Journal of Web Engineering, Vol. 8, No.1 (2009) 003-024

# ENRICHING INFORMATION RETRIEVAL RESULTS WITH WEB ACCESSIBILITY MEASUREMENT

MARKEL VIGO, MYRIAM ARRUE, and JULIO ABASCAL

*Informatika Fakultatea*

*University of the Basque Country, Donostia, Spain*
*{markel, myriam, julio}@si.ehu.es*

Received November 21, 2007
Revised May 21, 2008

Search engines are the most common gateway to search information in the WWW. Since Information Retrieval (IR) systems do not take Web accessibility issues into account, displayed results might not tailor to certain users' needs such as people with disabilities or mobile devices' users. In order to overcome this situation, we present a model aiming at considering Web accessibility as well as content relevance. The model consists of three components (Content Analysis Module, Accessibility Analysis Module and Results Collector Module) that carry out the following tasks: content analysis, automatic Web accessibility evaluation and accessibility measurement of results for re-ranking. Since criteria for ranking results provided by IR systems are necessary, quantitative metrics for accessibility have also been defined. Two prototypes that follow the specifications of the model have been developed in order to demonstrate the feasibility of this proposal. Finally, some case studies have been conducted aiming at discovering how traditional search engines deal with Web accessibility.

*Key words*: Web accessibility, Measurement, Information Retrieval, Quality Assurance, search engines

## 1 Introduction

Searching for information is a key activity when accessing to the Web. According to Kobayashi and Takeda [24], 85% of users make use of search engines when searching for information in the WWW. However, results are not tailored to the needs of users with disabilities. Therefore, they may find barriers when trying to access Web pages listed in results. For instance, the first result, which is the most relevant for a given query, can be completely inaccessible. Paradoxically, the quality of life of people with disabilities could be improved by the WWW and concretely search engines as there is a great potential to perform tasks they difficultly could accomplish by themselves in the physical world (i.e.: do shopping, buy tickets, attending class, etc.).

According to a study carried out by Andronico et al. with visually impaired users [2] only 38% of them found search engines results useful whereas 90% of the sighted users did not have any problem. This may be the reason why only the 23% of visually impaired users versus the 70% of sighted users in the mentioned study stated that used search engines habitually. This paper aims at tackling this problem by combining research done on Information Retrieval and Web Accessibility disciplines.

W3C Web Accessibility Initiative (WAI) is the most significant organization working in favour of Web accessibility. It released the Web Content Accessibility Guidelines (WCAG 1.0) [12] in 1999. WCAG 1.0 guidelines include specific testing techniques and checkpoints which refer to accessibility issues in a more accurate way. Depending on the way a checkpoint impacts on the accessibility of a Web page, each checkpoint has a priority assigned (1, 2 or 3 from more to less impact respectively) and three conformance levels are defined:

- Conformance level A. All priority 1 checkpoints are satisfied.
- Conformance level AA. All priority 1 and 2 checkpoints are satisfied.
- Conformance level AAA. All priority 1, 2 and 3 checkpoints are satisfied.

The WCAG 2.0 candidate recommendation was released in April 2008 [11] and proposes a new guideline concept. This set of guidelines describes Web accessibility in a novel way by defining the properties an accessible Web page has to accomplish. Similarly to the previous version of WCAG 1.0, each checkpoint defines three success criteria and analogous conformance levels. According to WCAG 2.0 an accessible Web page should meet the following properties:

- Make content PERCEIVABLE for any user.
- Ensure that interface elements in the content are OPERABLE by any user.
- Make content and controls UNDERSTANDABLE to as many users as possible.
- Use ROBUST Web technologies that maximize the ability of the content to work with current and future accessibility technologies and user agents.

This paper aims at exploring Web accessibility issues on traditional Information Retrieval (IR) in the Web, such as search engines. Since ranking algorithms in IR aim at matching a query with a list of the most suitable Web resources, quality issues are implicitly involved in this task. Thus, Section 2 analyzes the relationships between Web accessibility and Web quality. Section 3 describes related work on Web accessibility metrics, automatic Web accessibility evaluation and how some search engines deal with Web accessibility issues. In Section 4 a model that considers accessibility issues is proposed for IR systems while in Section 5 all the necessary components and processes that meet the requirements of the model are presented: a quantitative metric for measuring Web accessibility and two architectural approaches with their respective prototypes that follow the specifications of each model. Test cases are carried out in Section 6 as well as the analysis of the performance and finally, conclusions are drawn in Section 7.

## 2 Web Accessibility as a Quality Measure

The ISO 9126-1 standard [21] defines six software product quality characteristics: functionality, reliability, efficiency, usability, maintainability and portability. For evaluation purposes, it also defines a quality model for software product quality and it should be used in conjunction with the ISO 14598-1 [20] providing methods for the measurement, assessment and evaluation of software product quality. As far as Web applications are concerned, specific models such as 2QCV3Q by Mich et al. [27] have been proposed. Even if they include several aspects related to both usability and accessibility, Web accessibility is not considered as an important property of Web applications.

Nevertheless, people involved in the design and development of Web sites and researchers often make use of Web accessibility and Web usability terms indistinctly as they both enhance user satisfaction, effectiveness and efficiency. This may happen due to the guidelines and good practices overlap between both properties. Therefore, the boundary between them is quite fuzzy. Contributing to the confusion, some standards such as ISO/TS 16071 [19] define accessibility in terms of usability: "accessibility is the usability of a product, service, environment or facility by people with the widest range of capabilities". Thatcher et al. [35] state that Web accessibility is a subset of usability although both concepts tend to be dealt separately. Thus, a usable Web site should be accessible but an accessible Web application may not to be usable. In fact, while Web accessibility aims at making Web sites available to the broader spectrum of users, usability focuses on efficiency, learnability and satisfaction of such users, as stated by Gulliksen et al. [16]. On the other hand, Hoffman et al. [18] contrarily to an existing tendency stating that accessibility benefits all users (i.e. Pemberton [31]), claim that sometimes accessibility improves usability, while other times has no impact and in some other occasions it can even decrease general usability. Unfortunately, no empirical data are attached. Some empirical studies show that there is only a low correlation between both properties [32] while others conclude that there is not a significant correlation [4]. Therefore, it is necessary to consider both concepts separately when developing and evaluating Web applications.

However, the conceptual overlapping between accessibility and usability is only partially reflected in the methodology used for evaluating both properties. That is, although they are defined in similar terms, their evaluation is differentially covered. Web usability is usually evaluated by experts or testing methods with users while Web accessibility is evaluated by guidelines review even if theoretically user testing should be done more frequently. Although there are diverse methods and tools for Web usability evaluation [22], accessibility assessment has not been sufficiently developed, even though accessibility measurement, rating and assessment are essential for determining the quality of Web applications. The lack of such accurate measures and tools to automatically calculate them may be one of the reasons why accessibility is frequently forgotten in quality assurance processes.

All the approaches for measuring the quality of software products agree on the importance of creating adequate metrics in order to efficiently perform the quality evaluation process. The most accepted Web accessibility metric is the qualitative one proposed by the WAI in the WCAG 1.0 document. As previously mentioned, this metric assigns a 0, A, AA or AAA value to a Web page depending on the fulfilment of the WCAG 1.0 guidelines. This metric is not accurate enough in order to rate and classify Web applications according to their accessibility level. A Web page meeting all priority 1 checkpoints would obtain the same accessibility value as another Web page meeting all priority 1 checkpoints and almost all priority 2 checkpoints: both of them would get the A level conformance. This criterion seems to be based on the assumption that if a Web page does not meet one of the guidelines in a level, it is so inaccessible as if did not satisfy all of them. This might be true for some users, but in general it is essential to have not only a reject/accept validation, but a more accurate graduation for accessibility scores. Thus, as stated by Olsina and Rossi [28], defining quantitative accessibility metrics is essential. Moreover, it is a key factor in order to perform an adequate rating of Web sites and including accessibility scores in IR systems. This fact may encourage developers to consider accessibility in order to obtain higher ranks for their Web sites.

## 3 Related Work

There are two key tasks that should be considered in a model that would include Web accessibility scores into IR systems: accessibility evaluation and measurement. As stated in Section 2, the definition of accurate accessibility metrics is essential as a criterion to rate Web sites. Providing detailed accessibility evaluation reports is required so that metrics can be automatically applied in order to obtain accessibility scores. Both tasks should be automatically performed since the objective is to include them into other automatic processes such as IR systems. The following sections present the state-of-the-art regarding Web accessibility metrics, automatic accessibility evaluation and adaptation of search engines for accessibility.

### *3.1 Web Accessibility Metrics*

Sullivan and Matson [33] proposed to measure accessibility using the "failure-rate" (*fr*) between actual and potential points of failure, by evaluating a subset of 8 checkpoints from WCAG 1.0. For instance, 10 pictures missing an appropriate textual description out of 100 would lead to *fr*=0.1 while 5 images out of 25 lead to *fr*=0.2. Therefore, the normalized accessibility score would be 1-*fr*. The authors claim that the potential points of failure (or accessibility opportunities) should be penalized since they might contain accessibility barriers. Thus, a normalized incidence figure is calculated for these accessibility opportunities.

