Prediction of Web Page Accessibility Based on Structural and Textual Features

Sina Bahram, Debadeep Sen, Robert St. Amant
Department of Computer Science
North Carolina State University
Raleigh, NC, US
sbahram, dsen, stamant@ncsu.edu

ABSTRACT

In this paper we present an approach to assessing the accessibility of Web pages, based on machine learning techniques. We are interested in the question of whether there are structural and textual features of Web pages, independent of explicit accessibility concerns, that nevertheless influence their usability for people with vision impairment. We describe three datasets, each containing a set of features corresponding to Web pages that are "Accessible" or "Inaccessible". Three classifiers are used to predict the category of these Web pages. Preliminary results are promising; they suggest the possibility of automated classification of Web pages with respect to accessibility.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Evaluation/methodology; K.4.2 [**Computers and Society**]: Social Issues–Assistive technologies for persons with disabilities

General Terms

Human Factors, Measurement

Keywords

Web pages, accessibility, vision impairment, machine learning

1. INTRODUCTION

Assessment of the accessibility of Web pages for people with vision impairment (PWVI) has posed challenges for over a decade [7]. Analysis of Web pages in an accessibility context is typically driven by accessibility guidelines, such as those provided by the Section 508 of the U.S. Rehabilitation Act[1] and by the Web Accessibility Initiative of the World Wide Web Consortium[2]. Unfortunately, as Takagi et al. [12] observe, a focus on adherence to guidelines is not enough to ensure that Web pages are actually usable.

Metrics have been developed to assess Web page accessibility in quantitative terms [7]. Further, concentrating on accessibility guidelines leaves a related but (in our view) more basic question unaddressed: Are there structural and textual features of Web pages, independent of explicit accessibility concerns, that nevertheless influence their usability for PWVI?

In this paper we present an approach to assessing the accessibility of Web pages in an effort to give a preliminary answer to this question. The skeleton of our approach is straightforward: compile a set of Web pages, labeled as "Accessible" or "Inaccessible"; run machine learning classifiers on features of the Web pages to see how well the label can be predicted; examine the classifiers to try to gain insight into the prediction of the accessibility or inaccessibility of Web pages in general.

Our analysis is based on two compilations of Web pages, organized into three datasets: a selection of academic Web pages, a selection of non-academic, mostly commercial, websites, and The combination of these two.

A range of structural and textual features was automatically extracted from the Document Object Model (DOM) of each Web page. We chose three classification algorithms (a decision tree, a Bayesian network, and a support vector machine (SVM)) and generated classifiers for each of the datasets, using the accessibility category of the Web pages as the target.

We then performed a comparison of the results with respect to classification performance. Both compilations contained approximately the same number of "Accessible" and "Inaccessible" pages, which means that an uninformed decision rule should produce a prediction accuracy of about 50% for all three datasets. The best classifiers identified in our analysis performed better. For the academic dataset, the Bayesian network classifier averaged a true positive rate of 80%, with the others averaging around 60%. The non-academic dataset proved more difficult, with the SVM averaging 63.5%, the decision tree 60% and the Bayesian network 48%. On the combined dataset the Bayesian network performed the best, averaging 71%, with the decision tree at 64% and the support vector machine at 61%. Not all of these scores are high, but we believe they represent good progress and will improve as we collect more data.

[1]http://www.section508.gov

[2]http://www.w3.org/WAI/guid-tech.html

W4A2011 - Communications paper, March 28-29, 2011, Hyderabad, India.
Co-Located with the 20th International World Wide Web Conference.

We then applied another metric to the Web pages in each dataset: Takagi et al.'s "time to reach" function, part of the aDesigner suite [13]. Surprisingly, time-to-reach does not perform as well in predicting the accessibility of Web pages in general.

This paper provides three main contributions. The first is the new datasets containing the features of Web pages used in our analysis. We are unaware of a comparable public repository. Lists of accessible and inaccessible Web pages exist, but not with the associated information we provide. The second contribution is the identification of candidate features relevant to accessibility that go beyond those commonly included in accessibility guidelines. Further, our analysis identifies models based on these features that are predictive of accessibility. The last contribution is methodological: We present a new approach to assessing the accessibility of Web pages, with promising results in our preliminary testing. Given the data-driven nature of this work, we expect that the classification will improve as we collect more data and refine our classification algorithms.

