Quantitative Metrics for Measuring Web Accessibility

Markel Vigo
University of the Basque Country
Informatika Fakultatea
20018 Donostia, Spain
+34 943 015113
markel@si.ehu.es

Myriam Arrue
University of the Basque Country
Informatika Fakultatea
20018 Donostia, Spain
+34 943 015160
myriam@si.ehu.es

Giorgio Brajnik
Dip. di Matematica e Informatica
Università di Udine
Udine, Italy
+39 0432 558445
giorgio@dimi.uniud.it

Raffaella Lomuscio
Dip. di Matematica e Informatica
Università di Udine
Udine, Italy
+39 0432 558445
raffaella.lomuscio@gmail.com

Julio Abascal
University of the Basque Country
Informatika Fakultatea
20018 Donostia, Spain
+34 943 018067
julio@si.ehu.es

ABSTRACT

This paper raises the need for quantitative accessibility measurement and proposes three different application scenarios where quantitative accessibility metrics are useful: Quality Assurance within Web Engineering, Information Retrieval and accessibility monitoring. We propose a quantitative metric which is automatically calculated from reports of automatic evaluation tools. In order to prove the *reliability* of the metric, 15 websites (1363 web pages) are measured based on results yielded by 2 evaluation tools: EvalAccess and LIFT. Statistical analysis of results shows that the metric is dependent on the evaluation tool. However, Spearman's test produces high correlation between results of different tools. Therefore, we conclude that the metric is reliable for ranking purposes in Information Retrieval and accessibility monitoring scenarios and can also be partially applied in a Web Engineering scenario.

Categories and Subject Descriptors

H.3.5 [Online Information Services] *Web-based services;* I.7.5 [Document Capture] *Document analysis;* H.1.2 [User/Machine Systems] *Human factors*;

General Terms

Measurement, Human Factors, Verification.

Keywords

Web accessibility, automatic evaluation, quality assurance, metrics, measurement.

W4A2007, May 07–08, 2007, Banff, Canada. Co-Located with the 16th International World Wide Web Conference.

1. INTRODUCTION

In recent years, the interest in the development of metrics for measuring the accessibility level of web contents has significantly spread out. In this sense, different initiatives have been launched in the European Union such as the following research projects: European Internet Accessibility Observatory[1] (EIAO) and Supporting the creation of an e-Accessibility Mark[2] (Support-EAM). Both are related to the development of quantitative metrics for web accessibility since the former aims at creating rankings showing accessibility indicators of online European content whereas the latter's goal is to define an evaluation methodology and create an e-Accessibility Quality Mark.

Currently, the most broadly accepted measures for web accessibility are the qualitative levels proposed in WCAG 1.0 by the WAI: 0, A, AA and AAA. This approach is not precise enough since a website fulfilling only all priority 1 checkpoints would obtain the same accessibility value than another website fulfilling all priority 1 checkpoints and almost all priority 2 checkpoints: both of them would get the A level conformance. These criteria seem to be based in the assumption that if a webpage fails to accomplish one of the guidelines in a level, it is so un-accessible as if it fails to fulfil all of them. That is true for some users, but in general it is essential to have not only a reject/accept validation, but a more precise measurement scale of accessibility.

Thus, as stated in [9], defining quantitative metrics is necessary in order to overcome this situation. The following ones are some scenarios which would benefit from quantitative accessibility metrics.

1.1 Quantitative Assurance (QA) within Web Engineering

Web Engineering defines specific methodologies, models and techniques for web applications development. The final objective

[1] http://www.cerlim.ac.uk/projects/eiao/index.php

[2] http://www.support-eam.org/supporteam/default.asp

when engineering web applications is to obtain high quality products. Therefore, a Quality Assurance process is essential. This implies the necessity of applying metrics, methods and quality models throughout the development process.

In this sense, some quality models such as 2QCV3Q [8] and WebQEM [9] have been defined. The characteristics of web applications and the necessary metrics for their evaluation are included in these models. In both quality models, web accessibility is an attribute which should be measured in order to guarantee the quality of the product. For this reason, quantitative metrics are essential in order to accurately measure usability and accessibility properties since both are needed for QA. The measurement of web usability and accessibility should be performed in the different stages of the life cycle of a web application.

In addition, ranking prototypes of a product according to their accessibility level may be useful in order to assess the impact of changes, updates in functionalities, etc. in any iterative development process. Therefore, quantitative metrics which could be applied for measuring and ranking prototypes according to their accessibility level would be useful.

1.2 Information Retrieval

Ivory et al. [7] carried out a study with visual impaired users in order to determine the factors which would improve search engine results for those users. It is concluded that some users desire to know additional details in search results displays and it is suggested to have results sorted according to accessibility or usability criteria. Re-ranking results according to users' visual abilities would improve their search experience.

In this sense, Google has launched "Google Accessible Search" [5] where results are ordered according the criteria stated in their FAQ: "pages with few visual distractions and pages that are likely to render well with images turned off". However, this is not a comprehensive approach since only some guidelines for visually impaired users are being considered and users with other type of disabilities are not taken into account.

1.3 Accessibility Monitoring

Once a website has been developed, keeping track of the evolution of its accessibility level has a paramount importance since it may be framed by legal restrictions. Due to the nature of the WWW, updates in websites are quite frequent. Since updates can decrease the accessibility level, some websites find themselves in limbo situation which can result in administrative fines. Therefore, monitoring of the evolution of web accessibility requires accurate metrics in order to avoid those circumstances. These metrics should be used for ranking purposes so ordinal data may be enough.

