TWO-PHASE CATEGORIZATION OF WEB DOCUMENTS

Vladimír Bartík and Radek Burget
Faculty of Information Technology, Brno University of Technology, Božetěchova 2, Brno, Czech Republic

Keywords: Web Page Categorization, Visual Block Classification, Term Weighting, TF-IDF, Page Segmentation.

Abstract: The number of pages on the World Wide Web is permanently growing and there is a need to process pages efficiently and obtain some useful knowledge from them. Web page categorization is a very important issue in this area. The method proposed here takes both visual and textual information into consideration. It consists of two phases. In the first phase, web page areas obtained by segmentation are classified based on their visual properties, and in the second phase, pages are classified, based on information from the first phase and textual information. Several experiments with web pages taken from news web sites are presented in the final part of the paper.

1 INTRODUCTION

Due to the fact that the amount of data on the web is still increasing, it is necessary to find some interesting knowledge on the web by means of web mining techniques. Web page categorization (or classification) is a process of assigning some of pre-defined class labels to a web document.

There are two basic types of information contained on each web page – textual and visual information. In the contemporary methods, only the textual information is usually used to classify the page. In our categorization method, both visual and textual information are reflected in a document representation. Therefore, the process of document categorization is divided into two phases.

In the first phase, the web page is rendered and information about visual blocks of a page is obtained via a page segmentation algorithm. The information about visual blocks is stored and classification algorithm is applied.

For each visual block, also text contained in the block is stored to perform text-based categorization of pages.

In the method proposed here, the second phase includes creation of modified weight vectors. The modifications are based on an assumption that some parts (referred to as content blocks) of a page are more important for page content representation than others (non-content blocks). In these vectors, both textual and visual information are reflected. It is then used for the whole web page categorization.

In the following chapters, some related work in this area is introduced. Then, the whole process of two-phase categorization of web documents is described. In the final part, several experiments with both phases and the whole process are described.

2 RELATED WORK

There are two basic approaches to segmentation of web pages. The first one is based on a Document Object Model (DOM) representation of page content, while the second one is based on visual properties of web page blocks.

The first segmentation method proposed in (Lin and Ho, 2002) divides a page according to HTML tags, where the entropy is computed for each content block. This method is limited to pages using the *table* HTML tag. The segmentation method based on a DOM tree structure analysis, is proposed in (Chen et al., 2003). Here, in the first step, the DOM tree is used to find top-level content blocks, which are divided in the smaller ones in the next steps by finding the separators between the blocks.

The use of both methods is limited, because it does not reflect visual properties of web page blocks. The VIPS algorithm (Cai et al., 2003) uses page rendering and then, visual properties of rendered page are used to find visual blocks.

The method proposed in (Xiang et al., 2006) combines both approaches. It takes both visual

Bartik V. and Burget R..
TWO-PHASE CATEGORIZATION OF WEB DOCUMENTS.
DOI: 10.5220/0003096204580462
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 458-462
ISBN: 978-989-8425-28-7

properties and the most reliable properties from the DOM tree into consideration for segmentation.

Regarding page categorization, most of recent methods for web document classification is text-based. For basic text-based method, a set of text terms is extracted from the document and preprocessed. Then, each term is represented by a weight, while the most used are TF (term frequency) and TF-IDF (Salton and Buckley, 1999) weights (TF multiplied by inverse document frequency).

Several methods, which try to enrich TF-IDF weighting with some web page properties, have been proposed. In (Kwon and Lee, 2003), the information from HTML tags is reflected in the term weights. HTML tags are divided according to importance.

In some web page classification methods, graph representation instead of weight vectors is used. This allows to keep also the structural information about the document (Schenker et al., 2004).

3 2-PHASE CLASSIFICATION

As mentioned above, the process of web page classification consists of two separate classification phases. In Phase I, web page is rendered and visual text blocks and their visual properties are obtained by means of segmentation. The blocks are then classified according to their visual properties.

In Phase II, the text content is extracted from visual blocks and weight of each term is computed with respect to its frequency and importance of visual block, in which it is contained. The resultant weight vectors are input of the final classification.

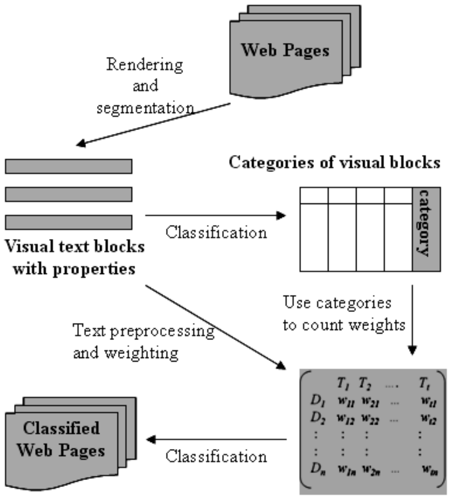
Web Pages
Rendering and segmentation
Categories of visual blocks
category
Classification
Visual text blocks with properties
Use categories to count weights
Text preprocessing and weighting
T1 T2 .... Tt
D1 w11 w21 ... wt1
D2 w12 w22 ... wt2
Dn w1n w2n ... wtn
Classified Web Pages
Classification

Figure 1: Outline of the method.

