HAL open science

Extraction of Web Image Information: Semantic or Visual Cues?

Georgina Tryfou, Nicolas Tsapatsoulis

To cite this version:

Georgina Tryfou, Nicolas Tsapatsoulis. Extraction of Web Image Information: Semantic or Visual Cues?. 8th International Conference on Artificial Intelligence Applications and Innovations (AIAI), Sep 2012, Halkidiki, Greece. pp.368-373, 10.1007/978-3-642-33409-2_38 . hal-01521418

HAL Id: hal-01521418

https://inria.hal.science/hal-01521418v1

Submitted on 11 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Extraction of Web Image Information: Semantic or Visual Cues?

Georgina Tryfou and Nicolas Tsapatsoulis

Cyprus University of Technology,
Department of Communication and Internet Studies,
Limassol, Cyprus

Abstract. Text based approaches for web image information retrieval have been exploited for many years, however the noisy textual content of the web pages makes their task challenging. Moreover, text based systems that retrieve information from textual sources such as image file names, anchor texts, existing keywords and, of course, surrounding text often share the inability to correctly assign all relevant text to an image and discard the irrelevant. A novel method for indexing web images is discussed in the present paper. The main concern of the proposed system is to overcome the obstacle of correctly assigning textual information to web images, while disregarding text that is unrelated to them. The proposed system uses visual cues in order to cluster a web page into several regions and compares this method to the use of semantic information and the realization of a k-means clustering. The evaluation reveals the advantages and disadvantages of the different clustering techniques and confirms the validity of the proposed method for web image indexing.

1 Introduction and Related Work

Numerous web image search engines have been developed as the amount of digital image collections on the web constantly increases [1]. These systems share the objective to minimize the necessary human interaction while offering an intuitive image search. The two main approaches that exist in the literature for content extraction and representation of web images are: (i) the text-based and (ii) the visual feature-based methods. The text-based approaches use associate text (*i.e* image file names, anchor texts, surrounding paragraphs) to derive the content of images. The text blocks that are used as concept sources for images may be extracted with several methods, with the following four being the most popular: (i) fixed-size sequence of terms [2], (ii) DOM tree structure [3, 4], (iii) Web page segmentation [5] and (iv) hybrid versions of the above [6]. The first approach is time-efficient but yields poor results since the extracted text may be irrelevant to the image, or on the other hand, important parts of the relevant text may be discarded. Approaches that use the DOM tree structure of the web page are in general not adaptive and they are designed for specific design patterns. Web page segmentation is a more adequate solution to the problem since it is adaptable to different web page styles. Most of the proposed algorithms in this

field though, are not designed specifically for the problem of image indexing and therefore often deliver poor results. The proposed system uses information obtained following the web page segmentation approach.

The paper is organized as follows: Section 2 presents the architecture of the proposed system. Section 3 presents the evaluation of the proposed system. Finally some conclusions and future perspectives are given in Sec. 4.

2 System Architecture

The general architecture of the proposed system is depicted in Fig. 1. As shown there the system consists of two main parts, which are described in the following sections.

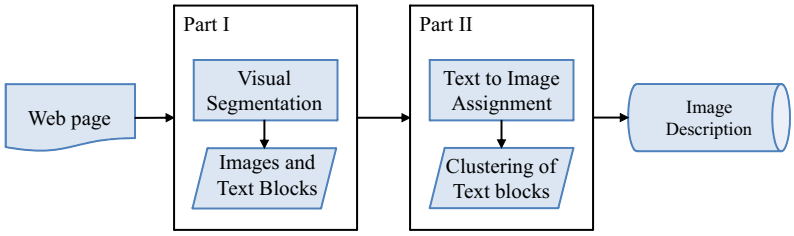

Fig. 1. The general architecture of the proposed system.

2.1 Visual Segmentation

The content extraction of web images is based on textual information that exists in the same web document and refers to this image. In order to determine to which image the various text parts of a web page refer to, we use the visual cues which are connected to the outline and the presentation of the hosting web page. In order to obtain the set of visual segments that form a web page, we use the Visual Based Page Segmentation (VIPS) algorithm [7], which extracts the semantic structure of a web page based on its visual representation. It attempts to make full use of the page layout structure by extracting blocks from the DOM tree structure of the web page and locating separators among these blocks. Each web page is represented as a set of blocks that bare similar Degree of Coherence (DOC). With the permitted DOC (pDOC) set to its maximum value, we obtain a set of visual blocks that consist of visually indivisible contents. An example of a web page segmentation using pDOC = 10 (*i.e.* maximum allowed value) is illustrated in Fig. 2.

