Measuring Web Accessibility by Estimating Severity of Barriers

Giorgio Brajnik

Dip. di Matematica e Informatica
Università di Udine
Italy
www.dimi.uniud.it/giorgio

Abstract. The paper addresses the issue of measuring web accessibility in such a way that differences in measurements reflect differences in the effectiveness experienced by disabled users. The paper presents the steps upon which a measuring methodology called MAMBO is based, and the data that are needed to compute the indexes, in addition to its conceptual rationale. An experimentation of MAMBO is then described, based on analysis of 14 accessibility reports; results are shown and discussed, including the effects that different severity judgments may have on the metric, how to estimate confidence intervals on the values, and how the metric can be used to estimate accessibility with respect to specific user groups.

1 Introduction

The importance of measuring web accessibility is increasing; many different activities are demanding it. For example, measuring takes place when quality assurance practitioners are monitoring accessibility of a website to ensure that it does not decrease as new content gets published. Similarly, measuring occurs when developing a new user interface of the website and comparing it to previous versions of the same website, or to websites of competitors (competitive analysis). Once accessibility defects are found, in order to set priorities, developers need to know which ones are more important in terms of negative impact on users experience: a measure of accessibility is again needed. Accessibility levels are also needed when end users want to know how accessible a website is before using it. This is the case, for example, when a search engine lists search results that are ranked also by accessibility level [2].

Many existing measurement processes are based on the number of violations of established requirements (for example, WCAG 1.0 checkpoints). While the conformance level of the website (*i.e.* the degree to which a website satisfies the requirements defined by a standard) is an important measure when formal regulations are in place, this is by no means the only way to determine the accessibility level. One limit of methods based on conformance is that it is difficult to relate the accessibility level to the actual hindrances that the website may raise against given user categories, such as blind users of screen readers,

S. Hartmann et al. (Eds.): WISE 2008, LNCS 5176, pp. 112–121, 2008.

low-vision users of screen magnifiers, motor-disabled users of a normal keyboard and/or mouse, deaf users, cognitively disabled users (with reading and learning disabilities and/or attention deficit disorders). Yet, this is what is often meant by *accessibility*: "a web site is accessible if people with some impairment can use it with the same effectiveness, security and safety as non-disabled people" (definition derived from [10]). If accessibility levels were determined with this definition in mind, and by applying a *valid* measurement process[1], then users could easily determine how much accessible the website is for them; developers and QA practitioners could estimate which user categories could be best or worst supported by the website; they could determine which parts of the website do a better (or worse) job in supporting these users; they could compare different versions of the website to determine how accessibility is evolving along the development; and they could set fixing priorities.

Several accessibility metrics have been discussed in the literature [1, 3, 11, 12, 14]. In many cases the measurement process is based on automated testing tools, capable of systematic application of an array of tests covering some or all of the requirements of a standard. The advantages behind such a solution is that tools are systematic scanners of web pages, efficient processors and reliable evaluators (in the sense that they produce repeatable results). However, tools are plagued with the problems of generating issues that are not accessibility problems (false positives), of missing certain true accessibility problems (false negatives), and of being incapable of estimating the severity of a requirement violation. Since the tools do conformance testing, they yield a measure of accessibility that is a function of the number of passed and failed requirements/tests. Often this function is the *failure rate* FR, defined as the number of violations of any checkpoint divided by the maximum number of violations of any of those checkpoints that can take place (*i.e.* by the number of *possible* violations). For example, two pages that include 10 and 20 images respectively, one with 2 properly defined "alt-text", the other with 8, have FR = 0.8 and FR = 0.6 respectively. Even though the second page has a larger number of violations, hence a larger number of potential obstacles to users, it has a smaller FR. Therefore, in addition to wrong estimates due to false positives and negatives, the values produced by accessibility metrics based on automatic tools cannot be directly be related to accessibility as defined above.

