Enriching Web Information Scent for Blind Users

Markel Vigo
University of the Basque Country
InformatikaFakultatea
20018 Donostia, Spain
markel@si.ehu.es

Barbara Leporini
HIIS Laboratory
ISTI-National Research Council
56124 Pisa, Italy
barbara.leporini@isti.cnr.it

Fabio Paternò
HIIS Laboratory
ISTI-National Research Council
56124 Pisa, Italy
fabio.paterno@isti.cnr.it

ABSTRACT
Link annotation with the accessibility level of the target Web page is an adaptive navigation support technique aimed at increasing blind users' orientation in Web sites. In this work, the accessibility level of a page is measured by exploiting data from evaluation reports produced by two automatic assessment tools. These tools support evaluation of accessibility and usability guideline-sets. As a result, links are annotated with a score that indicates the conformance of the target Web page to blind user accessibility and usability guidelines. A user test with 16 users was conducted in order to observe the strategies they followed when links were annotated with these scores. With annotated links, the navigation paradigm changed from sequential to browsing randomly through the subset of those links with high scores. Even if there was not a general agreement on the correspondence between scores and user perception of accessibility, users found annotations helpful when browsing through links related to a given topic.

Categories and Subject Descriptors
H.1.2 [**User/Machine Systems**]: Human factors. H.5.2. [**User Interfaces**]: Evaluation. H.5.4. [**Hypertext/Hypermedia**]: User issues. K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities.

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Information scent, web accessibility, blind users, adaptive navigation.

1. INTRODUCTION
The Web is fast growing with an enormous amount of information available and penetrating all facets of our life. Thus, this information abundance generates various orientation problems to Web users. In order to better understand browsing behaviour in the Web, Pirolli and Card [24] formulated the Information Foraging Theory as a way for modelling user decisions when traversing hypertext documents. This theory states that users will follow a determined hyperlink when the trade-off between information gain and access cost is low. The information scent, the underlying basis of Information Foraging Theory, predicts the hyperlink choices based on such trade-offs.

The growing unstructured amount of information is especially detrimental for people with disabilities. In the case of blind users, information overload and excessive sequencing are the main problems. Mostly because screen readers and Braille outputs render Web content in a linear way. Therefore, in order to get the overview of a page a blind user has to traverse the whole page diminishing browsing efficiency and increasing disorientation. By enriching hyperlinks (and thus the information scent) with additional information on the accessibility of the corresponding hypertext node our aim is to provide users with navigational cues that make user experience less problematic. This extra information consists of the accessibility score for the target page, which can be considered an indicator of how well people will be able to navigate it. We hypothesize that users will be more effective and satisfied in their navigation with this support. In addition, we want to observe user behaviour when relevance and accessibility information are provided together. Supporting our hypothesis, in an experiment carried out with blind and sighted users, Ivory et al. [16] found that when a Web page may not satisfy users' information needs, extra information features are preferred over relevance.

ASSETS'09, October 25–28, 2009, Pittsburgh, Pennsylvania, USA.

2. RELATED WORK
According to Goble et al. [13], visually impaired users need to be explicitly warned of obstacles since their reliance on environmental cues is higher than for sighted users. Similarly, Harper et al. [15] found that detecting and notifying users about barriers beforehand improves users' orientation at a web site and Bigham et al. [4] found that blind users are less likely to interact with non-accessible content. Therefore, warning blind users about forthcoming barriers may enhance their user experience. The above-mentioned outcomes lead us to provide mechanisms that diminish user disorientation by augmenting navigation mechanisms. Orientation and navigation are closely related since both refer to the user's navigational environment. Orientation is the user's understanding of current movements and the navigation context. Navigation is part of web browsing and consists in moving around in a hypertext document, deciding at each step where to go next [17]. The former answers the question "where can I go?" while the latter replies to "where am I?"

Since blind users strongly rely on navigation cues or landmarks, Takagi et al. [26] suggest the possible solutions for improving usability for blind users: improvement of XHTML specification in such a way that WAI-ARIA [9] statements can be adopted, simplification of the navigation interface, automatic suggestion of navigation methods and integrating transcoding functions. Other techniques, such as the one proposed by Harper and Patel [14], provide summaries for blind users so that they can ascertain in advance if a page's content is suitable for them. Since none of the previous contributions considers the task itself, Mahmud et al. [21] developed a method to capture the context of the selected link in order to guide the user directly to his target, thus removing

information overload. Furthermore, Bigham et al. [15] developed a method to make task completion less time consuming in interactive web applications by applying end-user development techniques. Regarding link augmentation for able-bodied users, Campbell and Maglio [7] explored how the tension of link relevance and link annotation (in this particular case links were annotated with the connection speed of the page beyond the link) determined user behaviour. They concluded that as link relevance decreases, users tend to rely more on annotations. However, when there was a conflict between relevance and annotations (e.g. the most relevant link to reach a target had a slow connection), users were able to ignore the annotations and relevance prevailed.

The purpose of this paper is to ascertain whether link annotation with accessibility scores improves usability in terms of efficiency and satisfaction. Similarly our goal is to observe the strategy that users follow. To this end, this paper proposes quantitative metrics for blind users as the criterion to enrich the information scent with accessibility scores in Section 3. In order to test our hypotheses a user testing was conducted and the experimental setting is explained in Section 4. Results and discussion in Section 5 lead us to interpret user behaviour and the implications for design. Lastly, some conclusions are drawn in Section 6.