2.2 Text to Image Assignment

Each text block found in the web page has to be assigned to an image block. In other words, we attempt to determine to which image, each textual block refers

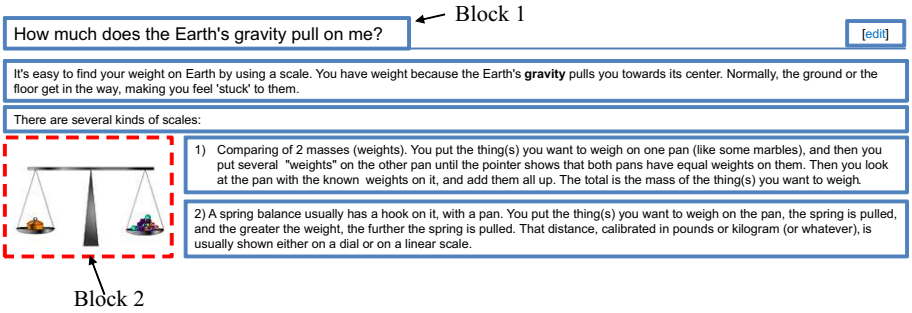

Fig. 2. The results of VIPS algorithm on a fragment of a web page. Each visual block is marked with a rectangle region: dashed-red when the block is an image and continuous-blue when it is text.

to. After this decision, the corresponding text blocks will be adequate to use for the extraction of the image information. The processing that takes place for this module is presented in Fig. 3 and each part is described in further details in the following paragraphs.

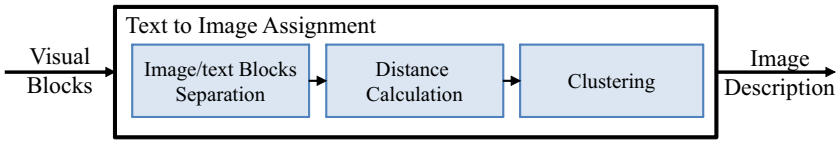

Fig. 3. The processing that takes place in the second part of the system: Text to Image Assignment

Image/text Blocks Separation. The first task towards the assignment of each text block to the image block it refers to, is to determine whether a block contains image content or text. The use of the maximum pDOC during VIPS execution certifies that for a well formed HTML document, each block will either be an image or a text block indicating that no blocks with mixed content will be returned. The HTML source code that corresponds to each one of these blocks, is returned by the VIPS algorithm and it is used in order to classify them into two categories: (i) image blocks and (ii) text blocks.

Distance Calculation. Once the blocks are separated into these categories, the Euclidean distance between every image/text block pair is calculated making use of the Cartesian coordinates returned by the VIPS algorithm. The distance calculation in this case is not a trivial problem since the goal is to quantify the intuitive understanding of how close two visual blocks are. This understanding depends not only on the distance of the centres of the two visual blocks, but also on their shape, size and relative position. In order to solve the distance calculation problem several approaches were taken into consideration. The calculation

of the distance between the closest edges of the blocks was found to offer the better representation of the block distance.

Clustering. After the distance calculation the web page has to be clustered into regions. Each region or cluster is defined by an image in its center and contains all the text blocks that have been found to refer to this image. For the clustering we took into consideration two different approaches. The first one is based only in visual cues and the location of the various blocks while the second approach mines semantic information and implements a k-means clustering.

Clustering Based on Visual Cues. In this approach, the text blocks are assigned to the corresponding cluster making use of the visual information that is available for them: the Euclidean coordinates which are returned be the VIPS algorithm are used for distance calculation and each text block is assigned to the cluster, whose center it is closest to. However, it is possible that one or more blocks of text do not refer to a certain image of the web page. For this reason, it is necessary to discard one or more text blocks from the calculated clusters. In order to determine which blocks of text are irrelevant to the image they are connected to, the blocks whose distance to the cluster center (*i. e.* corresponding web image) is longer than a defined threshold t, are discarded. In order to calculate this threshold, the distances d_i^c that appear in a cluster c are normalized to the maximum distance. Using the normalized values $\tilde{d_i^c}$ the threshold t is calculated as $t_{ed} = t' + m_{ed} - s_{ed}$, where t' is a static, predefined threshold (in our experiment is empirically set to 0.1), m_{ed} is the mean value of the euclidean distances found in the cluster c and s_{ed} is their standard deviation.

Clustering Based on Semantic Information. Semantic information, as obtained from the application of the Vector Space Model [8], is used in the second approach in order to create a clustering which is based on the content of each block rather than its location on the web page. The web page is considered the corpus from which the vocabulary is extracted. Each text block tb_i is one of the corpus' documents and it is expressed as a vector of term weights as $\boldsymbol{v}_i = [w_{1i} w_{2i} \ldots w_{Ni}]^T$. The term w_{ti} indicates the weight of text t in text block tb_i and is calculated using the **tf*idf** [8] statistic which expresses how important is each word for the representation of a text block. The k-means algorithm is then used on the vectors of term weights in order to cluster the text blocks into M regions. Since each text block may refer to any image but it may also be irrelevant to every image, the number of clusters M is equal to the total number of images found in the web page plus one for text blocks irrelevant to every image. Once the k-means algorithm is executed each text block is assigned to a specific cluster. However, it is not yet determined to which image each cluster refers to. To find the solution in this problem we considered the average distances among images and clusters as well as an initialization to the k-means algorithm based on the output of the first approach to the clustering procedure. The results presented in Sec. 3 are obtained using this initialization.

