HEADER DETECTION OF DATA TABLES
Towards the Improvement of the Web Navigation for Impared Visual People

Juan Manuel Fernández[1], Vicenç Soler[1,2]
[1]*Dept. Microelectrònica i Sistemes Electrònics, Universitat Autònoma de Barcelona, Cerdanyola del Vallés, Spain*
[2]*Ciber-BBN: Bioengineering, Biomaterials and Nanomedicine. Campus UAB, Bellaterra, Spain*

Jordi Roig
Dept. Microelectrònica i Sistemes Electrònics, Universitat Autònoma de Barccelona, Cerdanyola del Vallés, Spain

Keywords: Web Accessibility, Table navigation, e-integration, Repairing Tools, Disability, WAI, HTML.

Abstract: The advantages of the new era of information are out of the capacities of the people with visual impairments. This is due to by the lack of sensibility of the Web contents developers and the low use of the accessibility guidelines. These situations can be solved with technology, and the result of the research presented in this paper offers a new system of solving one of the most important problems: the data table navigation. We present a solution that offers excellent results to make the natural complexity of these elements easy. The system is based on the visualization of the table's content, in the same manner that uses a non-disabled person, but without the need made any image processing. The solution can help to us to obtain the total e-integration.

1 INTRODUCTION

The correct design of a Web site is an unresolved matter for the current developers. A correct use of the Word Wide Web Consortium (W3C) standards makes the navigation become easier and friendly. The access to the information stored in the Web can be done in a more efficiently way with the W3C standards. This lack of use of these standards and guidelines is a consequence of different causes, which we will see in this paper. We will talk about the observance of the specific normative of the World Content Accessibility Guidelines (WCAG) (W3C, 1999a) which was made by the World Accessibility Initiative (W3C, 2007a). This group, that is a part of the W3C, develops a set of standards and guidelines to provide a more accessible Web. The lack of use of the WAI's proposals makes the Web lose all its capacity to improve the labour and social integration of all kind of people.

One of the more important problems that exist in the field of Web accessibility is the navigation on tables. HTML tables have a complex nature that impedes to visual impaired people to obtain the information contained in these elements. Moreover, the use of stardards, like WCAG, is very low and therefore the navigation on tables gets worse.

Thus, we propose a solution to the correct data table's navigation, taking into account the stardards of the Web. Thanks to it, we avoid the use of new languages or specific software to read the information of a data table as the most of the solutions proposed (Yesilada, 2004) (Pontelli and Son, 2002) (Krüpl and Herzog, 2006) (K. Kottapally et al., 2003) (Filepp et al., 2002). So, the solution consists of a detection of header's rows and columns, that allows us to offer information regarding the relationships between the headers and cells of the table.

2 ACTUAL SITUATION

As we have already commented, the lack of standard's uses is a very important problem. The HTML grammar is the base of the Web, but it is ignored by the Web site's developers (Shan et al, 2005). This fact is reinforced by two business facts. On one hand, the design of Web sites thinking in a

Manuel Fernández J., Soler V. and Roig J. (2008).
HEADER DETECTION OF DATA TABLES - Towards the Improvement of the Web Navigation for Impared Visual People.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 241-246
DOI: 10.5220/0001529002410246

specific Web browser. And, on the other hand, the speed of the changes of the Web site's content, that is needed to be profitable. The latter, is well known in the business world: time-to-market. But in the Web that is more important because of its speed. The former is the result of a design based on the validation of the Web site with a browser but not with a system like the one offered by the W3C (Quality Assurance Activity, 2007).

One of the most important errors is the use of HTML to offer visual information instead of CSS. The use of both technologies allows us to offer, if necessary, a different visualization of the information without changing the structure of the HTML document. A good example is the use of tables to distribute the contents of the Web site. This use is larger than the expected one when we started the research. Figure 1 shows a comparison over a randomly group of 487 Web pages. The search of these Web documents has been done in an automatic way, totally random. The only condition was that the Web site had to use the table element.

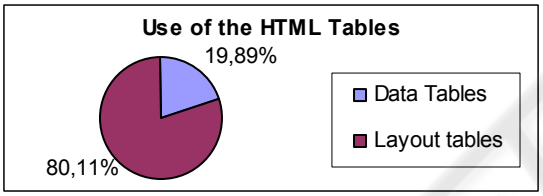

Figure 1: How the HTML tables are used.

Our research is not focused on this problem since there are alternatives to this problem. For instance, Chen (Chen and Shen, 2006) proposes an application to substitute the tables for CSS styles.