3. WEB GUIDELINES FOR BLIND USERS

Web accessibility guidelines define the requirements a Web page has to satisfy in order to provide accessible content. The most widely accepted sets of guidelines are the Web Content Accessibility Guidelines, WCAG (1.0 [8] and 2.0 [6]) by the W3C Web Accessibility Initiative. Since our purpose is to provide accessibility scores for enhancing browsing experience not only technical web accessibility is considered but also usability aspects for blind users. While accessibility guidelines tend to address the existence of mark-up issues that prevent users from accessing the information, usability guidelines for blind users focus on the adequacy of content and navigational mechanisms.

3.1 Technical Web Accessibility Guidelines

WCAG aim at giving guidance on how to build accessible sites for *all* users. This way, checkpoints, which are more specific best practices, provide guidance on how to remove barriers that may have an impact on several user groups. However, most checkpoints apply exclusively to a particular user group. Targeting blind users leads us to considering those guidelines that only affect this user group. In this sense, Brajnik [3] proposed a correspondence table between WCAG 1.0 guidelines and disabilities. This allows the identification of subsets of guidelines for determined groups and so we considered those guidelines that just impact on blind users. Therefore the subset of WCAG 1.0 that focuses on the blind users has been deployed in an automatic evaluation framework in which guidelines are independent of the evaluation engine [28] obtaining the Accessibility Checker for Blind users, ACB.

3.2 Web Usability Guidelines for Blind users

Usability plays a key role because even if pages meet accessibility standards they still can be difficult to traverse [18]. In this sense, Leporini and Paternò [20] proposed a set of guidelines for the usability of accessible pages. Usability Guidelines for Blind users (UGB) consist of usability criteria grouped in four principles: *structure and arrangement*, *content appropriateness*, *multimodal output* and *consistency*. Each principle contains several checkpoints that focus on specific usability issues for blind and visually impaired users. Checkpoints aim at describing and providing guidance in order to repair usability barriers that blind and visually impaired users may face while interacting with Web pages. Magenta [19] is a tool that evaluates Web pages against UGB, and its evaluation engine is independent of the representation of guidelines. Taking advantage of this feature, only those guidelines that just apply to blind users can be considered without changing the tool implementation.

3.3 Relationship Between Usability and Accessibility Guideline Sets

There is not a total correspondence between WCAG and UGB. Some checkpoints exclusively belong to WCAG set (e.g. "*do not use tables for layout*") while others belong to UGB (e.g. "*provide a consistent pathway to enable layout and terminological consistency*"). However, there is certainly an overlap of checkpoints. Depending on the relationship between checkpoints of both sets, the type of overlap can be categorized as follows:

- *Same*. Both guideline sets identify the same problem and suggest same techniques to check it.
- *Same but differently addressed by the tool*. Even if both sets aim at covering a certain guideline just one tool implements it. For instance, Magenta checks the existence of "*generic or ambiguous links*" while ACB cannot test it.
- *Precondition*. Techniques are complementary but they should be applied in a determined order so that some checkpoints are a precondition for others. WCAG tend to provide preconditions for UGB. For instance, while WCAG emphasizes to provide a summary for tables, UGB gives guidance on the content of the summary. In these cases UGB addresses the usability of the content thus extending the WCAG. This way, both tools complement each other.
- *Contradictory*. There is a contradiction between the statements in guidelines. For instance, while UGB states that frames should not be used, WCAG states how to label them. In this case UGB criteria will prevail.

3.4 Reporting Issues

Generally, checkpoints are stated in natural language entailing several interpretations for each evaluation rule. Thus, checkpoints tend to be divided into design techniques that tend to be technology dependent (in our case (X)HTML). For instance, ACB deals with the "*data tables without summary*" checkpoint dividing it into two techniques: (1) "*provide summaries*" and (2) "*provide abbreviations for headers in tables*". At the same time techniques can be divided into *test cases* which are (X)HTML element and attribute dependent statements. The former technique contains 4 *test cases*: 2 of them check the `summary` attribute of tables while the others check whether `caption` element is within `table` tags. *Test cases* are atomic rules that are evaluated against Web content and thus the content in accessibility reports is determined by such evaluations. If we adopt EARL [1] terminology so that ambiguity is removed, *test cases* are equivalent to `earl:TestCase` statements while techniques and checkpoints correspond to `earl:TestRequirement` cases. Both statements are subclasses of `earl:TestCriterion`, which is the way to refer to such terms in a generic way. Depending on the accessibility issue they produce, tools herein presented, classify evaluation techniques as follows:

- Issues that can be completely automatically checked (`earl:automatic`) yield the next issues:
  - errors (a_e): not satisfying this type of techniques raise accessibility barriers. They produce a *pass* (`earl:passed`) if the checkpoint is met and a *fail* (`earl:fail`) otherwise.
  - recommendations (a_r): techniques implementing these issues can automatically warn or make a recommendation in order to enhance accessibility. Violating this type of techniques does not have a strong impact on accessibility but maybe on usability. Sometimes it refers to those checkpoints that not all users perceive as an enhancement when they are implemented such as "*provide separation between subsequent links*". Other times, the fact that the interaction context strongly determines these kinds of techniques leads to not to be very strict on their fulfilment. For instance, users of older versions of Jaws 7 screen reader find problems when the content of `value` attribute in buttons is not meaningful.
- Issues that raise warnings (w) can only be checked partially in an automatic way (`earl:semiAuto`). For a complete evaluation, experts should verify whether it actually exists an accessibility barrier. For instance, for "*apply appropriate headings*", Magenta raises a warning if there are more than two headings. Afterwards an expert should manually check if headings were adequately placed.

