SAMBA: a Semi-Automatic Method for Measuring Barriers of Accessibility

Giorgio Brajnik and Raffaella Lomuscio
Dipartimento di Matematica e Informatica
Università di Udine
Via delle Scienze, 206 — 33100 Udine — Italy
giorgio@dimi.uniud.it, raffaellalomuscio@gmail.com

ABSTRACT

Although they play an important role in any assessment procedure, web accessibility metrics are not yet well developed and studied. In addition, most metrics are geared towards conformance, and therefore are not well suited to answer questions whether the web site has critical barriers with respect to some user group.

The paper addresses some open issues: how can accessibility be measured other than by conformance to certain guidelines? How can a metric merge results produced by accessibility evaluation tools and by expert reviewers? Does it consider error rates of the tool? How can a metric consider also severity of accessibility barriers? Can a metric tell us if a web site is more accessible for certain user groups rather than others?

The paper presents a new methodology and associated metric for measuring accessibility that efficiently combine expert reviews with automatic evaluation of web pages. Examples and data drawn from tests performed on 1500 web pages are also presented.

Categories and Subject Descriptors: H.1.2 Information systems: Online Information Services, Web-based Services; H.5.2 Information Interfaces and Presentation: User Interfaces, Evaluation/Methodology; K.6.4 Management of Computing and Information Systems: System Management, Quality Assurance, Management Audit.

General Terms: Human Factors, Measurement.

Keywords: Web Accessibility, Accessibility Metric, Accessibility Evaluation Method, Quality Assessment.

1. INTRODUCTION

"You can't control what you can't measure" [7] is a well known statement in software engineering: we think it applies very well to web accessibility, and that accessibility metrics, *i.e.* procedures to follow in order to represent the accessibility status of a web site as a single value, constitute an open research area. Although not explicitly mentioned, every time we evaluate accessibility we apply some metric (*e.g.* when carrying out a conformance test with respect to Section 508, we check if the number of violated requirements is 0).

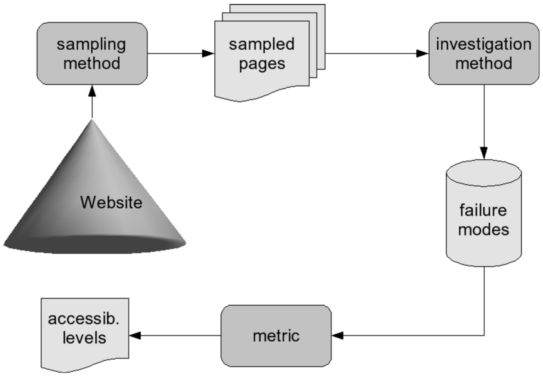

Figure 1: Processes taking place when evaluating accessibility.

Measuring accessibility requires several processes to take place (Figure 1): unless the web site is trivial in size and complexity, whenever we evaluate its accessibility a sampling process has to take place to select the pages to be analyzed, followed by application of some investigation method, and some way to determine the accessibility level. For example, the selection process may follow an *ad hoc* procedure (*e.g.* use the home page, site map, contact page, etc.), a random walk, a uniform random sampling, etc.; see [6] for a discussion and comparison of several sampling methods. The investigation method may be based on automatic testing, debugging with assistive technology, user testing, conformance testing, barrier walkthrough, etc. Finally, conclusions are drawn by summarizing data using a metric: *e.g.* counting violated checkpoints, or adopting the WAQM [1, 14] or UWEM [13] metric (see the survey in Section 2).

Each of the processes illustrated in Figure 1 have an impact on the final result. For example, in the Target legal case (see `www.jimthatcher.com/law-target.htm` for details), the National Federation for the Blind (NFB) claims that target.com is not accessible since some NFB's witnesses gave up when using the site; on the other hand, Target's witnesses testified that they were able to navigate, shop and that they actually enjoyed it; in addition, an NFB's expert declared in court that target.com fails to address accessibility since:

ASSETS'07, October 15–17, 2007, Tempe, Arizona, USA.

> ... 15 of the site's pages were analyzed: six top-level pages as well as nine pages that had to be navigated in order to complete a purchase. In those fifteen pages, alt-text was missing on 219 active images (links); none of the form controls were properly labeled; and there was no accommodation for screen reader or keyboard navigation, such as skip links or HTML headings.

Finally, the Court concluded that the question of the accessibility of target.com was not decided and so refused to grant a preliminary injunction.

We can see that there is substantial variability, and lack of standardization, in the way pages were selected, in the way accessibility was investigated, and in the way a conclusion was drawn.

This is one of the reasons we believe research in accessibility metrics should be pursued. Most metrics are based on automatically tested conformance with respect to checkpoints of some accessibility guidelines (very often WCAG 1.0), and they consider priority levels associated to checkpoints (if available). We think that these approaches carry some open issues that need to be investigated:

1. Can one measure accessibility rather than conformance? Would that be useful and viable? Would that yield accessibility levels that are substantially different than those obtained through conformance-based metrics?