González et al. developed KAI, which stands for "Kit for the Accessibility to the Internet", a set of applications aiming at enhancing the accessibility of Web pages for visually impaired users. In the context of KAI, an application to measure the accessibility level of Web pages was developed so that users could know its accessibility level beforehand [14]. Numerous metrics are defined with regard to WCAG 1.0 checkpoints. For instance, there are two metrics for checkpoint 5.3: percentage of tables with summaries and percentage of tables with descriptive summaries. Not only percentage terms are used to define a metric but also absolute number of items, such as the number of colours used as background as for checkpoint 2.2. In addition, a normalized overall accessibility value is calculated using the WebQEM method [28]. The fact that the metrics are automatically obtained and that visually impaired users provide feedback during the development of the project are the strong points of this approach.

Fukuda et al. [13] proposed two accessibility metrics for blind users: *navigability*, which measures how well Web content is structured and *listenability*, which relates to the appropriateness of alternative text. Both parameters are automatically calculated in a tool called Accessibility Designer by Takagi et al. [34]. Yet, there is no user testing that demonstrates the validity of the metrics and the way these metrics are calculated is not revealed.

Bailey and Burd [5] used tree-maps to represent the accessibility level of a Web site. They claim that this information visualization technique is more interactive and easier to comprehend for Web site maintenance. Each node within the tree represents a Web page and the size and colour of the node vary depending on its accessibility level, which is measured using the Overall Accessibility Metric (OAM),

$$OAM = \sum_{c} \frac{B_c W_c}{Attributes + Elements}$$

where *Bc* is the number of barriers found within confidence level *c* and $W_c$ corresponds to the weight of that confidence level. There are four confidence levels depending on how certain is an evaluation tool when evaluating a WCAG 1.0 checkpoint: checkpoints labelled as certain weigh 10, high certainty checkpoints weigh 8, while low ones get 4 and the most uncertain ones obtain 1. Therefore, the higher the certainty level is the more the corresponding barrier is penalized. This is divided by the sum of the total number of HTML attributes and elements in a Web page. The major drawback of this metric is that results obtained using OAM are unbounded.

The Web Accessibility Quantitative Metric (WAQM) by Arrue et al. [3] overcomes the limitations of the abovementioned metrics by automatically providing normalized results that consider the weights of the WCAG 1.0 priorities, exploiting the information in the reports produced by the evaluation tool EvalAccess [1]. Evaluation reports are based on WCAG 1.0 but the WAQM also provides an accessibility value for each WCAG 2.0 guideline (Perceivable, Operable, Understandable, Robust) since results are interfaced using a mapping function. Once WCAG 1.0 checkpoints are grouped by their WCAG 2.0 membership and their priorities in the WCAG 1.0, failure rates are computed for each subgroup. Since failure-rates tend to pile up close to 0, discrimination among failure rates is not very effective. Thus, a function to spread out these values is applied to the failure rates. As WAQM relies on reports yielded by automatic tools, checkpoints that can be automatically evaluated have a strong influence on the final scores even if the semi-automatic problems are also considered. Since this is the approach adopted for this paper, more details about the WAQM can be found in Section 5.1.

The previously mentioned metrics may consider the accessibility value for a whole Web site but are focused on single pages. Hackett et al. [17] and Parmanto and Zeng [29] proposed the Web Accessibility Barrier (WAB) score aiming at measuring quantitatively the accessibility of a Web site based on 25 WCAG 1.0 checkpoints. In each page *p*, a failure-rate (*fr*) between actual and potential errors is calculated for each checkpoint *c* and divided by the reciprocal of the priority of the checkpoint (1, 2 or 3).

$$WAB = \frac{1}{N_p} \sum_p \sum_c \frac{fr(p,c)}{priority_c}$$

The final result is divided by the total number of Web pages in a site. Higher values imply lower accessibility levels. The most important features of the metric are that it is automatically calculated using an automatic evaluation tool and the fact that the scope includes the whole Web site. On the other hand, the range of values is unbounded (not normalized) and checkpoint weighting has not solid empirical foundations.

In the context of the Unified Web Evaluation Methodology[a] (UWEM) several metrics have been proposed during its development process. The first public milestone was UWEM 0.5 and the last version to date (May 2008) is UWEM 1.2 by Velleman et al. [37]. An extension of the UWEM 0.5 metric [10] applies to a sample of pages *p* in a Web site to a given user group *u* and is defined by

$$A(p,u) = 1 - \prod_i \left(1 - R_{ib} S_{ub}\right) ,$$

[a] Available at http://www.wabcluster.org/

where *b* represents the barrier type and *i* is the identifier for a barrier. $R_{ib}$ reports whether the barrier *b* as been detected at location *i*, thus the value of $R_{ib}$ can be either 0 or 1. The incidence that a given barrier has on a user group is represented by the severity value $S_{ub}$, ranging from 0 to 1 depending on the impact. By using this metric, lower values indicate higher accessibility levels. Later, in UWEM 1.2 the metric is significantly simplified and the failure-rate is adopted for a single Web page. The metric applied to the Web site consists of the calculation of the mean value of every single page included in a sample of pages.

Brajnik and Lomuscio [9] proposed SAMBA, a methodology that involves not only evaluation tools but also expert reviewers in the context of the Barrier Walkthrough method discussed by Brajnik in [7, 8]. A sample of results provided by accessibility evaluation tools are used by a panel of experts in order to find accessibility barriers for different user groups; experts are asked to assign severities to these barriers and then appropriate generalizations can be inferred for the entire Web site. Within such a process, experts also consider the error-rate of the evaluation tool which affects the values computed by the metric. All these issues are put together in the last step of the method, the computation of accessibility indexes. Rather than measuring conformance to certain guidelines, SAMBA aims at measuring the accessibility level of a Web site for different user groups. As long as a test-to-barrier mapping function is provided, the SAMBA method is independent from the evaluation tool. However nothing is known regarding whether the values produced by SAMBA are tool independent.

Table 1. Properties of approaches for Web accessibility quantitative metrics

| Properties of the metric | Sullivan and Matson [33] | KAI [14] | Fukuda et al. [13] | OAM [5] | **WAQM [3]** | WAB [17, 29] | UWEM 0.5 [10] | UWEM 1.2 [37] | SAMBA [9] |
|---|---|---|---|---|---|---|---|---|---|
| Are potential errors considered? | Yes | not all metrics | N/A | No | **Yes** | Yes | No | Yes | Yes |
| Are the metrics normalized? | Yes | Yes | Yes | No | **Yes** | No | Yes | Yes | Yes |
| Do the metrics consider semi-automatic issues? | No | N/A | N/A | N/A | **Yes** | No | No | No | Yes |
| Are the metrics based on guidelines? | WCAG 1.0 | WCAG 1.0 | Some are based on WCAG 1.0 | WCAG 1.0 | **WCAG 1.0** | WCAG 1.0<br>Section 508 | WCAG 1.0 | WCAG 1.0 | WCAG 1.0 |
| How many tests are considered? | 8 (12%) | N/A | N/A | N/A | **44 (68%)** | 25 (38%) | N/A | N/A | 33 (51%) |
| Are guidelines weighted? How? | No | No | N/A | confidence levels | **WCAG 1.0 priorities** | WCAG 1.0 priorities | severity function | No | severity of the barrier |
| Is it automatically obtained? | No | Yes | Yes | Yes | **Yes** | Yes | N/A | N/A | partially |
| Are the metrics focused on a user group? Which one? | No | blind users | blind users | No | **No** | No | any user group | No | any user group |

A more schematic comparison among these approaches can be found in Table 1. It can be appreciated that the WAQM approach is the most comprehensive one as it meets most of the requirements stated in [3] and described in the first column of Table 1. In addition, as the WAQM makes some assumptions about the fulfilment of semi-automatic issues the number of tests that are considered is the highest. Moreover, further research [38] has proven that the WAQM is tool independent in scenarios such as the one concerning in this paper (Information Retrieval), where Web pages are ranked according to the scores obtained with the metrics. This entails that the behaviour of the metric will not change, no matter which evaluation tool is being used. Yet, to date it is not possible to obtain metrics for specific user groups. This is an open issue that future work will address.

### *3.2 Automatic Accessibility Evaluation*

In recent years, a great deal of tools for automatic accessibility evaluation has been developed[b]. Even if most of them evaluate predefined sets of general purpose accessibility guidelines such as WCAG 1.0 or Section 508, they vary in the number and type of test cases implemented. It should be emphasized that evaluation results returned by different tools may significantly vary [6]. For further information in this regard, Ivory et al. [22] carried out a comprehensive study on tools for automatic evaluation of usability. In 2004, Abascal et al. [1] proposed a novel approach for automatic accessibility evaluation: separation of guidelines from the evaluation engine. The usefulness of this approach relies on its flexibility and updating efficiency. Adaptation to new guideline versions does not imply re-designing the evaluation engine but guidelines updating. The guidelines specification language is based on XML and following this approach, in 2005, Vanderdonckt and Bereikdar [36] proposed the Guidelines Definition Language, GDL and later Leporini et al. [25] the Guidelines Abstraction Language, GAL.