Some automatic issues can raise either errors (a_e) or warnings (w) at the same time. For instance, when checking the appropriateness of summaries in tables, if `summary` is not provided or it is empty an error is produced whereas if it has content and it does not belong to a forbidden description list for table summaries, a warning is produced. Lists that contain forbidden words, such as "this is a summary" or "pic12" in the case of images, can be detected by Magenta. However, the tool cannot guarantee that all forbidden words are contained and besides they are natural language dependent.

3.5 Tool Coverage for Automatic Evaluation

UGB define 17 checkpoints grouped in 4 guidelines. Magenta can semi-automatically evaluate 11 of them implementing 29 *test cases*. 22 test cases produce automatic errors while 7 of them produce warnings. On the other hand, the subset of WCAG 1.0 for blind users consists of 33 checkpoints, 18 of which can be automatically evaluated to a certain extent. These 18 checkpoints are implemented in 32 techniques that at the same time specify lower level requirements in 101 *test cases*. 63 can be automatically verified (51 a_e and 12 a_r) while 38 produce warnings. Compared to UGB, WCAG aims at covering all accessibility barriers, providing numerous techniques to remove them. There is such a difference in the number of *test cases* (29 vs. 101) between the two guideline-sets because the UGB focuses on subtle usability issues.

When evaluating the conformance of a Web page with respect to the mentioned guideline-sets both tools can work independently except when there is a checkpoint dependency due to preconditioning issues. In such a case, a component to solve these dependencies has been introduced. As can be observed in Figure 1, the (X)HTML resource is retrieved and while ACB evaluates its conformance to technical accessibility, Magenta checks the usability. Each tool produces a report and depending on the type of issue raised, be it exclusively accessibility, exclusively usability or overlapping issue, the Metrics Calculation Module produces a quantitative score based on the metrics defined in the following section. The Dependencies Solver, which is encapsulated within the Metrics Calculation Module, deals with those checkpoints that complement each other or those that are one another's precondition.

3.6 Web Accessibility Quantitative Metric for Blind Users

Accessibility metrics that produce quantitative scores enable accurate discrimination among web pages as opposed to the WCAG conformance levels or success criteria. Quantitative scores are useful in those scenarios where accurate measurement is required such as in Web Engineering, Quality Assurance, accessibility monitoring observatories and Information Retrieval. In recent years, a good deal of research has been dedicated to Web accessibility metrics. Existing metrics provide a general approach for measuring accessibility as they do not consider specific user groups but rather general purpose guideline-sets. While some are automatically obtained [27] other require human judgement [2]. Even if there are some metrics for blind users [12] we believe our approach is more comprehensive since test typology and reporting particularities of each guideline set are considered in the process. In addition, results are normalized thus enabling interpretation of results in percentage terms.

Metrics are automatically computed exploiting evaluation reports produced by the ACB and Magenta. Based on the specifications of the WCAG 1.0 subset and UGB guidelines, evaluation test cases can produce the following metrics:

- *Failure-rate* (*fr*) measures the ratio between actual errors and potential errors (or accessibility opportunities) [25]. For example, the "*images lacking an alternative text"* test case, checks whether each picture has an alternative description. This way, 10 pictures out of 100 would obtain *fr*=0.1 while 5 images out of 25, *fr*=0.2. Therefore the normalized score in terms of conformance would be 1-*fr*.
- *Accept/reject*: whilst techniques to be measured by the failure-rate are checked every time a determined hypertext label or attribute appears some test cases are applied once. For instance, "*number of links*" test case in UGB and implemented by Magenta. This test produces one error if there are more

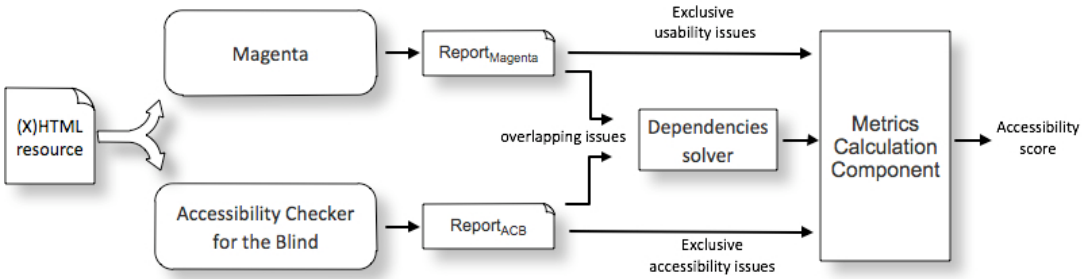

Figure 1. How evaluation tools take part in the process

than 30 links. The metric can be understood as a particular case of failure-rate when the range from 0 to 1 is covered just by integer values. If one test case, the conformance will decrease proportionately to the number of techniques in a guideline. For instance, if "*number of links*" fails, the overall usability of "*number of links and frames*" guideline will decrease in a 50%, as there are just two test cases.

In the case of those test cases that produce warnings it is not possible to know the number of actual errors. Due to the incompleteness of the evaluation and the uncertainty that it bears, for measurement purposes we will assume that all warnings are actual errors. This way, the final score will represent a low-bound accessibility score. Next section deals with the problem of joining all the scores produced by evaluation techniques in order to obtain a single overall score.