3 Evaluation

A set of manually labelled web images has been collected in order to create a corpus, based on which, the clustering of text blocks is evaluated. This corpus was obtained using an annotation tool that we designed in order to facilitate web image labelling. The annotator had to assign relevant text to each one of the images that exist in web pages which were rendered in the default browser. A dataset that consists of 40 web pages and their annotations was created in order to evaluate our method. Using the above described corpus and the evaluation measures *Precision, Recall* and *F-score* as defined in [9], we obtained two different sets of results using the clustering methods described earlier. In the first set of results, the whole processing is based on visual information as obtained from the application of the VIPS algorithm. In the second set, the results are based on semantic information, as it is obtained from the Vector Space Model applied on the content of the web page. As shown in Table 1 there is a significant variation on the efficiency of the two approaches with the first approach having the highest average F-score.

Table 1. The results for different Text Block Discarding Methods.

Clustering Method	*Precision*	*Recall*	*F-score*
Visual	0.8610	0.8836	0.8962
Semantic	0.3576	0.4528	0.4533

The execution of the proposed method yields an average *F-score* equal to **0.8610** for the total of the 131 annotated images. As shown in Fig. 4, more than 80% of the text blocks are identified with *Recall* and *Precision* values higher than 0.8, indicating that the system succeeds in retrieving most of the blocks that according to the annotation refer to a certain image.

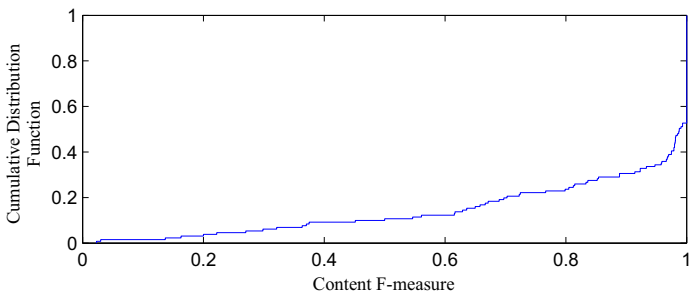

Fig. 4. Cumulative distribution function of F-measure.

4 Conclusions

In this paper we presented an image indexing system that uses textual information in order to extract the concept of the images that are found in a web page. The method uses visual cues in order to identify the segments of the web page and calculates euclidean distances among these segments. It delivers a semantic or euclidean clustering of the contents of a web page in order to assign textual information to the existing images. During the experimenting and the evaluation it was found that a clustering based on semantic information of the contextual information delivers poor results compared to the clustering that is based on visual information. This stresses out the importance of understanding and processing the web pages as a structured visual document when it comes to web image indexing rather than an unstructured bag of words.

The weight vectors that were used for the k-means clustering are in general sparse vectors, since the length of the vocabulary is not proportional to the size of each text block. Moreover, when two neighbouring text blocks refer to the same image and contain similar semantic content, the authors usually select synonyms to express the same meaning. The Vector Space Model does not account such connections among different terms. It is therefore expected that the use of an ontology that describes the semantic distances and relations among different words and phrases will improve the results of the k-means clustering and this is the direction of our future study.

References

[1] Sclaroff, S., La Cascia, M., Sethi, S., Taycher, L.: Unifying textual and visual cues for content-based image retrieval on the world wide web. Computer Vision and Image Understanding **75**(1–2) (1999) 86–98

[2] Feng, H., Shi, R., Chua, T.S.: A bootstrapping framework for annotating and retrieving www images. In: Proc. of the 12th ACM Int. Conf. on Multimedia. (2004)

[3] Hua, Z., Wang, X.J., Liu, Q., Lu, H.: Semantic knowledge extraction and annotation for web images. In: Proc. of the 13th ACM Int. Conf. on Multimedia. (2005) 467–470

[4] Fauzi, F., Hong, J.L., Belkhatir, M.: Webpage segmentation for extracting images and their surrounding contextual information. In: Proc. of the 17th ACM Int. Conf. on Multimedia. (2009)

[5] He, X., Cai, D., Wen, J.R., Ma, W.Y., Zhang, H.J.: Clustering and searching www images using link and page layout analysis. ACM Trans. on Multimedia Computing, Communications and Applications **3**(2) (June 2007)

[6] Alcic, S., Conrad, S.: A clustering-based approach to web image context extraction. In: Proc. of the 19th ACM Int. Conf. on Multimedia. (2011) 74–79

[7] Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Vips: a vision based page segmentation algorithm. Technical report, Microsoft Research (2003)

[8] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11) (1975) 613–620

[9] Alcic, S., Conrad, S.: Measuring performance of web image context extraction. In: Proc. of the 10th Int. Workshop on Multimedia Data Mining. (July 2010) 1–8