Image indexing based on web page segmentation and clustering

GEORGINA TRYFOU, NICOLAS TSAPATSOULIS
Technological University of Cyprus
Department of Communication and Internet Studies
Limassol, CYPRUS
{georgia.tryfou, nicolas.tsapatsoulis}@cut.ac.cy

Abstract: Thousands of images are nowadays available on the web. These images are accompanied by a wide range of textual descriptors, such as image file names, anchor texts and, of course, surrounding text. Existing systems that attempt to mine information for images using surrounding text suffer from several problems, such as the inability to correctly assign all relevant text to an image and discard the irrelevant. In this paper, we propose a novel method for indexing web images which is based on textual descriptors. The web document is segmented into visual blocks of text and then each block of text is assigned to the closet image. The text extraction is improved by assigning the text to an image following the intuitive understanding of how close two visual blocks are. The evaluation confirms the validity of the proposed method and demonstrates its possible extensions.

Key–Words: image indexing, automatic annotation, web page segmentation

1 Introduction

The development of cheap digital recording and storage devices has enabled the production of a huge amount of digital image collections. The millions of available images on the web, covering every conceivable topic, create the need for efficient image indexing and retrieval. Web search engines share the objective to offer to the user an intuitive image search by minimizing the necessary human interaction for the optimization of the results. Currently, two main approaches exist in the literature for content extraction and representation of web images: (i) text-based and (ii) visual feature-based methods.

The text-based approaches use the associate text to derive the content of image. Image file names, anchor texts, surrounding paragraphs, even the whole text of the hosting web page are examples of textual content that is often used in such systems. In the visual feature-based approaches image processing techniques are used in order to describe the content of a web image. Common features that are used in these methods are the texture, the shape and the colour histogram. A sample image is used in order to query for similar web images. The extraction of the visual features is a time consuming procedure, strongly related to the domain the query image belongs to and its characteristics.

The paper is organized as follows: Section 2 reviews related work. Section 3 presents the architecture while Section 4 presents the evaluation of the proposed system. Finally some conclusions and future perspectives are given in Section 5.

2 Related Work

In this work we propose a text based method for indexing web images. There are several systems that use textual information in order to search for images on the web: WebSeer [1], WebSeek [2], WebMARS [3] and ARTISTIC [4] are some representative examples. These systems share one or more of the following drawbacks: the text in the web page is only partially processed; only a few words are considered as textual features; it is not clear how textual information is used to support image indexing and retrieval; term lists or taxonomies that are built in the set-up phase of the systems demand high user intervention [4]. Opposed to that, the proposed system processes the web page as a whole and attempts to assign each text block to the image it refers to. Moreover, any word or phrase is a possible keyword or representation for an image and no user intervention is necessary once the indexing model is built.

In the problem of text blocks extraction from a web page and their use as concept sources for images, several approaches exist towards three main directions: (i) fixed-size sequence of terms [5], (ii) DOM tree structure [6], and (iii) Web page segmentation [7] or hybrid versions of the above [8].

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

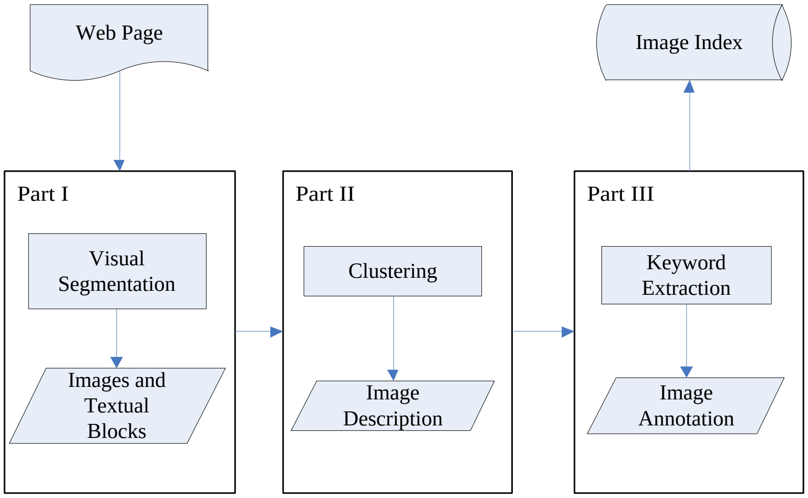


Figure 1: The general architecture of the proposed system.

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 third approach (web page segmentation).

3 System Architecture

The general architecture of the proposed system is depicted in Figure 1. The system consists of 3 main parts: the visual segmentation module (Part I), the clustering (Part II) and the keyword extraction module (Part III). The main tools used for each of these parts are the VIsion-based web Page Segmentation (VIPS) algorithm, 2D distance based text clustering using kNN algorithm and the Maui tool for keyword extraction. A more detailed description of these parts and the used tools follows.