3 DATA TABLES

The bi-dimensional nature of a table offers a lot of information. But this nature makes necessary to know the header of the row and the column of a cell to obtain the maximum information. Moreover, this search is only possible to be done visually and the helping navigation tools cannot offer this relationship, making the table to lose its bi-dimensional structure to a linear list of elements. In this list, usually the headers appear just at the beginning and they are very difficult to remember if the list has a great number of elements. For instance, if we get row by row the information of the white cells (not headers) from Table 1 (WebAIM, 2007), for sure that we will not remember what means "January 14" in the table. Visually it is clear that it is Beth's Birthday, but when read by a voice synthesizer it is quite difficult to notice it. For this reason a system that avoids the needing of converting the tables into lists is necessary.

Table 1: Example of a simple data table.

Shelly's Daughters

Name	Age	Birthday
Jackie	5	April 5
Beth	8	January 14

We propose a system that offers the relationship between the headers and the content of the cell. We use HTML standard elements to indicate these relationships. The HTML attribute *scope* (W3C, 2000) allows us to mark the relations of the header with all the cells in a column or row. It is recommended the use of this attribute in the TH tags, which mark the header, and the TD tags, which mark the cells, in simple tables.

In case of a complex table, like Table 2 (WebAIM, 2007), the use of the attributes *id* and *header* is recommended. In this case, the *id* attribute marks the headers, and the cells that have any relation. So, they must contain the value of this identifier in the value of the *header* attribute. Table 2 shows how the header "by birth" affects two rows.

Table 2: Example of a complex data table.

Shelly's Daughters

	Name	Age	Birthday
by birth	Jackie	5	April 5
	Beth	8	January 14
by marriage	Jenny	12	Feb 12

4 PREVIOUS PROPOSED SOLUTIONS

The solutions proposed nowadays can be classified in three basic groups. First of all, the group where the Web browsers adapted to offer correct navigation in tables is included. The second one contains proposals based on new languages. These languages allow a new organization of the information that marks the relationships among cells in an unequivocal way. At last, the proposals which try to modify the content of the document to mark these relationships.

4.1 Adapted Browsers

In this section we are not going to discuss about screen readers like JAWS (Freendom Scientific, 2007) or ZoomText (Ai Squared, 2007). With this software we can use not adapted Web browsers but the screen readers cannot access the structure of a table to offer this kind of information.

We are talking about software like a table browser called EVITA (Yesilada et al., 2004). This kind of solution has a limited field of use and the user has to learn how to use it. Our proposal is totally independent of the Web browser and the user does not need to learn to use new software.

4.2 New Languages

Enrico Pontelli and Tran Cao Son (2002) propose the use of Domain Specific Language to express the content of a table. This content is extracted thanks to the semantics of the information inside the table.

On the other hand, there exist other languages XML based that improve the interaction between the screen reader and the Web site. TTPML (Filepp et al., 2002) is a language of this kind that offers all the information to the screen reader in an easy way. Both of them have the same problem: they are not standards and the user needs specific software to obtain the information offered. Our proposal is based on the W3C standard.

4.3 Header Detection by Means of Visualization

The Web site developer offers information about the relationship of the table's content in a visual way. We can use this difference between cells to obtain the header of a table and to relate the different cells. There exist two approximations to this solution.

The first one is by means of a visual recognition after the Web page has been displayed by the Web browser (Krüpl and Herzog, 2006). This system has the inconvenient that it is strongly dependent of the Web browser. A Web page designed for another Web browser can cause problems to the system.

The second approximation, where we are, works with the source of the Web page. The visualization of the Web document is marked with HTML and CSS code and we can access to it independently from the browser. K. Kottapally et al. (2003) presented a system that implements this proposal. The application implements a logic system and a Hidden Markov Model system. The proposal has very good results but with a very poor test set which produces that the systems based on rules, like this one, can fell on a situation of memorization. On the contrary, our approximation is not based on rules to avoid this situation. It is based on a Bayes classifier and it will be explained in the next section.

5 HEADER DETECTION

As we have commented, it is possible to use the visualization of the different elements of a table to establish the existing relationships. To offer this visual information HTML has a group of tags and attributes that are specific to offer the visual layout. The tags structure the table, and they can offer information on the layout. On the other hand, it is important the use of the attributes of every tag. For instance, if we take the tag TH, this tag has 19 different attributes. We can think that there are a great number of possible combinations, but finally a subset is only used as we will see later.

In current state of our research, we can only work with simple data table, but the system can evolve easily with the method proposed. In this paper, we present the method for the detection of columns headers because of space limitations, but the system is portable to the rows headers.