This accessibility monitoring process may be also helpful for public institutions in order to keep track of their e-government websites' accessibility level. In addition, it may be useful for comparing the accessibility level of different websites and creating ranking lists as an accessibility observatory.

In this paper we propose and validate a quantitative metric for measuring accessibility of web sites. The proposed metric is automatically calculated from the output of evaluation tools. The goal of the validation is to demonstrate the independence of the metric results in respect of the output of different tools.

2. RELATED WORK

Some previous related work has been done as far as quantitative accessibility metrics are concerned. Sullivan and Matson [11] evaluate eight checkpoints from WCAG 1.0. As a result, the so-called "failure-rate" is a proportion between potential points of failure and checked real errors. Therefore, the result range goes from 0 to 1. It is a naive approximation since other factors such as error impact, error nature (whether checkpoints are automatic errors, warnings or generic problems) and other requirements explained in the following section are not taken into account.

$$failure\_rate = \frac{real\_errors}{potential\_failures}$$

Hackett et al. [6] proposed the WAB formula (Web Accessibility Barrier). This formula uses as input parameters the total pages of a website, total accessibility errors as well as potential errors in a web page and error priority. However, the returned ratings are not restricted to a limited range of values. Therefore, it can be useful only for ranking web pages according to their accessibility level. The drawback of this metric is that considering the result for a unique web page, it is not possible to have an accessibility reference since there are no boundaries for good or bad accessibility levels. The formula for a single web page is calculated for all WCAG checkpoints found in the page:

$$WAB\_score = \sum \frac{real\_errors}{potential\_errors \times priority}$$

Bühler et al. [4] propose a novel approach in order to adapt measurement to different disabilities groups. Therefore aggregation models are proposed. A simplification of that model is the following:

$$A(u) = 1 - \prod \left(1 - R_b S_{ub}\right)$$

Where R is the evaluation report and S is a severity value from 0 to 1 (for each barrier type b and user group u). However, these metric is still in a developing stage until better results are obtained. This metric is supposed to be integrated in the web accessibility benchmarking framework defined in [10].

Our quantitative metric proposal faces the challenge of overcoming the drawbacks that these metrics contain: they do not cover all aspects of web accessibility since some of them are focused on specific user groups or limited to some guidelines. In addition, they are not empirically validated nor contrasted with expert manual evaluation results. Finally, our metric is calculated automatically, which is an advantage over the other approaches. However, our metric adopts the good points of these metrics as can be observed in the following section.

3. WEB ACCESSIBILITY QUANTITATIVE METRIC

The Web Accessibility Quantitative Metric (WAQM) is calculated automatically from evaluation reports yielded by EvalAccess evaluation tool, a tool developed by the Laboratory of Human-Computer Interaction for Special Needs of the University of the Basque Country [1].

The main feature of EvalAccess is that the evaluation engine is independent of the accessibility guidelines. This implies that it can evaluate any guideline-set if the set follows a determined XML-Schema. As can be observed in Figure 1, it is a modular system consisting of a Guidelines Manager which retrieves guidelines from the Guidelines Repository and provides the Evaluation Module with them. Evaluation results are formatted in XML and yielded by the Reports Manager. In addition, it is implemented as a Web Service so that other applications can use it. Taking advantage of the modular architecture, a new layer which calculates the WAQM from the evaluation reports in XML has been deployed. As a result, accessibility quantitative metrics can be automatically calculated.

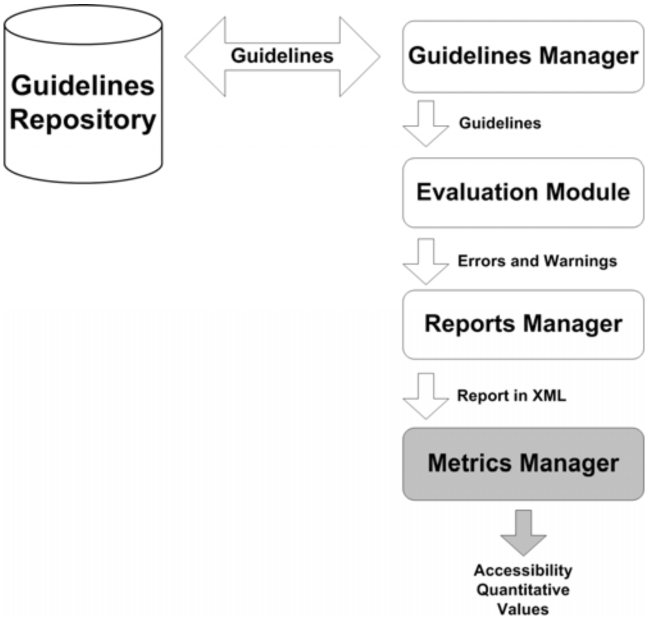

Figure 1. Architecture of EvalAccess

Accessibility problems in evaluation reports are classified by EvalAccess in three main groups:

- Automatic tests (errors): these problems should not require human judgment to check their validity.
- Manual or semi-automatic tests (warnings): human judgment is necessary to check potential problems associated to particular fragments of code implementing the page.
- Generic problems: human judgment is necessary to check potential problems that cannot be associated to any code fragments; these problems arise in every web page. E.g. WCAG 1.0 14.1 checkpoint: "Use the clearest and simplest language".

The requirements of the metric, the assumptions we made and the factual information found out during the research are described below.

Requirement 1: *The result of the metric should be normalized.*

In order to classify web sites according to their accessibility a limited ratio scale from 0 to 100 is chosen so that results of the final quantitative accessibility value are expressed in a percentage scale. The closer the result of the metric is to 0 the less accessible the web site is and the closer to 100 is the more accessible it is. This leads us to classify web pages according to their accessibility guidelines conformance percentage.