### *3.3 Adapting Search Engines for Accessibility*

Over the last few years, many applications aiming at improving the user experience of people with disabilities when interacting with search engines have been presented. Some of them are restricted to improve the user interface whereas others consider adapting the information retrieval process by adding new metrics or re-ranking the search results considering accessibility criteria.

Andronico et al. [2] and Yang and Hwang [39] enhanced the interface of Google search engine, for visually impaired users. The adaptations performed are specifically conceived in order to improve user experience when using a concrete screen reader such as Jaws or Big Eyes I. However, search engines results cannot be re-ranked and user experience might be discouraging. In order to tackle this problem, Ivory et al. [23] suggest providing additional page features and re-ranking search results according to users' visual abilities. In this context, Google has launched "Google Accessible Search"[c] where results are ranked by the criteria stated in their FAQ[d]: "page's simplicity, how much visual imagery it carries and whether or not its primary purpose is immediately viable with keyboard navigation". It is targeted at visually impaired users and the blind. In this sense, Masson and Michel [26] proposed a software agent which personalizes the search engine results by re-ranking them according to their accessibility level. Although the accessibility metrics used for this purpose are not clearly defined, it is stated that these metrics have been specifically developed for users with visual disabilities. Good and Jerrams-Smith [15] carried out an extensive study of the accessibility barriers which most affect to four users groups: blind, visually impaired, dyslexic and motor impaired. Then, a set of algorithms are defined for

[b] http://www.w3.org/WAI/ER/tools/

[c] http://labs.google.com/accessible/

[d] Accessible Search FAQ. Available at: http://labs.google.com/accessible/faq.html

selecting the most accessible Web pages for those groups of users in order to re-rank the search engines results. Nevertheless, there is not any tool that incorporates these algorithms and the metrics are quite naïve as they do not meet most of the properties stated in Table 1. Finally, Zhu and Gauch [40] integrated quality metrics into information retrieval systems. However, they did not consider accessibility as a quality attribute.

Table 2 outlines the strong and weak points of the systems that consider accessibility in their rankings. The last column corresponds to Evalbot which is the prototype that relies on the WAQM and has been developed following the model presented in this paper. Providing a tool that considers Web accessibility as a whole by considering WCAG 1.0 recommendations is one of the strong points of Evalbot. The approach of Good and Jerrams-Smith seems quite holistic but there is not a tool that computes the scores and besides of the simplicity of the metric, numerous guidelines are left out. The rest of the approaches just consider visually impaired users and accessibility criteria are neither accurate nor comprehensive.

Table 2. Properties of search engines that consider Web accessibility in their rankings

| Properties of the metric | Google Accessible Search | Masson and Michel [26] | Good and Jerrams- Smith [15] | **Evalbot** |
|---|---|---|---|---|
| Which are the accessibility criteria? | page's simplicity, amount of visual imagery and keyboard navigation feasibility | a subset of WCAG 1.0 and AccessiWeb[e] criteria | a subset of WCAG 1.0 | **WCAG 1.0** |
| Is there any metric? | N/A | N/A | their own algorithms | **WAQM** |
| Is targeted at any user group? | blind and visually impaired | visually impaired | blind, visually impaired, dyslexic and motor impaired | **all** |
| Is there any tool available? | Yes | Yes | No | **Yes** |

## 4 Proposed Model for Information Retrieval Systems

One of the aims of this paper is to present an architecture proposal where results provided by Information Retrieval systems are enriched with Web accessibility analysis. Thus, a model that produces results with the most suitable Web pages according to their content relevance and their accessibility level is proposed. Results provided by applications that follow this model should consider the ranking regarding content relevance as well as the accessibility score of each item. As it can be appreciated, the proposed model consists of three components (see Figure 1).

**Content Analysis Module (CAM)** performs the content analysis based on traditional Information Retrieval methods and techniques and returns a list of Web sites ranked according to their suitability for a specific query.

**Accessibility Analysis Module (AAM)** performs the accessibility evaluation of the Web resources returned by the CAM. Tools for automatic accessibility evaluation fit within this module.

**Results Collector Module (RCM)** has multiple purposes: on the one hand, it ensures that the information provided by the other two modules is adequately combined. On the other hand, it exploits evaluation reports returned by the AAM in order to obtain a quantitative value for each page. Finally, top ranked results provided by the CAM are labelled with their accessibility scores.

[e] http://www.accessiweb.org/fr/Label_Accessibilite/criteres_accessiweb/

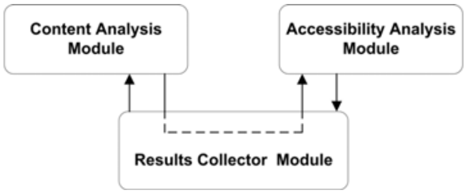


Fig 1. Model of the proposed architecture

Due to the modularity of the proposed model, automatic accessibility evaluation tools, libraries to access search engines results, Web crawlers etc. can easily interoperate. In addition, this modular architecture will guarantee a correct independent testing and adding new features becomes quite straightforward.

## 5 Implementation of a Prototype

Two prototypes have been developed following the proposed model. In this sense, the following sections describe the main tasks that have been carried out: definition of accurate quantitative metrics, automatic calculation of metrics using evaluation reports and the integration of accessibility evaluation, metrics calculation and content analysis.

### *5.1 The Web Accessibility Quantitative Metric, WAQM*

The WAQM [3, 38] aims at automatically calculating accessibility scores of Web pages exploiting the data stored in reports obtained from automatic evaluation tools. Thus, to the greater extent it considers problems that can be automatically found and to the lesser extent the semi-automatic errors (those that are automatically warned but have to be manually checked) that require human judgement. Other issues such as WCAG 14.1 checkpoint "use the clearest and simplest language" cannot be detected by evaluation tools and are thus discarded.

The WAQM assumes that the scores calculated with the metric should be normalized and rather than taking into account absolute number of accessibility problems a failure-rate (actual errors divided by potential errors) is calculated. The priority of a violated checkpoint in WCAG 1.0 is also considered and after empirical testing, values were assigned to these priorities: 0.80 for priority 1 checkpoints, 0.16 for priority 2 checkpoints and 0.04 for priority 3 checkpoints. It was observed that failure rates tend to pile up close to 0 making difficult the discrimination between those pages that obtained high accessibility scores. Therefore, we transform the failure rate according to a hyperbole function (see Figure 2) that spreads out values which are close to 0. The function in Figure 3, which is an approach to the function in Figure 2, changes depending on the values of variables *a* and *b* which have to be tuned depending on the evaluation tool.

Besides an average accessibility score, a value for each WCAG 2.0 guideline (Perceivable, Understandable, Operable, Robust) is also provided by the metric even if evaluations are carried out against WCAG 1.0 guidelines as a mapping function[f] allows interfacing the results.

[f] Available at http://www.w3.org/TR/2006/WD-WCAG20-20060427/appendixD.html

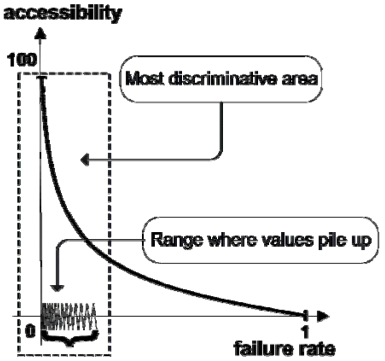


Fig 2. Hyperbole that discriminates failure rates close to 0

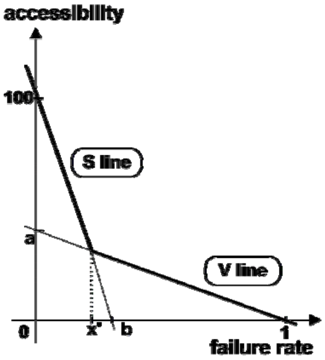


Fig 3. An approach to the hyperbole depicted in Figure 2

Using evaluation reports returned by automatic evaluation tools makes it easy to gather all the necessary data to compute the metric, such as checkpoint type (if it has been automatically found or it is a warning that should be checked by an expert), the number of potential errors (*t* value), the times each test fails to comply with the guidelines definition (*e* value), and its priority. All these parameters are grouped in 2 groups (*automatic* and *warning*). Each group contains 12 subgroups grouped by their priority in WCAG 1.0 (3 priorities) and their membership in each WCAG 2.0 guideline according to the previously mentioned mapping. The quantitative accessibility metric is calculated by the following algorithm:

```
    for i n each checkpoint in a guideline {P,O,U,R} loop
          for j in each type of checkpoint {automatic, warning} loop
                for k in each priority{1,2,3} loop
                      x'=calculate_x'_point(a,b)
                      if (failure_rate(e,t)<x’) then
                            Aijk=calculate_S_line(b, e ,t)
                      else
                            Aijk=calculate_V_line(a, e, t)
                end loop
(1)             $A_{ij} = \sum_{k=1}^{3} w_k \times A_{ijk}$
          end loop
(2)       $A_i = \frac{\sum_j N_{ij} \times A_{ij}}{N_i}$
    end loop
(3) $A = \frac{\sum_i N_i \times A_i}{N}$
```