The first step in the process was to reduce the number of elements involved in the definition of the tables. We have made a study to obtain the use of the different elements of a table. This study was made over a set of 107 random data tables and the first point to observe is the difference of the quantity

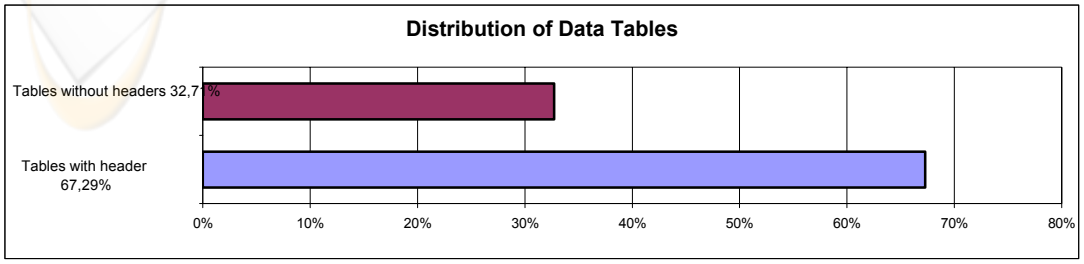

Figure 2: Number of tables with headers.

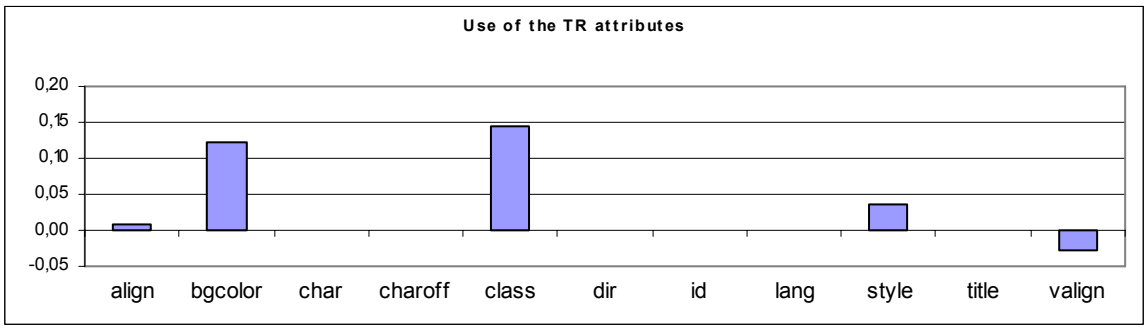

Figure 3: Distribution of the TR attributes.

of tables belonging to both classes. This effect is shown in Figure 2. The tables without header are the most of the part, and they are twice as the tables with headers. This fact makes, that the dataset was imbalanced and it is well known that it is a factor that adds more difficulty to the learning process (Kubat and Matwin, 1997).

If we continue with the study of the content of the tables, we can see that all of them have the header at the first row. This fact is produced because we focus our study on simple data tables. This study also offers interesting information like the very low use of the tags THEAD and TFOOT. These tags are the responsible ones for indicating the start and the end of the table, and all the information inside these tags is marked in an unequivocal way. Both elements only appear in one of the tables studied but the TBODY tag, which indicate the content of the table, appears in 105 tables. In opposition of the use of THEAD, we can work with the use of TH tag. This tag marks the cell like a header of row, and it appears in the 34% of the tables with header.

The study of the attributes has been made only for the tags which play an important role in the subset of the tables studied. These tags are TD, which mark the cells, and TR, which mark the rows.

Figure 3 shows the use of the attributes of the TR tag. The values are standardized, and a value of '1' is assigned if the value of the first row is completely different to the rest of values of the table. A value of '-1' means that the value of the first row is the same at the entire table. The values offered in the Figure 3 are the result of the difference between the average of the tables with and without headers. That means that a value far from the 0 is a good indicator because it means that the attribute is used in this way the most of the cases.

The same process applied to the rows can be used with the cells. We compare the horizontally adjacent cells. This result can be compared with the rest of the rows and we can know if the row is visually marked with every cell. We can see the results in Figure 4.

At last, it is necessary to analyze two situations really carefully: the use of *scope* and *id/header*. Both situations mark the header of the column, but the use of these elements is really low (W3C, 2007b) and none of them is appropriate for the learning system. Table 3 shows the attributes used in the classifier system.

Table 3: Results of Studied Subset

TR	TD	Others
Bgcolor	bgcolor	TH tag
Class	Class	
	valing	
	width	

We have focused on the attributes of the tags to obtain information about the visualization of the table, but it is possible to make these effects inside the cell. Moreover, this content can offer important information that marks if the cell is a header. We work with three elements:

- Tables: Nested tables are not allowed.
- List: If a cell has a list inside, it is a data cell, not a header cell.
- Headers: The text, inside a section marked with the header HTML tag can possibly mark a table header. It is true only if the whole text inside the cell is also inside the section of the header.