Requirement 2: *The metric should give one value for each accessibility attribute, as well as an overall value for each page.*

Although automatic accessibility evaluation reports returned by EvalAccess refer to WCAG 1.0 guidelines, these are mapped into WCAG 2.0 guidelines: Perceivable, Operable, Understandable and Robust. [http://www.w3.org/TR/WCAG20/appendixD.html]

Apart from an accessibility quantitative value for each guideline, an overall accessibility value based on POUR guidelines is also calculated. However, metric calculation according to these guidelines is useful to get a general idea of how accessible a page is.

Assumption 1: *Besides total number of errors for each checkpoint in the web page, the metric should also take into account the total number of times each checkpoint has been tested.*

The metric should not be based on the absolute number of found errors but in the relative number of found errors in relation to the number of tested cases [11]. That is, the ratio of errors and number of tested cases. For instance, if we analyze a web page that contains 5 images without text equivalent and another one containing 10 where 5 of them have a text equivalent, the second web page should obtain better accessibility score, since the failure percentage is 100% (5 of 5) and 50% (5 of 10) respectively.

Assumption 2: *The priority of an unfulfilled checkpoint should be reflected in the final result [6].*

Priority is an ordinal-scale qualitative variable of three levels: priority 1, priority 2 and priority 3. It is stated by the WAI that priority 1 checkpoints have more impact on the accessibility level of a web page than priority 2 checkpoints and so on. Consequently, their weight in the value obtained by the metric should be different. In order to empirically tune the weights, different values are assigned to the weights in some test files with different accessibility level. The unique restriction when selecting these weights is that 1 > *priority1_weight* > *priority2_weight* > *priority3_weight* > *0*.

These test files have a determined failure rate. In addition, they are simple enough to manually calculate a quantitative metric. This ensures that the tuning process is performed independently to any automatic accessibility tool.

Different values were given to weights to calculate the quantitative accessibility value of each file using the metric defined in section 3.2. The criterion for selecting the most appropriate weights was the similarity of the accessibility value to the failure rate on test files. The test files used are the following:

Low Accessibility level web page (LA)

This test file contains images without text equivalent, tables without summary, some links which open pop-up windows, auto-refreshing and wrong document language definition.

Accessible web page (A)

This test file contains the same potential errors but such that they do not cause any accessibility error: images have text equivalent, tables have summary, links do not open new windows, there is no auto-refresh and language is well defined.

Medium Accessibility level web page (MA)

Elements in this test file are the same than in Low Accessibility file but half of potential errors are actual errors.

Worse than MA

3/4 of the previously mentioned potential errors are actual errors.

Better than MA

This test file is composed of the same elements but 1/4 of them have an actual error.

Empty web page (E)

This test file only contains the necessary structural HTML tags without any content element.

Figure 2 shows the accessibility values obtained when using different weights in the metric calculation defined in section 3.2.

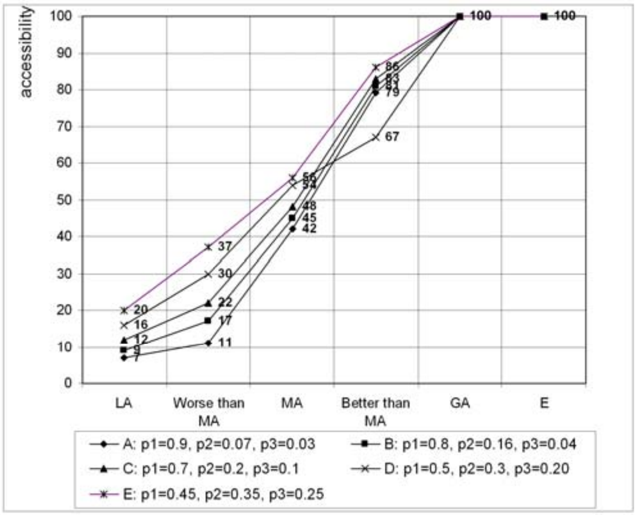

Figure 2. Resulting values of test files calculated with different weights

Discussion of results: no matter what weights we used, all approaches results were 100 for empty (E) and accessible (A) test files. D approach has been discarded since it deviates in "Better than MA" test file. The remainder approaches give similar results although results obtained with A and D approaches are quite far from the desired values (similar to the rate of errors and potential errors, 1/4=0.25, 1/2=0.5, 3/4=0.75). Therefore, either the ones in B (p1=0.8, p2=0.16, p3=0.04) and C (p1=0.7, p2=0.2, p3=0.1) approaches satisfy our criteria. From here on, weights used in B approach are applied.

Assumption 3: *Generic problems should not have influence on the final metric.*

When performing an automatic evaluation, all web pages get the same report of *generic problems* in order to manually check the referred checkpoints. Thus, a metric based on automatic evaluation should not take into account these checkpoints.

Fact 1: *The interval where the metric results for lowest ratios of errors and tested cases are situated has to be spread.*

We have empirically tested that in each POUR guideline, the ratio of errors over potential errors, the failure rate, tends to be very low. Thus, it is difficult to discriminate among different pages since they all get similar accessibility values. The function in Figure 4 would be an approach to the ideal hyperbole in Figure 3. In this hyperbole, the closer to 0 it is the error and tested cases ratio (*E/T*), the higher it will be discriminated. The advantage of this approach is that the value of x' can be empirically assigned, in order to easily control the height allocated to the failure rate *E/T*. This feature makes possible to increase or decrease the variability in any interval depending on the experimental results obtained modifying *a* and *b* variables. For this paper, we used *a*=20 and *b*=0.3 following an empirical approach similar to the one carried out in *Assumption 2*.

