Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

An HTML analyzer for the study of web usability

David Alonso-Ríos, Iván Luis-Vázquez, Eduardo Mosqueira-Rey, Vicente Moret-Bonillo, B. Baldonedo del Río
Department of Computer Science
University of A Coruña
A Coruña, Spain
dalonso@udc.es, iluisv@gmail.com, eduardo@udc.es, civmoret@udc.es, belenb@edu.xunta.es

***Abstract*—This paper presents an HTML analyzer for the study of web usability. The analyzer parses HTML code in order to extract usability information from web pages. For usability aspects that can be analyzed automatically, the analyzer draws conclusions and offers suggestions. For more subjective usability issues, it assists the expert by presenting relevant information in a convenient way. Many similar applications exist, but they mostly focus on well-known usability problems and pay little attention to subtler aspects. More alarmingly, they suffer from basic usability problems. Our results show that our analyzer examines several usability aspects – related to ease of navigation, understandability, flexibility, and compatibility – that are generally ignored by the other tools.**

Keywords—human-computer interaction, web usability, HTML analysis

I. INTRODUCTION[*]

The analysis of HyperText Markup Language (HTML) code is currently one of the most popular automated techniques for the study of web usability. The main reasons for this are the popularity of HTML – which is the standard programming language for the World Wide Web – and its inherent simplicity.

Even today, HTML is remarkable for its limited set of elements and its relatively linear and transparent structure. This means that HTML code is easy to parse automatically, that it is easy to identify common usability problems, and that most programming errors are fixed with little effort. This contrasts strongly with general-purpose programming languages like C++ and Java, which are much harder to parse and interpret.

Moreover, several accessibility guidelines for HTML code have been developed in the past ten years, such as the World Wide Web Consortium's WCAG [1].

The combination of these facts explains the current ubiquity of HTML analysis. Unfortunately, many usability issues are more ambiguous, and therefore harder to detect and interpret. In fact, the assessment of usability has an intrinsic subjective component that cannot be ignored. Sometimes, the opinion of a human expert is required. Thus, it is not enough to identify typical programming problems and guideline violations, it is also necessary to extract data on different types of usability aspects (both good and bad) and to present them in an easy to digest manner.

This paper presents an HTML analyzer for the study of web usability. Our goals are, firstly, to make the analyzer itself easy to use and understand, and, secondly, to go beyond the straightforward detection of well-known usability problems with the aim of detecting more complex usability issues.

II. BACKGROUND

A wide variety of HTML analyzers are currently available, many of them for free. However, most of these applications do not try to address all the aspects that the study of web usability typically involves. Most analyzers simply specialize in validating code syntax and checking compliance with accessibility guidelines, especially the previously mentioned WCAG [1] and the Section 508 of the government of the United States of America [2].

Popular tools include WAVE [3], Truwex [4], FAE [5], TAW [6], ATRC [7], Total Validator [8], CSE HTML Validator [9], Cynthia Says [10], the W3C Markup Validation Service [11], and the W3C CSS Validation Service [12].

Most analyzers examine a single HTML page at a time, but some are capable of performing an exhaustive inspection of an entire website.

A common flaw in all these tools is that they tend to detect simple usability problems and pay little attention to subtler issues. This will be explained in detail in section VI, which provides a comparative summary of specific usability aspects detected by our analyzer but generally ignored by the most prominent applications.

Another problem with these tools is that, ironically, they often have usability problems themselves. They tend to structure the information rather poorly, to overflow the user with repetitive messages, to force the user to hover the mouse over dozens of error icons in order to be able to read the associated messages, and to show messages that are simultaneously too verbose and too imprecise. For example, a common error message in WAVE is "ALERT: Problematic link text. Link text does not make sense out of context, contains extraneous text (such as "click here"), or is the same as another link on the page, but links to a different location". Therefore, the usability of the analyzer itself was one of our basic requirements.

[*] This research has been partly funded by the Xunta of Galicia Regional Government under Project 08SIN010CT.

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

III. Our HTML Analyzer

Our HTML analyzer examines HTML code to extract usability information about web pages. The analyzer identifies usability problems, checks compliance with usability guidelines, and interprets usability information to facilitate the task of human usability experts.