2. RELATED WORK

The work in this paper draws on influences in a few different research areas. The first area is machine learning and data mining techniques as applied in accessibility research. We rely on these techniques to provide guidance for identifying the relevant characteristics of usable Web pages, aside from the presence of accessibility-relevant mark-up. Machine learning has been suggested as a promising approach for improving accessibility [2, 10], and some systems have shown significant success. For example, Kottapally et al. [8] use inductive logic programming and hidden Markov models to infer information about HTML tables and frames. TrailBlazer [3] relies on a naive Bayes classifier to rank suggestions made to users about how to carry out tasks on the Web, using scripts from CoScripter [9]. The HearSay browser [11] uses support vector machines to identify relevant content on successive pages visited by a user, in service of context-directed browsing. HeadingHunter [4] uses a decision tree classifier to identify Web page elements as headings, based on visual and relational features. HeadingHunter is further notable in providing a practical way of converting such classifiers into JavaScript code appropriate for transcoding.

Considerable research has also been devoted to identifying properties of Web pages based on structural and textual features for the purposes of content extraction. Yi et al. [14] describe entropy-based measures for separating and identifying main content blocks on a page, distinguishing them from navigation links, advertising, and so forth. Part of the functionality of OntoMiner [6] is to identify main content, navigation panels, and advertising on a page, using hierarchical clustering techniques. Webstemmer [15] extracts the main text from Web news articles using inter-page clustering by layout. Such work has reached commercial software, as in Readability (lab.arc90.com/2009/03/02/readability), which strips out superfluous elements from some types of Web pages to leave only primary content.

A third area is work on quantitative metrics for Web accessibility. Freire et al. give a good survey of recent progress in the area [7]. Most such metrics are based on the concepts of points of failure and potential barriers. As an example, Freire et al. offer the case of an image without alternative text: this is an accessibility barrier, and all images are thus potential points of failure. The work of Bühler et al. [5] is particularly important in its validation of a specific metric through experimentation with users.

3. DATA COLLECTION

We collected Web pages from the sites listed in the appendix of this paper. The "academic" Web pages are a selection of those compiled by Jon Gunderson for the Chronicle of Higher Education[3]. We used the home pages of the top and bottom 26 sites in this list, labeled "Accessible" and "Inaccessible", respectively, giving 52 Web pages in total.

The other 52, "non-academic" Web pages were compiled from the results of an open-ended WebAIM survey[4] and a comparable open-ended survey sent to members of the gui-talk@nfbnet.org, ProgrammingBlind@freelists.org, and blindwebbers@yahoogroups.com mailing lists. Additional pages were identified by one of the co-authors of this paper. Labels of "Accessible" and "Inaccessible" were based on the respondents' judgments. total was also 52 pages.

Each Website in the appendix has superscripts indicating its source: m for the mailing list surveys, w for the WebAIM survey, s for our own collection, and c for the Chronicle of Higher Education compilation.

The procedure for converting each Web page to a set of features begins by loading the page into a specialized browser (based on the Microsoft .NET v3.5 framework). The generated DOM of the page is walked, to compute the features of the page. These features are recorded in an XML file, which is further converted into a form appropriate for further processing.

Three datasets were constructed through this process: an academic dataset, a non-academic dataset, and their union in a combined dataset.

4. ANALYSIS

Processing of the datasets was done in Weka, the open source collection of machine learning algorithms[5]. We applied three classification algorithms: the J48 decision tree classifier, the BayesNet classifier, and the Sequential Minimal Optimization support vector machine classifier. In each case we used a nominal feature with the two values "Accessible" and "Inaccessible" as the target label, and we did not alter the default parameters passed to each classification algorithm.

We chose these three classification techniques as being representative of successful approaches in machine learning as well as in accessibility. Decision trees are used in HeadingHunter [4]; they are simple but often effective in data mining. One attractive feature is their hierarchical decomposition of patterns in a dataset: in our dataset, it may be the case that different types of Web pages should be interpreted in different ways with respect to accessibility, and a decision tree may be able to capture such differences. Bayesian classifiers have also been used in accessibility research, as seen in TrailBlazer [3]. The potential here is that a Bayesian network may capture probabilistic causal relationships between Web page features, to predict accessibility.