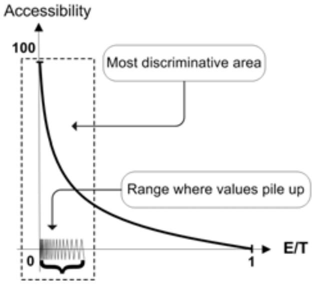

Figure 3. Ideal hyperbole

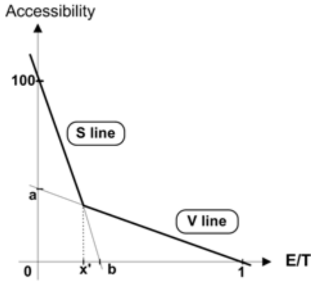

Figure 4. An approach to the hyperbole

According to the hyperbole approach, if *E/T* ratio is less than the intersection point *x'* the accessibility will be calculated using *S* line. Otherwise, *V* line is used. *x'* value depends on variables *a*, *b* and tested cases.

$$x' = \frac{a - 100}{\frac{a}{T} - \frac{100}{b}}$$

x' point calculation

$$A = E \times \left(\frac{-100}{b} \right) + 100 \qquad A = \left(\frac{-a}{T} \times E \right) + a$$

S line formula **V line formula**

Fact 2: *Manual tests (warnings) should be taken into account in the same way than errors.*

Our research concluded that the failure rate is highly correlated for errors and warnings when checkpoints were grouped by their guideline (POUR) and by their priority (1, 2, 3). Therefore, tested cases in warning checkpoints will fulfil the accessibility guidelines with the same ratio than their equivalent errors subgroup.

3.1 Variables, Constant Values and Final Metric

Table 1 contains a description of variables, constants and final values of the metric according to the requirements, assumptions and facts. Some constants are tool dependent while others are guideline-set dependent. These values have to be manually introduced. Since we used two evaluation tools to compare the metric (EvalAccess and LIFT) and one guideline-set (WCAG 1.0) it is necessary to know their characteristics beforehand in order to introduce these values in the Metrics Manager (see Figure 1).

Table 1. Variables, constants and metric for accessibility quantitative measurement

Variables	Description	range
E	number of accessibility **e**rrors in each checkpoint	0-∞
T	number of **t**ested cases in each checkpoint	0-∞
a	variable for hyperbole approach customization (y axis)	0-100
b	variable for hyperbole approach customization (x axis)	0-1
Constants		**value**
N	Total number of checkpoints (EvalAccess) / test cases (LIFT) evaluated	44/109
N_{xy}	Number of checkpoints in guideline $x \in \{P,O,U,R\}$, where *P* stands for Perceivable, *O* for Operable, *U* for Understandable and *R* for Robust, and type $y \in \{error, warning\}$	
N_x	number of checkpoints in guideline $x \in \{P,O,U,R\}$	
$N_{x,error}$	total number of automatic tests	18/55
$N_{x,warning}$	total number of manual tests	25/54
$N_{P,error}$	*error* checkpoints in Perceivable	4
$N_{O,error}$	*error* checkpoints in Operable	3
$N_{U,error}$	*error* checkpoints in Understandable	3
$N_{R,error}$	*error* checkpoints in Robust	7
$N_{P,warning}$	*warning* checkpoints in Perceivable	11
$N_{O,warning}$	*warning* checkpoints in Operable	1
$N_{U,warning}$	*warning* checkpoints in Understandable	6
$N_{R,warning}$	*warning* checkpoints in Robust	7
Weights		**value**
k_1	*priority 1* items	0.80
k_2	*priority 2* items	0.16
k_3	*priority 3* items	0.04
Metric		**range**
A_{xyz}	Accessibility of priority $z \in \{1,2,3\}$ in $x \in \{P,O,U,R\}$ guidelines and in $y \in \{error, warning\}$ type of checkpoints.	0-100
A_{xy}	Accessibility of $x \in \{P,O,U,R\}$ guidelines in $y \in \{error, warning\}$ type of checkpoints.	0-100
A_x	Accessibility of $x \in \{P,O,U,R\}$ guidelines	0-100
A	Mean accessibility value	0-100

3.2 Metric Calculation

Evaluation reports returned by EvalAccess simplify gathering of all the necessary data for metric calculation such as checkpoint type (*error* or *warning*), the times a checkpoint is tested (*T* variable), the times each test fails to be conformant with the guidelines definition (*E* variable), and its priority. All these parameters are grouped in 2 groups (errors and warnings). Each group contains 12 subgroups classified by their priority in WCAG 1.0 (3 priorities) and their membership in the WCAG 2.0 four POUR guidelines according to the previously mentioned mapping.