2. Given that automatically tested conformance is affected by errors (see [2, 11] for discussions on and comparisons of accessibility testing tools), how could human judgment be combined with automatic testing so that the measured values reflect also the error rate of the tool?

3. Could a metric be defined so that it considers also the effect of an accessibility barrier on end users (rather than considering only the priority level of checkpoints — which in some cases are not so closely related to impact, *e.g.* WCAG 2.0)?

4. Could we use a metric to understand how accessible a web site would be for a given user category (*e.g.* people with motor disabilities)? This could be useful in quality assurance processes when the accessibility policy favors certain user groups rather than others (*e.g.* because it is much more difficult to implement accessibility solutions that are effective for cognitively disabled users).

The purpose of this paper is to investigate these issues. In particular we propose a measuring methodology and a metric that merge human judgments with automatic testing, that go beyond conformance, and that provide values useful to understand how accessible a web site is with respect to some specific user category. The metric and methodology were experimentally tested on about 1500 pages belonging to several web sites.

2. ACCESSIBILITY METRICS

In order to distinguish accessibility from conformance, we adopt the following definition, derived from [9, 12]: "a web site is accessible if people with some impairment can use it with the same effectiveness, security and safety as non-disabled people". In this context, a *failure mode* of a web site is any accessibility hindrance that appears when somebody uses the web site.

Accessibility metrics, *i.e.* rules and procedures to analyze failure modes and yield a value, are needed to achieve several goals, including [14]: ranking web pages within search engines results according to their accessibility level; monitoring adoption of accessibility regulations and standards; monitoring penetration of accessibility in given areas or site genres; quality management and monitoring of accessibility levels of a single web site over time. More specifically the following relevant tasks can be identified.

Conformance testing Claiming that a web site has achieved a given conformance level (like "A", "AA", etc.) relies on the ability to count how many checkpoints failed on its pages, which is an example of an *ordinal* and *absolute* metric (since conformance levels are strictly ordered, and their values are independent from the specific web site being tested).

Monitoring trends of a given web site When developing or updating a web site (which occurs very often and with unpredictable changes in quality for "Web 2.0" web sites), or revamping it, an effective quality assurance process calls for ways to quantify accessibility. Such metrics need to be at least *ordinal* (or, even better, *quantitative* so that one has an idea of how much accessibility has changed); they can be *relative* to the specific web site as long as comparisons don't span different web sites.

Analysis of a single web site When analyzing a specific web site to identify critical areas, to compare its accessibility with respect to different disability groups, to rank pages according to their accessibility, then an *ordinal* and *relative* accessibility metric is needed.

Comparing different web sites When a comparison of different web sites has to be carried out, for example to generate a nation-wide or sector-wide ranking, or when doing a competitive analysis based also on accessibility, then a *quantitative* and *absolute* metric is necessary.

Over the past years a few accessibility metrics were defined. Sullivan and Matson [10] define the *failure rate* (FR) on the basis of a subset of WCAG 1.0 checkpoints. The FR of a page is defined as the number of violations of any of those checkpoints divided by the maximum number of violations of any of those checkpoints that can take place (*i.e.* by the number of *possible* violations). By doing so, two pages that include 10 images, one with 3 properly defined "alt-text", the other with 8, have FR = 0.7 and FR = 0.2 respectively. The advantage of such a metric lies in its simplicity: it's easily interpretable, it provides normalized, quantitative and absolute values. However, with a large set of checkpoints, FR values tend to be close to 0, reducing the ability to discriminate among web sites that are fairly accessible. Furthermore, such a metric is based on conformance rather than accessibility (as defined above) and it does not consider severity of detected violations.

[15] defines the *Web Accessibility Barrier Score* (WAB), for a web site constituted by N_P pages $\{p, ...\}$, analyzed by

WCAG 1.0 checkpoints $\{c, ...\}$ having priorities $priority_c$, as WAB $= \frac{1}{N_P} \sum_p \sum_c \left(\frac{FR(p,c)}{priority_c} \right)$, where $FR(p,c)$ is the failure rate of checkpoint c on page p.

Also in this case, a high WAB means a low accessibility level; WAB yields quantitative absolute values that are not normalized. This metric is based on conformance, and considers also the priority level of the checkpoints: higher priorities mitigate corresponding failure rates.

The Unified Web Evaluation Methodology 1.0 (UWEM) [13] is a methodology designed to assess accessibility by merging manual and automatic evaluations. Within such a context, the accessibility value of a page p is $AV_p = 1 - \prod_c (1 - R_{pc} F_c)$, where F_c is the probability that a violation of checkpoint c results in a failure mode (F_c is assumed to be constant, set to 0.05), whereas R_{pc} is a factor that depends on the agent who carries out the evaluation of the checkpoint. If it is a tool, then R_{pc} is the failure rate combined with the error rate of the tool; if it is an expert, then R_{pc} is the probability of error of the expert[1]. AV_p gives the probability that the page leads to a failure mode; its mean over all the pages gives the accessibility value for the entire web site.

The advantages of this conformance-based metric are that it yields a quantitative, normalized and absolute value, and that it is rooted on a clean mathematical background. In addition, the metric (in its more general definition) considers error rates, the impact on users, and supports integration of manual with automatic evaluations. However, from a practical viewpoint, no suggestions are given on how to reliably estimate these parameters.