The user interacts with the analyzer through a graphical interface. To analyze a web page, the user must enter its URL. The analyzer then creates a report with all the usability aspects analyzed by the application. These aspects are divided into six categories according to the type of element: the web page itself (global aspects), images, forms, tables, lists, and links. The user can then read the results for each category. Section IV provides a description of the specific aspects covered by each category.

For each aspect, the analyzer offers a brief description and classifies the results into “positive”, “negative”, and “warnings”. A warning is not necessarily positive or negative, but it is something that the user should be notified about. Of course, this distinction is not always clear-cut and depends largely on the context.

In order to extract and interpret usability data, the HTML analyzer creates a structured model of the elements of the page. This model is directly based on the HTML specification [13]. As an example, Fig. 1 shows one part of this model, namely, the links (which correspond to the “A” element of the HTML specification).

Additionally, the HTML analyzer forms part of a larger project consisting in a multi-agent system (MAS) for the usability analysis of websites. The MAS also includes several user agents that simulate the browsing process of the users and interact with the HTML analyzer with the aim of drawing conclusions and making suggestions about the usability of the website taken as a whole. This MAS was described in [14].

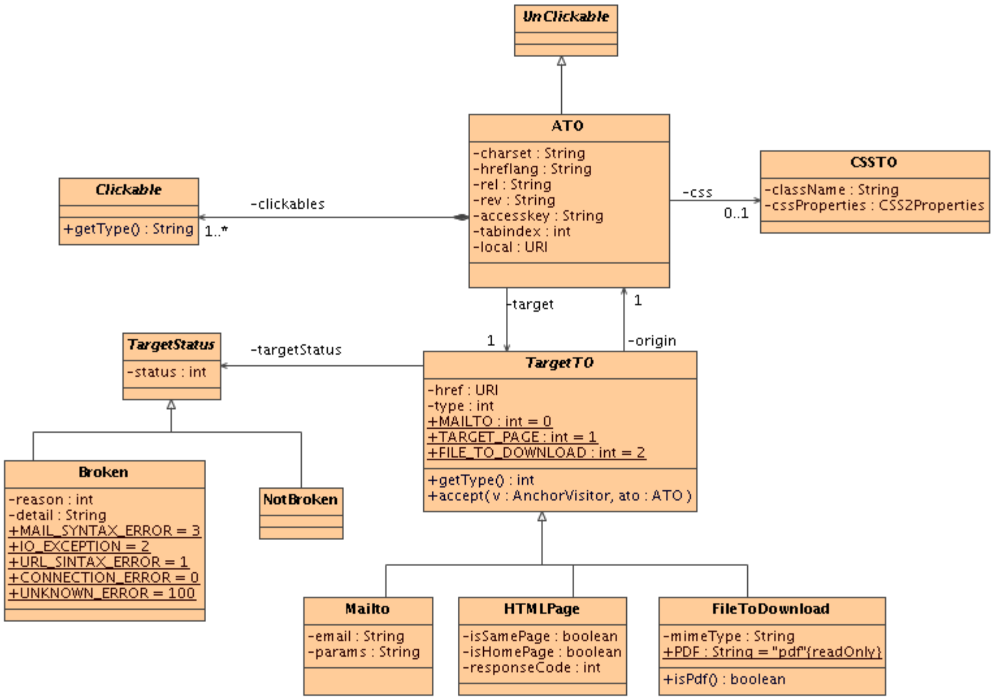


Figure 1. Class diagram representing the links of an HTML page.

IV. Usability Aspects Examined by the Analyzer

This section summarizes the main usability aspects that are examined by our HTML analyzer. Some aspects are always symptoms of usability problems, while others are significant issues (some positive, most negative) that should be notified to the usability expert. The aspects are classified into six categories: web page, images, forms, tables, lists, and links.

A. *Web page:*

- Estimated page size.
- Whether the page title is considered descriptive or not.
- Compatibility problems like browser-specific tags, deprecated tags, the lack of elements such as text encoding in HTTP headers and meta-tags, and the presence of special characters that should have been replaced by HTML entities.
- Flexibility problems, such as fixed font sizes and fixed-width elements, and forcing the use of scripts.
- The use of elements which can be problematic in certain contexts, such as frames, cookies, CSS formatting, scripts, events, and animations.
- The presence of search engines. These elements are very important in large websites. If search engines are used, they should preferably be present at the home page.
- Problems with the text content, such as date/time formats that do not specify the time zone.