Therefore, the quantitative accessibility metric takes into account the previously mentioned facts, assumptions and requirements. The quantitative accessibility metric is calculated by the following algorithm:

```
for x in each checkpoint in a guideline {P,O,U,R} loop
    for y in each type of checkpoint {error, warning} loop
        for z in each priority{1,2,3} loop
            x'=calculate_x'_point(a,b,T)
            if $\left(\frac{E}{T} < x'\right)$ then
                $A_{xyz}$=calculate_S_line(b, E)
            else
                $A_{xyz}$=calculate_V_line(a, E, T)
            end if
        end loop
        $A_{xy} = \sum_{z=1}^{3} k_z \times A_{xyz}$ ← Step a
    end loop
    $A_x = \frac{\sum_y N_{xy} \times A_{xy}}{N_x}$ ← Step b
end loop
$A = \frac{\sum_x N_x \times A_x}{N}$ ← Step c
```

In **Step a** we get all the A_{xy} values such as $A_{P,error}$. This means that we get values for *error* checkpoints in Perceivable guideline. In **Step b** an average value for each POUR guideline is calculated by weighting A_{xy} value with the number of *errors* and *warnings* in *x* guideline. Finally, we get an overall accessibility value in **Step c** weighting each POUR guideline with the number of checkpoints they contain. The last two steps take into account the number of guidelines in each category (guidelines and type) in order to distribute the weights in a well-balanced way.

This metric proved to correlate positively with a research carried out by experts on Spanish universities' websites classification according to their accessibility level as presented in [2].

4. EXPERIMENTAL ANALYSIS OF THE PROPOSED METRIC

In order to understand applicability of the proposed metric, we need to investigate how much it depends on the specific web accessibility testing tool. In this way we can characterize its *reliability*, i.e. the degree to which it yields comparable results when applied to the same websites but on the basis of results produced by different tools.

Our goal is to determine the degree of agreement of the results. The rationale is that if results were similar, then when using this metric for measuring accessibility, the particular tool being adopted is not an important decision for any of the scenarios we introduced at the beginning of the paper.

On the other hand, if results are not similar, but they do correlate, then the metric results obtained through different tools can be used only for scenarios that require ranking of websites at the most, since the metric would yield relative data rather than absolute values.

Finally, if results do not correlate at all, then the conclusion would be that the metric is closely tied to the tool being used to test the websites, and no comparisons are justified if different tools are used.

4.1 Experimental Design

In order to investigate these questions, we used two web accessibility tools, EvalAccess and LIFT. LIFT is a multi-user web-based system produced and sold by Usablenet for which a study was published on [3]. A user can specify one or more starting URL to be tested, and after downloading those pages and pages linked to them, LIFT applies a number of tests. The LIFT user can specify downloading preferences (how many pages, the depth of the site to be downloaded, the type of files to filter out, authentication data, data to be submitted to forms) and test profiles (which test sets to use – including WCAG, Section 508, usability – and possibly parameters affecting the behaviour of tests – for example typical text patterns that can be wrongly used as page titles, e.g. "Untitled document").

In our examples we used the default test profile based on WCAG 1.0. It consists of 55 automatic tests and 54 manual ones, covering 44 WCAG 1.0 checkpoints having priority 1 or 2. Some checkpoint is covered by more than one test; for example checkpoint 1.1 (provide text alternatives) is covered by a test that checks whether transparent gifs that are not used as links have an empty ALT; another test checks if buttons and clickable images have non empty and non trivial ALT; another one, based on heuristics on image size, check if potentially decorative images have empty ALT; etc. LIFT produces evaluation reports in several formats (HTML, PDF, XML); in the experiment we used the XML format, for which we wrote a simple XSLT filter that extracts the needed data.

Metrics Manager in EvalAccess was adapted to accommodate reports yielded by LIFT. As a result, an automatically calculated quantitative accessibility value can be obtained using reports of LIFT and EvalAccess. "Constant Values" (see Table 1) are manually introduced since each tool covers guidelines in a different way. We used both tools to test pages of several websites, in order to be able to determine if their behaviour, and more importantly, the behaviour of the metric is affected by the websites themselves.

To this end, we selected 10 websites belonging to European, United States and African universities, and 5 websites belonging to newspapers (see Appendix). Sites were chosen by the a-priori qualitative accessibility level of home page, that is, their accessibility level just by performing automatic evaluations with tools. We downloaded about 100 pages for each of them, starting from their home page and following a breadth-first strategy. This resulted in a total of 1363 pages, 918 pages belonging to universities, and 445 to newspapers. These pages were uploaded on an experimental web server and the two tools were then independently launched on these copies of the websites.

On each page we computed the value for A, based on results produced by EvalAccess and those by LIFT, yielding respectively two variables (A1, A2). For each website we computed the A value for the website by assuming that all its pages have the same weight, again producing two variables (A1*, A2*).

4.2 Experimental Results

The results we obtained for A1, A2 show that there is a marked difference between them. Figure 5 shows the boxplots of these variables, and their difference and average. A boxplot shows the distribution of the values by highlighting the median (the thick horizontal line), the 1st and 3rd quartile (the bottom and top of the box), and the outliers (values beyond the whiskers). It can be readily seen that A1 (EvalAccess) is much higher than A2 (LIFT): more than 75% of the pages evaluated by LIFT have an A score that is smaller than 42, whereas fewer than 25% of the pages evaluated by EvalAccess have an A score lower than 54. Median value for A1 is 69, for A2 it is 28. In addition we can see that the spread of A1 is larger than A2 (the distance between the quartiles for A1 is 37, whereas it is 16 – less than half – for A2). A1 reaches 100, whereas A2 reaches 82; both start from 11.

If we look at the difference (A1-A2), we can see that 50% of the pages have a difference that lies between 27 and 47 (median=40).