Another metric is Web Accessibility Quantitative Metric (WAQM) [1], which provides a solution to some of the problems discussed above. On the basis of automatic testing of WCAG 1.0 checkpoints, WAQM considers the failure rate of each checkpoint on each page: $FR(p,c)$. Such a failure rate is transformed, through a piecewise linear function that approximates a hyperbole, to values that are more spread out as they get close to 0. These values are then weighted by priority of the checkpoint, and finally weighted by $d_p = e^{-i}$, where p is a page and i is its depth level in number of links from the home page (for which $i = 0$). Weights associated to priorities, slopes and intercepts of the piecewise linear function need to be experimentally tuned. A further study [14] discusses the dependence of the WAQM metric on the specific tool being used, and shows that although the numeric values produced by WAQM are tool-dependent, the ranking of web pages and web sites does not depend significantly on the tool.

Advantages of WAQM include that it produces normalized quantitative values dependent on checkpoint priorities. On the other hand, it does not consider the error rate of the tool being used, and it does not provide means to combine automatic and manual evaluations.

The last metric we mention is the evaluation form used by the Accessibility Internet Rally (AIR) judges. In this competition among web developers (designed and managed by Knowbility), a web site is ranked according to points given to it by human judges. A spreadsheet[2] is used to collect the data and compute the score on the basis of penalty points associated to certain defects: for example, a deduction up to 20 points (out of 320) for images bearing information that have no proper alternative text. The criteria used by the judges include accessibility and usability aspects (*e.g.* aesthetics is also considered) and are not directly related to WCAG or Section 508 checkpoints, although addressing all important accessibility barriers.

While providing a structured way to compute an accessibility score, AIR is not based on automatic testing tools (although judges can use them) and it does not specify which pages of the web site should be considered. Therefore it is not clear how much AIR scales up when applied to web sites that are highly dynamic or very large.

3. SAMBA

We propose the *Semi-Automatic Method for measuring Barriers of Accessibility* (SAMBA), which is based on the following key points:

1. Using tools to automatically identify potential accessibility barriers;

2. Sampling results that are submitted to human judgment;

3. Statistically estimating — from the sample — false positives and severity of barriers for the entire web site;

4. Grouping barriers by disability types and deriving scores that represent non-accessibility with respect to disability type as well as a global non-accessibility level.

3.1 Barrier walkthrough method

A key element of SAMBA is the method for evaluating accessibility called *Barrier Walkthrough* (BW) [4, 5]: by applying a usability evaluation method called heuristic walkthrough [8], accessibility barriers identified by experts are contextualized within usage scenarios and receive a severity score.

A *barrier* is any condition caused by the web site that hinders user's progress towards achievement of a goal (*i.e.* a failure mode). A barrier is described in terms of: 1. the user category and the type of disability; 2. the type of assistive technology being used; 3. the failure mode (how the activity/task is hindered); 4. which features in the page raise the barrier.

In order to apply the BW method experts have: 1. to define user profiles (*i.e.* type of disability, experience level); 2. to define user scenarios (*i.e.* assistive technology, possible goals and user roles); 3. to select relevant types of barriers from existing lists; 4. to evaluate pages against barriers in the context of scenarios with respect to goals that users may achieve; 5. to estimate severity of detected barriers.

Severity of a barrier, expressed as *minor*, *major* or *critical*, is a function of its *impact* (the extent to which the user goal cannot be achieved) and its *frequency* (how often the barrier shows up when performing the task).

[1] Although [13] mentions the probabilities that the tool and the expert yield false negatives and false positives, no suggestions are given as to how to compute these probabilities, except for assuming that they are 0; in such a case R_{pc} becomes the failure rate $FR(p,c)$ when using a tool, 0 when the expert did not find any violation of c on p, 1 if the expert found at least one violation.

[2] See `www.knowbility.org/air-austin/?content=judgingFAQ` for details about the process and the actual judging form.

Experimental evaluation of the BW method [3] showed that this method is more effective than conformance testing in finding more severe problems and in reducing false positives; however, it is less effective in finding all the possible accessibility problems.

3.2 SAMBA phases

Phase 1: data collection.

To apply SAMBA, an accessibility testing tool needs to be run against a web site. From the set of checkpoint violations reported by the tool, and using a tool-specific correspondence table that maps checkpoints to barrier types listed in [4], a set of *potential barriers* can be computed. Since barrier types are tagged with one or more disability type (eight disabilities are considered in [4]), this classification gives rise to a subset of potential barriers for each disability. The *disability vector* gives the proportion of potential barriers for each disability.

In our experimentation we used LIFT, an accessibility testing web-based tool, configured to test WCAG 1.0 conformance. The checkpoint-to-barrier table maps most LIFT automatic tests to 24 barrier types, and manual tests[3] to additional 11 barrier types; 18 barrier types are not covered by LIFT (mostly associated to priority-3 checkpoints). In our study, LIFT was applied to approximately 1500 pages mirrored from 15 web sites.

