An Acceptability Predictor for Websites

Ray Adams, Anthony White, and Efe Ceylan

School of Engineering & Information Sciences, Middlesex University,
The Burroughs, Hendon, London NW4 4BT, UK
{ray.adams,a.white,e.ceylan}@mdx.ac.uk

Abstract. User acceptance is a high priority for website design and implementation. Two significant, but largely separate, approaches to acceptability are: First, the Web Accessibility Initiative (WAI) has explored the measurement of technical features of a website to gauge its accessibility. Second, human judgments about acceptability are obtained from intended users or experts. The present work explores the important question of how best to combine these two methods. Experiment One required new users to explore automatic website evaluation systems. They found two of four systems difficult or impossible to use and system outputs difficult to understand. Experiment Two combines formal properties and user judgments, using an automatic system to predict user judgments from formal website properties. A simple system was able to predict user judgments within 91% accuracy. Clearly, user judgments about websites can be predicted reliably, a result of value to designers.

Keywords: Prediction, acceptability, usability, accessibility, websites.

1 Introduction

Websites present a double-edged challenge [7]. First, there is the potential that a good quality, acceptable website can be achieved, but, second, web designs that overemphasize functionality and aesthetics at the expense of user acceptability may find that they are not very popular with users [1][2].

Furthermore, the World Wide Web is changing dramatically. We are moving on from the standard type of website, so that we now see social networking sites [3], avatar based systems [4], digital library portals [5] and semantic websites [6]. The danger is that, paradoxically, some innovative modes of human / system interactions, intended to provide users with new opportunities and so create new markets, may often create new usability and accessibility problems at the same time as they introduce new functions and novelties! Famously, the invention of the graphical user interface (GUI) introduced many users to innovative, new ways of interacting with computer systems, but locked out many users who were blind or who had limited vision. Screen reader software applications were particularly badly hit. They are still suffering from web sites that are designed without more than a moment's thought for screen readers. Innovations need careful and systematic introduction based on user-sensitive research. In our view, the GUI also made plagiarism much easier and more acceptable by supporting the "copy and paste" attitude to information.

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2009, LNCS 5616, pp. 628–634, 2009.

Current work is developing new design methods for systems innovations based on advanced user models, with systems having novel interfaces for maximum impact and accessibility, including work on a brain computer interface. It is ever more important to measure the acceptability, usability and accessibility of new or changing websites. There are at least two significant approaches to web accessibility. Whilst both are highly relevant and useful, yet they are rarely combined as they each require a substantial effort to implement.

Our interest is to explore the extent that we can bring together formal properties of websites with user judgments about user acceptance and the user experience. The questions asked here are not only "How can these two important approaches be integrated?", but also "How can this be accomplished in such a way that users, practitioners and researchers find to be accessible?" It would be an irony if systems designed to evaluate the acceptability of websites were presented themselves as unacceptable websites! In our time, we have used a number of free, online, automatic systems that aspire to measure the acceptability of web sites in general and accessibility in particular. Our subjective impression was that some such systems were often prone to failure, difficult to use, whilst providing feedback that only faculty members could fully appreciate. We set out to test the accessibility of free, online, automatic systems that measure accessibility in Experiment One, using relatively naïve participants. In addition, we were aware that artificial networks could be used to predict the values of one set of variables from the values of a second set of variables. If so, we ask "Can a user's judgment be predicted from a number of measurable attributes of a website, such as number of links?" Equally, we asked: " Can we use relatively simple software to do so?" That is the objective of Experiment Two.

2 Experiment One

The objective of this first experiment is to evaluate a sample of free, automatic, online website acceptability measurement systems.

2.1 Methods

Four systems were selected at random from the list of automatic accessibility measurement systems provided at http://www.w3.org/WAI/ER/tools/. (This URL was last checked for accessibility on Thursday, February 26, 2009). There were nine participants, aged between 20 and 30 years of age, seven males and two females. None had prior experience of automatic accessibility measurement systems. They were all first year students on a four-year degree programme in computing science. Each participant worked alone. They were asked to select any two websites that they wished to test with these systems. They were given a brief, non-technical explanation of accessibility and of automatic measurement system. They were then asked to use each system to evaluate their two websites. They were told that there were no time limits and to persevere if they encountered problems. However, if a system proved impossible to use, they could terminate their attempt. They were asked to comment on the accessibility of the outputs produced by the systems. These results were tested against the null hypothesis of zero failures as shown in table one.