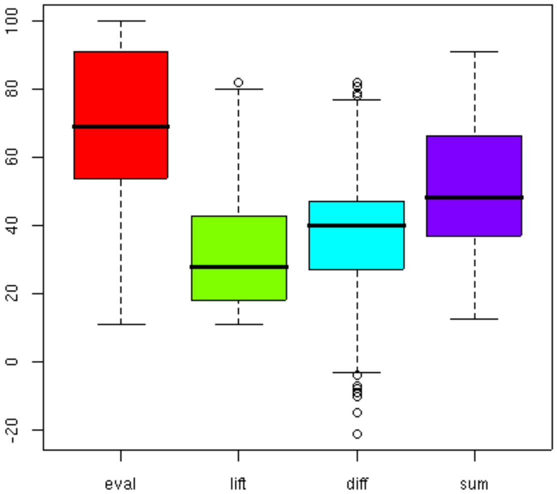

Figure 5. Boxplots of the distribution of A1, A2, A1-A2, (A1+A2)/2 on all pages.

Consider now Figures 6 and 7, showing the histograms of A1 and A2. Again it can be seen that there is a marked difference: both are multimodal, A1 has a peak on A1=100, whereas A2 has a peak on A2=16. A1 is negatively skewed, whereas A2 is positively skewed.

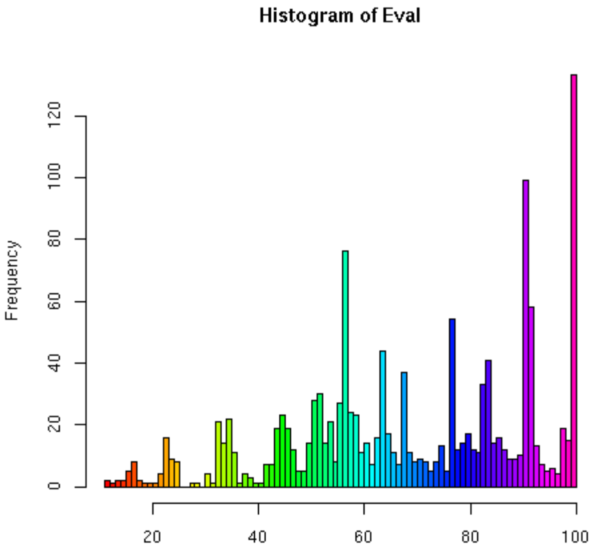

Figure 6. Histogram of absolute frequency of A1.

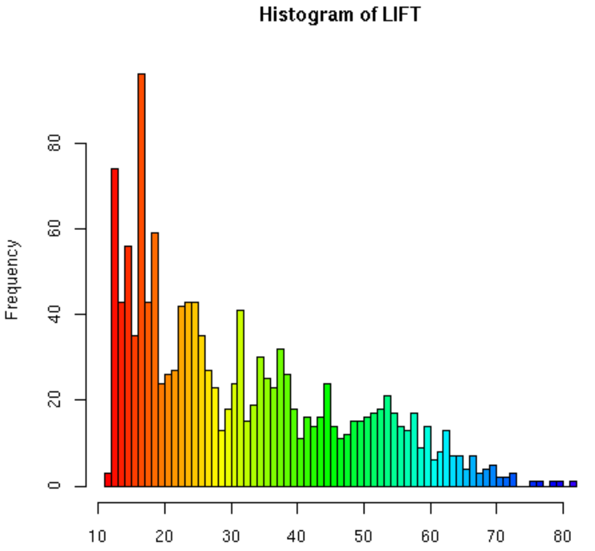

Figure 7. Histogram of absolute frequency of A2.

Let's look at correlation now. Figure 8 shows the scatter plot of the ranks of A2 against the ranks of A1. There are no outliers and there is a clear linear trend. In fact, Spearman's correlation is r(1363)=0.719, which is moderate-high, as it explains about 52% of the variance.

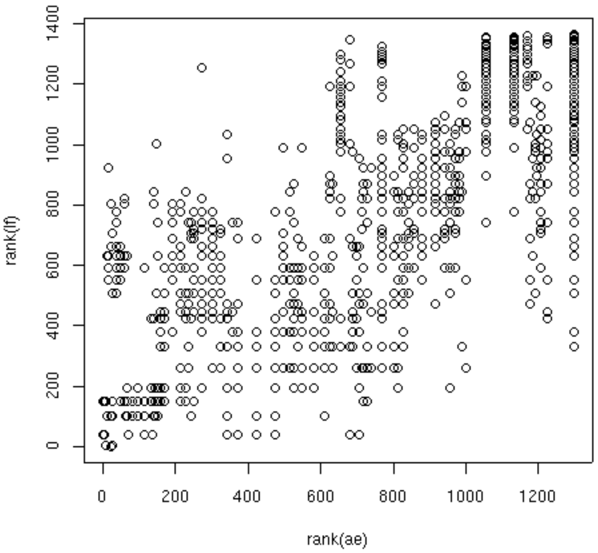

Figure 8. Scatter plot of ranks of A2 (rank(lf)) against ranks of A1 (rank((ae)).

A test to compare the means of A1 and A2 (respectively 69 and 32) yields a significant difference (p<0.1^15) with a confidence interval (at alpha=95%) around A1-A2 of [36, 38]. This means that A1 is larger than A2 by at least 36 and no more than 38 with 5% probability of error when EvalAccess and LIFT are applied to other web pages similar to the ones we tested.

By splitting the data site by site we can look at the distributions of A1 and A2, which are similar to the overall one: A1 is generally higher than A2 for all websites. Figure 9 shows the distribution of A1* and A2*, the values of the metric applied to entire websites. We can see that A1* is generally larger than A2*, and that while A1* spans a range from 34 to 95, A2* spans from 13 to 52.