B. *Images:*

- Image size (i.e., width and height), and whether it is explicitly declared in the HTML code or not.
- Weight (i.e., amount of kilobytes), which influences download time.
- Visual accessibility features such as alternative text (which should exist, but should not be too long) and the “long description” attribute (which should not be confused with alternative text).
- The use of image maps.

C. *Forms:*

- Number of elements.
- Submit script. This usually means that some kind of global validation is performed on the data before submitting them.
- Fields with an associated script. This serves to identify required fields or fields that are subject to some type of individual validation immediately after being filled. This is often annoying, and the global validation mechanism described above is generally preferable.
- Accessibility features such as easily clickable controls.

D. *Tables:*

- Size.
- Whether the tables are used for presenting data or for formatting the page. The former is their original purpose, whereas the latter is considered problematic.
- The existence of recommended elements such as headers, captions, and summaries.

E. *Lists:*

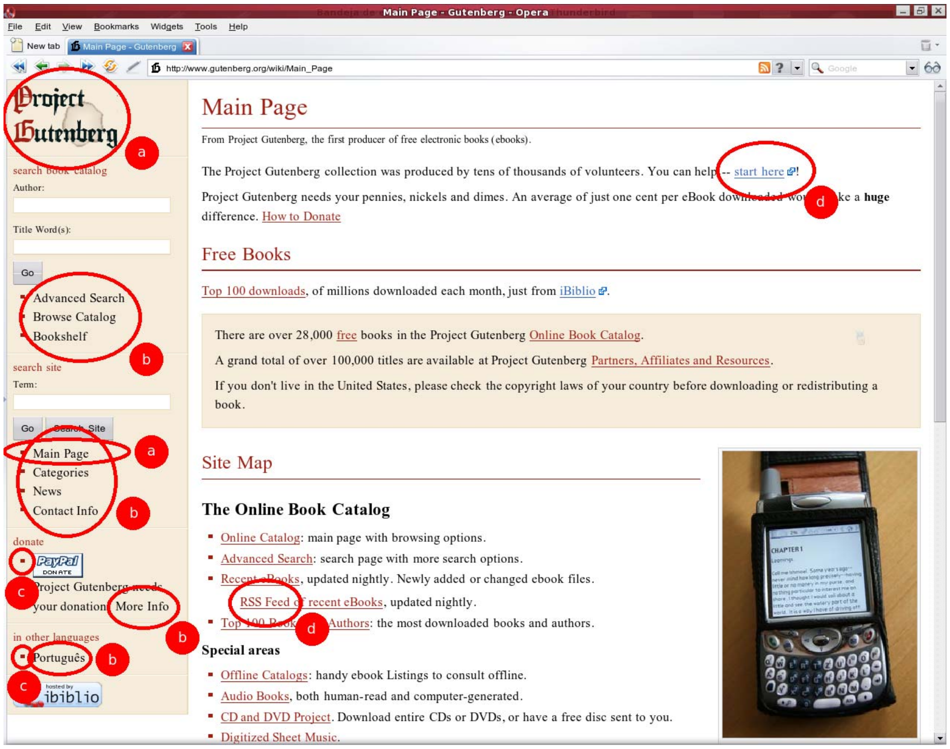
- Number of items.
- Type of lists (i.e., ordered, unordered, and definitions).

F. *Links:*

- Broken links.
- Badly constructed links.
- Pages that link to themselves. Links that point to the current page are at best a waste of time, at worst confusing.
- The use of anchors (i.e., links to a specific area of a page), distinguishing between same-page and different-page anchors.
- Non-standard representations that can cause the user to ignore the links (e.g., non-underlined links).
- Problematic link texts, such as non-descriptive phrases and jargon words. These types of texts do not represent actual content or actions, and may even be confusing. Another example of inappropriate link text is a page with links that have identical text but lead to different URLs.
- The existence of links to non-HTML files. The most common types are document files (especially PDF files, which are often fine for printing but not for reading), multimedia files, and archive files. While all these kinds of files are not necessarily bad in themselves, they can have a negative impact on usability in many ways. The main problems arise when these kinds of files are used as a substitute for text content that should have been available in the more flexible and manageable HTML format. Furthermore, lack of consistency in presentation and formats can have a negative effect on navigation. Many additional problems may occur: some formats are proprietary (e.g., Microsoft Word, MP3, and RAR), the necessary software may not be installed (and could even require the purchase of a commercial application), text browsers may not be able to display the images, and so on.