where *x' point*, *S line* and *V line* are calculated in the following way:

$$x' = \frac{a-100}{a-\frac{100}{b}}$$

x' calculation

$$A = \left(\frac{e}{t} \times \frac{-100}{b}\right) + 100$$

*S line*

$$A = \left(-a \times \frac{e}{t}\right) + a$$

*V line*

$A_{ijk}$ is the accessibility score given by either *S line* or *V line* while in **step (1)** $A_{ij}$ considers the priorities of violated checkpoints multiplying $A_{ijk}$ scores by their corresponding weightings $w_k$ where $w_1$=0.8, $w_2$=0.16 and $w_3$=0.04, obtaining values such as $A_{P,automatic}$. In this specific case, a score is calculated for those checkpoints that are automatically evaluated and are member of the Perceivable guideline. In **step (2)**, an average value for each POUR guideline is calculated by weighting *Aij* values with the number of automatable errors $N_{i,automatic}$ and warnings $N_{i,warning}$ in each *i* guideline. Finally, in **step (3)** an overall accessibility value is obtained weighting each POUR guideline with the number of checkpoints they contain, $N_i$.

### *5.2 Integrating Web Accessibility Evaluation and Content Relevance Analysis*

Since the model proposed in Section 4 can be quite vague it is more comprehensively developed in order to demonstrate how it can be deployed in real scenarios. From the point of view of the architecture two approaches are presented, each one consisting of an abstract representation and a prototype. Both approaches have the same components and what makes a difference is how they are deployed. The following paragraphs explain how the components have been implemented:

**Content Analysis Module**: search engines nowadays provide developers with APIs in order to make queries to their indexes from other applications. For instance, Google.com[g], Yahoo![h] and MSN Search[i] offer these services. Queries are made in a transparent way no matter what underlying technologies do these APIs implement (in these cases access to Web Services).

**Accessibility Analysis Module**: in order to carry out accessibility evaluations, EvalAccess [1] evaluation tool has been chosen. The fact that it is implemented as a Web Service is useful for our purposes since accessibility evaluation reports can be automatically obtained from client applications. Machine-understandable reports in XML make easier the exploitation of results and the calculation of the WAQM.

**Results Collector Module**: this module coordinates user requests with the abovementioned components as well as the coding and the decoding of the information flow between them. In addition, it is responsible for gathering evaluation reports and applying the metric in order to obtain accessibility scores. Since the Accessibility Analysis Module produces reports based on WCAG 1.0 guideline set and our metric is WCAG 2.0 guidelines oriented, evaluation reports are interfaced according to a mapping table. Afterwards, it manages the search results interface by labelling the URLs provided by search engines with their accessibility value and sorting them according to this score.

#### *5.2.1 Architectural Approaches*

Two prototypes that follow the specification of the proposed model have been developed. The objective is twofold: to show the flexibility of the model by developing two architectural approaches and observe the arguments for and against both prototypes.

[g] http://www.google.com/apis/

[h] http://developer.yahoo.com/search/

[i] http://www.microsoft.com/downloads/details.aspx?FamilyId=C271309B-02DE-42A7-B23E-E19F68667197&displaylang=en

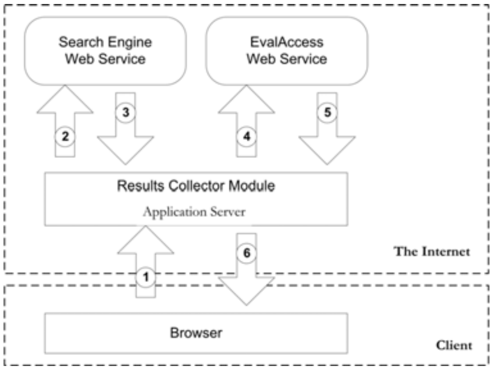


Fig 4. Fully server side approach

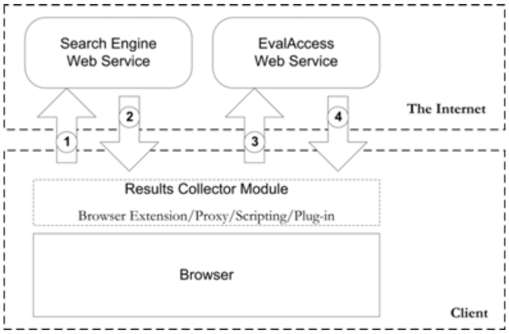
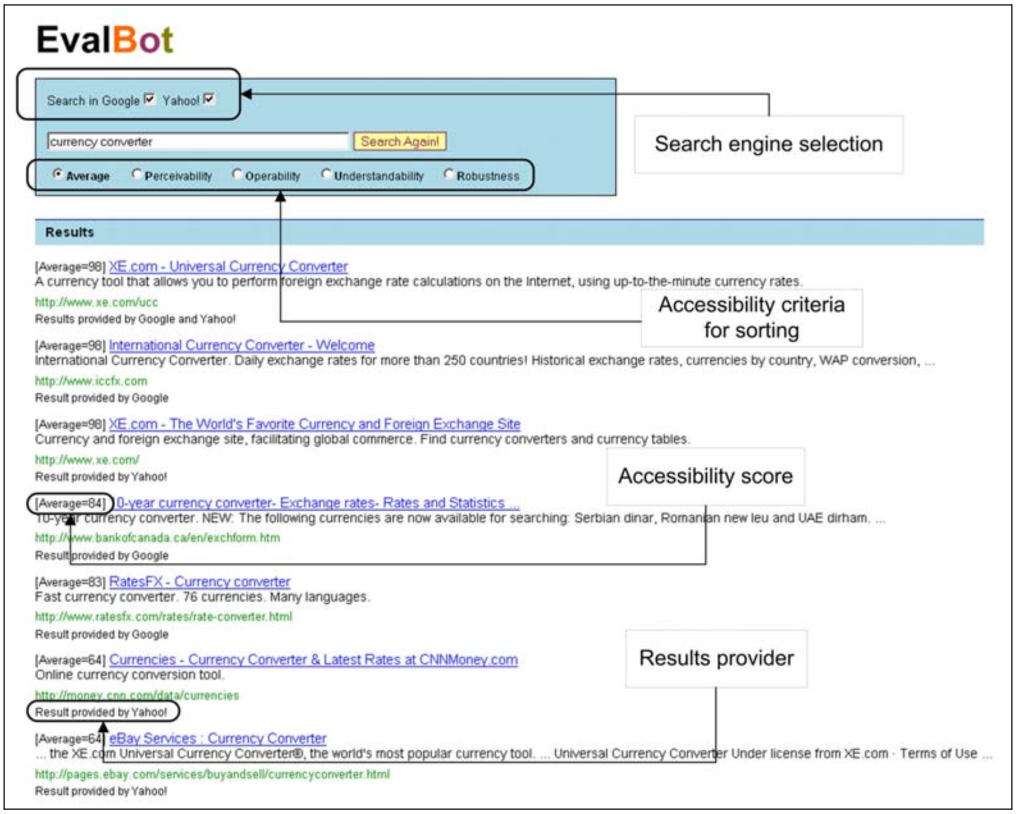


Fig 5. Shared tasks approach



EvalBot

Search in Google ☑ Yahoo! ☑

currency converter Search Again!

● Average ○ Perceivability ○ Operability ○ Understandability ○ Robustness

Results

[Average=98] XE.com - Universal Currency Converter
A currency tool that allows you to perform foreign exchange rate calculations on the Internet, using up-to-the-minute currency rates.
http://www.xe.com/ucc
Results provided by Google and Yahoo!

[Average=98] International Currency Converter - Welcome
International Currency Converter. Daily exchange rates for more than 250 countries! Historical exchange rates, currencies by country, WAP conversion, ...
http://www.iccfx.com
Result provided by Google

[Average=98] XE.com - The World's Favorite Currency and Foreign Exchange Site
Currency and foreign exchange site, facilitating global commerce. Find currency converters and currency tables.
http://www.xe.com/
Result provided by Yahoo!

[Average=84] 10-year currency converter- Exchange rates- Rates and Statistics ...
10-year currency converter. NEW: The following currencies are now available for searching: Serbian dinar, Romanian new leu and UAE dirham. ...
http://www.bankofcanada.ca/en/exchform.htm
Result provided by Google

[Average=83] RatesFX - Currency converter
Fast currency converter. 76 currencies. Many languages.
http://www.ratesfx.com/rates/rate-converter.html
Result provided by Google

[Average=64] Currencies - Currency Converter & Latest Rates at CNNMoney.com
Online currency conversion tool.
http://money.cnn.com/data/currencies
Result provided by Yahoo!

[Average=64] eBay Services : Currency Converter
... the XE.com Universal Currency Converter®, the world's most popular currency tool. ... Universal Currency Converter Under license from XE.com · Terms of Use ...
http://pages.ebay.com/services/buyandsell/currencyconverter.html
Result provided by Yahoo!

