A macroscopic Web accessibility evaluation at different processing phases

Nádia Fernandes, Luís Carriço
LaSIGE/University of Lisbon
Campo Grande, Edifício C6
1749-016 Lisboa, Portugal
{nadiaf,lmc}@di.fc.ul.pt

ABSTRACT

This paper details a comparative experimental study to understand the difference of the Web's accessibility properties regarding two different evaluation approaches: using WCAG 2.0 techniques and targeting the pages as they reach the browser; using WCAG 2.0, but evaluating the pages after the browser processing, thus as they will be delivered to the end-user. For that, we evaluated over 20000 Web pages using already established accessibility metrics. We then compared the results obtained from the WCAG 2.0 evaluation of the two processing phases. We observed some changes in the macroscopic properties of the evaluation. Regarding the comparison between the two phases, we observed a narrower distribution of quality, i.e., the worst pages are in fact not that bad, and the best ones not that good.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Measurement, Human Factors.

Keywords

Web accessibility, Web science, Web browser processing, Automated evaluation.

1. INTRODUCTION

The Web has became a universal platform and is constantly growing in size and complexity. It is increasingly in our everyday life, being now progressively indispensable for almost all our activities (i.e. public, business, personal efficiency or improvement of subsistent) [2]. Accessing the Web is nowadays such an important asset that is considered as a fundamental right.

W4A2012 - Communication, 16-17th April 2012 - Lyon, France.

In fact, the Web is used by people with diverse profiles and characteristics, with miscellaneous capabilities, including those with special needs. Thus, Web sites/pages should be designed so that information can be perceived by everyone, i.e., should be accessible. This way, it is important to guarantee that Web content is accessible for as many possible users, no matter their disabilities, and for that it is important that the evaluations are performed in the content that users use and perceive.

A framework [5] was proposed for the evaluation of Web pages, using WCAG 2.0 [4], at two different processing phases: before and after Web Browser processing. The first corresponds to the typical context for automated evaluation, which works on the result of the first HTTP request. The second is the context where users interact with the Web, after loading the initial scripts.

It was been proved [5] that the differences introduced by AJAX and other dynamic scripting features do influence the outcome of Web accessibility evaluation practices. The results of the automated Web accessibility evaluation before processing yield incorrect and incomplete results.

In this paper, we perform a deeper analysis of the differences between these two processing phases. We extend the sample to over 20000 Web pages thus reaching a macroscopic analysis. Moreover, we use the metrics proposed by *Lopes et al* [8] and assess some of the Web characteristics examined by that study, now in light of WCAG 2.0 techniques [3], and particularly at the after browsing processing phase.

2. RELATED WORK

Web accessibility evaluation (WAE) is an assessment procedure to analyse how well the Web can be used by people with different levels of disabilities [7]. Unfortunately, current studies show that many Web sites still cannot be accessed in the same conditions, by a large number of people [7, 9].

WCAG is one of the most used technical standards for accessibility evaluations, encouraging creators (e.g., designers, developers) in constructing Web pages according to a set of best practices. If this happens, a good level of accessibility can be guaranteed [7, 9]. Although these guidelines exist and are supposed to be followed by the creators, most Web sites still have accessibility barriers making its utilization very difficult or even impossible for many users [7]. Thus, WCAG can also be used as a benchmark for analysing the accessibility quality of a given Web page.

The results of an accessibility evaluation can be used to measure quantitatively the level of accessibility of a Web page. Furthermore, metrics are important to facilitate un-

derstanding, controlling, and improving products and processes in software development [6]. Additionally, in terms of accessibility metrics can also help the user to understand if a Web page/site can be used by them.

2.1 Automatic Evaluation

Automatic evaluation is performed by software, without the need of direct human intervention, and with expertise embedded in a software framework/tool. The major benefits are scalability and objectivity [9]. However, it has limitations that direct or users evaluations do not have (e.g., the depth and completeness of analysis). However, the automatic version of the evaluation allows to perform analysis with bigger sets of Web pages, enabling large-scale studies, like [8] for example. That study in particular revealed several macroscopic properties of web accessibility, with an evaluation using WCAG 1.0 checkpoints.

Traditionally, automated evaluations has been performed with source documents that are returned on the first HTTP request. However, nowadays, Web pages are mostly dynamic, the content presented to the user is often very different from what is obtained in that request. It has been showed that the differences of assessed elements, fails, warnings and passes between the evaluations done before and after browser processing are statistically significant [5]. Consequently, it is paramount that the WAE tools also evolve on the material that is assessed, targeting the rendered or transformed HTML.