2.2 Results

Each of nine participants evaluated two websites of their choice against four automatic assessment systems, giving seventy-two data points in all. For the first analysis, the system responses were classified as a success or a failure. Second, a stricter criterion of success was adopted, where the output was not understood. The results are summarized in table one below. The observed frequencies of row 2 were compared against the expected frequencies of the null hypothesis (row 4) by a chi-squared test. This comparison was statistically significant (χ = 39.56, df = 3, $p < 0.001$). Next, the observed frequencies of row 3 were compared against the expected frequencies of the null hypothesis (row 4) by a chi-squared test. This comparison was also statistically significant (χ = 44.28, df = 3, $p < 0.001$).

Table 1. Participant Successes with Automated Accessibility Systems

Row 1	Measure	System 1	System 2	System 3	System 4
Row 2	Successes (Out of 18)	18	0	0	10
Row 3	Strict successes	11	0	0	8
Row 4	Null hypothesis (No failures)	18	18	18	18

2.3 Discussion

We had expected that our sample of automated accessibility measurement would work every time. To our surprise, the "success" rate (row 2 of table one) was significantly worse than we had expected. What's more, of those who did manage to obtain significant output, only a smaller number (row 3 of table one) were able to consider which of their two websites could be judged to be better or, at least, the more acceptable. Both chi-squared statistics were highly significant, so we do not consider this a chance result. Neither, can this cannot be due to artifacts associated with the choice of specific websites, since each participant selected their own (different) pair of websites to evaluate. Thus, there was a general indication that system output was sometimes difficult to obtain, but when obtained, was often relatively inaccessible to new users, being couched in technical terms, with little or no narrative that they could use to understand the outputs better. Of course, we should point out that we are discussing the accessibility of these systems to *relatively naïve users.* We speculate that some problems were clearly down to inexperience, others to hardware and software installations that the software providers could not anticipate. We urge caution when applying these results to all levels of expertise. The present data are sufficient to raise the concern that automatic evaluation systems of accessibility may not automatically be accessible. But they do not provide further evidence. Clearly, we would need to extend the present work to larger and more diverse populations of users to delineate the different types of accessibility problem involved.

We now explore other ways to evaluate website acceptability. How can we predict the accessibility judgments of users? If we could do so, that would be very useful for evaluating new websites. We were aware that artificial networks could be used to predict the values of one set of variables from the values of a second set of variables. If so, we ask "Can a user's judgment be predicted from a number of measurable attributes of a website, such as number of links?" Equally, we asked: " Can we use relatively simple software to do so?" That is the objective of Experiment Two.

3 Experiment Two

The objective of the second experiment is to explore the possibility of using a simple off-the-shelf software application to predict the judgments of users about the acceptability of websites. This simple experiment was conducted as a proof-of-concept for the notion that human user acceptability judgments about websites can be predicted from key attributes (see below) of those websites. For this purpose, we have selected the Braincel software application. This software is a plug-in for Excel® spreadsheet application. Braincel acts rather like artificial neural networks, but without the attended complexity that goes along with their additional power. Braincel was interesting as it seemed relatively easy to learn. Braincel can be trained to learn the relationships between variables specified in the spreadsheet. The plug-in can be trained to predict one set of variable from the others. Thus the experiment consisted on a training phrase for Braincel, using human and website data together, followed by a Braincel performance phase in which its learning was put to the test.

3.1 Methods

As mentioned above, it is important to be clear that the use of Braincel requires a two-stage methodology. The first stage is the training phase and the second stage is the performance stage.

In the first stage, Braincel is trained to related one set of variables to another set of variables, so that, when trained, it can use values of the former variables to predict the values of the second set of variables. In this case, the variable to be predicted is website user acceptability ratings. The predictions are based on attributes of the website i.e. text cluster count, link count, page size, graphics count and colour count. For this stage, twenty-five participants were recruited to evaluate six websites. Each participant completed an acceptability questionnaire for each website, as shown in appendix one. These questionnaire scores could be broken down into four components; namely ease of use, efficiency of site, likeability of site and user experience / feelings. In the second phase, Braincel was used to predict user acceptance judgments on the basis of the five attributes of the websites i.e. text cluster count, link count, page size, graphics count and colour count.

3.2 Results

In the training phase, Braincel learns the relationships between the average scores of the four aspects of user, judged acceptability and the average scores of the five attributes of the websites i.e. text cluster count, link count, page size, graphics count and

colour count. As a measure of training efficiency of the system, we compared the actual scores of the participants against the scores predicted by Braincel. This led to a modest error rate of 6.30%. Next, the Braincel system was tested on a further set of six websites. Here the error rate was still low at 8.55%. Both error rates were significantly better than chance ($p<0.001$). Thus it is clear that the Braincel has been able to pick up the relationships between the variables considered here. Important features of these results are (a) that we have been able to combine formal web features with user judgments and (b) we have used a relatively simple application to do so.