Table 1 shows disability vectors for three of the 15 web sites.

Disability	outlook	smh	IrishTimes	range
blind	34	25	25	(1, 34)
cog. disabled	5	6	4	(1, 13)
deaf	1	0	1	(0, 1)
color blind	8	3	7	(0.2, 10)
low vision	27	16	24	(16, 96)
motor disabled	8	9	9	(1, 12)
no JavaScript	18	41	29	(0, 40)
epilepsy	0	0	0	(0,0)

Table 1: Three examples of disability vectors (values shown as percentages). The last column gives the range of the proportion for all the web sites.

Phase 2: human judgment.

At this point, human judgment is needed to get information on the error rate of the tool and on the severity of potential barriers. To achieve this, potential barriers are sampled using a non-proportional stratified sampling method with no replacement, using two strata: barrier types associated with manual tests and those associated with automated tests. Within each stratum a random sampling is performed.

A panel of judges is then asked to analyze selected potential barriers. More specifically, for each potential barrier, they are asked:

1. to view pages involved with the barrier;
2. to think of plausible goals achievable with those pages;
3. to assign a severity to the barrier, with respect to the scenario (goal and disability type).

Judgments are then merged by averaging, among judges, the severity associated to each sampled potential barrier, and using thresholds to reduce the average to values belonging to $\{0, 1, 2, 3\}$, corresponding to false positive (FP), minor, major, critical barriers. From each sample, the *sampled severity matrix* can be generated, giving the breakdown of all sampled potential barriers split by severity. Each element of the matrix $f_{d,s}$, that gives the proportion of sampled barriers associated with disability d and severity s, is defined by $f_{d,s} = \frac{N_A}{N} f_{A,d,s} + \frac{N_M}{N} f_{M,d,s}$, where $\frac{N_A}{N}$ and $\frac{N_M}{N}$ are the proportions of barrier types belonging to strata "automatic" and "manual", respectively, and $f_{A,d,s}$ and $f_{M,d,s}$ are the relative frequency of barriers related to (d, s) in strata "automatic" and "manual", respectively. $f_{d,s}$ is a correct estimator of the probability that a barrier associated with (d, s) occurs in the entire web site.

In our experimentation we computed a sample of 70 potential barriers for each web site, extracted from sets of potential barriers per site whose size varied from 6000 to 182000. A spreadsheet was generated with all the information needed by two judges to determine barrier severity. On the average, it took each judge about 90 minutes to analyze the sample on each web site. Globally, 288 sampled potential barriers were classified as false positive (out of 1050, *i.e.* 27.4%). Table 2 gives an example of the sampled severity matrix.

Disability	**Minor**	**Major**	**Critical**	**FP**	**Total**
blind	16	6	9	4	**35**
cog. disabled	2	2	3	0	**5**
deaf	0	0	0	0	**0**
color blind	0	0	0	0	**0**
low vision	17	5	6	4	**32**
motor disabled	1	1	3	3	**8**
no JavaScript	0	0	7	12	**19**
epilepsy	0	0	0	0	**0**
Total	**36**	**15**	**25**	**23**	**99**

Table 2: Sampled severity matrix for Outlook India (percentages add to 99% because of rounding errors).

A confidence interval is then computed for each relative frequency $f_{d,s}$ of the sampled severity matrix. Given the adopted sampling method and provided samples are large enough, the distribution of $f_{d,s}$ can be approximated by a normal distribution. Hence confidence intervals can be computed using the standard normal distribution, yielding the *confidence intervals severity matrix* which gives the range that $f_{d,s}$ can span due to chance variations.

The confidence interval severity matrix (with $\alpha = 0.05$) for the Outlook India web site is shown in Table 3. For example, $f_{blind,critical}$ is the interval $(6, 12)$, meaning that with probability 95% the percentage of barriers in the entire web site that are critical for blind people ranges from 6% to 12%. Similarly, between 1% and 7% of the barriers relevant for blind people are false positives.

Phase 3: computing accessibility indexes.

To compute the accessibility index we need the confidence interval severity matrix and F, the *barrier density* of a web site. F is defined as $\frac{\text{number of pot. barriers}}{\textit{number of HTML lines}}$, which can be interpreted as the probability that a line of HTML code of the site causes a barrier detected by the tool. In our experiment, F ranged from 0.142 to 0.9357.

[3] Manual tests are warnings suggesting that human judgment is necessary to determine if a violation occurs.

Disability	Minor	Major	Critical	FP
blind	(11,20)	(3,8)	(6,12)	(1,7)
cog. disabled	(0.04,4)	(1,4)	(1,5)	(0,0)
deaf	(0,0)	(0,0)	(0,0)	(0,0)
color blind	(0,0)	(0,0)	(0,0)	(0,0)
low vision	(12,22)	(03,8)	(3,8)	(1,7)
motor disabled	(0.01,2)	(0.01,2)	(1,4)	(1,5)
no JavaScript	(0,0)	(0,0)	(4,9)	(8,16)
epilepsy	(0,0)	(0,0)	(0,0)	(0,0)

Table 3: Confidence interval severity matrix for Outlook India (values as percentages).