3. EXPERIMENTAL STUDY

To perform the accessibility evaluation, we used the QualWeb evaluation framework [1], but we simulated the processing of the Web browser using *Phantom.js*[1].

Phantom.js is a command-line tool that uses WebKit [10], it works like a WebKit-based web browser. Besides, it can be controlled using Javascript, being consistent with our implementation. Moreover, it offers mechanisms to access both processing phases. We now issue only one request to get the evaluation target, the Web page, and use two different functions to simulate both processing phases.

This solution solves all the issues discovered in the previous work [1]: 1) it avoids data injection at the browser level; 2) it guarantees that the evaluated Web page before and after browser processing is exactly the same; and 3) it enables the integration of the evaluation framework with a crawler to perform intensive simulation.

After that, we performed an experimental study that aims at understanding the impact of the delivery processing phase, and of the usage of WCAG 2.0 on the accessibility characteristics of the Web, as they emerge from a large-scale study. We formulate the following research question:

> *How do the macroscopic properties emerging from Web accessibility change in respect to the processing phase of delivery?*

We used some Web pages from the Web sites provided by the Portuguese Web Archive (PWA) [2] (version of 2008), and 18 HTML WCAG 2.0 techniques already implemented in [5].

[1] http://www.phantomjs.org/
[2] http://www.fccn.pt/

The results of the evaluation are presented in terms of *PASS, WARN and FAIL*: *pass* or *fail*, if the elements verified by the techniques are in agreement or disagreement with the W3C recommendations for the techniques, respectively; and *warning* – if it is not possible to identify certain characteristic of an element as right or wrong, without the need of an expert intervention.

We used the metrics already applied in *Lopes et al* [8], namely: *conservative rate, optimistic rate* and *strict rate*. Each is normalized into a percentage, where the results are between *accessible* (100%) and *not accessible* (0%). These metrics use the *PASS, WARN and FAIL* as follows:

Conservative rate - The worst-case scenario on accessibility evaluation (WARN results are interpreted as failure):

$$rate_{conservative} = \frac{passed}{applicable} \tag{1}$$

Optimistic rate - A best-case scenario where warnings are dismissed as accessibility issues that were taken into account (WARN results are interpreted as passed):

$$rate_{optimistic} = \frac{passed + warnings}{applicable} \tag{2}$$

Strict rate - WARN results are dismissed (thus accounting only the actual FAIL results):

$$rate_{strict} = \frac{passed}{applicable - warnings} \tag{3}$$

4. RESULTS

The evaluation of the Web documents collection yield results for 24,462 Web pages. Our evaluation yielded differences in the HTML documents, in terms of the number of HTML elements for each of the analysed phases. Before processing the pages contained 24,918,720 elements, whereas after processing the number of elements raised to 41,967,072. This corresponds to a growth of approximate 68%. If we look at the average size of a page in terms of number of elements, the increase from 1010 to 1710 corresponds to a 69% growth.

We focused our study in two main sets of results: first, the difference of evaluation outcomes (*fail*, *pass*, *warning*) between the two phases; and second, the differences of distribution of the rates mentioned between evaluations.

4.1 Evaluation Outcomes

In Figure 1, we present the three evaluation outcomes, for each of the phase, as defined above:

- *Pass*: An average of approximately 9 elements passes their respective evaluation criteria before browser processing. This number highly increases after browser processing, to an average of 87 elements, a growth of approximately 867%;
- *Fail*: An average of approximately 46 elements fails their respective evaluation criteria before browser processing. After browser processing, this number rises to an average of approximately 176 elements, an increase of approximately 282%;
- *Warning*: An average of approximately 262 elements had warnings in their respective evaluation criteria before browser processing, a rise of approximately 72%.

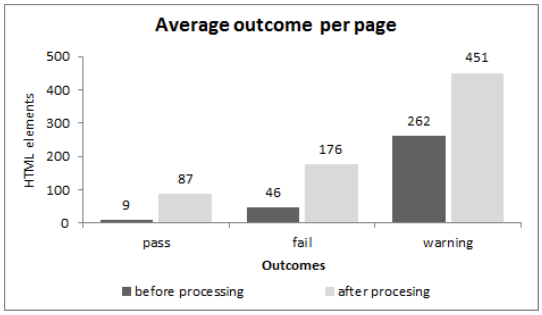

Figure 1: Average of outcome per page between evaluations.

It is interesting to notice that the biggest relative growth occurs for the elements that pass the criteria, amounting to 3 times the increase of fails and 12 times the one for warnings.