3.1 Phase I – Visual Block Classification

In the beginning of the process, to obtain a suitable representation of a document, it is necessary to make page rendering. This includes interpretation of the document's HTML code and creation of document representation in a form of a set of boxes. A box in this set represents a unit of the document content placed on some place of the page. This is a result of page rendering. Because these visual boxes can be nested, a tree of boxes is obtained.

This model contains all the information necessary for classification. This includes the content of the document (mainly text content) and visual style of each part of a page.

After rendering, page segmentation is performed. Its task is to obtain visual blocks contained in the document and their organization. A visual block represents a rectangular region in the page that is visually separated from other parts of the page.

Our segmentation algorithm (Burget, 2007), works in a bottom-up manner. Each block is represented by the coordinates of top-left and bottom-right point of a box. The task is to find the blocks that are visually separated from others by one of the following means:

- It contains a text or an image (represented by leaf block of a tree).
- It has some background color that is different from the background color of its parent block.
- It is separated by a visible frame.

The resultant tree is created recursively: if a block is visually separated, a new corresponding block is added to a tree. Then, the same is applied to the child boxes.

In the next step, the blocks are clustered – we find all visual blocks that are placed in the adjoining cells and are not visually separated from each other by the means mentioned above. Such areas are joined into a single area. This step corresponds to the detection of content blocks (for example text paragraphs) that consist of several boxes.

Lastly, we look for areas that are not separated but they are delimited with the visually separated areas around. These blocks are also clustered.

As we know, various visual blocks of a page have different importance for representation of the whole document. Therefore, we need to classify each block into one of categories from Table 1 according to visual properties shown in Table 2 obtained by the rendering machine.

If we have necessary information about visual properties and position of a block on a page, we can use arbitrary classification method to assign some of these categories for each web page block. As already mentioned, the visual properties are obtained with a web page rendering method.

Table 1: Categories of web page blocks.

Category	Description
Heading	The main heading and subheadings of a page (with higher font size).
Main text	The main text content.
Link area	Links to other related web pages – some of them may be irrelevant.
Navigation	Links to other sections of the web site – typically at the top of a page.
Date/Authors	Information about date and authors of a page – typically small font size.
Other contents	Other unimportant parts, such as advertisement, caption etc.

Table 2: Visual properties used to classify visual blocks.

Property	Description
Fontsize	Size of text in the block, expressed as percentage of average fontsize in the whole document.
Weight, Style	Dominating text weight / style in the block.
Above, Below, Right, Left	Number of visual blocks located above/below from the actual block or right/left from it.
Length, Digits	Text/digit characters in a block.
Lower, Upper, Spaces	Count of lower case / upper case / space characters that appear in the text inside the visual block.
Textbtns, Bgbtns	Average luminosity of the text / background of the visual block
Contrast	Contrast of the text – difference between luminosity of text and luminosity of background.

3.2 Phase II – Classification of Whole Web Pages

The knowledge about classes of visual blocks obtained in the first phase can be used in the second phase to improve the representation of text content. This is ensured by modifications of TF-IDF weights.

At first, two standard pre-processing procedures are applied on the text obtained from a web page. Stop words removal is performed to reduce the number of words in vectors. After that, stemming is executed to reduce the words into their stems.

Then, according to the knowledge about visual block categories, there are several possibilities to modify TF-IDF weights.

The basic idea is to set different weights for different page block categories by multiplying the weight by an importance coefficient, which is set for each category of visual block. Moreover, some non-content blocks (e.g. date/authors or navigation) can be omitted from representation.

Let us denote an input set of web documents as *D = {d₁, ... dₙ}* and a set of terms (words) *T = {t₁, ... tₘ}*, which are appearing in documents from *D*.

It is possible to divide each document into several visual blocks, each of which can be classified with some of pre-defined label. Let us denote a vector of all class labels as $C = (c_1, \ldots, c_k)$. Each class is evaluated by a coefficient according to its importance for representation of page contents. We can represent it as a vector of coefficients $V = (v_1, \ldots, v_k)$, where v_j is the coefficient of a class label assigned to visual block c_j.

Then, a modified document frequency of a term $t \in T$ in a web document $d \in D$ is defined as:

$$MTF(t,d) = \sum_{i=1}^{k} v_i * F(t,d,c_i), \qquad (1)$$

where $F(t, d, c_i)$ is a frequency of term t in all blocks with class label c_i in a document d. The resultant term weight is obtained as a summarization of all weights for component visual blocks, in which the term is present. The modified inverse document frequency (IDF) is computed as:

$$MIDF(t) = 1 + \log(\frac{n}{k_V}), \qquad (2)$$

where t is a term from the set of terms T, n is the count of documents, k_V is the count of documents, in which content blocks at least once contain the term t.

The resultant modified TF-IDF weight is obtained as a multiplication of MTF and MIDF.