If $\vec{D}$ is the disability vector of a web site, then $F \cdot \vec{D}$ is the barrier density split by disability type, *i.e.* the probability that a line of code causes a barrier for that disability. Similarly, if M is the sampled severity matrix, then $F \cdot M_{d,s}$ is the probability that a line of code causes a barrier for disability d with severity s.

The *Raw Accessibility Index* (AI_r) is computed by combining a disability vector $\vec{D}$ with the barrier density factor F: $AI_r = \prod_d (1 - F \cdot \vec{D}_d)^2$, where d is a disability type. Such an index can be read as the square of the probability that no line of code causes a potential barrier; squaring is needed in order to increase the weight of small values. Since $\vec{D}$ does not consider human judgment, AI_r can be easily computed.

If we combine the density factor F with the confidence interval severity matrix $\mathcal{M}$, we get the *Weighted Accessibility Index* (AI_w). Since it is based on confidence intervals, it is itself an interval $(\underline{AI_w}, \overline{AI_w})$, defined as follows:

$$\begin{aligned}
\text{let } \underline{\mathrm{H_d}} &= \frac{\underline{f}_{d,mnr}}{w_{mnr}} + \frac{\underline{f}_{d,maj}}{w_{maj}} + \underline{f}_{d,cri}, \\
\text{and } \overline{\mathrm{H_d}} &= \frac{\overline{f}_{d,mnr}}{w_{mnr}} + \frac{\overline{f}_{d,maj}}{w_{maj}} + \overline{f}_{d,cri}, \\
\text{then } \underline{\mathrm{AI_w}} &= \prod_d \left(1 - F \cdot min\left\{1, \overline{H_d}\right\}\right)^2, \\
\overline{AI_w} &= \prod_d \left(1 - F \cdot \underline{H_d}\right)^2
\end{aligned}$$

Weights associated with severity levels (*i.e.* w_{mnr} and w_{maj}) define the relative importance of severity levels: one critical barrier is equivalent, with respect to values measured by this metric, to w_{maj} major ones or to w_{mnr} minor ones. If weights were equal to 1, then such an interval could be interpreted as the range spanned by the square of the probability that no line of code contains any true barrier.

In our experimentation, we tested both with $w_{mnr} = w_{maj} = 1$ (unweighed AI, AI_u) and with $w_{mnr} = 9$ and $w_{maj} = 3$, *i.e.* where a critical barrier weighs as 9 minor ones and 3 major ones (weighed AI, AI_w).

4. EXPERIMENTAL RESULTS

The confidence interval severity matrix gives information on the disability group that is more or less likely to hit barriers with a given severity. For example, Table 3 shows that blind persons have a 12% chance to hit a critical barrier, whereas motor disabled ones have a 4% chance (in the worst case). These values are relative ones, *i.e.* they are specific for that web site, and cannot be used to compare two different web sites.

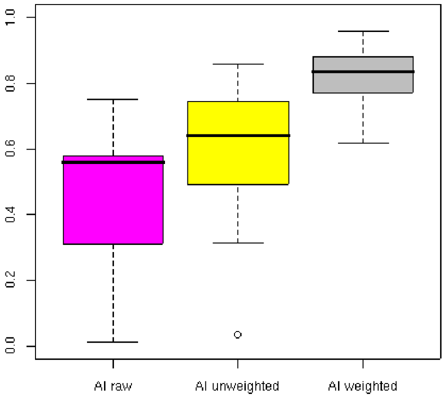

Figure 2: Boxplots of the AI's on 15 web sites; boxplots for AI_u and AI_w are computed on the basis of the middle point of the confidence interval.

In addition, the matrix gives an estimation of the error rate in terms of false positives of the tool, split by disability group. These numbers can be used to monitor correct application of the tool, to fine tune the tool or to compare different tools, for example. From Table 3 one can see that an error rate up to 16% is possible for the "no JavaScript disability", whereas that rate drops to 5% for motor disabilities (in the worst case).

Web site	AI_r	AI_u	AI_w
The Belfast Telegraph	0.56	(0.50,0.78)	(0.77,0.92)
University of Bolton	0.75	(0.74,0.91)	(0.9,0.97)
University of Calgary	0.56	(0.61,0.87)	(0.84,0.96)
University of Cambridge	0.57	(0.61,0.84)	(0.81,0.92)
Daily Express	0.16	(0.13,0.50)	(0.55,0.85)
University of Dundee	0.32	(0.29,0.64)	(0.68,0.9)
The Irish Times	0.15	(0.16,0.56)	(0.5,0.79)
The University of Kansas	0.58	(0.57,0.81)	(0.79,0.92)
Lancaster University	0.52	(0.62,0.88)	(0.75,0.92)
City University London	0.73	(0.78,0.94)	(0.94,0.98)
University of Nigeria	0.01	(0,0.07)	(0.52,0.72)
Outlook India	0.42	(0.39,0.68)	(0.65,0.85)
University of Pretoria	0.70	(0.74,0.90)	(0.90,0.97)
The SMH	0.30	(0.33,0.71)	(0.72,0.92)
Berkeley University	0.58	(0.48,0.75)	(0.74,0.9)

Table 4: Accessibility indexes for tested web site.

