IMPROVING HTML DATA TABLES NAVIGATION
A Method to obtain Information for Visually Impaired 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: Nowadays the broad use of the new technologies based on the Web gives facilities to people all over the world, but for impaired people. This leads us to the field of Web Accessibility and one of the biggest problems in it is the use of data tables on HTML documents. For disabled users, elements such as these and their natural bi-dimensional structure make more difficult to navigate than for the rest of the users. In this paper we present a solution to avoid those difficulties that disabled users find while navigating. The system we propose is based on the way a non-disabled person visualizes the table contents, but avoiding the processing of the images which is the natural procedure.

1 INTRODUCTION

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 standards, like World Content Accessibility Guidelines (WCAG), is very low and therefore the navigation on tables gets worse.

The correct design of a Web site is an unresolved matter for the current Web content developers. A correct use of the Word Wide Web Consortium 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. We will talk about the observance of the specific normative of the World Content Accessibility Guidelines (WCAG) which was made by the World Accessibility Initiative. The lack of use of the WAI's proposals and guidelines makes the Web to lose all its capacity to improve the labour and social integration of all kind of people.

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. So, the solution consists of a detection of header's rows and columns, that allows us offer information regarding the relationships between the headers and the cells of the table.

2 USE OF HTML TABLES

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. One of the most important errors is the use of HTML to offer visual information rather than 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 of that is the use of tables to distribute the contents of the Web site.

The layout use of tables 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. The only condition was that the Web site had to use the table element.

Manuel Fernández J., Soler V. and Roig J. (2008).
IMPROVING HTML DATA TABLES NAVIGATION - A Method to obtain Information for Visually Impaired People.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - HCI*, pages 397-400
DOI: 10.5220/0001714703970400

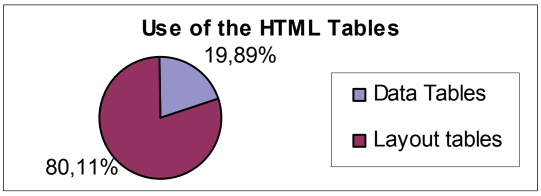

Figure 1: How the HTML tables are used.

3 HTML 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. The helping tools offer the information contained on a table in a linear list of elements. In this list, the headers usually appear just at the beginning and they are very difficult to remember. Our system offers the relationship between the headers and the content of the cell. We use HTML standard elements to indicate these relationships.

4 PREVIOUS PROPOSED SOLUTIONS

We can classify the current proposed solutions in three basic groups: the group where the Web browsers adapted to offer correct navigation in tables, new languages and the proposals which try to modify the content of the document to mark these relationships.

4.1 Adapted Browsers

This kind of solution has a limited field of use, because this sort of browsers can only solve one problem. And also, the user has to learn how to use it as well as other Web browsers. A very good example of that is the table browser called EVITA (Yesilada et al., 2004). 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

All the solutions of this kind 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. In this way we can see, for instance, the proposals of Pontelli and Filepp.

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.

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.

The first approximation 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

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.

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 of

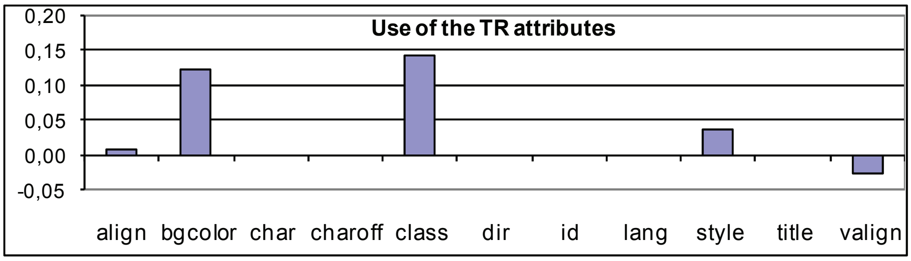

Figure 2: Distribution of the TR attributes.

tables belonging to both classes. 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 adds more difficulty to the learning process.

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 2 are the result of the difference between the average of the tables with and without headers.

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 (see Figure 3).

We rejected the idea of comparing all the content of a cell to obtain difference 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.

Once discussed the attributes we are going to talk about the system of learning used. We use Naive 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. Also, this method offers the advantage of not being affected by the unknown values of the elements.

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 fact that allows us to use the class, that implements the method Naive Bayes, in the ACTAW platform (Fernández et al., 2007). This repairing tool allows us to obtain all the information contained in a HTML document and modifies it in an easy way.

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. We can see the results on Table 1.

Table 1: 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 and without noise. 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 2.

Table 2: 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 correct classification is 87.12%, it is really high and the false positive is so low enough. In Table 3 we can see a comparative of the results

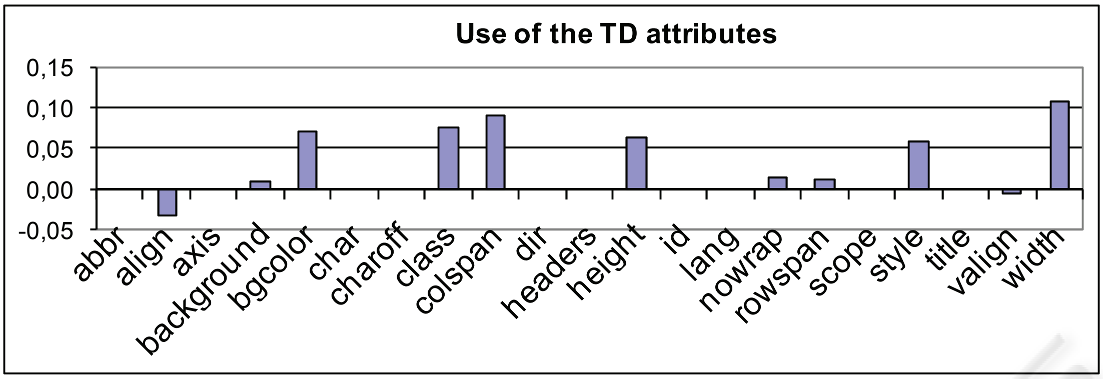

Figure 3: Distribution of the TD attributes.

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 3: 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 us to detect the headers of a table. This tool is independent of Web browsers and compliant with the W3C standards. With the presented method, 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 system developed up to now.

REFERENCES

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.

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.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learnign Tools and Techniques* 2^{nd} *Edition*.Elsevier, San Francisco, USA 2005. ISBN: 0-12-088407-0