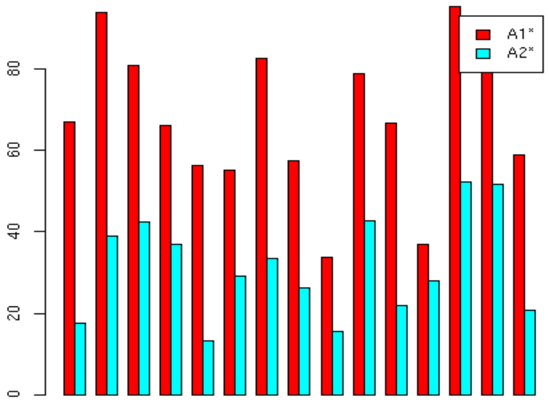

Figure 9. Each pair of bars refers to a website

If we compute the Spearman's correlation index on A1* and A2* we get r(15)=0.735, again an index of moderate-high correlation which explains 53% of variance. If we use the T test to compare the means of A1* and A2* (respectively 68 and 31) we get a 95% confidence interval of A1*-A2* of [24.5, 48.6] (p<0.1^5). This again shows that A1* is larger than A2* by at least 24.5 and no more than 48.6 with 5% probability of error when EvalAccess and LIFT are applied to other web pages similar to the ones we tested.

Finally, if we compare A1 and A2 computed on university websites and on newspapers websites, we can see that:

1. For universities, A1 and A2 correlate moderately (Spearman's r(918)=0.66, which explains 43% of the variance);
2. For newspapers, A1 and A2 correlate more weakly (Spearman's r(445)=0.57, explaining 32% of variance);
3. For universities, T test of the means of A1 and A2 (respectively 75.2 and 38.8) yields a significant difference (p<0.1^15) whose 95% confidence interval is [34, 38];
4. For newspapers, T test of the means of A1 and A2 (57.5 and 18.1) yields a significant difference (p<0.1^15) whose 95% confidence interval is [38, 41].

4.3 Interpretation of Results

These results lead to two definite conclusions:

1. A1 and A2 are definitely different. This is true for the specific sample of (1363) pages. However this conclusion can be also generalized to similar pages in other websites (surely for other universities and newspapers). The difference between A1 and A2 is not only statistically significant but it has also a non negligible effect, showing that the metric applied to results produced by EvalAccess is between 36 and 38 points higher than that applied to LIFT results. This is because LIFT covers automatically detectable checkpoints more widely than EvalAccess and therefore yields lower results. This difference can be seen overall, site by site, and within the two groupings of pages (universities and newspapers). Even though for universities the difference between A1 and A2 is slightly smaller, the lower bound of the confidence interval is still quite large, at 34 points. A similar conclusion can be reached for A1* ad A2*.
2. The second conclusion is that A1 and A2 do correlate, as do A1* and A2*. The lowest correlation index was obtained when comparing A1 and A2 on newspapers and the index explains only 1/3 of the variance of the data. But correlation of unrestricted A1 and A2, and more importantly, of A1* and A2*, is no less than 0.719, which is a moderate to strong correlation.

The conclusive interpretation of the results is that the metric can be used for ranking websites (at least, when using EvalAccess and LIFT and with respect to websites similar to the ones we tested). This is the case for the Information Retrieval and the accessibility monitoring scenarios, where only an ordinal scale of the accessibility level is needed. In fact, given the correlation existing between A1 and A2 (and A1* and A2*), despite the large quantitative difference between the variables, we would expect that similar results would hold also for tools other than EvalAccess and LIFT.

Secondly, for tasks where accessibility levels needs to be measured on a quantitative scale, like in the QA within Web Engineering scenario, then the choice of the tool constrains the evaluator. Once a tool is chosen, the QA process needs to continue using the same tool. No comparisons are allowed with data produced by different tools. However, as shown in Figure 9, the metric is capable of producing data that can be used to finely compare different websites, and producing therefore a ranked list.

5. CONCLUSIONS AND FUTURE WORK

We have described three scenarios where a quantitative accessibility metric is useful: Quality Assurance within Web Engineering, Information Retrieval and accessibility monitoring. Afterwards, we have proposed a quantitative accessibility metric which is automatically obtained from reports of automatic evaluation tools. As far as the metric proposal is concerned, after reviewing and considering related work and literature, requirements and assumptions for the metric are stated. Then, facts regarding empirical data are also provided.

In order to investigate the *reliability* of the metric we measured the accessibility of 1363 web pages based on the evaluation reports yielded by EvalAccess and LIFT tools. Results of the statistical analysis lead to the conclusion that the metric is tool dependent. Tools have a different coverage of guidelines in terms of quantity and accuracy. EvalAccess covers fewer checkpoints than LIFT does. In addition, LIFT checks more tests in a checkpoint than EvalAccess does. Therefore LIFT discovers more potential errors and has a higher *E/T* ratio (error and potential error ratio, see Fact 1) than EvalAccess has. This is the main reason why LIFT yields lower values.

However, there is a high correlation between results yielded by EvalAccess and LIFT. Therefore, this metric can be used in ranking scenarios such as Information Retrieval, accessibility monitoring and more limitedly in QA within Web Engineering.

As can be observed in Figure 10 and Figure 11 other conclusions can also be drawn. We determined that some websites have a very low variability of accessibility value (*A*) among their pages. This behaviour has been observed in most online newspapers and few universities. These sites have in common a heavy use of templates. Therefore, given that the template is faulty, there is not a substantial change among different pages design in these "patterned" sites, since only content changes. We can say that if web sites templates were accessible, the maintenance of the accessibility of a whole site would be easier.