Table 4 gives the values for the three indexes over all the web sites involved in our study.

Figure 2 shows that AI_w has a much smaller range than the other two indexes (range: 0.62 to 0.96; 1.st and 3.rd quartile: 0.77 and 0.88). This is due to the smaller variability of critical barriers, which weigh a lot. In addition, AI_w ranges higher than the other ones; again, this is due to the higher weight given to critical barriers.

AI_u is more spread than AI_w (range: 0.035 to 0.86; 1.st and 3.rd quartiles: 0.49 and 0.74) and slightly lower than AI_w because all barriers are weighed the same.

Finally, AI_r covers a wider range (from 0.01 to 0.75), has a spread similar to AI_u (1.st and 3.rd quartiles: 0.31 and 0.58) and is smaller than AI_u. This is because AI_r is also

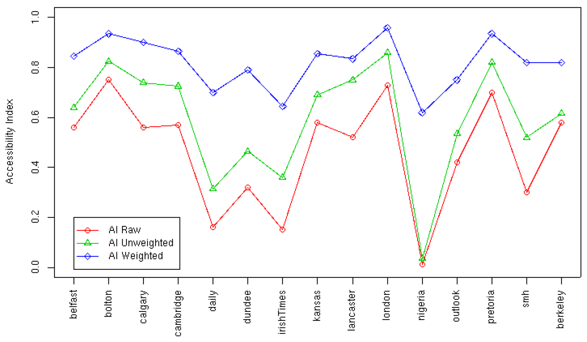

Figure 3: Distribution of the AI's over 15 web sites; AI_u and AI_w are computed on the basis of the mid point of intervals.

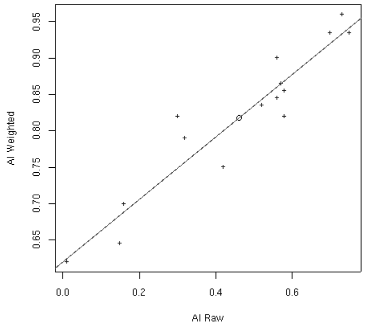

Figure 4: Scatterplot of AI_r against mid points of AI_w.

a function of false positives. Its distribution is negatively skewed (median: 0.56), whereas in the other cases it is quite symmetrical. This indicates that, in terms of *potential* barriers, 25% of the sites have scores between 0.56 and 0.58, and 50% of the sites have scores between 0.56 and 0.75.

Therefore, with AI_w we loose in resolution but gain in usefulness and validity: AI_w is more valid than AI_r since it excludes false positives; it is more useful than AI_u because barriers are weighted.

Figure 3 confirms such interpretations. The distance between the two bottom lines (*i.e.* $AI_u - AI_r$) gives an idea of the proportion of false positives that were found on each web site: it is highest for `lancaster` and lowest for `nigeria`. The distance between the top and middle lines (*i.e.* $AI_w - AI_u$) gives an idea of the effect of weights on each web site: the higher the distance, the higher the proportion of minor and major barriers; it is highest for `nigeria` and lowest for `lancaster`.

Figure 4 shows that there is a strong linear correlation between AI_r and AI_w (Pearson's coefficient is 0.945). Therefore, one can use a linear model to predict AI_w from AI_r, with a substantial reduction of effort since no additional judging would be needed: $f(x) = 0.62068 + 0.42907x$ gives an accurate estimate of AI_w midpoints starting from AI_r, whose maximum error is 7%, provided that this model is applied to pages and scenarios similar to the ones we tested. Application of this model would be then particularly useful to predict changes of AI_w when analysis of the same web site(s) is repeated.

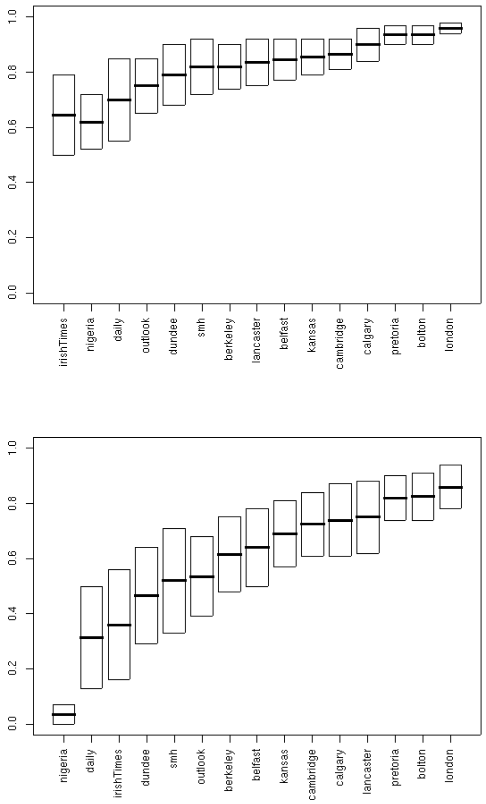

Figure 5: Distribution of intervals for AI_w (top) and AI_u (bottom).