3.7 Adapting Logic Scoring Preferences (LSP)

Traditional scoring techniques work as follow: a number of n components are independently evaluated, in this particular case is n=50, which is the number of techniques that tools verify semi-automatically (32 by ACB and 18 by Magenta). Evaluation results are a set of normalized scores $E_1,..,E_n$ where $0 \leq E_i \leq 1$. When evaluated components have a different impact on the measurement, positive normalized weights are associated to each evaluation result $W_1,..,W_n$ where $0<W_i<1$ and $\sum_i W_i = 1$. As a result, the global score is $E=W_1E_1+..+W_iE_i+..+W_nE_n$, $0 \leq E \leq 1$. However, these traditional techniques have the following limitations:

- Mandatory requirements cannot be modelled. If $E_i=0$, E will never be equals to zero.
- If the number of components is very high the impact of a low score of a component is not very significant.
- If components are significant and thus have a high weight, the impact of low-weighted components is irrelevant.

Logic Scoring Preferences (LSP)[11] is an aggregation model that overcomes the above-mentioned limitations. LSP can also be understood as a preferential neural network model. Its strength relies on the capacity of evaluating complex systems that at the same time include numerous subsystems that can be composed by more subsystems and elements. Similarly, UGB and WCAG are composed by general guidelines that at the same time contain numerous checkpoints, which at the same time are decomposed into several techniques for evaluation purposes. The high number of subcomponents and the fact that they can be grouped according to guideline/checkpoint membership leads us to believe that LSP appropriately fits with our purpose of aggregating numerous scores.

$$E = \left(W_1 E_1^{\rho(d)} + \ldots + W_i E_i^{\rho(d)} + \ldots + W_n E_n^{\rho(d)} \right)^{1/\rho(d)}$$

Besides, LSP was successfully applied in the context of the measurement of Web applications usability [22]. LSP overcomes the drawbacks of traditional aggregation systems by applying the weighted power mean. Values of $\rho(d)$ are predefined elsewhere [10] and they are selected upon the required logical relationship between elements of the system, be different levels of conjunction and disjunction. The output of the $\rho(d)$ function changes depending on the number of elements to measure and d, which is the degree of disjunction. The value of d ranges from total disjunction (d=1), arithmetic mean (d=0.5), to conjunction (d=0) depending on the logical relationship to be applied. When simultaneity in satisfying the requirements is necessary, conjunction and similarity are applied. In this case low scores heavily determine the final results. Contrarily, if the objective is to penalize the main component only if all subcomponents fail, the disjunction is applied. This way, only if most scores are low there will be an impact on the final result. Intermediate values are preferred, as extreme cases do not apply. This intermediate range of values is ($0<d<0.5$) for *quasiconjunctions* ($0.5<d<1$) and for *quasidisjunctions*. Depending on the value of d, relationships between elements can be weak, medium or strong. More details on the mathematical background can be found in [11].

LSP is useful when components in a system are hierarchically shaped and there are numerous items. The four UGB principles can be decomposed into 18 checkpoints and at the same time, several techniques implement checkpoints. The subset of WCAG has been classified in seven groups: images, tables, scripting, content, navigation, structure and forms. The relationships between the components in the system are determined by their typology (be the technique automatically or semi-automatically testable) and the location within the hierarchy. These relationships are described below:

- **Relationship between evaluation techniques that implement a checkpoint**. Evaluation techniques are understood as the basic, minimum requirements that describe a particular accessibility attribute. As for automatic evaluation purposes, techniques are often decomposed into *test cases*. It is thus required that all the techniques are met in order to satisfy a checkpoint. In other words, it is mandatory satisfying all techniques simultaneously. This way, low input values will strongly determine the final result. Regarding LSP, this idea of simultaneity fits with the conjunction logical relationship and can be clearly explained by the "a chain is as strong as its weakest link" statement. However, as the typology of techniques may vary regarding their fulfilment certainty, it is crucial to define their relationship and the degree or conjunction or disjunction applied:
  - **Case 1**: a_e vs. a_e. It is assumed that there is strong certainty for each score since they have been automatically obtained. Thus, as simultaneity is required to meet the whole checkpoint, the strong *quasiconjunction* (C+, d=0.125) is applied.
  - **Case 2**: a_e vs. w. As we assume that warnings will fail to meet the specific technique in a checkpoint their value is equals to 0. This may seem a pessimistic approach that is counteracted by applying the medium *quasidisjunction* (DA), where simultaneity is not required and d=0.75.
  - **Case 3**: a_r vs. w. This relationship is similar to the above, which entails that there is some uncertainty due to w. In order to not to determine final score by low values produced by w, the presence of a_r will increase the final score when applying the weak *quasiconjunction* (D-), where d=0.625.
  - **Case 4**: a_e vs. a_r. Both issues are fully automatable tests but failing a a_r technique will not be considered a problem, as it is not believed to produce a serious loss in the interaction quality. Therefore, in order to not penalize its presence, the strong *quasidisjunction* (D+, d=0.875) is applied, where low values do not have a strong impact in the results.