Future work will lead us to carry out experiments with users in order to investigate on the validity of the metric. User testing will useful to know if the metric reflects the accessibility level experienced by users. In other words, if the metric really measures what user perceive. In addition, a possible future work is to see how the metric works (with EvalAccess and LIFT) when different sets of checkpoints such as US Section 508 are used.

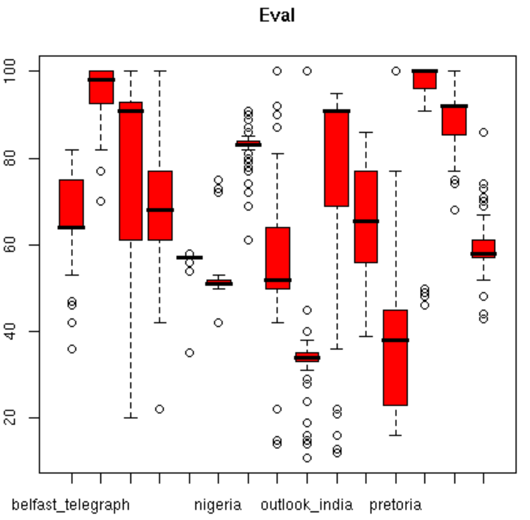

Figure 10. Boxplots of websites measured with EvalAccess results

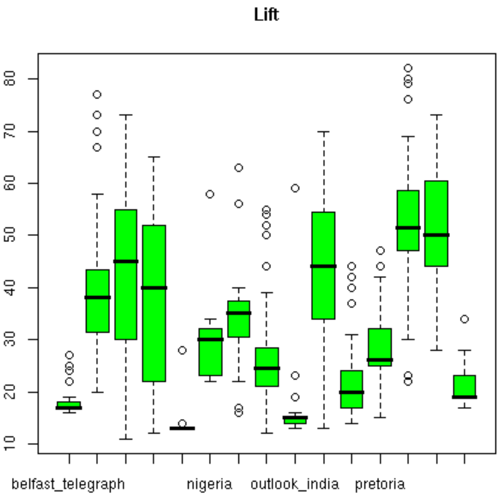

Figure 11. Boxplots of websites measured with LIFT results

6. ACKNOWLEDGMENTS

Markel Vigo's work is funded by the Department of Education, Universities and Research of Basque Government.

7. REFERENCES

[1] Abascal, J., Arrue, M., Fajardo, I., Garay, N., and Tomás, J. Use of Guidelines to automatically verify web accessibility. *International Journal on Universal Access in the Information Society (UAIS)* 3(1), 71-79, 2004, Springer.

[2] Arrue, M., Vigo, M., and Abascal, J. Quantitative Metrics for Web Accessibility Evaluation. In Proceedings of *ICWE 2005 Workshop on Web Metrics and Measurement*. University of Wollongong School of IT and Computer Science.

[3] Brajnik, G. Comparing accessibility evaluation tools: a method for tool effectiveness. *Universal Access in the Information Society (UAIS)* 3(3-4), 252-263, 2004, Springer.

[4] Bühler, C., Heck, H., Perlick, O., Nietzio, A., and Ullveit-Moe, N. Interpreting Results from Large Scale Automatic Evaluation of Web Accessibility. In K. Miesenberger et. al (Eds.). *Computers Helping People with Special Needs 2006*. Lecture Notes in Computer Science 4061. Springer, 184-191.

[5] Google Accessible Search. Available at http://labs.google.com/accessible/

[6] Hackett, S., Parmanto, B., and Zeng, X. Accessibility of Internet websites through time. In Proceedings of *6th International ACM SIGACCESS Conference on Computers and Accessibility 2004*, pp. 32-39.

[7] Ivory, M., Yu, S., and Gronemyer, K. Search result exploration: a preliminary study of blind and sighted users' decision making and performance. *CHI Extended Abstracts 2004*, pp. 1453-1456.

[8] Mich, L., Franch, and M. Gaio, L. Evaluating and Designing the Quality of Web Sites. *IEEE MultiMedia* 10(1), 34-43, 2003.

[9] Olsina, L. and Rossi, G. Measuring Web Application Quality with WebQEM. *IEEE Multimedia* 9(4), 20-29, 2002.

[10] Snaprud, M.H., Ulltveit-Moe, N., Pillai, A.B., and Olsen M.G. A Proposed Architecture for Large Scale Web Accessibility Assessment. In K. Miesenberger et al. (Eds.). *Computers Helping People with Special Needs 2006*. Lecture Notes in Computer Science 4061, pp. 234-241, Springer.

[11] Sullivan, T. and Matson, R. Barriers to use: usability and content accessibility on the Web's most popular sites. Proceedings of *ACM Conference on Universal Usability 2000*, pp. 139-144.

APPENDIX

Website, URL, and a-priori qualitative accessibility of homepage

City University London, http://www.city.ac.uk/, AA
Lancaster University, http://www.lancs.ac.uk, A
The University of Kansas, http://www.ku.edu, A
University of Cambridge, http://www.cam.ac.uk, A
University of Dundee, http://www.dundee.ac.uk, A
University of California, Berkeley, http://www.berkeley.edu, 0
The Irish Times, http://www.ireland.com, 0
University of Calgary, http://www.ucalgary.ca, 0
University of Bolton, http://www.bolton.ac.uk, 0
Outlook india.com, http://www.outlookindia.com, 0
University of Nigeria NSUKKA, http://www.outlookindia.com, 0
The Sydney Morning Herald, http://www.smh.com.au, 0
University of Pretoria, http://www.up.ac.za/, 0
Daily Express, http://www.express.co.uk, 0
The Belfast Telegraph, www.belfasttelegraph.co.uk, 0