[3]http://www.chronicle.com/article/BestWorst-College-Web/125642
[4]http://www.webaim.org/projects/screenreadersurvey
[5]http://www.cs.waikato.ac.nz/ml/weka/

The dataset of non-academic Web pages contained 27 "Accessible" and 25 "Inaccessible" Web pages. The simplest classification rule is to choose the majority element, which always classifies a Web page as "Accessible", with an accuracy of about 52%.[6] The dataset of academic Web pages is evenly divided between "Accessible" and "Inaccessible" pages, for which the majority element decision rule will be accurate 50% of the time.

For each classification algorithm and each dataset we performed a stratified, 10-fold cross-validation, using 80% of the data for training and the remaining 20% for testing. The decision tree classified 59.6% of the academic instances correctly, the same for the non-academic instances, and 64.4% of the combined instances correctly. Precision and recall are moderate for the combined dataset, both at 64.4%, producing an F measure of 0.644. These numbers are not high, but they are better than we might expect by using a majority rule classifier.

The Bayesian network performed better overall, correctly classifying 80.0% of the academic instances, 48.1% of the non-academic instances, and 71.2% of the combined instances. Precision and recall for the combined dataset are 71.3% and 71.2%, with an F measure of 0.711.

The SVM produced results comparable to those of the decision tree. The SVM correctly classified 59.6% of the academic instances, 63.5% of the non-academic instances, and 60.6% of the combined instances. Precision and recall for the combined dataset are 61.1% and 60.6%, with an F measure of 0.599.

The best results are generated by the Bayesian network classifier, with an average accuracy over all datasets of 71.2% in correct predictions, with 28.7% false positives. This is far from perfect, but it is respectable. However, the classifier generated a flat network in which all 26 attributes are used to directly predict accessibility. This gives us relatively little insight into the features that influence accessibility, in general.

Notice that performance on the academic and non-academic datasets is below the performance on the combined dataset. There are two potential explanations for this pattern.

First, there may be differences between the two datasets; their combination may allow the decision tree to take advantage of different structural/textual features in different ways within different partitions.

A second part of the explanation for the improved performance of the decision tree classifier on the combined dataset is that there is simply more data to work with: 104 instances, the sum of instances in the academic and non-academic datasets. We constructed new combined datasets of size 52 by sampling equally from the academic and non-academic instances. When we ran the decision tree classifier on these new combined datasets, the performance improvement disappeared. This indicates that even though the decision tree is able to generalize over differing patterns in the academic and non-academic datasets, it's dependent on the size of the combined dataset: more data, for this classifier, produces better performance.

5. EXTERNAL COMPARISON

It is reasonable to ask whether our results replicate work that could have been carried out more easily with existing tools. It seems intuitively obvious that the structure and content of Web pages should influence their accessibility, independent of their adherence to accessibility guidelines. While a number of quantitative metrics for Web accessibility exist, such as those surveyed by Freire et al. [7], one drawback in using them in this paper is their lack of validation with users; instead, they mainly rely on coupling with guidelines.

An alternative analysis could rely on the time-to-reach heuristic of Takagi et al. [12, 13], which is more closely tied to the dynamics of interaction than other metrics, relies on the structure of Web pages, and has been evaluated (and within the context of aDesigner, validated) on a sample of Web pages comparable in size to the datasets in this paper.

We were surprised to find, with one strong exception, time-to-reach alone provides little predictive power in categorizing the Web pages in our datasets. We computed time-to-reach using aDesigner[7], for all the elements in each of the Web pages in our datasets. For each Web page, we calculated three time-to-reach values: the mean, median, and maximum. While the differences between the values match our expectations—time-to-reach is in the main lower for accessible Web pages—there is sufficient spread to make the use of time-to-reach alone problematic as a predictor.