Figure 2 shows where all the afore-mentioned cases are located in the range of logic relationships that LSP provides:

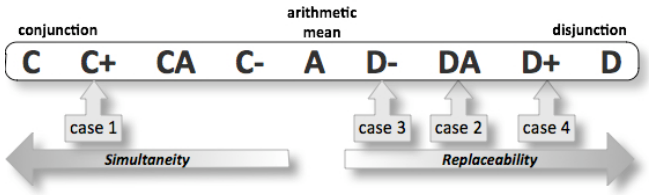

Figure 2. Subtest cases location in the LSP value range

We can consider the checkpoint "*proper form layout*" in Magenta to illustrate the method. Let's consider the following scenario without considering priorities of techniques:

- Technique 1, "*adequate control matching*" (a_e), finds that out of 4 input elements 3 of them lack the required `for` tag which entails that failure rate is fr=0.75 and $score_1$=0.25.
- Technique 2, "*insert mandatory elements*" (w), checks the existence of "*" character for mandatory values in forms. Since 3 `label` elements do not contain it 3 warnings are produced.
- Technique 3, "*scripting issues*" (a_e), penalizes the *OnClick* event. None is found so $score_3$=1.
- Technique 4, "*label buttons*" (a_r), finds that out of 3 buttons 1 is adequately labelled, thus fr=0.66 and $score_4$=0.33.
- Technique 5, "*groping elements*" (a_e), find that `fieldset` or `legend` tags are missing. Thus, the $score_4$=0.

Figure 3 depicts the application of the metric for "*proper form layout*" checkpoint.

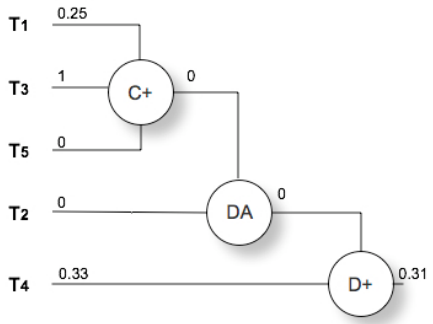

Figure 3. Example of LSP application to "*proper form layout*"

C+ is applied because T_1, T_3 and T_5 are automatically obtained values, that is, they are a_e type. As C+ logical relationship models simultaneity among scores, 0 score produced by T_5 heavily determines the result, which is equal to zero. Next, DA is applied between the previous result and score obtained with T_2 obtaining 0 as an intermediate value. After applying this logical function the type of this result becomes a_e. Finally D+ is applied between T_4 and the previous intermediate result scoring 0.31, a quite low accessibility score.

- **Relationship between checkpoints that implement a guideline**. Single checkpoints can be considered elemental attributes since they target a particular usability or accessibility issue. In addition, sets of these checkpoints are grouped in order to satisfy higher-level usability principles, in the particular case that concerns this paper: *structure and arrangement*, *content appropriateness* and *consistency* for the UGB and *images*, *tables*, *scripting*, *content*, *navigation*, *structure* and *forms* for the subset of WCAG. Since simultaneity is also a requirement among those test cases within a guideline, C- logical function is applied.
- **Relationship between guidelines**. Among all the above principles, each guideline is weighted by the number of checkpoints it contains divided by the total number of checkpoints where the weight, $0 \le w_i < 1$ and $\sum_i W_i = 1$. In this particular case, the overall accessibility score is $\sum_i^{principles} W_i E_i$ where $0 \le E \le 1$.

4. USER TESTING

A user test with 16 blind users with a mean age of 43 (sd=11) was conducted in order to analyse the navigation strategies they followed with annotated links[1]. All of them used JAWS screen reader on Internet Explorer 7 and 8 except one that used JAWS jointly with a screen magnifier. All of them were experienced Internet users since the 43% spent more than two hours a day browsing the Web, the 36% between 1-2 hours and the 21% less than one hour.

4.1 Test Environment

We conducted a remote user test in order to let participants use their own environment (i.e. their PC with the personal settings of the screen reader) allowing them to work in a more comfortable situation. This way, people from numerous geographical locations could get involved, which is a particularly important aspect when involving users with special needs and characteristics, such as blind users. The remote user testing was composed of (1) a set of tasks to be carried out with the remote environment and (2) a questionnaire aimed at collecting subjective opinions.

Each user had to connect to a Web site designed to assist users to carry out the assigned tasks. The system is able to manage task control, timings, and capture main user actions carried out via keyboard and mouse. All collected information was captured and recorded in a log file by using a remote logging tool already used for other analogous tests [20]. Specifically, for each user the URL of the opened pages, the name and the timestamp data were automatically stored in the log file. The URLs of the pages provided the sequence of the links chosen by the user. Efficiency and effectiveness was inferred from log files data. In addition, users were able to write their comments in a form after each task and were told to fill in a post-test questionnaire in order to gather demographic data and collect their impressions and suggestions as well.

4.2 Tasks

When referring to navigation as a general term, Jul and Furnas [17] distinguished between two tasks (browsing and searching) and two tactics for accomplishing these tasks (navigation and querying). In this user test, we focused on navigation tactics for both tasks.

4.2.1 Scenario 1: Browsing by Navigating

According to the terminology proposed above [17] this task is defined as *looking for what is available in a web page by moving oneself sequentially deciding at each step where to go next*. The