Fig 6. Search engine results enriched with Web accessibility scores: fully server-side approach

**Fully server side approach** (Figure 4): In this approach, the three components of our system are hosted in the server side. There are several architectural advantages in this approach: it is useful to collect statistics, do load balancing and do server side caching. From the end user point of view only a browser is needed to access to the system. This is the sequence of actions:

1. User makes the query
2. The RCM translates the query and invokes the search engine Web service
3. Search engine results are obtained by the RCM
4. URLs from these results are extracted and sent to the accessibility evaluation service
5. Accessibility evaluation reports are produced
6. After calculating the accessibility scores with the metric, results from step 3 are labelled with their respective accessibility values. Results can be sorted according to their accessibility value or can be ranked according to the criteria of the search engine.

Figure 6 is a screenshot of the prototype implemented following this approach. It is the results page and in this case results are ranked according to their accessibility score. The features that make it different from traditional search interfaces are the following: the user can select the search engines to which the query is made, results can be sorted according to one of the five accessibility scores that the metric implements, accessibility scores are explicitly displayed and the URL providers are also displayed.

**Shared tasks approach** (Figure 5): The second approach requires the installation of an application that plays the part of the Results Collector Module. This component is a mediator between the browser and the rest of the components located at the server-side. It can be implemented in many ways: extending the functionalities of the browser, embedding a plug-in, using a proxy based solution etc. The network latency decreases compared with the previous approach since 2 queries and responses are required while in the full server approach 3 queries and responses are necessary:

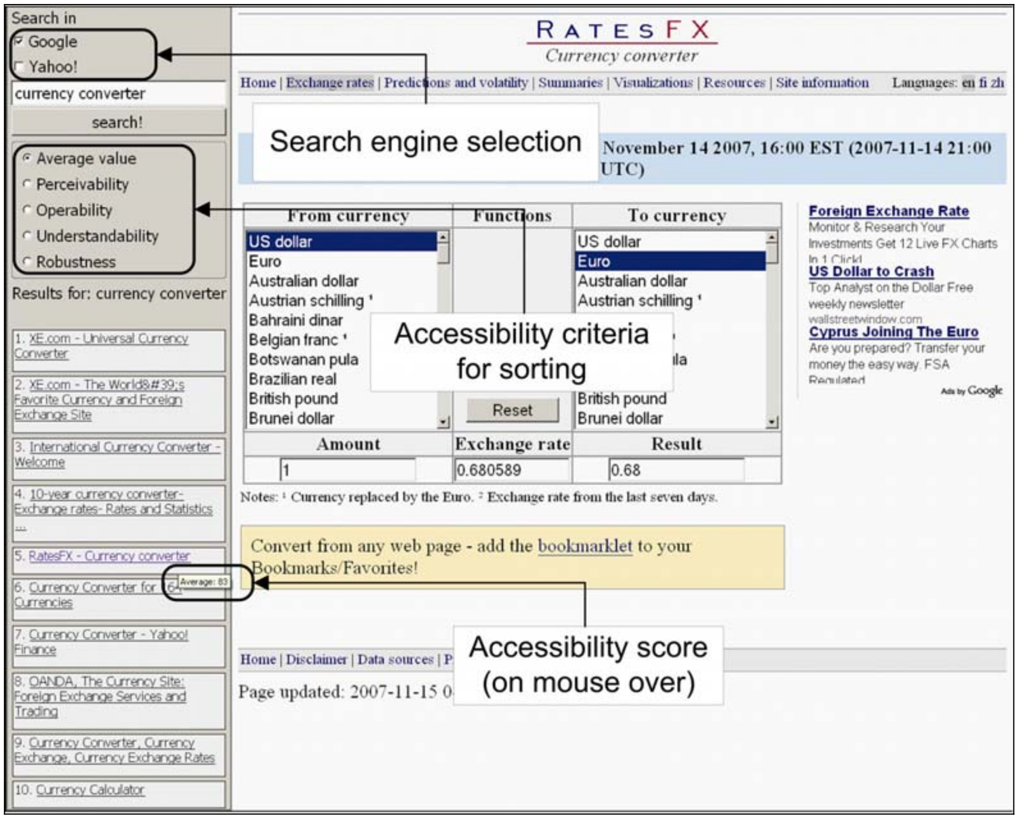
1. The query is handled by the RCM that works jointly with the browser and invokes the search engine service
2. Search engine results are obtained by the RCM
3. URLs from these results are extracted and sent to the accessibility evaluation service
4. After calculating the accessibility scores with the metric, results from step 2 are labelled with their respective accessibility values. The browser updates its content showing the results sorted according to the selected accessibility criteria

In addition, CPU cycles are also saved using this approach. On the other hand, the user is required to install the software component locally.

A solution implemented as an extension of the Mozilla Firefox browser and the underlying XUL technology[j] is proposed (see Figure 7). XUL is a language to create XML based user interfaces that can be deployed in several Mozilla applications such as Thunderbird, Firefox or Sunbird. Its easiness to use makes it practical for rapid prototyping developments. In our case the extension allows the user to

[j] http://www.mozilla.org/projects/xul/

select the search engine to which the query is made as well as results sorting criteria. Results are labelled with their accessibility score which appears when the mouse is over the link.



Search in

☑ Google

☐ Yahoo!

currency converter

● Average value

○ Perceivability

○ Operability

○ Understandability

○ Robustness

Results for: currency converter

1. XE.com - Universal Currency Converter
2. XE.com - The World&#39;s Favorite Currency and Foreign Exchange Site
3. International Currency Converter - Welcome
4. 10-year currency converter- Exchange rates- Rates and Statistics ...
5. RatesFX - Currency converter
6. Currency Converter for 164 Currencies
7. Currency Converter - Yahoo! Finance
8. OANDA, The Currency Site: Foreign Exchange Services and Trading
9. Currency Converter, Currency Exchange, Currency Exchange Rates
10. Currency Calculator

Average: 83

Search engine selection

Accessibility criteria for sorting

Accessibility score (on mouse over)

RATESFX
Currency converter
Home | Exchange rates | Predictions and volatility | Summaries | Visualizations | Resources | Site information
Languages: en fi zh
November 14 2007, 16:00 EST (2007-11-14 21:00 UTC)
From currency
Functions
To currency
US dollar
Euro
Australian dollar
Austrian schilling ¹
Bahraini dinar
Belgian franc ¹
Botswanan pula
Brazilian real
British pound
Brunei dollar
Reset
Amount
Exchange rate
Result
1
0.680589
0.68
Notes: ¹ Currency replaced by the Euro. ² Exchange rate from the last seven days.
Convert from any web page - add the bookmarklet to your Bookmarks/Favorites!
Foreign Exchange Rate
Monitor & Research Your Investments Get 12 Live FX Charts In 1 Click!
US Dollar to Crash
Top Analyst on the Dollar Free weekly newsletter
wallstreetwindow.com
Cyprus Joining The Euro
Are you prepared? Transfer your money the easy way. FSA Regulated
Ads by Google
Home | Disclaimer | Data sources | P
Page updated: 2007-11-15 0

Fig 7. Search engine results enriched with Web accessibility scores: shared tasks approach

As can be observed in Figures 6 and 7 the implementation of the former approach is more traditional search engine alike while the later one is a more complex search bar which is always present while the user is browsing and can make use of it whenever they want. Both approaches make use of the services offered by Google.com and Yahoo!

Not only these two approaches are valid but other approaches such as the one including the accessibility evaluation in the client side are not excluded. The multiple combinations that can be done prove the flexibility of the model. The developer should be aware of the limitations and assume the trade-offs of each approach when developing the model.

## 6 Case Study and Discussion

In order to check the behaviour of the metric and its application in the developed prototypes a comparison of results in different search engines has been carried out. The rankings of the top ten results provided by Google.com, Google Accessible Search (GAS), Yahoo!, and our prototype called Evalbot have been compared. For these queries, Evalbot makes use of the services provided by Google.com and Yahoo! Note that results obtained by querying these indexes tend not to be permutations of the results provided by the search engines, especially in the case of Google as the API works with a different index.

Table 3 shows the URLs returned by the system when making the "homeopathic medicine" query, in May the 14th, 2008. The second column contains the URL while the next three columns refer to their accessibility errors grouped by their priority, priority 1, priority 2 and priority 3 respectively. Next, the accessibility score obtained with the WAQM and the ranks in each search engine.

