Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 9 (2014) 31 – 36

AASRI Procedia

www.elsevier.com/locate/procedia

2014 AASRI Conference on Circuits and Signal Processing (CSP 2014)

A Novel Imaging Approach of Web Documents Based on Semantic Inclusion of Textual and Non – Textual Information

Martina Zachariasova*, Patrik Kamencay, Robert Hudec, Miroslav Benco, Slavomir Matuska

Department of Telecommunications and Multimedia, University of Zilina, Zilina, Slovakia

Abstract

This paper deals with research in the area of a novel imaging approach of web documents based on semantic inclusion of textual and non-textual informations. The main idea was to create a robust method for relevant display results into search engine based on search by keywords or images. Thus, we proposed method called Semantic Inclusion of Images and Textual (SIIT) segments. The output SIIT method is short web document. It contains image and textual segments, which are semantic linked. Creation of short web document to possible three steps was divided. Firstly, the all images and textual segments from main content web document were extracted. Secondly, extraction images were analyzed in order to obtain of semantic description objects into image. Finally, linked images and textual segments using linguistic analysis.

Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute

Keywords: digital images; image classification; support vector machine; descriptors

1. Introduction

In the last year's, there are several mechanisms for semantic inclusion of web objects using text analysis. Sh. Behnami [1] described design of Filimage system, which is intended for the images automatic extraction

* Corresponding author. Tel.: +421 41 513 2239.
E-mail address: martina.zachariasova@fel.uniza.sk.

Peer-review under responsibility of Scientific Committee of American Applied Science Research Institute
doi:10.1016/j.aasri.2014.09.007

and their textual comments available on the web. Mulendra Parag Joshi and Sam Liu in [2] described a technique for extracting text and images from a web document using the Document Object Model analysis and natural language processing. In the first phase, the article body from a Web document is extracted. Then, the semantic similarity based on NLP (Natural Language Processing) is used to find relevant images corresponding to the article. J. Pasternack and D. Roth [3] introduce maximum subsequence segmentation, method of global optimization over token-level local classifiers, and applied it to the domain of news websites. L. P. Florence [4] described an image and text mining tool named TNT. This tool is based on Contextual Exploration and work on different points of view. This tool offers a reorganization of the text guided by the images and annotated segments that are associated.

2. Proposed Method

Recently, the university research in the area of processing web pages for creation short web documents focuses on the analysis of text segments around images. Research in image processing is currently at a high level, would have been wrong not to use this fact to creation short web document. Our proposed method is based on existing systems, which are described in previous section. In this system we put the emphasis on the minimization or elimination of potential deficiencies (see Tab. 1). The proposed method allows an automatic processing of various components of web pages. The system architecture called Semantic Inclusion of Image using Textual segments is shown in Fig. 1. Automatic extraction takes place along two axes, one oriented towards the text and the other oriented image.

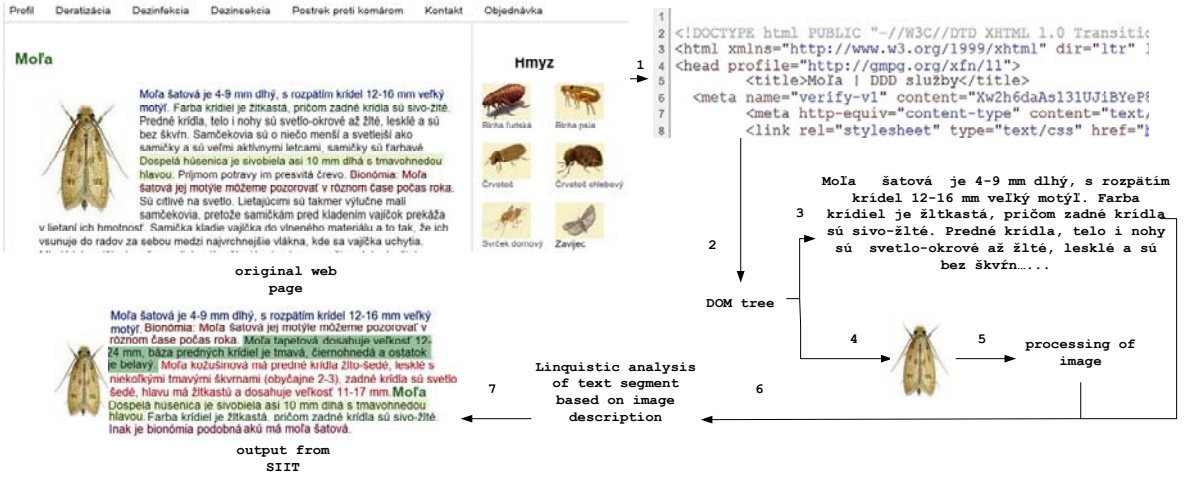

Fig.1. Sample web page processing of our proposed method

The important steps for automatic extraction:

- Loading web page
- Reading source code of web page and creating Document Object Model tree
- Identification and extracting of textual segments from the main content of web page
- Extracting images from around textual segments
- Processing of images using extraction features and its classification
- The textual segments are coming from semantic analysis according to semantic description
- A connection between the two modules allows automatic matching and associating operation

Table 1. Advantages and disadvantages of methods to obtain coherent semantic blocks

		Parameter				
		Description of image obtained image analysis	Description of image obtained text analysis	Unwanted objects removed form web page	Creation abstract	Ontology
Methods	Filimage system	-	+	-	+	-
	Method based on browsed context (TNT tool)	-	+	-	+	+
	Method SIIT based on ontology	+	+	+	+	+

3. Experimental Results

In this section, the proposed method for creating a short web document is presented. The experiment has been done on Caltech 101 image database and 100 offline web documents. The proposed approach using programming languages JAVA and MATLAB was implemented.

The process of creation short web document:

- Automatic extraction of web objects
- Analysis of images
- Semantic inclusion of web objects

3.1. Automatic extraction of the main web objects

Identify the main block to the desired web document content using the Document Object Model [5] is now possible in several ways. As principal, we can mention:

- Known name identifier element: (for example)
- Counter signs
- Counter sentences

In practice we identifier element and simple character counter not work. The amount of Web documents in question contained multiple blocks with more than 500 characters. Dynamic sites is a long text strings used to create the drop-down menu or similar interactive blocks, from one point in the chain is presented as part of using JavaScript as moving text or list.

In this case is preferable to use a natural language description to distinguish the relevant text from the auxiliary strings. Instead of a single counter of characters or a word that has the same drawbacks, we chose counter full sentences. In natural language, we assume the characteristics of structure of sentences - each sentence is followed by a terminating punctuation character of the preceding sentence (period, question mark, exclamation mark) and a space.

Observation, we found that the vast majority of normal cells do not use the terminating question mark or exclamation sentences to the extent that it affected the identification of the main article. Therefore we use as a rule separating the new sentence string "dot + loophole."

After reading a text document, test the contents of each element of the number of sentences that rule. The relevant content will be such a text block that contains at least 5 sentences. Elements that hide the appropriate text with relevant content are also tested for the presence of images. We assume that the images associated with text can be found in these blocks. If confirmed their presence, their extraction algorithm ensures together with the text. To identify the main text in the form of a flowchart is shown in Fig. 2.

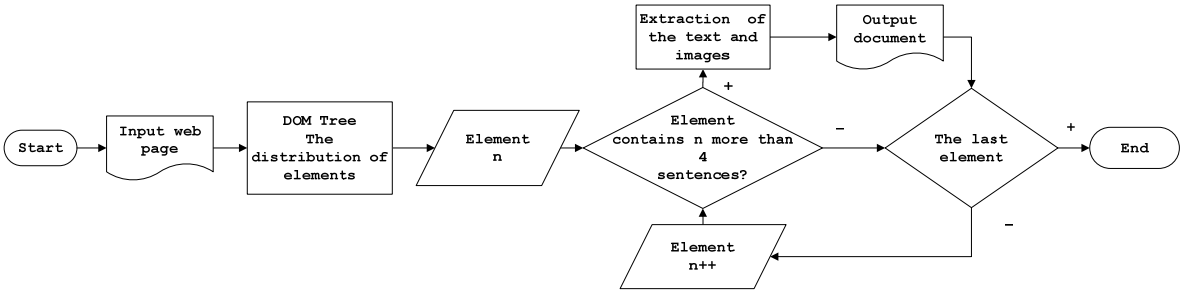

Fig.2. Workflow for web article extraction pipeline

For comparison, the counter of characters was used. This part of the algorithm was implemented in the JAVA programming language. First, all spaces from the input document using counter were removed. Subsequently, all the characters of text in application were counted.