For these reasons, we defined and experimented SAMBA [5], a method for measuring accessibility by using the output of testing tools *coupled with* focused opinions of experienced human evaluators, so that correct estimates of tool errors can be assessed, and appropriate estimations of severities of barriers are used. When adopting SAMBA, an accessibility testing tool is used to automatically test many web pages; this generates a large number of checkpoint violations, that are automatically mapped to potential accessibility barriers and then sampled

[1] *Validity* is "the extent to which the problems detected during an evaluation are also those that show up during real-world use of the system" whereas *reliability*, often called *reproducibility*, is "the extent to which independent evaluations produce the same result" [8].

randomly. The sample of potential barriers is submitted to a panel of judges that assign them a severity level, including 0 ("not-a-problem").

This paper describes a manual method for measuring barriers of accessibility (MAMBO) that derives from SAMBA and that can be used when manually evaluating accessibility of a website. The accessibility indexes defined by MAMBO are standardized (and therefore can be used to compare accessibility levels of different websites and/or obtained by different evaluators). In addition, MAMBO offers more than one accessibility index, which can be used to measure accessibility with respect to different user groups, and to estimate the uncertainty due to having analyzed only a fraction of the available web pages. To adopt MAMBO evaluators follow an assessment method called *barrier walkthrough* [4]; computing the accessibility indexes requires little additional effort.

2 Barrier Walkthrough

The barrier walkthrough (BW) method [4, 6] is an accessibility inspection technique. An evaluator has to consider a number of predefined barriers which are interpretations and extensions of well known accessibility principles; they are linked to user characteristics, user activities, and situation patterns so that appropriate conclusions about user effectiveness, productivity, satisfaction and safety can be drawn, and appropriate severity scores can be consequently derived. The method is rooted on heuristics walkthrough [9] which considers the context of use of the website. For BW, context comprises certain user categories (like blind persons), usage scenarios (like using a given screen reader), and user goals (corresponding to *use cases*, like submitting an IRS form).

An *accessibility barrier* is any condition that makes it difficult for people to achieve a goal when using the web site through specified assistive technology (see figure 1 for an example). A barrier is a failure mode of the web site, described in terms of (i) the user category involved, (ii) the type of assistive technology being used, (iii) the goal that is being hindered, (iv) the features of the pages that raise the barrier, and (v) further effects of the barrier.

Notice that several barriers can depend on the same cause: *e.g.* for a missing *skip-links* link (defect) a barrier for a blind user of a screen reader is that s/he cannot get quickly to the relevant content of the page; the barrier for a keyboard user is that s/he cannot move the focus directly to the relevant controls in the page; the barrier for a low vision person is that s/he cannot move directly the field of vision on the relevant content.

Severity of a barrier depends on the context of the analysis (type of user, usage scenario, user goal). The BW method prescribes that severity is graded on a 1–2–3 scale (minor, major, critical), and is a function of *impact* (the degree to which the user goal cannot be achieved within the considered context) and *frequency* (the number of times the barrier shows up while a user is trying to achieve that goal). Therefore the same type of barrier may be rated with different severities in different contexts; for example, the missing *skip-links* link may turn out to be a nuisance for a blind user reading a page that has few preliminary

stuff, while the same defect may show a higher severity within a page that does a server refresh whenever the user interacts with links or select boxes.

Potential barriers to be considered are derived by interpretation of relevant guidelines and principles [7, 13]. A complete list can be found in [6].

barrier	users cannot perceive nor understand the information conveyed by an information rich image (*e.g.* a diagram, a histogram)
defect	an image that does not have accompanying text (as an ALT attribute, content of the OBJECT tag, as running text close to the picture or as a linked separate page)
users affected	blind users of screen readers, users of small devices
consequences	users try to look around for more explanations, spending time and effort; effectiveness, productivity, satisfaction are severely affected

Fig. 1. Example of barrier

3 MAMBO

MAMBO (MAnually Measuring Barriers Of accessibility) is an accessibility metric not based on conformance. It can be computed directly by scanning a barrier walkthrough report, by highlighting reported barriers, and by considering their severity and the kind of user groups to which they refer (blind persons, motor disabled ones, etc.).