Table 3. Results obtained in different search engines for the "homeopathic medicine" query as well as accessibility errors grouped by their priority

| item | URL | $P_1$ | $P_2$ | $P_3$ | WAQM score | Evalbot rank | Google .com rank | GAS rank | Yahoo! Search rank |
|---|---|---|---|---|---|---|---|---|---|
| 1 | http://en.wikipedia.org/wiki/Homeopathy | 0 | 0 | 12 | 97 | 1 | 2 | 1 | 3 |
| 2 | http://www.quackwatch.org/01QuackeryRelatedTopics/homeo.html | 0 | 13 | 5 | 93 | 2 | 10 | 6 | - |
| 3 | http://www.hmedicine.com | 0 | 19 | 27 | 88 | 3 | 3 | 3 | 5 |
| 4 | http://www.myhomeopathic.com | 0 | 20 | 8 | 88 | 3 | - | - | 6 |
| 5 | http://abchomeopathy.com/taking.htm | 11 | 18 | 6 | 64 | 4 | 1 | 2 | - |
| 6 | http://www.herbalremedies.com/homeopathics.html | 13 | 1141 | 19 | 64 | 4 | 9 | - | - |
| 7 | http://www.homeopathic.com | 52 | 10 | 5 | 63 | 5 | - | - | 4 |
| 8 | http://www.canismajor.com/dog/altern2.html | 4 | 41 | 8 | 53 | 6 | - | - | 8 |
| 9 | http://lyghtforce.com/HomeopathyOnline | 6 | 32 | 3 | 52 | 7 | - | - | - |
| 10 | http://homeopathyusa.org/faq.html | 2 | 10 | 1 | 52 | 7 | - | 8 | - |
| 11 | http://kulisz.com/homeopathic_medicine.htm | 1 | 84 | 8 | 52 | 7 | - | - | 10 |
| 12 | http://www.drweil.com/drw/u/id/ART00470 | 5 | 19 | 5 | 45 | 8 | - | - | 7 |
| 13 | http://nationalcenterforhomeopathy.org | 26 | 83 | 29 | 45 | 8 | - | - | 9 |
| 14 | http://www.hpathy.com | 16 | 155 | 16 | 43 | 9 | 5 | - | 2 |
| 15 | http://www.holisticonline.com/Homeopathy/hol_homeopathy.htm | 1 | 123 | 6 | 41 | 10 | 6 | 4 | - |
| 16 | http://www.ritecare.com/homeopathic.asp | 20 | 345 | 63 | 38 | 11 | 8 | 9 | 1 |

Due to the weights assigned to different priorities it can be observed that the less priority 1 errors a URL has the higher is ranked. The metric behaves similarly with priority 2 and 3 errors. However, this is not the rule of the thumb since the metric also takes into account the failure-rate rather than absolute number of errors that is, accessibility opportunities are rewarded by the WAQM. For instance, this is the reason why item number 10 ($P_1$=2, $P_2$=10, $P_3$=1) is ranked below item number 5 ($P_1$=11, $P_2$=18, $P_3$=6). As this sample is not significant enough to draw a solid conclusion, it has been carefully observed the behaviour of 12 randomly-chosen queries extracted from the TREC 2004 Web Track topics[k]. It consists of 5 queries of three terms, 5 queries of two terms and 2 queries of one term (see first column in Table 4). Correlations tests (Sperman's ρ and Kendall's τ) were applied between all rankings in order to shed more light on the rationale of rankings. Table 4 contains just significant correlations between pairs of rankings which consist of Google Accessible Search vs. Google.com (column 4), results rearranged by Evalbot with the URLs provided by Google.com and Yahoo! APIs (column 5) and finally, column 6 contains the significant correlations found between Yahoo! and the

[k] These topics are used in Information Retrieval experiments in specific corpuses. Available at http://trec.nist.gov/data/web/Web2004.query.stream.trecformat

rearranged URLs of Yahoo! The second and third column contain statistical data about the accessibility scores of the rearranged URLs. These values were collected in the 13th and 14th of May, 2008.

Table 4. Statistical data of accessibility scores provided by Evalbot (columns 2 and 3) and relevant ranking correlations for 12 topics

| term | Evalbot$_{Google.com}$ | Evalbot$_{Yahoo!}$ | GAS vs. Google.com | Evalbot$_{Yahoo!}$ vs. Evalbot$_{Google.com}$ | Yahoo! vs. Evalbot$_{Yahoo!}$ |
|---|---|---|---|---|---|
| Groundhog day Punxsutawney | min= 20<br>max= 64<br>$Q_1$= 31.75<br>median= 50<br>$Q_3$= 51.75<br>mean= 43<br>sd= 15.3 | min= 28<br>max= 98<br>$Q_1$= 50<br>median= 52<br>$Q_3$= 54.75<br>mean= 58.5<br>sd= 22 | ($\rho$=0.79, p=0.04) | ($\tau$ =1, p=0.04) | |
| human genome research | min= 45<br>max= 97<br>$Q_1$= 51<br>median= 80<br>$Q_3$= 92.75<br>mean= 73.5<br>sd= 22 | min= 27<br>max= 97<br>$Q_1$= 50<br>median= 69.5<br>$Q_3$= 93<br>mean= 69.1<br>sd= 24.6 | ($\tau$ =0.62, p=0.05)<br>($\rho$=0.75, p=0.05) | | |
| white house fellowships | min= 28<br>max= 100<br>$Q_1$= 50.75<br>median= 72.5<br>$Q_3$= 93.5<br>mean= 70.2<br>sd= 24.8 | min= 28<br>max= 100<br>$Q_1$= 50<br>median= 59.5<br>$Q_3$= 65.75<br>mean= 58.6<br>sd= 21.9 | | ($\tau$ =1, p=0.04) | |
| Mojave desert ecology | min= 50<br>max= 100<br>$Q_1$= 53<br>median= 63<br>$Q_3$= 92.25<br>mean= 71.5<br>sd=20.9 | min= 38<br>max= 100<br>$Q_1$= 60.75<br>median= 87.5<br>$Q_3$= 96.25<br>mean= 77.8<br>sd= 22.4 | | ($\tau$ =1, p<0.02) | |
| Why study comets? | min= 44<br>max= 100<br>$Q_1$= 55.5<br>median= 67<br>$Q_3$= 72<br>mean= 65.7<br>sd= 16.6 | min= 15<br>max= 100<br>$Q_1$= 47.5<br>median= 61<br>$Q_3$= 69<br>mean= 61<br>sd= 24.5 | ($\rho$=0.75, p<0.003)<br>($\tau$ =0.81, p=0.01) | ($\tau$ =1, p<0.02) | |
| career information | min= 50<br>max= 98<br>$Q_1$= 55.75<br>median= 72<br>$Q_3$= 89.5<br>mean= 72.2<br>sd= 18.9 | min= 34<br>max= 98<br>$Q_1$= 44.75<br>median= 58.5<br>$Q_3$= 82.75<br>mean= 63.5<br>sd= 24.2 | | ($\tau$ =1, p=0.04) | ($\rho$=0.62, p=0.05) |
| homeopathic medicine | min= 38<br>max= 97<br>$Q_1$= 45.25<br>median= 58<br>$Q_3$= 82<br>mean= 63.2<br>sd= 22.2 | min= 38<br>max= 97<br>$Q_1$= 45<br>median= 52.5<br>$Q_3$= 81.75<br>mean= 61.2<br>sd= 21.8 | ($\rho$=0.86, p<0.02)<br>($\tau$ =0.73, p<0.04) | ($\tau$ =1, p<0.05) | |

| | | | | | |
|---|---|---|---|---|---|
| drunk driving | min= 11<br>max= 100<br>$Q_1$= 42.75<br>median= 75<br>$Q_3$= 93.5<br>mean= 66.4<br>sd= 33.6 | min= 11<br>max= 100<br>$Q_1$= 40<br>median=<br>$Q_3$= 97<br>mean= 68.8<br>sd= 35.4 | ($\rho$=0.76, p<0.03)<br>($\tau$ =0.6, p=0.04) | ($\rho$=1, p=0)<br>($\tau$ =1, p<$10^{-4}$) | |
| Vietnam war | min= 28<br>max= 98<br>$Q_1$= 42.25<br>median= 56<br>$Q_3$= 78<br>mean= 61.3<br>sd= 24.5 | min= 18<br>max= 98<br>$Q_1$= 40.75<br>median= 55<br>$Q_3$= 95.75<br>mean= 62.4<br>sd= 31.1 | ($\rho$=0.86, p<0.02)<br>($\tau$ =0.71, p<0.03) | ($\tau$ =1, p=0.04) | ($\rho$=0.72, p<0.02)<br>($\tau$ =0.54, p<0.03) |
| endangered species | min= 45<br>max= 97<br>$Q_1$= 51.25<br>median= 60<br>$Q_3$= 80<br>mean= 65.8<br>sd= 17.8 | min= 48<br>max= 98<br>$Q_1$= 53.75<br>median= 72.5<br>$Q_3$= 86<br>mean= 71.8<br>sd= 19.1 | | | |
| salmon | min= 16<br>max= 97<br>$Q_1$= 44.25<br>median= 68<br>$Q_3$= 89.5<br>mean= 97<br>sd= 27.8 | min= 14<br>max= 97<br>$Q_1$= 52.75<br>median= 80<br>$Q_3$= 96<br>mean= 71.8<br>sd= 27.8 | ($\tau$ =0.81, p=0.01) | | ($\tau$ =0.60, p=0.01) |
| recycling | min= 39<br>max= 97<br>$Q_1$= 61.5<br>median= 79.5<br>$Q_3$= 90.75<br>mean= 74.3<br>sd= 19.2 | min= 13<br>max= 97<br>$Q_1$= 56.5<br>median= 82.5<br>$Q_3$= 93.75<br>mean= 69.6<br>sd= 30.5 | ($\rho$=0.86, p<0.02)<br>($\tau$ =0.71, p=0.02) | ($\tau$ =1, p<0.005) | |