Metrics for image size, table size, the number of elements in a form, and the number of items in a list can be directly obtained from HTML code. Page size, however, has to be estimated from the data contained in the page – using, for example, the number of words, the size of the images, etc.

More information on these and other well-known problems in web usability can be found in [15].


Main Page - Gutenberg - Opera
File Edit View Bookmarks Widgets Tools Help
New tab
Main Page - Gutenberg
http://www.gutenberg.org/wiki/Main_Page
Google
Project Gutenberg
a
b
a
b
c
b
b
c
d
d
hosted by ibiblio
PayPal DONATE

search book catalog

Author:

Title Word(s):

Go

- Advanced Search
- Browse Catalog
- Bookshelf

search site

Term:

Go Search Site

- Main Page
- Categories
- News
- Contact Info

donate

- PayPal DONATE

Project Gutenberg needs your donation! More Info

in other languages

- Português

Main Page

From Project Gutenberg, the first producer of free electronic books (ebooks).

The Project Gutenberg collection was produced by tens of thousands of volunteers. You can help -- start here!

Project Gutenberg needs your pennies, nickels and dimes. An average of just one cent per eBook downloaded would make a **huge** difference. How to Donate

Free Books

Top 100 downloads, of millions downloaded each month, just from iBiblio.

There are over 28,000 free books in the Project Gutenberg Online Book Catalog.

A grand total of over 100,000 titles are available at Project Gutenberg Partners, Affiliates and Resources.

If you don't live in the United States, please check the copyright laws of your country before downloading or redistributing a book.

Site Map

The Online Book Catalog

- Online Catalog: main page with browsing options.
- Advanced Search: search page with more search options.
- Recent eBooks, updated nightly. Newly added or changed ebook files.
  RSS Feed of recent eBooks, updated nightly.
- Top 100 Books and Authors: the most downloaded books and authors.

Special areas

- Offline Catalogs: handy ebook Listings to consult offline.
- Audio Books, both human-read and computer-generated.
- CD and DVD Project. Download entire CDs or DVDs, or have a free disc sent to you.
- Digitized Sheet Music.

Figure 2. Some usability problems found in the home page of the Project Gutenberg, including: a) links to the current page; b) undecorated links; c) lists with only one item; d) problematic link texts.

V. Example

In order to test our tool, we decided to analyze the Project Gutenberg website [16]. Project Gutenberg was created in 1971 and is currently one of the best-known archives of free electronic books in the web. Its interface is based on the popular MediaWiki software for the development of collaborative websites. Below, we list and briefly describe the most relevant usability issues found in the home page. Fig. 2 shows a screenshot of the home page and visually highlights some usability problems.

A. *Web page:*

1) *Positive:*

- The page title is considered descriptive.
- Appropriate use of meta-tags, such as page type and charset.
- Language and charset encoding are specified in the HTTP headers and are consistent with the meta-tags.
- No deprecated tags are used.
- No browser-specific tags are used.
- No frames are used.
- No cookies are used.

2) *Negative:*

- Scripts are used. More problematically, there is no “noscript” element. That is, the page offers no alternative to the use of scripts.
- The page uses date/time formats that do not specify the time zone.
- The page is somewhat cluttered. In particular, there is a high number of links (63).

B. *Images:*

1) *Positive:*

- “Height” and “width” formatting attributes are declared.

2) *Negative:*

- One image lacks the “alt” attribute (alternative text).
- One image has an overly long “alt” attribute.
- No image has the “longdesc” attribute (long description).

C. *Forms:*

1) *Positive:*

- There are no fields with an associated script.
- Text fields are easily selectable by clicking on the label.

2) *Warnings:*

- Forms do not have an associated submit script.

D. Tables:

- Not applicable. There are no tables.

E. Lists:

1) Negative:

- Some lists have only one item.

2) Warnings:

- Some lists are unordered.