The basic computation of the accessibility index (AI) is similar to SAMBA [5]. In particular (Table 1 shows some example):

1. By tabulating the number of barriers split by user types against each severity value, we obtain the *severity matrix*; each element of the matrix M gives the proportion of sampled barriers associated with disability d and severity s.
2. The *confidence intervals severity matrix* $\mathcal{M}$ can then be generated, by computing the 95% confidence interval around each proportion $M_{d,s}$.
3. The *barrier density* of a web site needs to be computed. It is defined as $F = k\frac{\text{number of barriers}}{num.\ of\ bytes}$, which can be interpreted as the probability that k bytes of HTML code of the site causes a barrier; if M is the severity matrix, then $F \cdot M_{d,s}$ is the probability that k bytes of code causes a barrier for disability d with severity s; the scale factor k is used to tune the values produced by MAMBO.
4. If we combine the density factor F with the confidence interval severity matrix $\mathcal{M}$ we obtain $\mathcal{F}$; after using appropriate weights to balance different severity levels, we get the *Weighted Accessibility Index* ($\underline{AI_w}$). Since it is based on confidence intervals, it is itself an interval ($\underline{AI_w}, \overline{AI_w}$), defined as follows:

$$\text{let } \underline{\mathrm{H_d}} = \frac{\underline{f}_{d,1}}{w_1} + \frac{\underline{f}_{d,2}}{w_2} + \underline{f}_{d,3}$$

$$\text{and } \overline{\mathrm{H_d}} = \frac{\overline{f}_{d,1}}{w_1} + \frac{\overline{f}_{d,2}}{w_2} + \overline{f}_{d,3}$$

$$\text{then } \underline{\mathrm{AI_w}} = \prod_d \left(1 - F \cdot min\left\{1, \overline{H_d}\right\}\right)^2$$

$$\text{and } \overline{\mathrm{AI_w}} = \prod_d \left(1 - F \cdot min\left\{1, \underline{H_d}\right\}\right)^2$$

where $\mathcal{F}[d,s] = f_{d,s}$, and $\frac{1}{w_s}$ is the weight to be given to minor and major barriers (*i.e.* $s = 1, 2$). Each term in the product defining AI_w can be interpreted as the probability that no barriers for disability d are raised, and the resulting value is related to the probability that there are no barriers at all. Squaring the terms further amplify the contribution of each disability.

For example, the severity matrix illustrated in Table 1 shows that 35 barriers for blind users were found; 14% were minor ones, 23% major and 63% were found to be critical. The table shows also the confidence intervals around these proportions; for example, it is safe to assume that critical barriers for blind users range between 45% to 78%.

Table 1. (Left) Severity matrix obtained from a barrier walkthrough report. Columns 1 to 3 show the proportion of barriers that were given severity 1 to 3 (*minor* to *critical*); the last column gives the total number of barriers. (Right) Confidence interval matrix from the same report ($\alpha = 0.05$).

	Severity			
category	1	2	3	**tot**
cb (color blind)	0.00	0.00	1.00	**2**
md (motor disab.)	0.11	0.47	0.42	**19**
lv (low vision)	0.00	0.00	0.00	**0**
nh (no hearing)	0.00	0.00	0.00	**0**
nv (no vision)	0.14	0.23	0.63	**35**
cd (cogn. disab.)	0.36	0.32	0.32	**25**
js (no javascript)	0.00	0.33	0.67	**9**

	Severity					
category	1		2		3	
cb	0.00	0.80	0.00	0.80	0.20	1.00
md	0.02	0.35	0.25	0.71	0.21	0.66
lv	-		-		-	
nh	-		-		-	
nv	0.05	0.31	0.11	0.41	0.45	0.78
cd	0.10	0.46	0.28	0.68	0.10	0.46
js	0.00	0.37	0.09	0.69	0.31	0.91

For the same report the barrier density factor is 0.039 (barriers/k bytes of code, with $k = 20$); using weights $1/9, 1/3$ (one critical barrier weighs as much as 9 minor and 3 major ones), if we restrict to the *no-vision* category, we obtain $AI = 0.94$, and an interval of $(0.92, 0.96)$. After combining all the disability types we get $AI = 0.88$ and an interval of $(0.68, 0.75)$.