Relevant correlations have not been found between the ranks provided by Google.com and Yahoo! and their respective rearrangements performed by Evalbot. This leads us to state that Web accessibility does not play a relevant role in the rankings of these search engines and shows the necessity of a tool such as the one presented in this paper. There is not even a correlation between Google Accessible Search and the rearrangements made by Evalbot to the results provided by Google.com API. In addition, as it can be appreciated in column 4, there is a strong correlation between the accessible version (GAS) and the one that does not consider accessibility in the rankings (Google.com). This entails that GAS just makes a few rearrangements in order to rank its results according to their accessibility level but the rank of correlation between GAS and Evalbot suggest that GAS is not a holistic approach to Web accessibility as stated in the related work section. In addition, the last column in Table 4 may suggest that results provided by Yahoo! tend to consider Web accessibility in their rankings. However, this correlation only occurs in 3 cases out of 12. In order to ascertain whether this statement can be generalized, large-scale studies should be conducted. The 5th column shows that there is a correlation between the re-ranks of Google.com and Yahoo! performed by Evalbot. This is logical to some extent since there is an overlap of URLs. Statistical data in columns 2 and 3 show that results provided by Yahoo! cover the 0-100 range more widely than those provided by Google.com. The frequencies of accessibility scores are depicted by Figure 7.

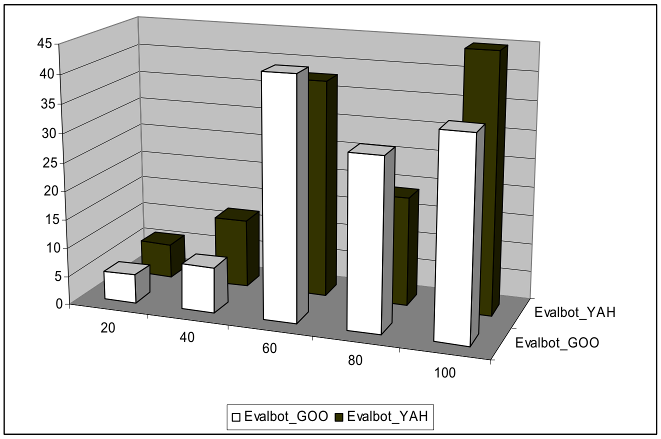


Fig 7. Frequencies of accessibility scores provided by Google.com and Yahoo!

Figure 7 shows that both histograms are positively skewed but those URLs provided by Yahoo! concentrate more in the 80-100 range while those provided by Google concentrate in the 40-60 range. Also, URLs provided by Yahoo! were more highly ranked than those provided by Google. It can be concluded that while there is a tendency to provide accessible URLs these are not ranked according to their accessibility score. These results support the statements by Pemberton [30] regarding the higher visibility of accessible Web pages for crawlers.

If content relevance is prioritized rather than Web accessibility, search results can be labelled with their respective accessibility score so that the user can decide whether access to it or not. This would be an intermediate solution that may suit better to all audiences.

### *6.1 Analysis of the Performance*

In order to explore the performance of the proposed model, 111 queries were selected (11 of one term, 41 of two terms and 59 of three terms) from the TREC 2004 Web Track topics. The performance of was measured following the *fully server-side approach* (see Section 5.2.1) where the Content Analysis Module was the Google search engine and the Accessibility Analysis Module consisted of the EvalAccess Web Service. The Accessibility Analysis Module and the Results Collector Module were deployed in an Apache Tomcat server in a Fedora 7 operative system that ran in a 2GB RAM and 2.8 GHz dual CPU server. Note that this server manages remote user accounts and hosts several Web Services that may slow down the performance.

Search task is divided in two main tasks, retrieving URLs from a search engine and accessibility evaluation. In Figure 8, the box plot on the left depicts the time taken to retrieve the URLs from search engines while the following box plots refer to the evaluation stage and the whole task (evaluation+URLs retrieval from search engines) respectively. These box plots show 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). Obtained values can be observed more carefully in Table 5.

Table 5. Properties of search engines that consider Web accessibility in their rankings

| Task | min | max | mean | 1st Q | 3rd Q |
|---|---|---|---|---|---|
| URLs retrieval | 0.23s | 0.73s | 0.36s | 0.33s | 0.42s |
| Evaluation | 9.42s | 21.19s | 16.73s | 15.21s | 18.27s |
| Total | 9.94s | 21.67s | 17.13s | 15.65s | 18.65s |

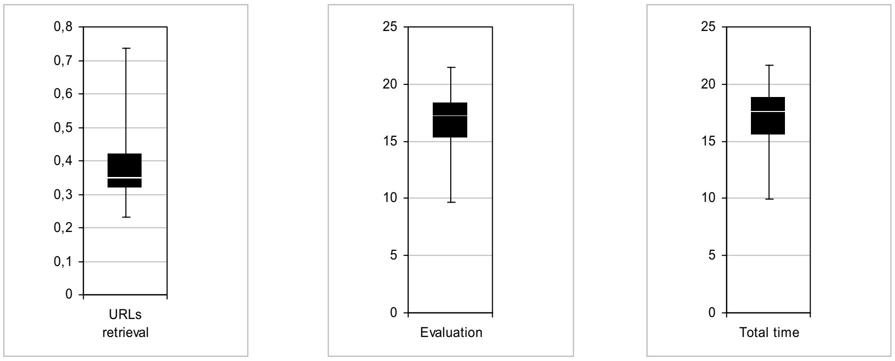


Fig 8. Box plots for each subtask

It is concluded that the evaluation task takes most of the time (98%) while URLs retrieval is insignificant (2%). Once the URLs are obtained, it should be noted that the evaluation task consists of the source code retrieval and the accessibility evaluation itself. The fact that some resources are not always available slows down the evaluation stage as the Accessibility Analysis Module keeps on trying to download the source code. On average, it takes 17 seconds to provide the list of results ranked according to their accessibility level on-the-fly. However, if at the same time than crawling, search engines measured Web resources in a background task, it would not take so long to provide the results since scores would be ready. It goes without saying that the hardware platforms where measures were taken are not the optimal and that search engines companies have an infrastructure that would minimize this latency.

## 7 Conclusions

This paper proposes a model for enriching search engine results with their respective accessibility score. It has been demonstrated that by following this model, customized search engines that consider accessibility can be developed. It is believed that this will significantly enhance user experience satisfaction when searching for information in the WWW [23], as users could access to results ranked by relevance as well as their accessibility level. Future work in this regard contemplates user-testing of the presented approach so that current metrics and framework can be polished up. User testing with other accessibility metrics will also be considered.

The necessity of a quantitative metric for Web accessibility assessment has been demonstrated in this paper. If accurate discrimination among Web pages is required, these measures are necessary. The proposed metric aims at being a general approach to accessibility awareness in search engines since it does not take into account specific users grouped by their disabilities (hearing, visually or physically impaired). This can be understood as a strong point since accessibility evaluations take into account the WCAG 1.0 recommendations as a whole, which is a significant contribution in respect of the related work. Besides an average value, POUR values can also be provided because the system performs a mapping between WCAG 1.0 and WCAG 2.0 draft. It is foreseen that the WAQM will be extended in order to define user adapted metrics and including them in the prototypes so that user-testing can be performed.

Two prototypes based on the proposed model have been developed. They have been deployed following different architectures demonstrating the flexibility of the model and its components: on the one hand, a fully server-side prototype and on the other hand, a shared task prototype relying on Mozilla XUL technology. Web pages provided by search engines such as Google or Yahoo! feed the accessibility evaluation tool with URLs. EvalAccess [1], an automatic evaluation tool for accessibility has been used to obtain accessibility evaluation reports of Web documents. Since reports produced by EvalAccess Web Service are XML-based, necessary data to calculate the metric are automatically obtained. In this sense, the component called "Results Collector Module" deals with the calculation of the metric and shows the results to the user.

Detailed analysis of the results proves that even if to a certain extent search engines consider accessibility in their results, these results are not ranked according to their accessibility level. Therefore, ranking URLs considering accessibility scores would enhance user interaction. It could be discussed whether the trade-off of content ranking versus the accessibility ranking is really worthwhile. If content relevance is prioritized, results can be displayed in the order provided by search engines and each item can be labelled with its accessibility score. Thus, the user would decide to access to a URL.

The most significant disadvantage of these prototypes is the increase in the latency compared to traditional search engines which could discourage users to do search tasks. However, if information retrieval processes in search engines evaluated Web pages for accessibility while crawling the WWW the accessibility scores would be stored beforehand and the response time would not be affected.