F. Links:

1) Positive:

- There are no broken links.
- File formats are non-proprietary (there is one RSS/XML text file).

2) Negative:

- The page links to itself.
- Anchors are used.
- Non-standard link styles are used. Even worse, the left side of the screen has several links that are indistinguishable from plain text because they are undecorated (i.e., they are black and are not underlined). The user can sometimes infer from the context that they are links, but not always. For example, a text states: "Project Gutenberg needs your donation! More Info". The "More Info" part is a link, and the rest is plain text, but they both look the same.
- In addition to this, link styles are inconsistent.
- Problematic link texts are used. For example, vague, non-descriptive phrases such as "start here". Jargon words such as "RSS Feed" are also used, which can be problematic for some users.
- Links are missing some recommended attributes such as "rel", "rev", "hreflang", "charset", "accesskey", and "tabindex".

VI. Comparison with Other Analyzers

Table 1 offers a brief comparison between our results and those obtained by other analyzers. Obvious usability issues that are covered by most applications (e.g., broken links, the absence of "alt" attributes, and image height and width) have been omitted.

The results show that our analyzer addresses several usability issues that are generally ignored by the other tools. For example:

- The presence of elements that make the website easier to navigate, such as search engines or descriptive page titles.
- Usability problems that have a negative impact on navigation, like inappropriate link styles or pages that link to themselves.
- Flexibility problems, such as using tables for formatting.
- Compatibility problems, such as browser-specific tags or deprecated tags.
- Suggestions on readability and understandability, such as problematic link texts, unspecified time zones, length of "alt" attributes, page/link clutter, and lists with only one item.
- Information on the types of files that are linked and the necessary software.

TABLE I. Summary of usability aspects detected by our analyzer but ignored by other analyzers

Usability aspect	Category	WAVE	FAE	TAW	ATRC	Our analyzer
Descriptive page title	Web page	Only if problematic	-	-	Yes	Yes
Browser-specific tags	Web page	Partially	-	-	-	Yes
Deprecated tags or attributes	Web page	Partially	-	Yes	-	Yes
HTTP headers	Web page	-	-	-	-	Yes
Meta-tags	Web page	-	Yes	-	-	Yes
Noscript elements	Web page	Only if used	Only if used	Yes	-	Yes
Cookies	Web page	Only if used	Only if used	-	-	Yes
Frames	Web page	Yes	Only if frame lacks a title	-	-	Yes
Search engines	Web page	-	-	-	-	Yes
Time zone	Web page	-	-	-	-	Yes
Number of links	Web page	-	-	-	-	Yes
Length of "alt" text	Images	-	-	-	-	Yes
Using tables for formatting	Tables	-	Only if used	-	Yes	Yes
Table tags	Tables	Yes	Partially	Yes	Partially	Yes
Lists with less than two items	Lists	-	-	-	-	Yes
Ordered and unordered lists	Lists	Yes	Yes	-	-	Yes
Links to current page	Links	-	-	-	-	Yes
Inappropriate link styles	Links	-	-	-	-	Yes
Problematic link texts containing non-descriptive words	Links	Yes	-	-	-	Yes
Problematic link texts containing jargon	Links	-	-	-	-	Yes
File type of link target	Links	Yes	-	-	-	Yes
Required software for link target	Links	-	-	-	-	Yes

It should be pointed out that most of these issues appeared in the example described previously.

It should also be noted that there are some elements, such as cookies and noscripts, that are analyzed by the other tools but are only mentioned in their reports in some cases. For example, when they are used in the web page in question or when they are problematic. When they are not present, the reports do not inform of the fact that the issue has been analyzed and that no problems have been detected, leaving the user unsure as to whether the issue has been examined or not. In contrast, our analyzer informs the user about positive and negative aspects and makes warnings and suggestions.

VII. Discussion and conclusions

In this paper we have presented an HTML analyzer for the study of web usability. The analyzer automatically draws conclusions and makes suggestions by analyzing both the HTML code of web pages and their content.

HTML analysis is a very popular automated technique for studying web usability, mainly because HTML is simple to parse and has become the standard language for web pages. Regardless of the simplicity of the language, inferring non-trivial usability issues from HTML code can still be a very complicated task. Sometimes a computer program cannot interpret all the ambiguities, which means that the opinion of a human expert is necessary.

