous HTML errors, such as missing closing tags, broken links, and invalid tag attributes. For measurement purposes, Weblint is configured to support Netscape extensions and to disable checks for minor HTML errors, such as the use of quotations around attribute values and specifying alternative text for images. The specific command line used is below.

Measure	Standardized Coefficient	Significance
Graphic Bytes	1.011	0.000
HTML Bytes	0.646	0.000
Script Bytes	0.164	0.005
Object Bytes	0.396	0.000
Graphic File Count	-0.559	0.000
HTML File Count	0.380	0.000
Script File Count	-0.184	0.001
Object File Count	-0.236	0.004
Table Count	-0.555	0.000

Table 5.20: Standardized coefficients indicating the contribution of each measure to download speed predictions along with t-test results (2-tailed significance).

weblint -x Netscape -i -d quote-attribute-value,extension-attribute,extension-markup,img-alt,attribute-delimiter filename

5.12.7 Page Performance Measures: Scent Quality

Much has been said about the use of "scent" – hints to help the user decide where to go next – as a way to improve navigation [Chi *et al.* 2000; Furnas 1997; Larson and Czerwinski 1998; Miller and Remington 2000; Nielsen 2000; Rosenfeld and Morville 1998; Sawyer *et al.* 2000; Spool *et al.* 1999; Spool *et al.* 2000]. Specific guidance includes the following.

- Use clear headings with related links (i.e., link clustering) to enhance scent [Spool *et al.* 2000].
- Effective navigation requires small pages (views), few clicks between pages, and strong scent [Furnas 1997].
- Weak scent (i.e., ambiguous link text) impedes navigation [Chi *et al.* 2000; Miller and Remington 2000; Spool *et al.* 1999; Spool *et al.* 2000].
- Avoid using 'Click Here' for link text [Nielsen 2000].
- Use breadcrumbs (i.e., displaying a navigation trail) rather than long navigation bars [Nielsen 2000].
- Use link titles to help users predict what will happen if they follow a link [Nielsen 1998b].

Many of these suggestions are difficult to measure in an automated manner. Furthermore, it is difficult to gauge users' understanding of link text. Nonetheless, several research efforts have yielded computational models of scent [Chi *et al.* 2000; Chi *et al.* 2001; Miller and Remington 2000; Pirolli and Card 1995; Pirolli *et al.* 1996; Pirolli 1997]. For Web navigation, the most promising approach compares proximal cues on the source page (link text, text surrounding the link, graphics related to a link, and the position of the link on the page) to text on the destination page [Chi *et al.* 2000; Chi *et al.* 2001; Pirolli *et al.* 1996]; this approach also considers the site's linkage topology and usage patterns from server log data. The authors have used this approach for predicting how users will navigate a site given some information need as well as for inferring users' information needs from navigation patterns. Several information retrieval approaches, such as TF.IDF (Term Frequency by Inverse Document Frequency) and simple word overlap [Baeza-Yates and Ribeiro-Neto 1999], have been used for determining similarity between the link text and destination page text.

The approach discussed in [Chi *et al.* 2000; Chi *et al.* 2001; Pirolli *et al.* 1996] was used as a starting point for developing eighteen scent quality measures. Given the nature of the Metrics Computation Tool, it was not possible to incorporate site topology and usage data into these measures; this would require exhaustive site crawling and access to server logs. Furthermore, simple word overlap is used as opposed to TF.IDF or other information retrieval algorithms, since it does not require a large collection of pages for each site. The measures focus on assessing content similarity as follows.

Source Page Text vs. Destination Page Text: This comparison establishes whether or not the content is similar between the two pages. If there is a high similarity, one would expect the link text similarity to reflect this.

Link Text vs. Destination Page Text: Link text includes the actual words in the link (including image alt attributes for graphic links) as well as twelve good words before and after the link text; if an image precedes or follows a link, text is extracted from the image's alt attribute if specified. If there is a high similarity between text on the source and destination pages, then this comparison should reflect this similarity. If not, this may indicate poor scent quality.

