Chapter 1

Introduction

Despite the abundance of design recommendations, recipes, and guidelines for building a usable Web site [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], usability, especially for information-centric Web sites, continues to be a pressing problem. Given that an estimated 90% of sites provide inadequate usability [Forrester Research 1999], steady growth in new sites [Internet Software Consortium 2001], and a severe shortage of user interface professionals to ensure usable sites [Nielsen 1999b], tools and methodologies are needed to accelerate and improve the Web site design process.

One way to ensure the usability of Web sites is via formal testing with users. Nielsen [1998a] claims that it takes 39 hours to usability test a Web site the first time, including planning the test, defining the test tasks, recruiting test participants, conducting a test with five participants, analyzing the results, and writing the report; with experience, this time can be reduced to 16 hours. Nielsen further claims that a usability test with five participants will typically reveal 80% of the site-level usability problems (e.g., home page, information architecture, navigation and search, linking strategy, etc.) and 50% of the page-level problems (e.g., understandability of headings, links, and graphics). The author advocates increasing page-level usability through other methods such as heuristic evaluation. Contrary to these findings, Spool and Schroeder [2001] have shown that five participants only find 35% of usability problems when the participants do not complete the same tasks. Thus, it appears that usability testing may not be a viable method for accelerating and improving the Web design process.

As a complement to usability testing, many detailed usability guidelines have been developed for both general user interfaces [Open Software Foundation 1991; Smith and Mosier 1986] and for Web page design [Comber 1995; Lynch and Horton 1999]. However, designers have historically experienced difficulties following design guidelines [Borges *et al.* 1996; de Souza and Bevan 1990; Lowgren and Nordqvist 1992; Smith 1986]. Guidelines are often stated at such a high level that it is unclear how to operationalize them. A typical example can be found in Fleming's book [Fleming 1998], which suggests ten principles of successful navigation design including: be easily learned, remain consistent, provide feedback, provide clear visual messages, and support users' goals and behaviors. Fleming also suggests differentiating design among sites intended for community, learning, information, shopping, identity, and entertainment. Although these goals align well with common sense, they are not justified with empirical evidence and are mute on actual implementation.

Other Web-based guidelines are more straightforward to implement. For example, Nielsen [1996] (updated in 1999 [Nielsen 1999a]) claims that the top ten mistakes of Web site design include using frames, long pages, non-standard link colors, and overly long download times. These are

apparently based on anecdotal observational evidence. Another essay by Nielsen [1997] provides guidelines on how to write for the Web, asserting that since users scan Web pages rather than read them, Web page design should aid scannability by using headlines, using colored text for emphasis, and using 50% less text (less than what is not stated) since it is more difficult to read on the screen than on paper. Although reasonable, guidelines like these are not usually supported with empirical evidence.

Furthermore, there is no general agreement about which Web design guidelines are correct. A survey of 21 Web design guidelines found little consistency among them [Ratner *et al.* 1996]. This might result from the fact that there is a lack of empirical validation for such guidelines. Surprisingly, no studies have derived Web design guidelines directly from Web sites that have been evaluated in some way, such as usability testing or heuristic evaluation. This dissertation presents the first attempt to analyze a large collection of example interfaces to develop statistical models for evaluating the quality of new interfaces (applied specifically to Web sites). Such an automated evaluation approach can potentially accelerate and improve the Web design process.

As background for the methodology and tools presented in this dissertation, Chapter 2 summarizes an extensive survey of usability evaluation methods for Web and graphical interfaces. It shows that automated methods are greatly underexplored in the Web domain and that existing methods require some form of usability testing to employ.

It is natural to question what existing and new automated approaches evaluate and furthermore, do they improve interface design. Chapter 2 discusses evaluation methods under the assumption that they all support usability evaluation to some degree. According to ISO9241, usability is the extent to which users can use a computer system to achieve specified goals effectively and efficiently while promoting feelings of satisfaction in a given context of use [International Standards Organization 1999]. Only methods that solicit user input, such as usability testing and surveys, enable the assessment of whether a site is usable according to this definition; thus, other methods (non-automated and automated) may not actually evaluate usability. However, what these methods may evaluate is conformance to usability principles, predicted usability, and possibly other aspects related to the usability of the interface; all of these aspects are important and can potentially increase the likelihood that an interface will be usable.