Many HTML analyzers have been developed in the past few years, and some of them are even available for free. However, we have found that some problems still remain.

Firstly, most tools do not address all the aspects that the study of usability typically involves. Many applications are, by their own admission, just accessibility checkers or HTML validators. Even when usability problems are detected, this is done very straightforwardly, with little attention paid to subtler issues. Instead, they tend to limit themselves to checking compliance with well-known usability guidelines.

Another significant problem we found is that these applications have usability problems. They tend to organize the information poorly, to be overly repetitive, and to force the user to perform awkward movements in order to read the information.

Thus, our intended goals for the HTML analyzer were, firstly, to apply the principles of usability and to try and make the application itself easy to use and understand, and, secondly, to try and detect subtler usability issues. These aspects are where the novelty of our approach lies.

Our results show that our analyzer examines several usability aspects that are generally ignored by the other tools. These aspects are mainly centered on ease of navigation, understandability, flexibility, and compatibility.

The HTML analyzer also forms part of a larger project consisting in a multi-agent system for the usability analysis of websites. In this MAS, several user agents interact with the HTML analyzer in order to draw conclusions about the browsing process of the real users and the usability of the website as a whole.

As for future work, we plan to add more functionalities and to address more technically complex aspects. For example, JavaScript parsing would allow us to analyze events such as mouse clicks. Another key goal is to keep improving the usability of the graphical interface. That is, in addition to following common usability guidelines for interfaces, we intend to offer alternative ways to organize the information. We have found that the descriptions given by a typical HTML analyzer explain usability problems in terms of HTML elements, which in practice is too technical for most users, even usability experts. It would be interesting then to organize usability information in a more conceptual way. We have recently developed a taxonomy to describe the concept of usability [17]. This taxonomy has six first-level attributes, namely, knowability, operability, efficiency, robustness, safety, and subjective satisfaction, which in turn are subdivided into several subattributes. For example, a descriptive page title would contribute to clarity, which, in our taxonomy, is a subattribute of knowability; not forcing the use of cookies would contribute to both confidentialy (safety) and flexibility (operability), and so on. This taxonomy will serve as the basis for organizing our usability reports.

Acknowledgment

This research has been partly funded by the Xunta of Galicia Regional Government under Project 08SIN010CT.

References

[1] W3C: Web Content Accessibility Guidelines 2.0 (WCAG 2.0). In Caldwell, B., Cooper, M., Guarino Reid, L., Vanderheiden, G. (eds.), http://www.w3.org/TR/WCAG20/

[2] Section 508, http://www.section508.gov/

[3] WAVE: Web Accessibility Evaluation Tool, http://wave.webaim.org/

[4] Truwex Online, http://checkwebsite.erigami.com/accessibility.html

[5] FAE: Functional Accessibility Evaluator, http://devserv.rehab.uiuc.edu/fae/

[6] TAW: Web Accessibility Test, http://www.tawdis.net/taw3/cms/en

[7] ATRC Web Checker, http://checker.atrc.utoronto.ca/index.html

[8] Total Validator, http://www.totalvalidator.com/

[9] CSE HTML Validator, http://www.htmlvalidator.com/

[10] Cynthia Says, http://www.cynthiasays.com/

[11] The W3C Markup Validation Service, http://validator.w3.org/

[12] The W3C CSS Validation Service, http://jigsaw.w3.org/css-validator/

[13] HTML 4.01 Specification, http://www.w3.org/TR/html401/

[14] E. Mosqueira-Rey, D. Alonso-Ríos, A. Vázquez-García, B. Baldonedo del Río, and V. Moret-Bonillo, "A multi-agent system based on evolutionary learning for the usability analysis of websites," in Intelligent Agents in the Evolution of Web and Applications (Studies in Computational Intelligence, Vol. 167), N. T. Nguyen, and L. C. Jain, Eds. Berlin, Heidelberg : Springer, 2009, pp. 11-34.

[15] J. Nielsen and H. Loranger, Prioritizing Web Usability. Berkeley, CA: New Riders, 2006

[16] Project Gutenberg, http://www.gutenberg.org/

[17] D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo. "Usability: a critical analysis and a taxonomy," International Journal of Human-Computer Interaction. United Kingdom: Taylor & Francis Group, in press.