In the following sections, we detail the differences arising from the comparison of the application of rates already mentioned.

4.2 Rates Before and After Processing

For the comparison of rate values between the evaluation before and after browser processing we will present in detail the results for each of the rates.

Figures 2, 3 and 4 present the distributions of before processing and after processing for *conservative rate*, *optimistic rate* and *strict rate*, respectively.

For the *conservative rate* is clear (Figure 2), that the average quality assessed increases in the after processing. We can also notice that the majority of the results with quality between 60% and 90% disappeared in after processing.

For the *conservative rate* (Figure 3), the average quality assessed decreases in the after processing, from approximately 80% to around 70%. We can observe that there is a decrease of the number of pages that have accessibility quality lower than 20%. Thus worst pages tend to be assessed with higher quality after processing. There is also a decrease on the number of pages with higher accessibility quality, leading to a lower accessibility average.

For the *strict rate* (Figure 4), the average is approximately 0.11 before processing and after processing is approximately 0.35. However, it can be seen that the majority of the results higher than 85% disappeared in after processing. Generally the trend is similar to the other rates: worst pages before processing get higher scores in after processing evaluation and better pages are ranked lower.

5. DISCUSSION

Our study addresses aspects related with emerging macroscopic properties deriving from Web accessibility evaluation, exploring the perceived differences on those properties depending on the browser processing phase, i.e., before or after browser processing.

Regarding the differences between *processing phases*, the results obtained for this experimental study confirm and emphasize, considering the scale, the previous results of [5]. This reaffirms that evaluation should be done *after browser processing*.

When looking at accessibility distribution *after processing*, we can verify that the best accessibility quality values get worse and the worst accessibility quality values get better. Thus, we can conclude that Web pages can possess higher uniformity. We believe the reason for that emerges from the usage of templates and particularly, reusable code, such as libraries. Considering that the number of elements increases between before and after processing for each page, and that the accessibility rates get more uniform, then the browser processing phase must introduce more uniformly code with a medium accessibility rate, i.e., not that bad and not that good. If some of that that code is reused than the uniformity of results would occur. This observation concurs with the argument that the usage of templates and libraries will improve the accessibility of the Web, provided that they are accessible. Of course our results only confirmed uniformity. The goal is that since reusable code is usually done by non-novice programmers and designers then the novices' accessibility mistakes will be minimized. Moreover, the accessibility quality of templates/libraries will certainly improve, with frequent revisions of the reused material, as well as with the improvement of the designers' accessibility expertise.

The aforementioned arguments answer the research question, showing the differences on the macroscopic properties of web accessibility between evaluation settings. They also stresses the advantages of using WCAG 2.0 and of performing accessibility evaluation after browser processing.

5.1 Impact on Designing Accessible Web Pages

The results of the experiment can also be used to improve the design of accessibility Web pages. We can conclude that Web pages which reuse code tend to have not as bad (or relatively better) accessibility quality, so it can be positive to use libraries or templates. However, it is important that developers/designers with expertise in accessibility design those templates/libraries, to ensure their correction.

Additionally, to capitalize the corrections of the templates/libraries is important to share reusable code. This way, the accessibility problems are being discovered and corrected by the multiple users.

5.2 Impact on the Perception of Accessibility

The current study allowed us to detect differences between rates. This might indicate that using WCAG 2.0 evaluator may provide a better perception of accessibility problems, which may lead to the creation of Web pages with better accessible quality.

Regarding the processing phases, the impact in perception is also relevant. Considering that users interact with Web pages after processing, it is important that evaluations assess that phase, especially because results are different, as this study shows. Note that these results, and thus the users' perception of accessibility will tend to be more pronounced as the Web evolves towards dynamic content and HTML 5.

5.3 Limitations of the Experiment

Our experiment has faced some limitations on the type of results that can be extrapolated, including:

- *Techniques coverage:* we used the 18 HTML WCAG 2.0 implemented techniques of QualWeb (corresponding to 15 success criteria). However, it would be important to have CSS techniques implemented and adopt new emerging techniques that will, for sure, emerge from the conformance to HTML 5 new features;

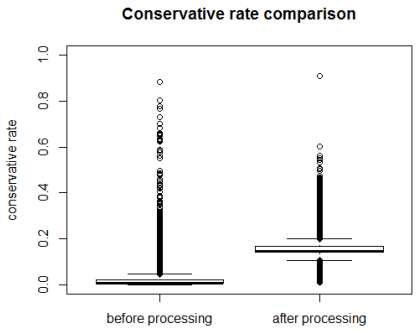

Figure 2: Box-plot comparing *conservative rate* before and after browser processing.