## Acknowledgements

Markel Vigo is the recipient of a pre-doctoral fellowship by the Department of Education, Universities and Research of Basque Government. "Gipuzkoako Foru Aldundia-The County Council of Gipuzkoa" has also supported this research in the context of the *AdaptAccess* project. Finally, the authors would like to thank the reviewers for their valuable recommendations.

## References

1. Abascal, J., Arrue, M., Fajardo, I., Garay, N., and Tomás, J. Use of Guidelines to automatically verify Web accessibility. International Journal of Universal Access in the Information Society 3(1), 71-79, Springer, 2004.
2. Andronico, P., Buzzi, M., Castillo, C., and Leporini, B. Improving search engine interfaces for blind users: a case study. International Journal of Universal Access in the Information Society 5(1), 23-40, Springer, 2006.
3. Arrue, M., Vigo, M., and Abascal, J. Quantitative Metrics for Web Accessibility Evaluation. Proc. of Workshop on Web Metrics and Measurement, co-located with 5th Intl. Conf. on Web Engineering, ICWE 2005. University of Wollongong School of IT and Computer Science, 2005.

4. Arrue, M., Fajardo, I., López, J.M. and Vigo, M. Interdependence between technical Web accessibility and usability: its influence on Web quality models. International Journal of Web Engineering and Technology 3(3), 307-328, Inderscience, 2007.
5. Bailey, J. and Burd, E. Tree-Map Visualisation for Web Accessibility. 29th Computer Software and Applications Conference, COMPSAC'05, 275-280, IEEE Computer Society Press, 2005.
6. Brajnik, G. Comparing accessibility evaluation tools: a method for tool effectiveness. International Journal of Universal Access in the Information Society 3(3-4), 252-263, Springer, 2004.
7. Brajnik, G. Web Accessibility Testing with Barriers Walkthrough. 2006. Available at www.dimi.uniud.it/giorgio/projects/bw
8. Brajnik, G. Ranking Web sites through Prioritized Web Accessibility Barriers. Technology and Persons with Disabilities Conference, CSUN, California State University Northridge, 2007.
9. Brajnik, G. and Lomuscio, R. SAMBA: a Semi-Automatic Method for Measuring Barriers of Accessibility. Proc. of 9th ACM SIGACCESS Conf. on Computers and Accessibility, ASSETS'07, pages 43-49, ACM Press, 2007.
10. Bühler, C., Heck, H., Perlick, O., Nietzio, A., and Ullveit-Moe, N. Interpreting Results from Large Scale Automatic Evaluation of Web Accessibility. Computers Helping People with Special Needs, ICCHP 2006. Lecture Notes in Computer Science 4061, 184-191, Springer, 2006.
11. Caldwell, B., Cooper, M., Guarino Reid, L. and Vanderheiden, G. (eds). Web Content Accessibility Guidelines 2.0. (W3C Candidate Recommendation). Available at http://www.w3.org/TR/WCAG20/ (2008, April 30)
12. Chisholm, W., Vanderheiden, G., and Jacobs, I. (eds.). Web Content Accessibility Guidelines 1.0. (W3C Recommendation). Available at http://www.w3.org/TR/WAI-WEBCONTENT/ 1999.
13. Fukuda, K., Saito, S., Takagi, H. and Asakawa, C. Proposing new metrics to evaluate Web usability for the blind. Extended Abstracts Proc. of 2005 Conf. on Human Factors in Computing Systems, CHI 2005, 1387-1390, ACM Press, 2005.
14. González, J., Macías, M., Rodríguez, R. and Sánchez, F. Accessibility Metrics of Web Pages for Blind End-Users. In J.M. Cueva Lovelle et al. (eds.), Web Engineering. Lecture Notes in Computer Science 2722, pages 374-383, Springer, 2003.
15. Good, A. and Jerrams-Smith, J. Enabling Accessibility and Enhancing Web Experience: Ordering Search Results According to User Needs. Proc. of 4th Intl. Conf. on Universal Access in Human-Computer Interaction UAHCI 2007. Universal Access in Human-Computer Interaction. Applications and Services. Lecture Notes in Computer Science 4556, 34-44, Springer, 2007.
16. Gulliksen, J., Andersson, H. and Lundgren, P. Accomplishing universal access through system reachability - a management perspective. International Journal Universal Access in the Information Society 3(1), 96-101, Springer, 2004.
17. Hackett, S., Parmanto, B., and Zeng, X. (2004). Accessibility of Internet Web sites through time. Proc. of 6th Intl. ACM SIGACCESS Conf. on Computers and Accessibility, 32-39. ACM Press, 2004.
18. Hoffman, D., Grivel, E. and Battle, L. Designing software architectures to facilitate accessible Web applications. IBM Systems Journal 44(3), 467-483, 2005.
19. International Organization of Standardization (ISO). Ergonomics of human-system interaction: Guidance on accessibility for human computer interfaces (ISO/TS 16071). Geneva, Switzerland, 2003.
20. International Organization of Standardization (ISO). Information Technology - Software Product Evaluation (ISO 14598). Geneva, Switzerland, 1999.
21. International Organization of Standardization (ISO). Software Engineering - Product Quality - Part1: Quality Model (ISO 9126-1). Geneva, Switzerland, 2001.
22. Ivory, M.Y. and Hearst M.A. The state of art in automating usability evaluations of user interfaces. ACM Computing Surveys 33(4), 470-516, ACM Press, 2001.

23. Ivory, M.Y., Yu, S., and Gronemyer, K. Search result exploration: a preliminary study of blind and sighted users' decision making and performance. CHI 2004 Extended Abstracts, 1453-1456, ACM Press, 2004.
24. Kobayashi, M. and Takeda, K. Information Retrieval on the Web. ACM Computing Surveys 32(2), 144-173, ACM Press, 2000.
25. Leporini, B., Paternò, F. and Scorcia, A. Flexible tool support for accessibility evaluation. Interacting with Computers 18(5), 869-890, Elsevier, 2006.
26. Masson, R. and Michel, G. A search engine for the visually impaired. Proc. of 8th Conf. for the Advancement of Assistive Technology in Europe (AAATE), 573-577, 2005.
27. Mich, L., Franch, M. and Gaio, L. Evaluating and Designing Web Site Quality. IEEE Multimedia 10(1), 34-43, 2003.
28. Olsina, L. and Rossi, G. Measuring Web Application Quality with WebQEM. IEEE Multimedia 9(4), 20-29, 2002.
29. Parmanto, B. and Zeng, X. Metric for Web Accessibility Evaluation. Journal of the American Society for Information Science and Technology 56(13), pages 1394-1404, Wiley, 2005.
30. Pemberton, S. The kiss of the spiderbot. interactions 10(1), 44, ACM Press, 2003.
31. Pemberton, S. Accessibility is for Everyone. interactions 10(6), 4-5, ACM Press, 2003.
32. Petrie, H. and Kheir, O. Relationship between Accessibility and Usability of Web sites. Proc. of 2007 Conf. on Human Factors in Computing Systems, CHI 2007, 397-406, ACM Press, 2007.
33. Sullivan, T. and Matson, R. Barriers to use: usability and content accessibility on the Web's most popular sites. Proc. of ACM Conf. on Universal Usability 2000, CUU'2000, 139-144, ACM Press, 2000.
34. Takagi, H., Asakawa, C., Fukuda, K. and Maeda, J. Accessibility Designer: Visualizing Usability for the Blind. Proc. of 6th ACM SIGACCESS Conference on Computers and Disability, ASSETS 2004, pages 177-184, ACM Press, 2005.
35. Thatcher, J., Burks, M.R., Heilmann, C., Lawton Henry, S., Kirkpatrick, A., Lauke, P.H., Lawson, B., Regan, B., Rutter, R., Urban, M. and Waddell, C.D. Web Accessibility: Web Standards and Regulatory Compliance. Springer-Verlag, New York, NY, 2006.
36. Vanderdonckt, J. and Bereikdar, A. Automated Web Evaluation by Guideline Review. Journal of Web Engineering 4(2), 102-117, 2005.
37. Velleman, E., Meerbeld, C., Strobbe, C., Koch, J. Velasco, C.A., Snaprud, M. and Nietzio, A. D-WAB4, Unified Web Evaluation Methodology (UWEM 1.2 Core). 2007. Available at http://www.wabcluster.org/uwem1_2/
38. Vigo, M., Arrue, M., Brajnik, G., Lomuscio, R. and Abascal, J. Quantitative Metrics for Measuring Web Accessibility. Proc. of 2007 Intl. Cross-Disciplinary Conf. on Web accessibility, W4A 2007, pages 99-107, ACM Press, 2007.
39. Yang, Y.F. and Hwang, S.L. Proc. of 4th Intl Conf on Universal Access in Human-Computer Interaction UAHCI 2007. Universal Access in Human-Computer Interaction. Applications and Services. Lecture Notes in Computer Science 4556, 997-1005, Springer, 2007.
40. Zhu, X. and Gauch, S. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Proc. of 23rd Annual Intl ACM SIGIR Conference on Research and Development in Information Retrieval, 288-295, ACM Press, 2000.