[1] Further information on the experimental settings and results is available at http://sipt07.si.ehu.es/assets09/exp_settings.html

objective of this scenario was to check how users behaved when they do not have a specific target in mind or when the target is too vague and thus changeable. In this particular case, in addition to personal preferences, link relevance may also play an important role. We designed a 10-link scenario where each link pointed to a real page that had previously been stored in a local server. Two similar sites were created in order to compare the behaviour of users and the strategies they adopted. Each site included pages with links collected from the top ten Google Search results for the keywords "Pisa" and "Firenze". Search results for each query were very heterogeneous because of the vagueness of the query and the results followed a pattern in their first 10 results: Wikipedia page, city council, local football team, local university, and so on. Thus, both pages followed almost the same structure with regard to topics and rankings while only the content changed. One of the sites was manually annotated with the accessibility scores according to the method in Section 3.6, and relevance scores, while the other site had no annotations at all. The purpose was to observe if any difference occurs in the user behaviour in pages with annotated links or not. In addition to the two tasks, the users were asked to fill in a very short form to report some comments regarding free navigation. The relevance of each link was measured according to its ranking in search results. First and second results were considered "very relevant", third and forth were "relevant", fifth and sixth had "medium relevance", seventh and eighth had "low relevance" and ninth and tenth were "irrelevant". Therefore relevance was measured with a scale of 5. Users were told to freely browse each site for 5 minutes bearing in mind that after browsing they had to fill out a form explaining what they had learned in each session. After 5 minutes the user testing environment instructed them to proceed to the second site. The purpose of such objective was to motivate users to perform the task.

4.2.2 Scenario 2: Searching by Navigating

In the second scenario each user had to *look for a known target* (searching) *by moving oneself sequentially deciding at each step where to go next* (navigation). The objective of this scenario is to ascertain how annotating links with accessibility scores affects users behaviour when they have a specific target in mind. Again, a site with 10 links was built, although this time relevance was not considered because otherwise users would have been directed towards the target. In contrast to Scenario 1, this time the site was information-neutral and homogeneous as the list of links was obtained with a more specific keyword, "accommodation in Pisa". This way, the site contained 10 links pointing to the home page of hotels in Pisa and two tasks were devised for this site: (1) given a determined telephone number users should find a hotel and (2) given a street name users had to find the name of the hotel. Users were asked to write the answer in a form. This allowed us to understand if the user successfully accomplished the task. We performed a within-subject test. Therefore, all users performed tasks with annotated links and without annotations in the same site. In order to remove the learning effect, task order was inverted for the two scenarios.

Where appropriate in both scenarios, hyperlinks were manually annotated with the numerical accessibility score of the page they pointed to right after the content of the link. Accessibility scores of annotated links are depicted in Figure 4; in the *browsing by navigating* scenario the maximum score was 75 out of 100, the minimum 45, the median 52 and the mean 56; while for the *searching by navigating* scenario, the data were the following max=54, min=34, median=41, mean=43.

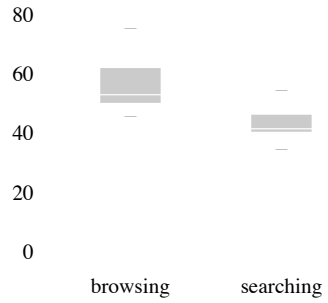

Figure 4. Box-plots for accessibility scores in annotated sites

5. RESULTS AND DISCUSSION

In order to observe the strategy adopted by users, browsing paths were analyzed. In scenario 1, for the page with no annotations, when the paths taken by users were compared with the link sequence in the page, applying the Kendall τ correlation 9 users obtain values for τ that range from 0.8 to 1 (at most $p<0.03$). This indicates that they proceeded sequentially. For the annotated page only 2 users proceeded similarly, with $\tau=1$ (at most $p<0.05$). Regarding accessibility scores in the annotated page, none of the users followed the sequence of links based on their accessibility scores. However, when aggregating accessibility scores of visited pages, a mean of 59 is obtained, which is 7 points over the median. This indicates that even if users did not follow an accessibility-based path, they only browsed in those pages that were annotated as highly accessible. Within accessible pages they might have selected links according to their particular likings. It was also observed that few of them proceeded in a dichotomous way as if they were comparing the accessible results with the less accessible ones. Finally, relevance was not considered by any of the users.

In the non-annotated page of the second scenario, 2 users followed a linear sequence of links $\tau=1$, both $p<0.05$ and one of them proceeded the other way around $\tau=-1$, $p<0.02$ that is, from the last link to the first. The rest of users proceeded randomly. In the annotated version two people proceeded following the link sequence $\tau=1$, $p<0.05$ and only one followed the accessible link path $\tau=1$, $p<0.02$. However, as in the previous scenario users preferred the accessible sites to the non-accessible ones, since the mean for their aggregated accessibility values was 47, as before 6 points over the median. This time they also chose in favour of the accessible links, but they might have been guided in these choices by their intuition. In addition, users performed better in the annotated page as efficiency (task execution time) was higher: it took users a mean of 108 seconds against the 132 seconds in the non-annotated one to complete a task. Effectiveness (successfully completed task rate) was also higher for the annotated page 100% while for the page without annotations was 87%.

Even if they were not asked to, users were able to make some comments after each task in the first scenario. 8 of them appreciated the links annotated with accessibility scores. However, the suitability of the scores was more controversial: 5 users were satisfied with scores ("*I think the values adequately reflect the accessibility level*", "*scores are useful*", "*scores are interesting*", "*accessibility values seem correct*", "*navigation is better if scores are included*"); 3 participants, who performed first

the browsing task with annotated links, stated that "*I'm doubtful about the accessibility criteria*", "*I find it surprising there is so much difference in the accessibility level when scores are similar*", "*scores seem random*". However, after browsing the page without annotations, the three of them changed their minds: "*links with accessibility scores are useful*", "*accessibility scores make navigation more instinctive and smoother*", "*I missed the accessibility scores for this task*". The reason for this may be that the users are very demanding about the accuracy of the scores, but nevertheless they appreciate them. If scores were misleading they would not miss them. There were also criticisms such as "*strange validation*", "*interesting annotation even if some scores are not very coherent*". In general, it seems that scores make navigation less difficult, and most users find annotations useful.