The proposed algorithm on selected sample of documents has been tested. The overall efficiency for text extraction was 88.99%. Extraction of the images depends on the identification of the main block of the document therefore follows the success of the state of text extraction.

3.2. Analysis of images

In this section the extracted images were processed. The main goal of this step was obtained automatic semantic description of static images. This description is important for matching objects of web document. The overall procedure of automatic semantic description of image is shown in Fig. 1. The basic steps of automatic semantic description can be divided into the following steps:

- Loading extracted images
- Loading SVM model
- Division of the image into segments
- Feature extraction using SIFT method
- Find the closest conformity between extracted descriptors and SVM model
- The output of the analysis of images is a semantic description of objects

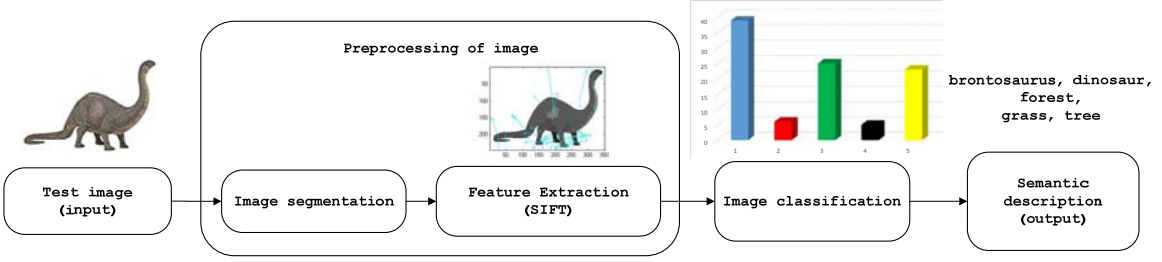

Fig.3. General schematic diagram for analysis of image

Firstly, we loaded of extracted images and SVM model. In the next phase, we defined number of segments for segmentation images using K-means [6]. The extracted images from web pages contains only three segments (e.i. water, tree, sky, ...). For each input image we set three segments using K-Means. Next, we used

filter for holes filling in the binary segment into image. A hole is a set of background pixels that cannot be reached by filling in the background from the edge of the image.

In the third phase, features from segments using SIFT descriptors was extracted. These descriptors by threshold 5 or 8 pixels have been modified. This threshold was selected on the basis of previous publications [7]. In the test image, the vector length than eight pixels is not obtained. For this reason, the threshold greater than eight will not be used.

In the following phase, we classify extraction of features using SVM models. Models contain 9 classes. Every extraction feature of segment was evaluated and description was allocated by using SVM model. In the last phase of our experimental part, description of image was assigned (shown in Fig. 3).

3.3. Semantic inclusion of web objects

Finally, an algorithm for semantic inclusion of web objects was proposed. Semantic inclusion is based on description of extracted images. The textual segments of the web page are searched and extracted. The sentences, which contents identical words description are extracted into short web document. Then, the image with the sentences which have same index as the description were associated. The general block diagram for semantic inclusion is shown in Fig. 4.