3.3 Discussion

The results indicate that a simple software application can be used successfully to predict user judgments about the acceptability of websites. Excel® spreadsheet plug-in called Braincel was used in this study and the resulting evidence from this simple experiment has provided a proof-of-concept for the notion that human user acceptability judgments about websites can be predicted from key attributes of those websites. In the present study, we used a significant number of participants (n=25) to generate our user judgments of acceptability of the chosen (n=6) websites. This may not always be a realistic option for busy designers. In that case, we would recommend a smaller sample or the use of realistic user models. Despite the statistical significance of the present data, we see the need to explore further both the effect of sample sizes, user judgments and the chosen website parameters. In our case, background research suggested the value of the following user judgments: namely ease of use, efficiency of site, likeability of site and user experience / feelings. Further work is needed to investigate to see if our present choice of user judgments would be specific to these exact judgments or is generalisable to other types of user judgments. Similarly, background research suggested the following website parameters: i.e. text cluster count, link count, page size, graphics count and colour count. Here again, further work is needed to investigate to see if our present choice of website parameters would be specific to these exact judgments or is generalisable to other types of website parameters. Nevertheless, the present results, from experiment two particularly, are showing promising indications of the way forward for the synthesis of formal web attributes as a basis for user acceptability and metrics based on user judgments about user accessibility.

4 Discussion and Conclusions

Both sets of data were surprising and, thus, informative. In Experiment One, naïve users found the automatic evaluation systems sufficiently more difficult than we would have predicted. Clearly more work with other samples of websites would be helpful. For naïve users, it may be necessary to provide built in training for them or even consider redesigning the interfaces to make them more accessible. The statistical significance of the present results would make the probability of these being chance results to be seen as unlikely. In the second experiment, the data were also surprising, such that the error rates associated with both the learning phase and the performance phase were encouragingly low. The appeal is that as the system becomes more experienced with additional websites, more website parameters and more

user judgments, it should show a significant learning curve and become better and more relevant for website design for acceptability.

In conclusion, this research has generated two important conclusions. First, automatic assessment tools are not for naïve users, unless either the users are given training and / or the interfaces and contents are made much more accessible. Second, I simple predictive system can apparently predict human judgments from web-site attributes with an acceptable level of accuracy without requiring extensive training. These demonstrations are notable. Now that they have been established at the proof-of-concept level, we need to explore the parameters of these findings further.

References

1. Adams, R.: Natural computing and interactive system design. Pearson, London (2005)
2. Adams, R., Smith-Atakan, A.S.: Human-computer interaction: users, tasks and designs. Middlesex University Press, London (2005)
3. Adams, R., Gill, S. (eds.): HCI 2007. LNCS, vol. 4554, pp. 584–592. Springer, Heidelberg (2007)
4. Adams, R.: Evaluating the next generation of multimedia software. In: Tsihrintzis, G.A., Virvou, M., Howlett, R.J., Jain, L.C. (eds.) New Directions in intelligent interactive multimedia. Studies in computational intelligence, vol. 142. Springer, Heidelberg (2008)
5. Adams, R., Granić, A.: Technology Enabled Learning Worlds. In: Lazinica, A. (ed.) Advances in Human-Computer Interaction, I-Tech., Vienna (2008)
6. Adams, R.: User modelling & monitoring. In: Stephanidis, C. (ed.) The Universal Access Handbook (LEA) (in press)
7. Moseley, R.: Developing Web Applications. John Wiley and Sons, London (2007)

Appendix

Questionnaire items
Please rate your disagreement / agreement with the following on the 10 point scales provided.
1. Annoying
2. Confusing
3. Frustrating
4. Interesting
5. Stimulating
6. Tiresome
7. Useable
8. Unpleasant
9. I feel in control when I am using this site
10. This site uses terms understandable and familiar to me.
11. This site needs more introductory explanations
12. I find this site useful
13. Everything on this site is easy to understand
14. This site is too slow
15. I get what I expect when I click on objects on the site

16. I find it difficult to move around this site
17. I feel efficient when using this site
18. Compared to what I expected, the tasks did go really quickly
19.I will characterize this site as an innovative one
20.Overall, I am quite satisfied with this site.