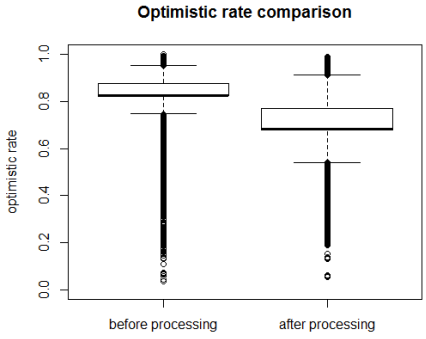

Figure 3: Box-plot comparing *optimistic rate* before and after browser processing.

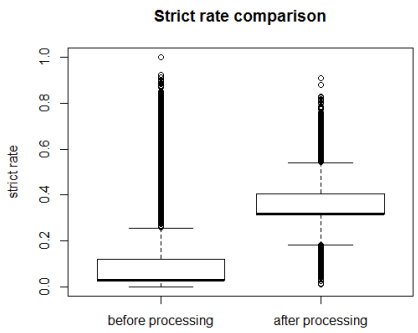

Figure 4: Box-plot comparing *strict rate* before and after browser processing.

- *Dynamic content:* to be able to pick the dynamic trigger action in the browser processing;
- *Automated evaluation:* since this experiment is centred on automated evaluation of Web accessibility quality, it shares all of the inherent pitfalls.

6. CONCLUSIONS AND FUTURE WORK

This article presented a large-scale study of accessibility on the Web, using a Web document collection from 2008, collected from the Portuguese Web Archive (PWA). The results allowed us to characterize some accessibility properties of the Web, pointing some differences: between the common way of assessing Web pages -, *before browser processing*,- and a recent proposal of assessment -, *after browser processing*- and closer to what the end-users perceive.

The results obtained on the evaluation of pages after browser processing tend to be more homogeneous than before, i.e., bad pages tend to be assessed as better and good pages are actually worse. Definitely, there is a difference between what is available before and after processing. Considering that the end-user interacts with the "after processed pages" then most studies about Web quality should be redone.

Facing with the obtained results, on-going work is being conducted in the following directions:

1. Enlarge the coverage of WCAG 2.0 implemented tests, in particular considering the analysis of CSS;
2. Evaluate Web applications, considering that they are conquering the Web;
3. Perform a comparative set of studies that relate the evolution of accessibility in PWA for recent years.

7. ACKNOWLEDGMENTS

This work was funded by Fundação para a Ciência e Tecnologia (FCT) through the *QualWeb* national research project PTDC/EIA-EIA/105079/2008, the Multiannual Funding Programme, and POSC/EU.

8. REFERENCES

[1] Omitted for blind review purposes.

[2] R. Baguma and J. T. Lubega. A web design framework for improved accessibility for people with disabilities (wdfad). In *Proceedings of the 2008 international cross-disciplinary conference on Web accessibility (W4A)*, W4A '08, pages 134–140, New York, NY, USA, 2008. ACM.

[3] M. Cooper, G. Loretta Reid, G. Vanderheiden, and B. Caldwell. Techniques for WCAG 2.0 - Techniques and Failures for Web Content Accessibility Guidelines 2.0. W3C Note, World Wide Web Consortium (W3C), October 2010. Last accessed on November 26th, 2010, from `http://www.w3.org/TR/WCAG-TECHS/`.

[4] M. Cooper, L. G. Reid, G. Vanderheiden, and B. Caldwell. Techniques for WCAG 2.0 - Techniques and Failures for Web Content Accessibility Guidelines 2.0. W3C Note, World Wide Web Consortium (W3C), October 2010. Last accessed on November 26th, 2010, from `http://www.w3.org/TR/WCAG-TECHS/`.

[5] N. Fernandes, R. Lopes, and L. Carriço. On web accessibility evaluation environments. In *W4A '11: Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A)*, New York, NY, USA, 2011. ACM.

[6] 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.

[7] S. Harper and Y. Yesilada. *Web Accessibility.* Springer, London, United Kingdom, 2008.

[8] R. Lopes and L. Carriço. Macroscopic characterisations of web accessibility. *New Review of Hypermedia and Multimedia*, 2010.

[9] R. Lopes, K. V. Isacker, and L. Carriço. Redefining assumptions: Accessibility and its stakeholders. In *The 12th International Conference on Computers Helping People with Special Needs (ICCHP)*. Springer, 2010.

[10] Webkit. The webkit open source project, 2011. Available from: http://www.webkit.org/.