14 users filled out a post-test questionnaire answering the following questions in a 5-pointLikert scale (1-totally disagree, 3-quite agree, 5-totally agree).

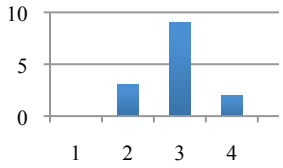

Figure 5.1. Scores are useful for the browsing task.

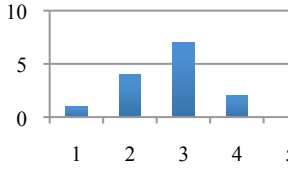

Figure 5.2. Scores are useful for the searching task.

It can be observed there is a peak in 3 for both scenarios entailing that they agree on the usefulness of scores to a certain extent.

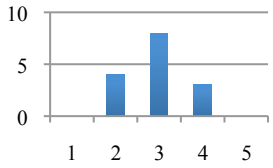

Figure 6.1. Scores are correlated with the actual accessibility perception in the browsing task.

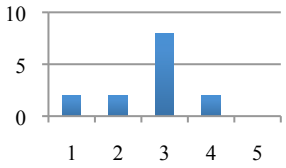

Figure 6.2. Scores are correlated with the actual accessibility perception in the search task.

There is a peak in 3 again for both scenarios even if Figure 6.1 is more balanced. Hence, there is not an agreement on the usefulness of annotations and on the actual perception of scores. As one user remarked this may happen because "*the perception of accessibility depends on each user and his/her particular computer settings*". We can thus conclude that the presented metric is useful even if it does not suit to all users. User perception of metrics is less balanced in the search scenario as the pictures above show. This may happen because the scores in this task do not span the whole value range (see Figure 4), that is, scores might not have enough discriminative power to differentiate different links. When asked if they would make use of annotated links in a regular browsing scenario, 11 stated they would do it. Yet, when they were asked more specifically about the usage scenario they responded that it should be in a site where links target pages containing content regarding similar topic. There was not an agreement if users prefer qualitative (8 users) or quantitative scores (6 users). In any case, quantitative metrics are required to define qualitative scores.

In the browsing scenario the majority of users accessed links sequentially. However, when links were annotated with accessibility scores, users lean towards browsing through the subset of the most accessible. The choice among the accessible links did not follow any particular criterion. It seems that personal preferences or intuition influence their choice among the subset of accessible links. It can be concluded that links annotated with accessibility scores changed the sequential way users browse since they focus on the accessible ones by setting themselves a lower accessibility numeric threshold, which was 7 points over the median in our user test. As for the searching task, only a few users followed a sequential path (not even in the non-annotated version) or the accessibility scores based one. In this scenario, annotated links outperformed regular links when it comes to efficiency and effectiveness and also browse through the subset of those pages scoring over 6 points over the median. Therefore, we can conclude that links annotated with accessibility change blind user navigation paradigm, which is usually sequential in the *browsing by navigating* scenario. In addition, user preference for the annotation technique and perception of accessibility scores is more balanced than in the *searching by navigating* scenario. Users show their preferences for using the annotation technique when links are homogeneous with regard to the topic addressed by the target pages. For example, when browsing through Web directories (say a directory of travel guides) users narrow down their choices according to their likings until they get to the desired node which is usually related to a given topic (e.g. a page containing a set of links pointing to different travel guides of Morocco). Once this leaf node is reached annotating the outgoing links with accessibility scores will enhance user experience.

This technique can be complementary to search engine results rankings as the provided link list can be annotated with the respective accessibility score. As it was demonstrated elsewhere [29], top ten results provided by search engines such as Google and Yahoo! are accessible to the greater extent supporting Pemberton's [23] statements on the similarities between blind users and web crawler behaviour. However, accessibility is not considered as the ranking criterion, not even by Google Accessible Search [labs.google.com/accessible/].

6. CONCLUSIONS and FUTURE WORK

This paper presents a novel technique for adaptive navigation support of blind users: information scent augmented by annotating hyperlinks with the numeric quantitative accessibility score of the target Web page. Accessibility scores are automatically obtained with the assessment framework herein presented. The framework consists of two automatic guidelines review tools: the ACB and Magenta, which support accessibility and usability guidelines respectively. The Metrics Calculation Component exploits data from evaluation reports and yields quantitative accessibility scores for web pages that are calculated by adapting the Logic Scoring Preferences measurement method for the specific reporting issues of this assessment framework.

In order to analyze user behaviour with annotated links two experimental settings were devised: *browsing by navigating* and *searching by navigating*. In the former scenario it was observed that users browse in a sequential way. However, when links are annotated with accessibility scores they browse according to their particular likings through the subset of the most accessible pages. Regarding accessibility scores there was not an agreement between scores and actual accessibility perception although users state that scores convey the perceived accessibility to a certain extent. This may happen because the metric yields lower bound accessibility scores caused by the assumption that all warnings are actual errors. Comments made by users lead us to conclude that the annotation technique might prevail over the scores themselves.

That is, even if sometimes scores are not as accurate as expected, users find them helpful as they make the navigation easier. As one user remarked "*I was sceptical about scores but finally I led to the most accessible links*". Finally, it is concluded that annotations can play an important role when links are homogeneous regarding the topic of the target page. The outcomes can be extrapolated in order to be applied in other adaptive navigation support techniques such as adaptive sorting or local orientation. Since accessibility scores are automatically obtained future work foresees the encapsulation of the assessment framework into a browser plug-in. This way, links will be automatically annotated.