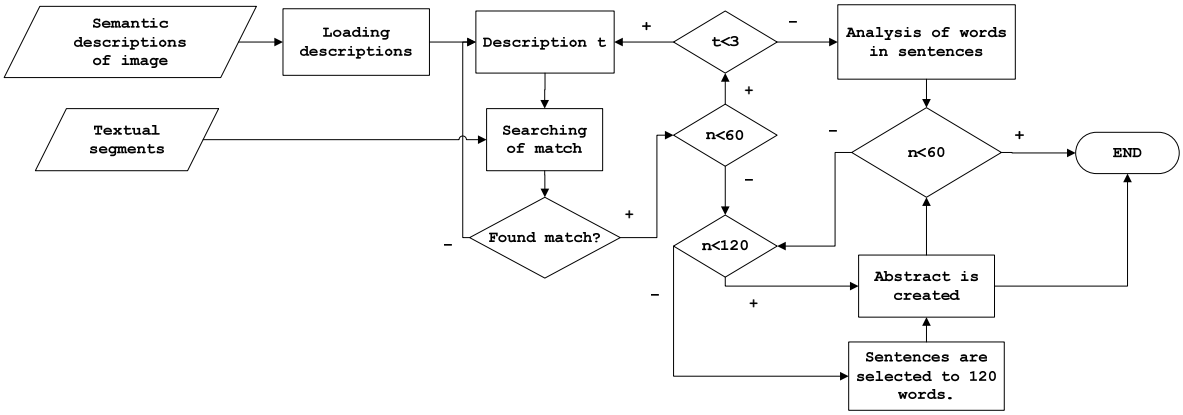

Fig.4. Block diagram for semantic inclusion of textual and non-textual objects

Design of algorithm for semantic inclusion of web objects is based on three steps. Firstly, descriptions of image and textual segments were loaded. Secondly, linguistics analysis of textual segments was made. Finally, all experimental results are displayed. The main part of semantic inclusion is linguistics analysis of textual segments. It is consists of three parts (lexical analysis, morphological analysis, syntactic analysis). The all words whose length is less than three phonemes using lexical analysis are removed. The complexity of the Slovak language is main reason for setting this value. This is mainly due to fact that in the Slovak language different prepositions are used. Furthermore, in lexical analysis the frequency of sentences with the identical words are counted. These words are identical to the description. Finally, the total number of words in sentences that were extracted as relevant text was counted. Morphological analysis of text segments is based on the creation of word-formation base of words and phrases analysis. The average number of words and length of sentences was counted. In the last step, duplicate sentences in syntactic analysis were reduced. The experiments have been done on 100 Web documents. The best results for threshold 8 pixels were achieved (79.59%).

4. Conclusion

Growing amount of web documents leads to necessity of effective system management for automatic annotation the web pages based on semantic inclusion image and text. The best results for 8 pixels are achieved. The accuracy for creating short web documents was 79.59 %. The proposed method is based on SIIT (see Fig. 1). The input is original web page and the output is modified web page. As can be seen, only one abstract was created. Namely image, which was into the main of block text. Other images for extracting information from a Web document were unwanted (contain links to other websites). The text from the main body of the web page was extracted and the banners or navigation menu were removed. Only relevant text for an abstract was used. Since the original image - text abstract contained text segment less than 60 words, morphological analysis was performed. For the better visualization of text segments a different color marking was used (see Fig. 1).

In future work, a semantic map will be implemented. This map should help with more accurately description of non-textual information and improve accuracy of the animal's description in the image. Based on the recognition efficiency of image objects, it could be assumed, that the presented classification is sufficient for evidence of animal presence in the vicinity of a road, and for the corridors build-up.

Acknowledgements

The work presented in the paper has been supported by the Slovak Science project Grant Agency, Project No.1/0705/13 "Image elements classification for semantic image description" and EUREKA project no. E! 6752 – DETECTGAME: R&D for Integrated Artificial Intelligent System for Detecting the Wildlife Migration.

References

[1] Behnami Sh. Filimage System: Webs Images and Texts Automatic Extraction. Word Scientific and Engineering Academy and Society (WSEAS). Izmir. Turkey. 2004.
[2] Parag M. J., Sam L. Web document text and images extraction using DOM analysis and natural language processing. In Proceedings of the 9th ACM symposium on Document engineering DocEng 09 (2009).
[3] Pasternack J., Roth D. Extracting article text from the web with maximum subsequence segmentation. In Proceedings of the 9th ACM symposium on Document engineering DocEng 09 (2009). 971-980.
[4] Florence L. P. Image and Text Mining Based on Contextual Exploration from Multiple Points of View. Twenty-Fourth International FLAIRS Conference 2011. Palm Beach, Florida, 18-20 May.
[5] Joshi A.K., Thomas N., Shetty S.; Thomas, N. VEDD- a visual wrapper for extraction of data using DOM tree. Communication, Information & Computing Technology (ICCICT). 2012 International Conference on, vol., no., pp.1-6, 19-20 Oct. 2012. doi: 10.1109/ICCICT.2012.6398114
[6] Hung Ch. Ch., Sun M., Ant colony optimization for the K-means algorithm in image segmentation, Published in: Proceeding ACM SE '10 Proceedings of the 48th Annual Southeast Regional Conference. Article No. 48, ACM New York, NY, USA ©2010. ISBN: 978-1-4503-0064-3.
[7] Zachariasova M., Kamencay P., Hudec R., Benco M., Matuska S. A new approach to short web document creation based on textual and visual information. Telecommunications and Signal Processing (TSP). 2013.