VISUALIZATION OF DOCUMENT ELEMENTS CLASSIFICATION

Michael Kunc
Doctoral Degree Programme (2), FIT BUT
E-mail: kunc@fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

This paper deals with visualization of segmented web-pages. Segmentation is done by visual properties of the document areas. Classification and visualization was done by open-source data mining tool *Weka* and *J48* algorithm. Results show properties that are necessary to decide to which category the area belongs.

1 INTRODUCTION

On the Web there is today huge amount of electronic documents. Most of them contain embedded textual information. When computers are interacting with these documets, they use text information only. But for people, there are different aspect they are looking for. For example, visual information of the document should be as important as textual information, and in some cases, even more important. This paper describes visual properties of (X)HTML documents and their applications.

2 INPUT DOCUMENT

Input documents for detection of visual properties are web-pages of czech and foreign news-servers. These pages contain vast amount of data so they are good example for visual segmentation of a web-page. All principles shown in this paper are able to work also on another types of documents, for example PDF (Portable Document Format), PostScript, DOC, (Microsof Word), RTF (Rich Text Format), ODF (Open Document Format) and others, that contain textual information. There are another problems to solve in raster image, so this paper doesn't deal with these image documents.

3 PREPROCESSING

The first step of preprocessing is reconstruction of final document representation. There is used rendering machine constructed on FIT BUT. The rendering machine is software that interprets input document (HTML) and shows the final form of the document. Than, the output from the rendering machine is used as an input for visual segmentation algorithm, that spreads out the document into hierarchical structure of visual areas.

4 PROPERTIES OF VISUAL AREAS

Investigated properties of visual areas detected in the segmentation algorithm are as follows: Font size (average font size in the document is 100%), font weight (normal or bold), font style (normal or italic), number of areas that are above, below, on the left and on the right with regard to the area, number of characters, spaces, numbers, lower case and capital letters in text areas, average luminosity of text and backgroud and average contrast. All these parametres were under examination of the experiment.

5 EXPERIMENTS

In 16 documents from czech and foreign news servers were detected 5778 visual areas, that were manually assigned to following pre-defined classes:

- *h1* – main heading of the document
- *h2* – heading of an article
- *h3* – heading of short article
- *aktualita* – newsreel or short article
- *menu* – area of navigation
- *date* – date of publication
- *none* – other (unannotated) areas

Figure 1 shows counts of each categories detected and annotated in the documents.

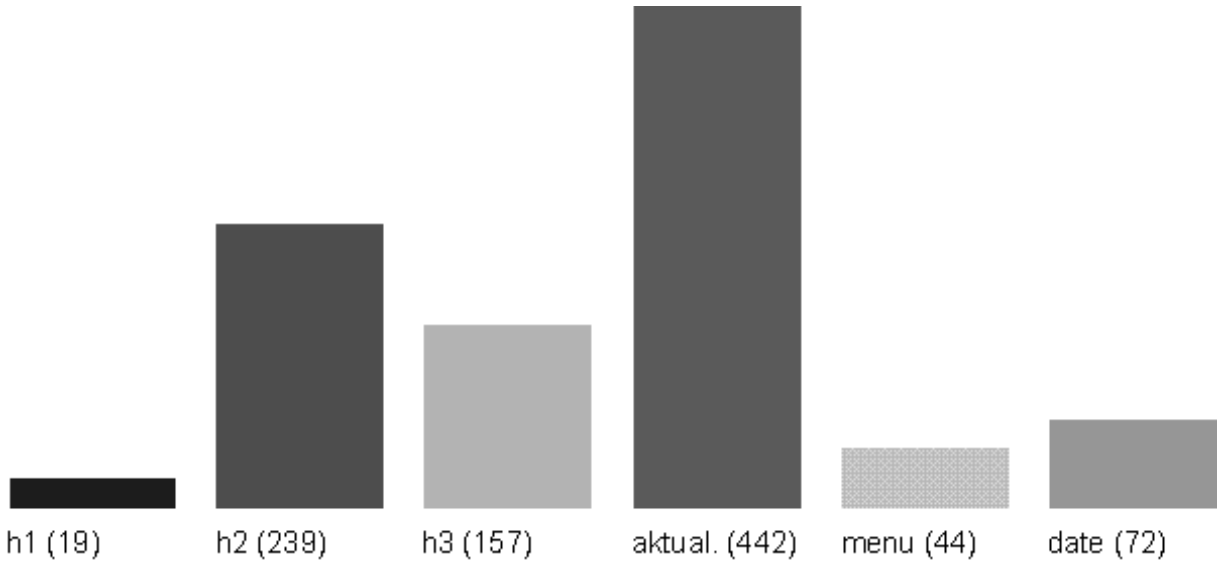

Figure 1: Annotated visual areas

5.1 MACHINE LEARNING SOFTWARE

Classification was done by open-source machine learning software *Weka*. This software contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

The format of input data for *Weka* is following:

```
@RELATION page
@ATTRIBUTE class {h1,h2,h3,aktualita,menu,date,none}
@ATTRIBUTE fontsize NUMERIC
@ATTRIBUTE weight {bold,normal}
@ATTRIBUTE style {italic,normal}
@ATTRIBUTE aabove NUMERIC
@ATTRIBUTE abelow NUMERIC
@ATTRIBUTE aleft NUMERIC
@ATTRIBUTE aright NUMERIC
@ATTRIBUTE tlength NUMERIC
@ATTRIBUTE tdigits NUMERIC
@ATTRIBUTE tlower NUMERIC
@ATTRIBUTE tupper NUMERIC
@ATTRIBUTE tspaces NUMERIC
@ATTRIBUTE textbtns NUMERIC
@ATTRIBUTE bgbtns NUMERIC
@ATTRIBUTE contrast NUMERIC
@data
h1, 156, bold, normal, 0, 2, 0, 0, 15, 0, 10, 2, 2, 0.07793828458751126,
0.8506198007899897, 7.039486293674317
```

5.2 CLASSIFICATION ALGORITHM

All classification algorithms were tested on different inputs and *J48* was chosen as the best algorithm for classification of detected areas. The algorithm is derived from *C4.5* algorithm, that was developed by Ross Quinlan [1]. Classification is done by decision trees, where are used concepts of information entropy. Each attribute can be used to make a decision and to split the data into some subsets.