Source Page Title vs. Destination Page Title: This comparison is mainly to assess whether page titles vary between pages in the site. As discussed in Section 5.4.2, Nielsen [2000] suggests that designers use different page titles for each page. A similar page title consistency measure was developed to assess variation in page titles throughout the site; this measure will be discussed in Section 5.13.1.

For the first two comparisons, two forms of page text are considered: visible words (page title, headings, and body text) and all words (page title, headings, body text, meta tags, invisible text, and image alt text). The first type reflects content that can be seen by users during browsing, while the latter type reflects text used for searching, specifically indexing documents for searching.

A Java program adapted from one developed by Marti Hearst is used for the comparisons. The program employs simple word overlap and reports several measures for each comparison: 1. the maximum number of terms (good words; Section 5.4.12) considered between the source and destination text (link, title, or page contents); 2. the number of unique terms in the source text (link, title, or page contents); 3. the number of terms that appear in both the source and destination text (hits); and 4. the weighted score for the common terms. For the weighted score, each term in the source text is assigned a weight equal to the number of times it appeared in the source text; frequently-used terms will have a higher weight. When the term appears in the destination text, this weighted score is used for each hit. The actual word overlap is a ratio of these measures ($\frac{hits}{unique\ terms}$ and $\frac{score}{terms}$) and as such is not considered. The terms and unique terms are reported simply for reference. The eighteen scent quality measures are summarized below; these measures are associated with the destination page as determined by site crawling order.

Source Page Text vs. Destination Page Text Measures: visible text (browsing): terms, unique terms, hits, and score; and all text (searching): terms, unique terms, hits, and score.

Link Text vs. Destination Page Text Measures: visible text (browsing): terms, unique terms, hits, and score; and all text (searching): terms, unique terms, hits, and score.

Source Page Title vs. Destination Page Title Measures: all text (browsing and searching): terms, unique terms, hits, and score.

There are many ways to express similar concepts; however, term expansion is not used. Another major limitation of these measures is that they do not reflect subjective preferences or whether users understand the terms. Although the measures are computed with high accuracy, this is not a reflection of how accurately they align with users' perceptions. Ideally, a controlled study focusing on content similarity would be conducted. The study would capture user ratings of the similarity between link text and text on destination pages. Study results could then be used to develop better measures of scent quality. This will be addressed with future work.

One major limitation of the link text measures is that the proximity of the associated link text (i.e., twelve good words before and after a link) is not considered. For example, it is possible for associated link text to be taken from a different paragraph or even a navigation bar, since the order that text appears in the HTML is used. Image processing is needed to ensure that the associated link text is actually part of the content surrounding a link.

5.13 Site Architecture Measures

Tables 5.21 and 5.22 summarize sixteen site architecture measures for assessing the following aspects of Web interfaces.

1. How consistent are page elements?
2. How consistent is the formatting of page elements?
3. How consistent is page formatting?
4. How consistent is page performance?
5. What is the overall consistency?
6. How big is the site? Big in this context refers to the breadth and depth of pages as well as the total number of pages based on crawling with the Site Crawler Tool.

5.13.1 Site Architecture Measures: Consistency

The consistency of page layout across the site has been discussed extensively in the literature [Flanders and Willis 1998; Fleming 1998; Nielsen 2000; Mahajan and Shneiderman 1997; Sano 1996; Sawyer *et al.* 2000]. Specific guidance includes the following.

- Consistent layout of graphical interfaces result in a 10–25% speedup in performance [Mahajan and Shneiderman 1997].
- Use consistent navigational elements [Flanders and Willis 1998; Fleming 1998].
- Use several layouts (e.g., one for each page type) for variation within the site [Sano 1996].
- Consistent elements become invisible [Sawyer *et al.* 2000].

These guidelines are obviously contradictory; hence, several measures for assessing site consistency were developed. Specifically, the variation in text elements and formatting, page formatting, page performance, and other aspects are computed. Variation is an inverse measure of consistency; larger variation indicates less consistency and vice versa. The variation measures are based on the Web site aspects presented in Figure 5.5. The average variation for measures within each block of Figure 5.5 is computed, as well as overall element variation (across the bottom row of measures), overall formatting variation (2nd and 3rd row), and overall variation (all rows except the top one). Recall that the scent quality measures include the page title hits and score to assess the similarity of page titles between pairs of pages (see Section 5.12.7); variation in these measures is also computed. The following process was followed in developing these site consistency measures.