Figure 5 shows the distribution and width of the intervals associated with AI_w and AI_u. Widths are smaller for AI_w, due to smaller variability in the number of critical and major barriers. They are, however, more overlapping. This is another factor that reduces resolution of the AI_w metric, since overlapping intervals mean that two web sites cannot be soundly compared.

Intervals of AI_w induce a partial order relationship on web sites, based on whether intervals overlap or not. Table 5 shows the pair-wise distance between AI_s for some of the sites. The distance between two intervals is 0 if they overlap or just touch, and it is the difference of adjacent endpoints otherwise. From the table it's easy to see that, for example, `london` is more accessible than most other web sites, except for `calgary`, `bolton` and `pretoria`. It is also very close to five others, while being 22 points more accessible than `nigeria`. Since AI_w is computed from the confidence intervals severity matrix, these differences are valid (because free from tool errors), appropriate for accessibility (because they consider severities), significant (because they are conservative) and reliable (because they are statistically inferred).

Figure 6 shows the values of the three accessibility indexes coupled with the values of the WAQM metric and computed on the same web sites using checkpoint violations computed

	...	belfast	kansas	cambridge	calgary	bolton	pretoria	london
irishTimes		0	0	0.02	0.05	0.11	0.11	0.15
nigeria		0.05	0.07	0.09	0.12	0.18	0.18	0.22
daily		0	0	0	0	0.05	0.05	0.09
outlook		0	0	0	0	0.05	0.05	0.09
dundee		0	0	0	0	0	0	0.04
smh		0	0	0	0	0	0	0.02
berkeley		0	0	0	0	0	0	0.04
lancaster		0	0	0	0	0	0	0.02
belfast			0	0	0	0	0	0.02
kansas				0	0	0	0	0.02
cambridge					0	0	0	0.02
calgary						0	0	0
bolton							0	0
pretoria								0
london								

Table 5: Portion of the distance table between sites; for example, to go from `nigeria` (row) to `belfast` (column), AI_w increases by 0.05**.**

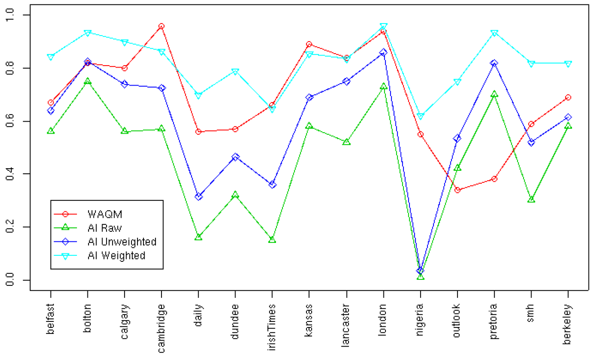

Figure 6: Comparison of AI's with respect to WAQM (using midpoints for AI_u and AI_w).

by LIFT, whilst disregarding the final weighting by page level (*i.e.* the WAQM accessibility value of the site is the mean value of its pages).

Correlation between WAQM and AI_r, AI_u and AI_u is rather weak: Pearson's coefficient is 0.43 for AI_r, 0.47 and 0.44 for AI_w. Spearman's rank coefficient is slightly higher: 0.54, 0.59 and 0.58. Such moderate correlations suggest that WAQM on the one hand and AI_r, AI_u and AI_w on the other, measure different things, as we expected.

Finally, figures 7 and 8 illustrate AI_u and AI_w intervals upon which values of WAQM are superimposed. It can be readily seen that not only values differ numerically, but also that the rankings of web sites induced by the metrics differ.

A preliminary sensitivity analysis showed that slight perturbations of human judgments, of web site size, and of weights used in AI_w yield limited changes in the metrics values. It appears therefore that AI_w is also robust.

5. CONCLUSION

The SAMBA methodology effectively and efficiently integrates automatic evaluations of accessibility of large and dynamic web sites with human judgments applied in the con-

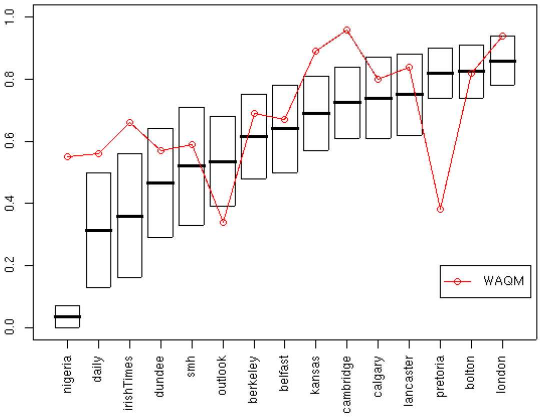

Figure 7: Comparison of AI_u with respect to WAQM.