For the non-academic and combined datasets, logistic regressions using each of these metrics individually produce χ^2 values close to zero in each case. If we use these statistics (mean, median, and maximum time-to-reach) as features of the Web pages in those datasets, we find that the decision tree and Bayesian network classifiers are no different from a majority element predictor, and the SVM classifier performs worse than this. All this holds for the academic dataset as well, with one interesting exception: a decision tree with a single feature, the median time-to-reach (split at 76) provides a higher true positive rate with an average TP Rate of 0.615, FP Rate of 0.385, Precision of 0.686, Recall of 0.615, and F-Measure 0.575.

In other words, time-to-reach as a predictor of accessibility shows mixed results. Our findings are potentially due to the small size of our datasets, but this issue deserves more study. One natural direction for future work is to fold time-to-reach into our set of features for classification.

6. DISCUSSION

The work in this paper is preliminary and has a number of limitations. First and most obviously, we have examined only a small number of Web pages and relied on analysis of only a small number of their structural and textual features. Our next step, having established a baseline for our expectations, is to extend our data collection and feature generation efforts in both directions.

Another issue arises in the situation where we limit our consideration to a specific classifier, and we identify a set of predictive features, it is not yet known whether the relationships we might find are causal or simply associational. In other words, if we were to change the properties of an accessible Web page such that it would be reclassified as accessible, would we actually have improved its accessibility? This direction will require direct user studies.

[6] Our intention was an even split between accessible and inaccessible Web pages, but we discovered an irregularity in our data collection late in the analysis process.

[7] http://www.alphaworks.ibm.com/tech/adesigner

Finally, we believe that there is significant benefit to be gained from using results from machine learning techniques to drive transcoding [1]. There is good potential for further work in this direction.

7. ACKNOWLEDGMENTS

The authors would like to thank the participants in the online survey. In addition, assistance from Kentarou Fukuda, from the Accessibility Tools Framework (ACTF) project community was valuable in gathering results from ADesigner.

8. REFERENCES

[1] C. Asakawa and H. Takagi. Transcoding. *Web Accessibility*, pages 231–260, 2008.

[2] J. Bigham, R. Kaminsky, R. Ladner, O. Danielsson, and G. Hempton. WebInSight:: making web images accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 181–188. ACM, 2006.

[3] J. Bigham, T. Lau, and J. Nichols. Trailblazer: enabling blind users to blaze trails through the web. In *Proceedings of the 13th International conference on Intelligent user interfaces*, pages 177–186. ACM, 2009.

[4] J. Brudvik, J. Bigham, A. Cavender, and R. Ladner. Hunting for headings: sighted labeling vs. automatic classification of headings. In *Proceedings of the 10th International ACM conference on Computers and Accessibility*, pages 201–208. ACM, 2008.

[5] C. Bühler, H. Heck, O. Perlick, A. Nietzio, and N. Ulltveit-Moe. Interpreting results from large scale automatic evaluation of web accessibility. *Computers Helping People with Special Needs*, pages 184–191, 2006.

[6] H. Davalcu, S. Vadrevu, S. Nagarajan, and I. Ramakrishnan. Ontominer: Bootstrapping and populating ontologies from domain-specific web sites. *Intelligent Systems, IEEE*, 18(5):24–33, 2005.

[7] A. P. Freire, R. P. M. Fortes, M. A. S. Turine, and D. M. B. Paiva. An evaluation of web accessibility metrics based on their attributes. In *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, pages 73–80, New York, NY, USA, 2008. ACM.

[8] K. Kottapally, C. Ngo, R. Reddy, E. Pontelli, T. Son, and D. Gillan. Towards the creation of accessibility agents for non-visual navigation of the web. In *Proceedings of the 2003 Conference on Universal usability*, pages 134–141. ACM, 2003.

[9] G. Leshed, E. Haber, T. Matthews, and T. Lau. CoScripter: automating & sharing how-to knowledge in the enterprise. In *Proceeding of the twenty-sixth annual ACM Conference on Human Factors in Computing Systems*, pages 1719–1728. ACM, 2008.

[10] J. Mahmud, Y. Borodin, D. Das, and I. Ramakrishnan. Combating information overload in non-visual web access using context. In *Proceedings of the 12th International conference on Intelligent user interfaces*, pages 341–344. ACM, 2007.