Figure 2 shows visualization of J48 algorithm results. In ovals, there are examined attributes. These attributes generates the decision tree. Values of the attributes splits each attribute to two sub-trees. Finally, in the boxes, there are pre-defined classes.

5.3 RESULTS

As we can see, there are many interesting results. The main heading of the document (class *h1*) is the easiest assignable class of the document. Only one condition is required – the size of font greater than 135% with regard to average text size, that is 100%. As a heading *h2* are assigned areas, where the size of font is greater than 130%, less than or equal to 135% and the number of areas that are on the right of these area is equal to 0. The class *date* was recognized as an area, where the size of font is less than or equal to 130%, luminosity of text is greater than 0.133209,

there are no other areas on the right, number of digits is greater than 3, font is bold and number of all characters in the area is less than or equal to 39.

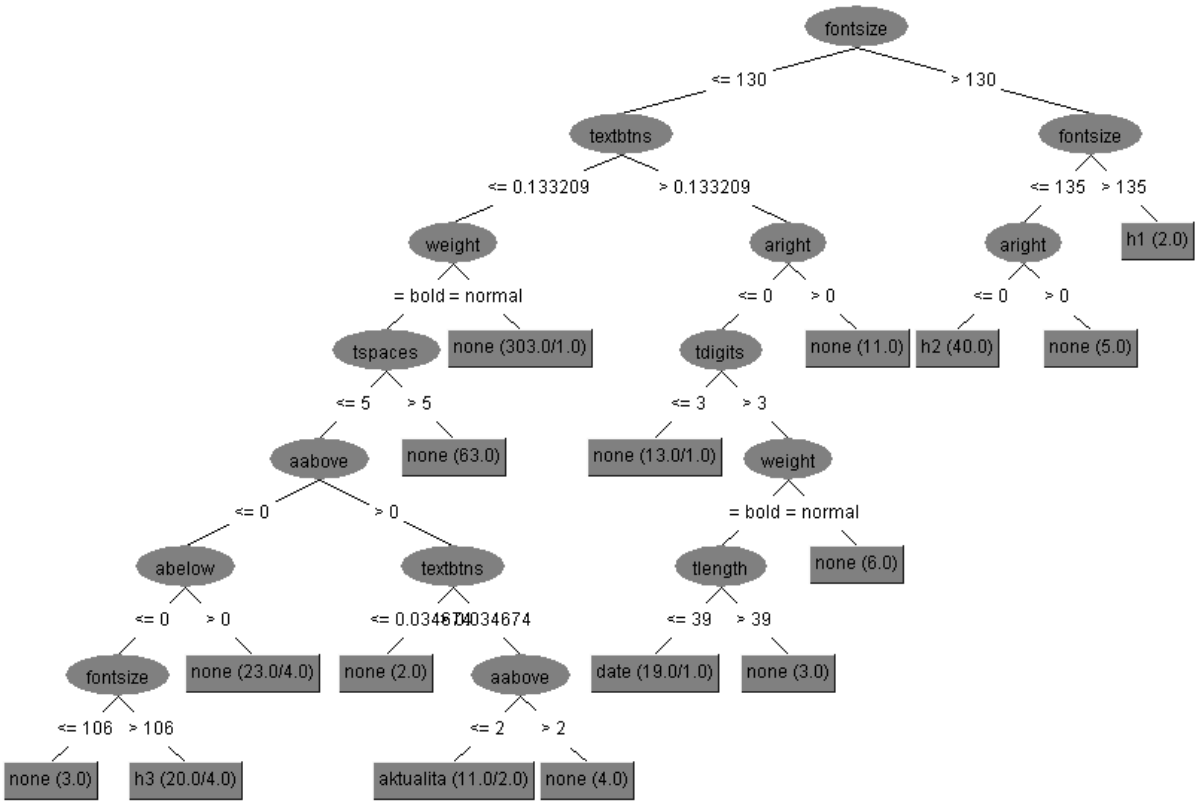

Figure 2: Visualization of annotated areas

6 CONCLUSION

This paper shows results of *J48* classification algorithm on a group of electronic documents. Segmented web-pages are input of the algorithm and graphic form of decision tree is the result. There are shown properties that are important for assignment of each area to some class. We can summarize that the most important properties are font size, text luminosity and text weight.

REFERENCES

[1] Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[2] Burget, R.: Automatic Document Structure Detection for Data Integration, In: Lecture Notes in Computer Science, 2007, nr. 4439, p. 391-397, ISSN 0302-9743

[3] Kunc,M., Burget, R.: Klasifikace prvků dokumentu na základě vizuálních rysů, In: Znalosti 2008, Bratislava, STU, 2008, p. 347-350, ISBN 978-80-227-2827-0