1. Compute the Coefficient of Variation (CoV, $100 * \frac{\sigma}{\bar{x}}$, where σ is the standard deviation, and $\bar{x}$ is the mean) [Easton and McColl 1997] for each measure across all of the pages in a site. The CoV is a standard, unitless measure of the amount of variance in a data set.
2. Compute the median CoV for relevant measures within a category (text element, graphic formatting, and so on). The median does not require data that is normally distributed with equal variances; thus, it is more appropriate in this case than the mean. The median CoV is reported as the variation measure (i.e., text element variation, graphic formatting variation, etc.).

Measure	Description	Aspects Assessed				Accuracy		
		ID	ND	GD	ED	Hit	Miss	Avg.
How consistent are page elements?								
Text Element Variation	Variation in text elements across pages				√	–	–	100.0%
Page Title Variation	Variation in page titles across pages		√		√	–	–	100.0%
Link Element Variation	Variation in link elements across pages				√	–	–	100.0%
Graphic Element Variation	Variation in graphic elements across pages				√	–	–	100.0%
How consistent is formatting of page elements?								
Text Formatting Variation	Variation in text formatting across pages				√	–	–	100.0%
Link Formatting Variation	Variation in link formatting across pages				√	–	–	100.0%
Graphic Formatting Variation	Variation in graphic formatting across pages				√	–	–	100.0%
How consistent is page formatting?								
Page Formatting Variation	Variation in page formatting across pages				√	–	–	100.0%
How consistent is page performance?								
Page Performance Variation	Variation in page performance across pages				√	–	–	100.0%
What is the overall consistency?								
Overall Element Variation	Variation in all elements across pages				√	–	–	100.0%
Overall Formatting Variation	Variation in all formatting across pages				√	–	–	100.0%

Table 5.21: Summary of site architecture measures (Table 1 of 2). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

		Aspects Assessed				Accuracy		
Measure	**Description**	ID	ND	GD	ED	Hit	Miss	Avg.
What is the overall consistency?								
Overall Variation	Variation in elements, formatting, and performance across pages				√	–	–	100.0%
How big is the site?								
Page Count	Number of crawled pages				√	–	–	100.0%
Maximum Page Depth	Maximum crawl depth				√	–	–	100.0%
Maximum Page Breadth	Maximum pages crawled at a level				√	–	–	100.0%
Median Page Breadth	Median pages crawled across levels				√	–	–	100.0%

Table 5.22: Summary of site architecture measures (Table 2 of 2). The aspects assessed – information design (ID), navigation design (ND), graphic design (GD), and experience design (ED) – are denoted with a √. Hit and miss accuracies are only reported for discriminating measures.

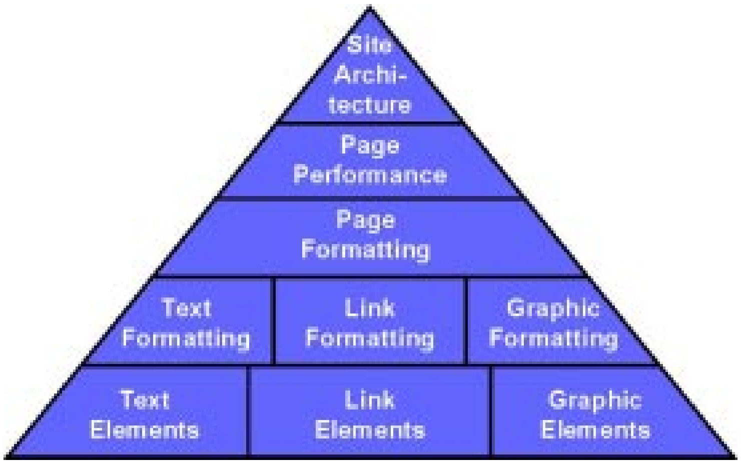

Figure 5.5: Aspects associated with Web interface structure. This is a repetition of Figure 5.2.