As discussed above, usability testing may require considerable time, effort, and money and may not reveal all of the problems with a site. Chapter 3 proposes new methods of automated usability evaluation based on measurement, analytical modeling, and simulation methods used in the performance evaluation domain, in particular for evaluating the performance of computer systems. The outcome of these methods is typically quantitative data that can be used to objectively compare systems.

The automated evaluation methodology developed in this dissertation is a synthesis of both performance evaluation and usability evaluation. In particular, the methodology consists of computing an extensive set of quantitative page-level and site-level measures for sites that have been rated by Internet professionals. These measures in conjunction with the expert ratings are used to derive statistical models of highly-rated Web interfaces. As is done with guideline review methods in the usability evaluation domain, the models are then used in the automated analysis of Web pages and sites. However, unlike other guideline review methods, the guidelines in this case are in essence derived from empirical data.

Chapter 4 summarizes the methodology and tools. Chapter 5 describes the page-level and site-level quantitative measures developed as the result of an extensive survey of the Web design literature, including texts written by experts and usability studies. The survey revealed a set of usability aspects, and Chapter 5 describes a total of 141 page-level and 16 site-level measures for assessing many of these aspects, such as the amount of text on a page, the number of colors used,

the download speed, and the consistency of pages in the site.

Chapter 6 describes the analysis of data from over 5300 pages and 330 sites to develop several statistical models for evaluating Web page and site quality. The models distinguish pages and sites that are rated as good, average, and poor by experts and make it possible to take into consideration the context (e.g., the page's functional style or the site's topical category) in which pages and sites are designed. These models include one to assess whether a page is a good home page and one to assess whether a site is a good health information site. The accuracy of page-level models range from 93%–96%, while the accuracy of site-level models range from 68%–88%; the site-level accuracy is considerably less, possibly due to inadequate data.

The methodology developed in this dissertation differs from other automated evaluation methods for Web interfaces in a major way – it is based on empirical data. In essence, the statistical models can be viewed as a reverse engineering of design decisions that went into producing highly-rated interfaces; presumably, these design decisions were informed by user input (e.g., usability testing or surveys). Nonetheless, two studies were conducted to provide insight about the questions posed above (what is evaluated and are designs improved). Chapter 7 describes a study that focused on linking expert ratings to usability ratings. Although the results suggest some relationship between expert and end user ratings, strong conclusions could not be drawn from the study due to problems with the study design.

Chapter 8 demonstrates use of the statistical models for assessing and improving an example site and shows that the model output does inform design improvements. It also provides more insight into what the profiles actually represent and the type of design changes informed by them. Chapter 9 presents findings from a study of this site and four others similarly modified based on model output; site modifications were made by the author and three students – two undergraduates and one graduate. The study focused on determining whether the statistical models help to improve Web designs. Thirteen participants (four professional Web designers, three non-professional designers who had built Web sites, and six participants who had no experience building Web sites) completed two types of tasks during this study: 1. explore alternative versions of Web pages and select the ones exhibiting the highest quality; and 2. explore pages from sites and rate the quality of the site. The results show that participants preferred pages modified based on the Web interface profiles over the original versions, and participants rated modified sites (including the example site) higher than the original sites; the differences were significant in both cases.

Chapter 10 demonstrates use of the statistical models for exploring existing Web design guidelines. The chapter examines contradictory or vague guidelines for nine aspects of Web interfaces, including the amount of text, font styles and sizes, colors, and consistency. The statistical models reveal quantitative thresholds that validate and in some cases invalidate advice in the literature. This examination shows that the methodology makes it possible to derive Web design guidelines directly from empirical data.