4 Practical Examples and Discussion

A practical analysis of MAMBO was carried out by analyzing 14 barrier walkthrough reports produced by students of my course (user centered web design).

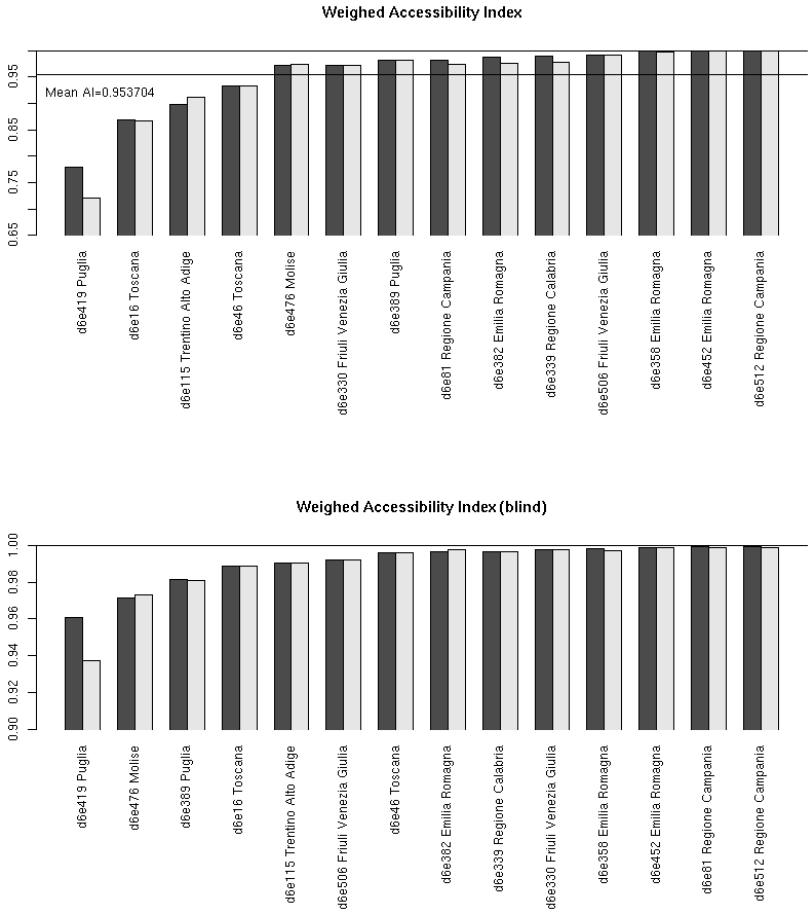

Fig. 2. Web Accessibility Index for the 14 reports (some refer to the same web site, Italian local government web sites; for example, "d6e382 Emilia Romagna" and "d6e358 Emilia Romagna" are two different reports about the same web site). (*Top*) The dark area is the index derived from the judged severity; the other one derives from the original severity scores. The horizontal line gives the mean index (0.9537). (*Bottom*) Same indexes, but with data restricted to barriers for blind users.

Students were exposed to web accessibility, conformance testing and barrier walk-through for about 15 lecture hours, after which they were asked to analyze given web sites and write corresponding reports[2].

These BW reports were analyzed by a judge, who was asked to validate the severity judgments made by the authors of reports. The judge had to give her own severity level to each of the barriers mentioned in the report. In this way we

[2] The entire set of Italian reports is available at `www.dimi.uniud.it/giorgio/dida/psw/galleria/galleria.html`