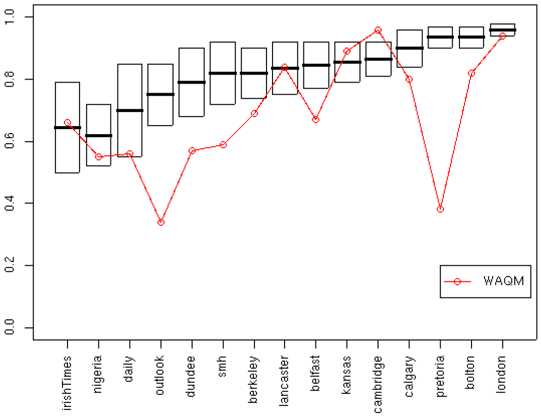

Figure 8: Comparison of AI_w with respect to WAQM.

text of the Barrier Walkthrough analysis method. Human judgments identify the error rate, in terms of false positives, of the tool being used and assign severity scores to sampled barriers. SAMBA should scale up well with increasing sizes of web sites, offering a viable solution for measuring and monitoring web accessibility.

Although the Weighted Accessibility Index has a limited resolution when compared with other known metrics, it is free from tool errors, it is focused on accessibility rather than conformance, it considers severity of detected barriers, it is valid since reported differences are significant, and it is robust with respect to slight changes in the data. The Weighted Accessibility Index is just one of the several indexes that can be generated through SAMBA; other ones can be used to compare the accessibility of a web site with respect to several user groups. Under certain circumstances, values of the Weighted Accessibility Index can be easily predicted through a linear model from row data produced by the testing tool, without requiring further human judgments.

SAMBA is independent from the adopted testing tool, provided that the test-to-barrier map is defined. We did not investigate how much the accessibility indexes change when tools are changed. Another study showed that numeric values of WAQM when applied to data produced by different tools cannot be compared, whereas induced rankings are more reliable [14]. A similar effect is expected for SAMBA since the failure density factor and the error rate will be highly affected by different tools.

Although SAMBA considers only false positives as a source of error produced by a tool, this is only one side of the coin. Tools can lead to wrong data also because of false negatives, *i.e.* problems that are missed. Extending SAMBA to cope with false negatives requires that judges are asked to analyze pages also to determine how many undetected barriers can be found; then, the sampled pool of barriers needs to be extended with these new barriers, and appropriate changes to the accessibility indexes need to be formulated. This is one of the directions we will pursue.

Finally, the Weighted Accessibility Index has not been independently validated against web sites whose accessibility is known. This is another research direction we will pursue.

6. ACKNOWLEDGMENTS

We would like to thank the team of the Laboratory of Human-Computer Interaction for Special Needs of the University of the Basque Country, and especially Markel Vigo, for their help in setting up the test web sites and the WAQM calculations. Many thanks also to Jim Thatcher for suggesting us to consider the AIR metric and process.

7. REFERENCES

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

[2] G. Brajnik. Comparing accessibility evaluation tools: a method for tool effectiveness. *Universal Access in the Information Society*, 3(3-4):252–263, Oct 2004. `www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10209-004-0105-y`.

[3] G. Brajnik. Web Accessibility Testing: When the Method is the Culprit. In K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *ICCHP 2006, 10th International Conference on Computers Helping People with Special Needs*, Lecture Notes in Computer Science 4061, Linz, Austria, July 2006. Springer Verlag.

[4] G. Brajnik. Web accessibility testing with barriers walkthrough. `www.dimi.uniud.it/giorgio/projects/bw`, March 2006. Visited: May 2007.

[5] G. Brajnik. Ranking websites through prioritized web accessibility barriers. In *Technology and Persons with Disabilities Conference*, Los Angeles, March 2007. CSUN, California State University Northridge. `www.dimi.uniud.it/giorgio/papers/csun07.pdf`.

[6] G. Brajnik, A. Mulas, and C. Pitton. Effects of sampling methods on web accessibility evaluations. In S. Trewin, editor, *9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, Tempe, AZ, Oct 2007. ACM Press.

[7] T. DeMarco. *Controlling Software Projects: Management, Measurement, and Estimates*. Prentice Hall PTR, 1986.

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

[9] J. Slatin and S. Rush. *Maximum Accessibility: Making Your Web Site More Usable for Everyone*. Addison-Wesley, 2003.

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

[11] J. Thatcher, M. Burks, C. Heilmann, S. Henry, A. Kirkpatrick, P. Lauke, B. Lawson, B. Regan, R. Rutter, M. Urban, and C. Waddell. *Web Accessibility: Web Standards and Regulatory Compliance*. FriendsofED, 2006.

[12] U.S. Government. SEC. 508. ELECTRONIC AND INFORMATION TECHNOLOGY, 1998 Amendment to Section 508 of the Rehabilitation Act. `www.section508.gov/index.cfm?FuseAction=Content&ID=14`, 1998. Visited: January 2007.

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

[14] M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio, and J. Abascal. Quantitative metrics for measuring web accessibility. In *W4A '07: Proceedings of the 2007 International Cross-Disciplinary Workshop on Web Accessibility (W4A)*, pages 99–107, New York, NY, USA, 2007. ACM Press. `www.dimi.uniud.it/giorgio/papers/w4a07.pdf`.

[15] X. Zeng. *Evaluation of Enhancement of Web Content Accessibility for Persons with Disabilities*. PhD thesis, University of Pittsburgh, 2004.