Chapter 11

Conclusions and Future Work

This dissertation makes several key contributions to the advancement of automated Web interface evaluation, including the following.

- It summarizes the current state of automated evaluation methods for graphical and Web interfaces and proposes ways to expand existing support.
- It describes how methods used for evaluating the performance of computer systems can be applied towards the problem of automated interface evaluation.
- It presents a methodology and tools for automated evaluation of Web interfaces that is a synthesis of approaches used in the usability and performance evaluation domains and based on empirical data.
- It describes an extensive set of quantitative Web interface measures for assessing many aspects discussed in the Web design literature.
- It describes the development of highly-accurate statistical models for assessing Web interface quality.
- It documents the efficacy of the statistical models for improving Web interface quality and for validating established Web design guidelines.

This dissertation presents the first research in which a large collection of expert-rated Web sites was analyzed to develop statistical models to support automated evaluation of new sites. This method represents a synthesis of approaches employed in the usability evaluation and performance evaluation domains. It entails computing an extensive set of 157 highly-accurate, quantitative page-level and site-level measures for sites that have been rated by Internet experts. 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.

This dissertation shows that highly accurate models can be developed to assess Web page and site quality while taking into consideration the context in which pages and sites are designed. For example, models were developed to assess whether a page is a good home page or whether a page is consistent with pages on good health sites; site-level models were developed to enable similar assessments. A usability study suggests that the expert ratings used to derive the models are somewhat consistent with usability ratings; however, concrete conclusions cannot be drawn

from the study due to the time difference between expert and usability ratings. It was also shown that the models could be used to guide the modification of an example Web site, and another study showed that participants rated the modified version of this site slightly higher than the original version, although the difference was not significant. However, the study showed that for a larger set of sites, participants preferred pages modified based on the Web interface profiles over the original versions, and participants rated modified sites higher than the original sites; the differences were significant in both cases. Thus, this dissertation shows the first steps towards using the models to provide concrete Web design guidance.

Even though it was possible to find correlations between values for measures and expert ratings, no claim is being made about the profiles representing causal links. It is possible that the highly-rated sites are highly rated for reasons other than what is assessed with the measures, such as the quality of the content of the site. The current models and tools cannot improve on poor content. However, the study of modified sites provided preliminary evidence that they can provide insight on how to take good content that is poorly presented and improve its presentation, thus improving users' experience in accessing that content. And, because it is possible to empirically find commonalities among the presentation elements of the highly-rated sites, this provides strong evidence that the presentational aspects of highly-rated sites that differ from those of poorly-rated sites are in fact important for good design.

This dissertation represents an important first step towards enabling non-profes- sional designers to iteratively improve the quality of their Web designs. The methodology and tools are in their infancy and only provide support for refining an implemented site; thus, there are many ways in which they can be improved. Ideally, predictive models would be developed based on usability test results instead of expert ratings; however, a large effort needs to be launched within the HCI community to secure a large sample of usability tested sites. Developing and deploying models is very time and resource intensive; thus, automating some aspects of these activities would be extremely helpful. Even after developing and deploying models, considerable work needs to be done to fully understand and validate design principles gleaned from them.

The set of quantitative measures can also be expanded to measure other aspects discussed in this dissertation, such as the reuse of Web interface elements across pages in a site. A major limitation of the current set of measures is that they do not assess content quality. Future work will explore using text analysis techniques to possibly derive other measures of content quality. Another limitation of the measures is that they do not gauge accessibility for the disabled; it was shown that the good Web pages tended not to be accessible as determined by the Bobby tool. Future work will examine other ways to measure accessibility, for instance computing the nesting level of tables. (Although tables may help sighted users scan pages, they may impede blind users.)

Image processing could be used to improve the accuracy of existing measures, to enable the development of new ones, and to enable support for non-HTML pages and early design representations. Supporting early design representations also requires adjustments to be made to the profiles, such as ignoring certain measures during analysis.

All of the developed tools need to be reimplemented as part of a robust, open source browser, such as Mozilla or Opera; this will enable support for framesets, scripts, applets, and other objects as well as real-time analysis. Real-time analysis is crucial for developing an interactive evaluation tool to support iterative design and evaluation.

Other key components of the interactive evaluation tool include: recommending design improvements based on model predictions; applying recommendations so users can preview the changes; and showing comparable designs for exploration. Some of the model deviations are easy to correct, such as removing or changing text formatting, using good color combinations, and resizing images; it is possible to automatically modify the HTML to incorporate these types of changes.

Other changes, such as reducing vertical scrolling, adding text columns, improving readability, and adding links are not as straightforward. More work needs to be done to better understand the profiles before interactive evaluation can be supported. For example, factor analysis or multidimensional scaling techniques could be used to reduce the number of measures and to gain more insight about relationships among measures. This would enable recommendations to be based on combinations of measures versus individual measures.

A natural extension of this work is to enable profiles to be developed to capture effective design practices in other cultures. This would require support for non-English languages, mainly using language-appropriate dictionaries and text analysis algorithms. This may also require changes to the assessment of good color usage, since colors have specific meanings in some cultures.

Another extension of this work would be to develop a quality summary for Web pages and sites. This summary would represent a quick overview or characterization (possibly in graphical format) that could be used by search engines for ordering search results or possibly in visualizations of clickstreams through sites. The summary for search engine ordering may consist of a single rating, while the summary for the latter case may consist of a rating in addition to other details, such as the predicted page type, the closest good page cluster, and download speed.

Another application of the corpus of Web pages and sites is to enable designers to explore the collection to inform design. Ideally, characteristics of Web pages and sites could be represented in a way that facilitates easily identifying pages and sites that satisfy some design criteria, such as effective navigation schemes within health sites or good page layouts, color palettes, and site maps. Task-based search techniques that exploit metadata [Elliott 2001; English *et al.* 2001; Hearst 2000] should be helpful; metadata for Web interfaces could consist of the quantitative measures developed in this dissertation as well as others that describe for instance the size of the site, the type of site, page size, a page's functional type, elements on a page (e.g., navigation bars), as well as site ratings.

Another extension of this work is to use the profiles to derive parameters for the Web interface simulator proposed in this dissertation. Monte Carlo simulation was suggested as a way to automate Web site evaluation by mimicking users' information-seeking behavior and estimating navigation time, errors, etc. Similar to the guideline derivation, profiles could be used to derive simulation model parameters, such as thinking and reading times for pages. The cluster models could be used to derive timing estimates for these activities based on the average number of links, words, and other measures that reflect page complexity. User studies would be conducted to validate the baselines before incorporating them into the simulator.

This dissertation lays the foundation for a new approach to automated evaluation of interfaces wherein the interfaces themselves are analyzed as data. Continuous, ongoing analysis of interfaces will enable the models embedded in the approach to evolve and constantly reflect the current state of effective design. This approach is not intended to be used as a substitute for user input. Automated methods do not capture important qualitative and subjective information that can only be unveiled via usability testing and other inquiry methods. Furthermore, it is not the case that the issues identified by automated tools are true usability issues. Several studies, such as the one conducted by Bailey *et al.* [1992], have contrasted expert reviews and usability testing and found little overlap in findings between the two methods. Nonetheless, this approach should be a useful complement to non-automated evaluation techniques.