[11] I. Ramakrishnan, J. Mahmud, Y. Borodin, M. Islam, and F. Ahmed. Bridging the Web Accessibility Divide. *Electronic Notes in Theoretical Computer Science*, 235:107–124, 2009.

[12] H. Takagi, C. Asakawa, K. Fukuda, and J. Maeda. Accessibility designer: visualizing usability for the blind. In *Proceedings of the 6th International ACM SIGACCESS conference on Computers and accessibility*, pages 177–184. ACM, 2004.

[13] H. Takagi, S. Saito, K. Fukuda, and C. Asakawa. Analysis of navigability of web applications for improving blind usability. *ACM Transactions on Computer-Human Interaction*, 14(3):13, 2007.

[14] L. Yi, B. Liu, and X. Li. Eliminating noisy information in web pages for data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 296–305. ACM, 2003.

[15] M. Yoshioka. IR interface for contrasting multiple news sites. *Information Retrieval Technology*, pages 508–513, 2008.

9. APPENDIX

The accessible academic Web sites are: missouristate.edu^c, csun.edu^c, calpoly.edu^c, illinois.edu^c, iub.edu^c, csuci.edu^c, oit.edu^c, nsc.nevada.edu^c, evansville.edu^c, csuchico.edu^c, msu.edu^c, utulsa.edu^c, holycross.edu^c, umn.edu^c, uic.edu^c, luc.edu^c, csueastbay.edu^c, utexas.edu^c, siu.edu^c, ucsf.edu^c, ku.edu^c, psu.edu^c, duke.edu^c, uh.edu^c, ttu.edu^c, csufresno.edu^c, csusb.edu^c, sfsu.edu^c. The inaccessible academic websites are: usafa.af.mil^c, fordham.edu^c, ysu.edu^c, weber.edu^c, wright.edu^c, usna.edu^c, gbcnv.edu^c, providence.edu^c, uc.edu^c, lasalle.edu^c, usma.edu^c, montana.edu^c, smu.edu^c, georgetown.edu^c, ucsb.edu^c, lafayette.edu^c, okstate.edu^c, ucsc.edu^c, marshall.edu^c, csustan.edu^c, home.uncc.edu^c, sou.edu^c, usm.edu^c, ecu.edu^c, uab.edu^c, tcu.edu^c, valpo.edu^c, harvard.edu^c.

The accessible non-academic websites are: developer.android.com/reference/android/content/ContentProvider.html^s, en.scientificcommons.org^m, en.wikipedia.org/wiki/Main˙Page^{m,w}, mobile.hertz.com^m, notepad-plus-plus.org^m, suddendebt.blogspot.com^s, woot.com^s, aa.com/homePage.do^s, arstechnica.com^s, audible.com^w, bbc.co.uk^w, bestbuy.com^s, blindbargains.com^m, bookshare.org^s, brandonsanderson.com^m, cbc.ca/news^m, cpan.org^s, gnome.org^m, google.com/search?q=ford+raptor^{m,w}, nokia.com^m, nytimes.com^s, orbitz.com^m, prolificliving.com^s, sennheiserusa.com/home^s, usconstitution.net/const.html^s, w3.org^s, zorrolegend.blogspot.com/search/label/New%20World%20Zorro%20DVD%20Information^s. The inaccessible non-academic websites are: blogs.barrons.com/techtraderdaily^s, fittv.discovery.com/fansites/blaine/recipes/recipes.html^s, jetblue.com^s, lenovo.com/us/en^s, msdn.microsoft.com/en-us/library/microsoft.csharp.csharpcodeprovider.aspx^s, travisa.com^s, anandtech.com^s, apple.com^s, aupeo.com^{m,w}, christinefeehan.com^m, cnet.com^s, dell.com^s, delta.com^s, dv.is^m, facebook.com^s, fmbrewery.com^m, genaw.com/lowcarb/burger˙recipes.html^s, germanwings.com/en^m, hertz.com^m, mbl.is^m, myspace.com^s, nivea.de^m, pandora.com^s, rogersthankyou.com^s, tdameritrade.com^s.