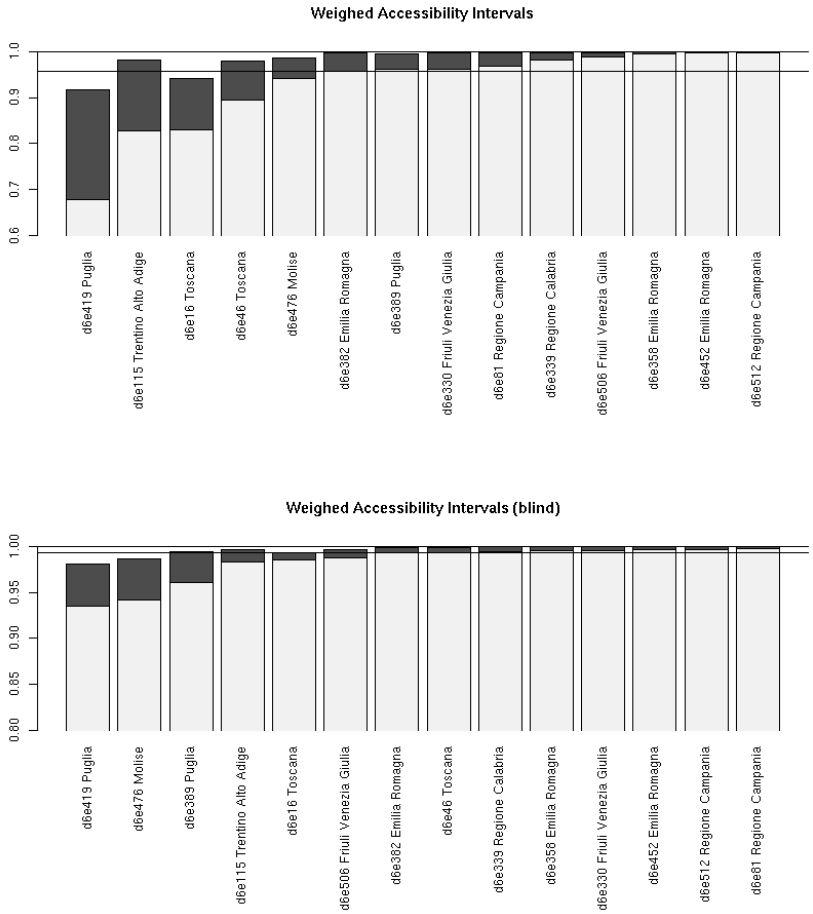

Fig. 3. (*Top*) The intervals for AI on the same reports (the dark area represent lower and upper bounds of AI), based on judge's severity. (*Bottom*) The intervals for AI on the same reports restricted to barriers for blind users.

can estimate what is the effect of *false positives* on the metric. The values given below were computed using the severity that the judge assigned to problems, including 0 for what was deemed to be a "not-a-problem".

Figure 2 (top) shows that AI spans a relatively small range (0.72 to 0.99); this is due to the magnitude of the density factor: the smaller it is, the wider is the range spanned by AI. Therefore, by using an appropriate scale factor k, results can be tuned to the desired level of resolution. More importantly, Figure 2 shows that different judgments of barriers severity have limited effects on the overall AI value, even though over 327 analyzed barriers, there were 59 disagreements in severity (18%).

When restricting to barriers relevant for a given user group, for example blind users (bottom part of Figure 2), the AI range narrows (since fewer barrier types

are considered, and fewer disability-related contributions to AI are included). Also in this case the effect of misjudged severity is marginal.

Figure 3 (top) shows the intervals around AI, computed on the basis of the 5% confidence intervals of the severity matrix. When two AI intervals overlap then no comparison can be made on the corresponding websites (or reports), since the level of uncertainty determined by the number of found barriers and their splitting into different severities is too high. However, when two intervals do not overlap a comparison between web sites/reports can be stated with relatively high certainty. For example, only two sites are less accessible than the sixth one (d6e382 Emilia Romagna): those whose upper bound is below the horizontal line. The certainty level of this statement is higher than 95%.

Obviously, when restricting to fewer disability types, the range narrows; this is shown at the bottom part of the Figure. But also in this case non-overlapping intervals can be used to compare web sites. For example, two websites are less accessible than the seventh one (d6e382 Emilia Romagna).

Although reducing the precision of the metric, intervals are useful to represent accessibility indexes when we know before hand that only a fraction of the website pages were analyzed. Intervals, in such cases, lead to comparison statements that have a measurable level of certainty.

5 Conclusions

MAMBO is a metric that can be used with data gathered from barrier walkthrough accessibility reports, containing estimates of the severity of accessibility barriers. Using these estimates, the probabilistically-based schema used by MAMBO leads to sound accessibility indexes.