Site variation measures are only computed for sites with at least 5 pages of metrics data. Data from 333 sites discussed in Chapter 6 was explored to determine the median CoV for each page-level measure. For each group of measures, metrics that exhibited median Coefficients of Variation smaller than 250% were eliminated, since they could potentially dampen the variation in the remaining measures. This cutoff eliminated nine text element measures, including the Word, Page Title, and Good Body Word Counts discussed in Sections 5.4.1 and 5.4.2. Page depth and type as well as the term measures used for scent quality assessment (e.g., Visible Page Text Terms and All Unique Link Text Terms; Section 5.12.7) were also excluded, since they are intended to be used for reference. Page Title Variation only considers the page title hits and scores; this provides some insight about the use of different page titles throughout the site [Nielsen 2000].

Although the site variation measures are computed with high accuracy, this is not a reflection of how accurately the measures align with users' perceptions. How users gauge site consistency will be verified with a future study. Analysis in Chapter 6 should identify a subset of measures for these studies. Future work will also entail measuring the consistent use of image, script, stylesheet, and other files across pages in a site.

Another limitation of the site variation measures is that all types of pages (home, link, and content) are grouped together in computations. For example, link pages may have a substantially larger number of links than content pages, which could inflate the link element variation. A better approach would be to assess the variation within home, link, content, form, and other pages separately and then use a weighting factor to aggregate across page types. Unfortunately, this computation would require at least five pages of each type, which may not always be possible.

5.13.2 Site Architecture Measures: Size

The size of a site (i.e., the number of pages or documents) is used in the literature for classifying sites into genres, such as newspaper, course outline, and book [Bauer and Scharl 2000; Shneiderman 1997]. In addition, the breadth (how many links are presented on a page) and depth (how many levels must be traversed to find information) of pages within the site has been associated with the quality of the information architecture [Chi *et al.* 2000; Fleming 1998; Furnas 1997; Larson and Czerwinski 1998; Miller and Remington 2000; Nielsen 2000; Rosenfeld and Morville 1998; Sawyer *et al.* 2000; Spool *et al.* 1999; Spool *et al.* 2000]. Several usability studies have been conducted to provide guidance about the breadth and depth of sites. For example, Larson and Czerwinski [1998] suggest that Web designers use moderate levels of breadth with minimal depth (e.g., two levels) in the information architecture. As another example, Rosenfeld and Morville [1998] suggest that Web designers minimize the number of options on the home page to ten and minimize depth to less than five levels.

Several measures were developed to possibly provide some insight into the size, breadth, and depth of a site, including the Page Count, Maximum Page Depth, and Maximum and Median Page Breadths. These measures are based on the configuration of the crawler at the time of data collection as well as the number of pages for which page-level metrics are computed. Hence, they may not accurately reflect actual site characteristics even though they are computed accurately. If the crawler is configured for unlimited crawling and page-level metrics are computed for all of the downloaded pages, then these measures will actually reflect the degree to which the site conformed to crawler restrictions (e.g., pages at subsequent levels were not visible at the previous level, pages are not advertisements, guestbook, or chat room pages, and pages are not pdf, Word, or other documents). Future work will involve capturing more definitive measures of site size.

5.14 Summary

This chapter presented a view of Web interface structure and 157 highly-accurate, page-level and site-level measures for quantifying many aspects, including the amount and type of text on a page, page download speed, colors and fonts used on the page, similarity in content between pairs of pages, and the amount of variation in measures across pages in a site. Quantifying Web interface aspects makes it possible to develop statistical models for distinguishing good interfaces; this is the subject of the next chapter. As shown in this chapter, design guidance can sometimes be contradictory; quantifying Web interface aspects also makes it possible to provide concrete design guidance.

Subsequent chapters use quantitative measures computed for a large collection of expert-rated Web sites to develop several statistical models for assessing Web interface quality. Chapter 7 describes a study on linking the profiles to usability. Chapter 8 demonstrates the use of profiles to assess an example site, and Chapter 9 describes a study that examines the efficacy of the profiles in improving the design of the example site and four others. Finally, Chapter 10 demonstrates the use of statistical models to examine many Web design guidelines presented in this chapter.