7. REFERENCES

[1] Abou-Zahra, S., and Squillace, M. (Eds.). Evaluation and Report Language (EARL) 1.0 Schema. http://www.w3.org/TR/EARL10-Schema

[2] Brajnik, G., and Lomuscio, R. (2007). SAMBA: a Semi-Automatic Method for Measuring Barriers of Accessibility. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS'07, 43-49.

[3] Brajnik, G. (2009). Barrier Walkthrough: Heuristic Evaluation Guided by Accessibility Barriers. http://www.dimi.uniud.it/giorgio/projects/bw/bw.html

[4] Bigham, J., Cavender, A., Brudvik, J., Wobbrock, J., and Ladner, R. (2007). WebinSitu: a comparative analysis of blind and sighted browsing behavior. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS'07, 51-58.

[5] Bigham, J., Lau, T., and Nichols, J. (2009). TrailBlazer: Enabling Blind Users to Blaze Trails Through the Web. Intelligent User Interfaces, IUI'09, 177-186.

[6] Caldwell, B., Cooper, M., Guarino Reid, L., and Vanderheiden, G. Web Content Accessibility Guidelines 2.0. Available at http://www.w3.org/TR/WCAG20/

[7] Campbell C.S., and Maglio, P.P. (1999). Facilitating Navigation in Information Spaces: Road-signs on the World Wide Web. International Journal of Human-Computer Studies 50(4), 309-327.

[8] Chisholm, W., Vanderheiden, G., and Jacobs, I. (1999). Web Content Accessibility Guidelines 1.0. http://www.w3.org/TR/WAI-WEBCONTENT/

[9] Craig, J., Cooper, M., Pappas, L., Schwerdtfeger, R., and Seeman, L. (2009). Accessible Rich Internet Applications (WAI-ARIA) 1.0. http://www.w3.org/TR/wai-aria/

[10] Dujmovic, J.J. (1991). Neural Networks - Concepts, Applications, and Implementations. Preferential Neural Networks, 155-206. Prentice Hall.

[11] Dujmovic, J.J. (1996). A Method for Evaluation and Selection of Complex Hardware and Software Systems. International Computer Measurement Group Conference, 368-378.

[12] Fukuda, K., Saito, S., Takagi, H., and Asakawa, C. (2005). Proposing new metrics to evaluate Web usability for the blind. Extended Abstracts of CHI'05, 1387-1390.

[13] Goble, C., Harper, S. and Stevens R. (2000). The Travails of Visually Impaired Web Travellers. ACM Conference on Hypertext and Hypermedia, Hypertext'00, 1-10.

[14] Harper, S., and Patel, N. (2005). Gist summaries for visually impaired surfers. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS'05. 90-97.

[15] Harper, S., Goble, C., and Stevens, R. (2005). Augmenting the mobility of profoundly blind Web travellers. New Review of Hypermedia and Multimedia 11(1), 103-128.

[16] Ivory, M.Y., Yu, S., and Gronemyer, K. (2004). Search Result Exploration: A Preliminary Study of Blind and Sighted Users' Decision Making and Performance. Extended Abstracts of CHI'04, 1453-1456.

[17] Jul, S., and Furnas, G.W. (1997). Navigation in electronic worlds: a CHI 97 Workshop. ACM SIGCHI Bulletin, 44-49.

[18] Lazar, J., Allen, A., Kleinman, J., and Malarkey, C. (2007). What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. International Journal of Human-Computer Interaction 22(3), 247-269.

[19] Leporini, B., Paternò, F., and Scorcia, A. (2006). Flexible tool support for accessibility evaluation. Interacting with Computers 18(5), 869-890.

[20] Leporini, B., and Paternò, F. (2008). Applying Web Usability Criteria for Vision-Impaired Users: does it really improve task performance? International Journal of Human-Computer Interaction 24(1), 17-47.

[21] Mahmud, J., Borodin, Y., and Ramakrishnan, I.V. (2007). CSurf: A Context-Driven Non-Visual Web-Browser. World Wide Web Conference, WWW'07, 31-40.

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

[23] Pemberton, S. (2003). The kiss of the spiderbot. interactions 10 (1), 44.

[24] Pirolli, P., and Card, S. (1999). Information foraging. Psychological Review 106(4), 643-675.

[25] Sullivan, T., and Matson, R. (2000). Barriers to use: usability and content accessibility on the Web's most popular sites. ACM Conference on Universal Usability, CUU'00, 139-144.

[26] Takagi, H., Saito, S., Fukuda, K., and Asakawa, C. (2007). Analysis of navigability of Web applications for improving blind usability. ACM Transactions on Computer-Human Interaction 14(3), article 13.

[27] Vigo, M., Arrue, M., Brajnik, G., Lomuscio, R., and Abascal, J. (2007). Quantitative Metrics for Measuring Web Accessibility. International Cross-Disciplinary Conference on Web accessibility, W4A'07, 99-107.

[28] Vigo, M., Kobsa, A., Arrue, M., and Abascal, J. (2007). User-Tailored Web Accessibility Evaluations. ACM Conference on Hypertext and Hypermedia, Hypertext'07, 95-104.

[29] Vigo, M., Arrue, M., and Abascal, J. (2009). Enriching Information Retrieval Results with Web Accessibility Measurement. Journal of Web Engineering 8(1), 3-24.