MAMBO is flexible: it can be used for generically comparing websites or accessibility reports; for numerically estimating the effects of judgment errors; and for estimating the uncertainty levels due to an accessibility investigation that was limited to few pages. Provided that appropriate severity judgments are applied also on conformance reviews (like those based on WCAG 2.0), then MAMBO can be used on those reports as well.

The level of experience in accessibility of web technologies, in accessibility evaluations, and in assistive technologies obviously affect the outcome of MAMBO. Although appearing robust with respect to judging mistakes, by providing incorrect ratings of severities any results can be produced, and MAMBO has no intrinsic correction mechanism. Incorrect rating could reflect more false negatives, more false positives, different distributions of these among user categories, or unbalanced judgment of severities for the true positives. However, the study we reported here was performed on reports written by novice evaluators where a substantial number of judging mistakes were made, and nevertheless the confidence intervals produced through MAMBO were shown to be relatively small.

But this study was limited to false positives; a limit of MAMBO is its inability to cope with false negatives, *i.e.* accessibility barriers that are missed by evaluators. We plan to investigate if appropriate merging of semi-automatic and

manual evaluation techniques can provide some reliable estimate of false negatives. Another future research opening is to set up a web accessibility observatory based on MAMBO and SAMBA and to determine how MAMBO correlates with SAMBA and with failure-rate based metrics, such as WAQM.

Acknowledgements

I would like to thank Chiara Cecotti for her help in judging students' reports, and the anonymous reviewers for their helpful suggestions.

References

[1] Arrue, M., Vigo, M., Abascal, J.: Quantitative metrics for web accessibility evaluation. In: Proc. of the ICWE 2005 Workshop on Web Metrics and Measurement (2005)

[2] Arrue, M., Vigo, M., Abascal, J.: Web accessibility awareness in search engines results. Int. Journal on Universal Access in the Information Society (2008); In press

[3] Bailey, J., Burd, E.: Tree-map visualization for web accessibility. In: COMPSAC 2005: Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC 2005), vol. 1, pp. 275–280. IEEE Computer Society, Washington (2005)

[4] Brajnik, G.: Web Accessibility Testing: When the Method is the Culprit. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 156–163. Springer, Heidelberg (2006)

[5] Brajnik, G., Lomuscio, R.: SAMBA: a semi-automatic method for measuring barriers of accessibility. In: Trewin, S., Pontelli, E. (eds.) 9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS, Tempe, AZ, October 2007. ACM Press, New York (2007)

[6] Brajnik, G.: Web accessibility testing with barriers walkthrough (March 2006) (Visited May 2008),`www.dimi.uniud.it/giorgio/projects/bw`

[7] DRC. Formal investigation report: web accessibility. Disability Rights Commission (March 2006) (Visited January 2006),
`www.drc-gb.org/publicationsandreports/report.asp`

[8] Gray, W.D., Salzman, M.C.: Damaged merchandise: a review of experiments that compare usability evaluation methods. Human–Computer Interaction 13(3), 203–261 (1998)

[9] Sears, A.: Heuristic walkthroughs: finding the problems without the noise. Int. Journal of Human-Computer Interaction 9(3), 213–234 (1997)

[10] Slatin, J., Rush, S.: Maximum Accessibility: Making Your Web Site More Usable for Everyone. Addison-Wesley, Reading (2003)

[11] Sullivan, T., Matson, R.: Barriers to use: usability and content accessibility on the web's most popular sites. In: Proc. of ACM Conference on Universal Usability, pp. 139–144 (2000)

[12] Velleman, E., Velasco, C.A., Snaprud, M., Burger, D.: D-WAB4 Unified Web Evaluation Methodology (UWEM 1.0). Technical report, WAB Cluster (2006)

[13] W3C/WAI. How people with disabilities use the web. World Wide Web Consortium — Web Accessibility Initiative (March 2004), http://w3.org/WAI/EO/Drafts/PWD-Use-Web/20040302.html
[14] Zeng, X.: Evaluation of Enhancement of Web Content Accessibility for Persons with Disabilities. PhD thesis, University of Pittsburgh (2004)