We rejected the idea of comparing all the content of a cell to obtain differences between cells because the same visual effect can be done with different combinations of HTML elements. The lack of improvement does not justify the growth of complexity which would suppose the adding in the system. Therefore, we only use the data of the structure and visualization of the table.

Once discussed the attributes we are going to talk about the system of learning used. We use Naive

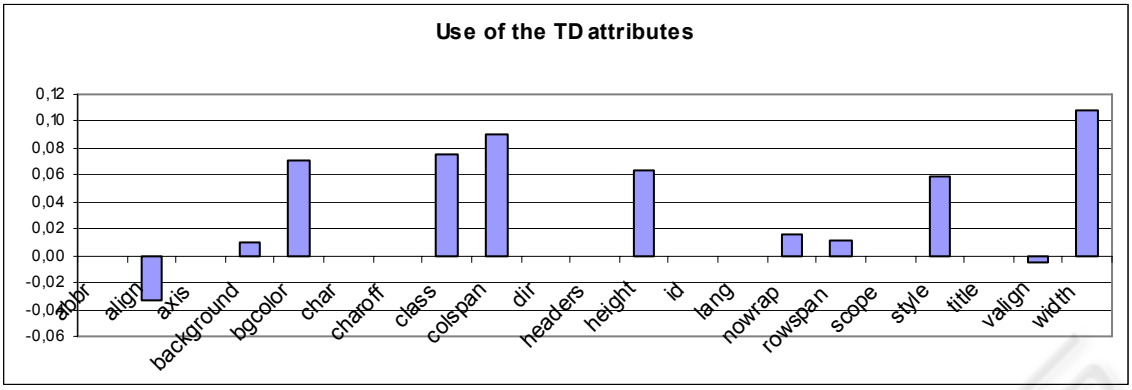

Figure 4: Distribution of the TD attributes.

Bayes, a method of supervised leaning, because it is very powerful but at the same time it has a temporal cost and the complexity is very low. Moreover, this method offers the advantage of not being affected by the unknown values of the elements (Witten and Frank, 2005). It is a situation which appears in our case, because is really difficult that all the attributes appear in the table's definition.

6 LEARNING PROCESS AND RESULTS

The learning process and the tests have been done using WEKA (Witten and Frank, 2005). This application allows us to test our set of tables and selection of attributes in an easy and quickly way. Furthermore, it is developed in Java and licensed under a General Public License (GPL) (Free Software Foundation, 2007), fact allows us to use the class, that implements the method Naive Bayes, in the ACTAW platform (Fernández et al., 2007) which is also GPL licensed. This repairing tool allows us to obtain all the information contained in a HTML document and modifies it in an easy way.

After analyzing the results offered by WEKA, we saw the great number of noise that the set of tables contain, caused by the tables belonging to the subset of tables without headers. This fact made us to decide to classify all the tables without information, like tables without header. It also complies with one of the premises of the WCAG, which is not offer information that can confuse the user, and this classification does not offer information of this kind.

It is very important to avoid offering incorrect information because it can disturb to the user. That means that the header only will be marked in those tables which contained them. On the contrary those tables which were not clearly marked visually, will be classified as if the would have not headers. With the selected elements we obtain excellent results as we can see in Table 4.

Table 4: Results of the studied Subset.

Class	Positive	Negative	% OK	% Wrong
Positive	37	3	92.5%	7.5%
Negative	1	10	90.91%	9.09%
		Total	92.16%	7.84%

This classification was made with a set of learning of 50 elements and a set of test composed by 51 elements. This subset does not offer noise to the system. And, in both of the subsets, the percentage of tables with header respect to tables without them is the same.

With this studied distribution, the number of tables classified as tables with header when they really are not, i.e. false positive, is very low. Only the 7.84% was false positive, and the correct classification is 92.15%, a very high result.

On the other hand, and to corroborate the good results, the system has been tested by using Cross Validation. We use 5 subsets for the validation, and we can see the results in Table 5.

Table 5: Results of Cross Validation.

Class	Positive	Negative	%Ok	%Wrong
Positive	68	4	94.45%	5.55%
Negative	9	20	68.97%	31.03%
		Total	87.13%	12.87%

The results are really good. The correct classification is 87.12%, it is really high and the false positive is so low enough. In Table 6 we can see a comparative of the results offered by the cross validation and the first test In both of them the

values of g-means (Kubat and Matwin, 1997), which is the most used measure to evaluate results in imbalanced datasets, and the area under the ROC curve are really good.