If the two-phase classification is used, it is important to decide, if the blocks of different categories are content or non-content.

Another task is to set the importance coefficients for each block class. According to experiments, we have found a suitable setting described later.

4 RESULTS OF EXPERIMENTS

Some experiments showing the classification accuracy are presented here. We have used the WEKA tool (Holmes et al., 1994) for experiments.

A dataset with web pages from 5 news web sites (CNN.com, Reuters.com, nytimes.com, boston.com and usatoday.com) has been used. This dataset contains almost 500 web pages, which have been manually annotated. Each of the web pages belongs to one of six possible categories: politics, business, sport, art, health and science.

For Phase I, the classification accuracy depends on a condition, if all pages from the dataset are from one web site or from various web sites. Best results are provided by one of lazy classification methods (k-nearest neighbours) and by two decision tree based methods (J48, based on C4.5, and RandomTree method). The results of classification accuracy for this experiment are shown in Table 3.

The table shows that classification achieves good accuracy above 90%; it is slightly lower for the whole dataset of pages from 5 different sites. If we perform classification only for one web site only (CNN.com), the accuracy is above 95%.

Table 3: Results of visual blocks classification.

Method	All web sites	One site only
J48	92.4%	97.5%
k-NN	90.5%	94.9%
R_Tree	91.1%	96.0%

From these results we can see that the visual blocks classification can be a good foundation for successful two-phase classification.

For the second phase, best results are provided by a Bayes Net classifier, a tree-based classifier (Functional Trees) and SMO method (Sequential Minimal Optimization). We will take the results of J48 algorithm in the Phase I as input for Phase II.

The setting of coefficients for individual content visual block categories is following: for links, the value is 1; for main text: 2; for heading: 5. Detailed experiments with different setting of coefficients and detailed description of possibilities to modify weights are presented in (Bartik, 2010).

The classification accuracy is compared with a classical text-based classification (extracted text weighted by TF-IDF). Again, results on dataset of pages from one web site and from all five web sites are compared.

The results in Table 4 lead to a conclusion that using modified weights causes better web page classification accuracy, even if all the visual blocks are not classified correctly in the first phase.

Table 4: Results of the whole two-phase categorization.

Method	Text Based	All sites (2-phase)	One site (2-phase)
Bayes Net	90.6%	93.3%	94.7%
FT	88.6%	91.8%	92.9%
SMO	88.0%	91.9%	93.9%

4 CONCLUSIONS

We have presented a method of two-phase classification of web pages. The basic idea is that the text, which is extracted, is represented by modified term weights. They are modified according to importance of a visual block, in which the term is present. The information about category of visual blocks is obtained by classification in the first phase.

The second phase makes classification of whole web pages with use of vectors of modified weights.

The experiments with data from news sites showed that the accuracy of two-phase classification exceeds 90%.

In the future research, it is possible to try using results of visual block classification and modified term weights in some other web mining problems, for example finding similar documents within some dataset or in some information retrieval tasks.

ACKNOWLEDGEMENTS

This research was supported by the Research Plan No. MSM0021630528 – "Security-Oriented Research in Information Technology" and by the BUT FIT grant No. FIT-10-S-2.

REFERENCES

Lin, S. H., Ho, J. M., 2002. Discovering Informative Content Blocks from Web Documents. In *SIGKDD 2002, 8th Conference on Knowledge Discovery and Data Mining,* ACM.

Chen, Y., Ma, W. Y., Zhang, H. J., 2003. Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices. In *WWW 2003, Twelfth International World Wide Web Conference,* ACM.

Cai, D., Yu, S., Wen, J. R., Ma, W. Y., 2003. VIPS: a Vision-based Page Segmentation Algorithm. *Microsoft Research.*

Xiang, P., Yang, X., Shi, Y., 2006. Effective Page Segmentation Combining Pattern Analysis and Visual Separators for Browsing on Small Screens. In *International Conference on Web Intelligence*, ACM.

Salton, G., Buckley, C., 1999. Term Weighting Approaches in Automatic Text Retrieval. In *Information Processing and Management, Vol. 24*, Elsevier.

Kwon, O. W., Lee, J. H., 2003. Text Categorization Based on K-nearest Neighbor Approach for Web Site Classification. In *Information Processing and Management, Vol. 39*, Elsevier.

Schenker, A., Last, M., Burke, H., Kandel, A., 2004. Classification of Web Documents Using Graph Matching. In *International Journal of Pattern Recognition and Artificial Intelligence, Vol. 18*, World Scientific.

Burget, R., 2007. Layout Based Information Extraction from HTML Documents. In *ICDAR 2007, Ninth International Conference on Document Analysis and Recognition*, IEEE.

Holmes, G., Donkin, A., Witten, I.H., 1994. WEKA: A Machine Learning Workbench. In *Second Australia and New Zealand Conference on Intelligent Information Systems.*

Bartik, V., 2010: Text-Based Web Page Classification with Use of Visual Information. In *OSINT-WM 2010, International Symposium on Open Source Intelligence & Web Mining*, IEEE (accepted).