Table 6: Results of C. Validation and the studied Subset.

		C.Validation	Studied subset
G-means		0.8071	0.9170
ROC Area	Positive	0.869	0.941
	Negative	0.869	0.941

7 CONCLUSIONS

We have presented a system of learning that allows detecting the headers of a table. We can offer the relationship between the header and the cells under its scope. This is an important improvement because it means that the content of the table is not only a list of elements. The table recovers the bi-dimensional nature and allows the impaired user to obtain all the information inside the table.

The proposed solution has been tested with a heterogeneous set of real Web pages. The selection of this set was completely random and with it we can assure that the system does not offer good results for only a concrete situation. The system obtains excellent results and improves the results of the systems developed up to now. Hence, the study done in Section 4 regarding the previous work has allowed us to improve some lacks detected in this subject.

The next step to do will be to include this system in the ACTAW platform. With this application the feature can offer easily its help to all kind of people.

REFERENCES

World Wide Web Consortium (W3C) (1999a) Web Content Accessibility Guidelines 1.0, http://www.w3.org/TR/WCAG10/ (Retrieved on June 2007)

World Wide Web Consortium (W3C) (2007a) World Accessibility Initiative (WAI). http://www.w3.org/ WAI/ (Retrieved on June 2007)

Chen Shan, Hong Dan, Vicent Shen. An experimental study on Validation Problems with existing HTML Webpages. In *Proceeding of International Conference on Internet Computing*, pages 373-379, Las Vegas, EUA, 2005.

Quality Assurance Activity (W3C) (2007) The W3C Markup Validation Service. http://validator.w3.org/ (Retrieved on September 2007)

Benfeng Chen, Vicent Y. Shen. Transforming Web Pages to Become Standard-Compliant through Reverse Engineering. In *Proceeding of Workshop Web for All in International World Wide Web Conference*, pages 14-22, Edinburgh, UK, 2006.

World Wide Web Consortium (W3C) (2000) HTML Techniques for Web Content Accessibility Guidelines 1.0, http://www.w3.org/TR/WCAG10-HTML-ECHS/ (Retrieved on June 2007)

Creating Accessible Tables – Data Tables (2007) Web Accessibility in Mind (WebAIM) http://www.webaim.org/techniques/tables/data.php (Retrieved on September 2007)

Kubat, M, Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of the 14th International Conference on Machine Learning*, 1997.

Freedom Scientific (2007). JAWS for Windows. http://www.freedomscientific.com/fs_products/software_jaws.asp (Retrieved on June 2007)

Ai Squared (2007) Ai Squared site. http://www.aisquared.com/index.cfm (Retrieved on June 2007)

Yeliz Yesilada, Robert Stevens, Carole Goble and Shazad Hussein. Rendering Tables in Audio: The Interaction of Structure and Reading Styles. In *Proceeding ASSETS'04*, pages 16-23, Atlanta,Georgia, USA, 2004.

Juan Manuel Fernández, Vicenç Soler and Jordi Roig. Automatic Conversion Tool for Accessible Web. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies,* pages 459-462. Barcelona, Spain, 2007.

Enrico Pontelli and Tran Cao Son. Planning, Reasoning, and Agents for Non-visual Navigation of Tables and Frames. In *International ACM SIGCAPH Conference on Assistive Technologies* pages 73-80. Edinburgh, UK, 2002.

Robet Filepp, James Challenger and Daniela Rosu. Improving the accessibility of aurally rendered HTML tables. In *International ACM SIGCAPH Conference on Assistive Technologies* pages 9-16. Edinburgh, UK, 2002.

Bernhard Krüpl and Marcus Herzog. Visually Guided Bottom-Up Table Detection and Segmentation in Web Documents. In *Proceeding of International World Wide Web Conference*, pages 933-934, Edinburgh, UK, 2006.

K Kottapally, C. Ngo, R. Reddy, E. Pontelli, T.C.Son and D.Gillan. Towards the Creation of Accessibility Agents for Non-visual Navigation of the Web. In *ACM Conference Universal Usability*, pages 134-141, Vancouver, Canada, 2003.

World Wide Web Consortium (W3C) (2007b). www-html@w3.org Mail Archives. http://lists.w3.org/Archives/Public/www-html/2007May/0416.html (Retrieved on June 2007)

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learnign Tools and Techniques 2nd Edition*. Elsevier, San Francisco, USA 2005.

Free Software Foundation (2007). GNU General Public License Version 3. http://www.gnu.org/copyleft/gpl.html (Retrieved on September 2007)