3.1 Web page segmentation using VIPS

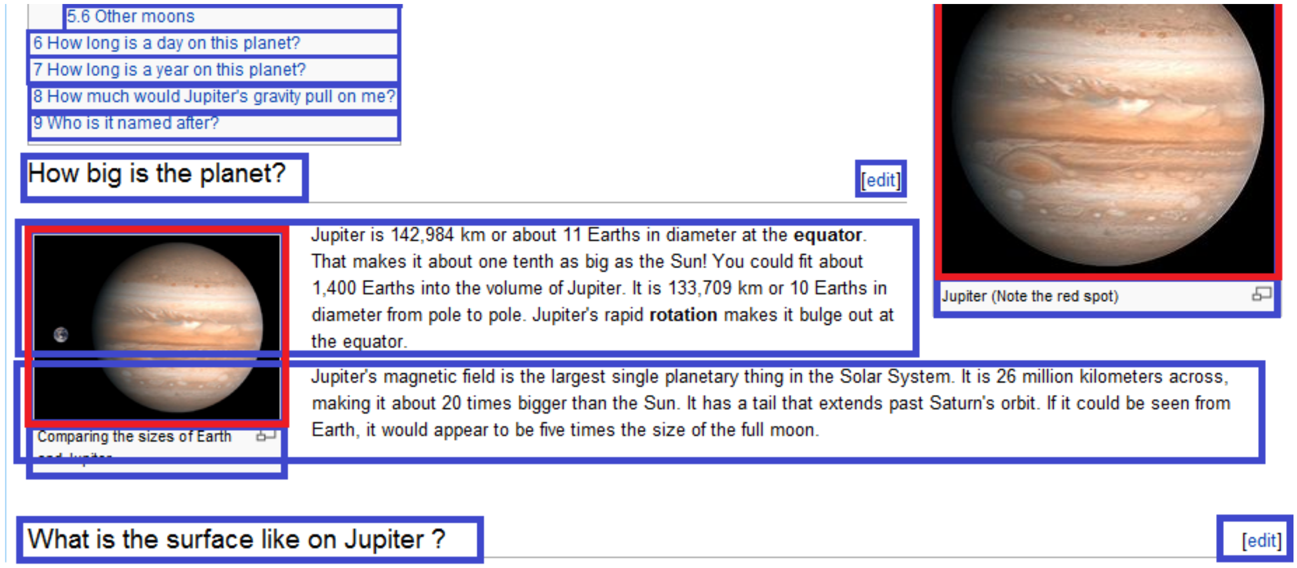
In the proposed algorithm the content extraction of each web image is based on textual information that exists in the same web document and refers to this image. Initially both image and text blocks must be identified. In order to obtain the set of visual segments that form a web page, we use the Visual Based Page Segmentation (VIPS) algorithm [9]. The VIPS algorithm 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. Therefore, a 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 is illustrated in Figure 2.

3.2 Distance based text clustering

For each visual block, obtained in the previous step, the VIPS algorithm returns the two dimensional Cartesian coordinates of its location in the web page. The HTML source code that corresponds to each one of these blocks is used in order to classify them into two categories: (i) image blocks, and (ii) text blocks.

The objective of the second module of the proposed system is to assign each text block to an image block. In other words, we attempt to determine to which image, each textual block refers to. To achieve this task, the Euclidean distance between every image/text block pair is calculated using Cartesian coordinates. 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 two visual blocks, but also on their size and relative position.

In order to solve the distance calculation prob-



5.6 Other moons

6 How long is a day on this planet?

7 How long is a year on this planet?

8 How much would Jupiter's gravity pull on me?

9 Who is it named after?

How big is the planet?

[edit]

Jupiter is 142,984 km or about 11 Earths in diameter at the **equator**. That makes it about one tenth as big as the Sun! You could fit about 1,400 Earths into the volume of Jupiter. It is 133,709 km or 10 Earths in diameter from pole to pole. Jupiter's rapid **rotation** makes it bulge out at the equator.

Jupiter's magnetic field is the largest single planetary thing in the Solar System. It is 26 million kilometers across, making it about 20 times bigger than the Sun. It has a tail that extends past Saturn's orbit. If it could be seen from Earth, it would appear to be five times the size of the full moon.

Comparing the sizes of Earth and Jupiter

Jupiter (Note the red spot)

What is the surface like on Jupiter ?

[edit]

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

lem we took into consideration several approaches like: (i) Euclidean distance between the centres of the blocks, (ii) Euclidean distance between the starting points (*i. e* location) of the blocks, (iii) Euclidean distance between the closest edges of the blocks, and (iv) Approach (i) or (ii) normalized to the surface of the blocks.

The first was found to offer a very poor representation of the block distance since, the largest the surface of the blocks is, the bigger the distance is calculated even though the blocks may be adjacent. The second approach only takes into consideration the starting point (point (0,0)) of the blocks and yields poor results, especially when the two blocks have big surfaces. In the third approach, only the closest edges of the two blocks is taken into consideration. Moreover, it is not affected by the orientation and exact positioning of two blocks or the width to height ratio of the blocks as the results from the fourth approach. The third approach is the one selected for the solution of the distance calculation problem between two visual blocks.

After the distance calculation the web page has to be clustered into regions. Each region is defined by a web image which is also considered to be the center of it. Each text block is assigned to the cluster, whose center is closest to it. A kNN algorithm with $k = 1$ is therefore implemented in order to cluster the web page into 1 or more clusters.

Until here each web page is separated into one or more clusters and all blocks of text that exist in the document are assigned to their closest image. However, it is possible that several blocks of text do not refer to a certain image; meaning that it is necessary to discard these blocks from the calculated clusters. A text block is discarded when its distance to the cluster center (*i. e.* corresponding web image) is bigger than a defined threshold t.In order to calculate this threshold the distances d_i^c that appear in a cluster c are normalized as follows:

$$\tilde{d_i^c} = \frac{d_i^c}{\max_i d_i^c}. \qquad (1)$$

Using the normalized values $\tilde{d_i^c}$ the threshold t is calculated as follows:

$$t = t' + m_d - s_d, \qquad (2)$$

where $t' = 0.1$ a static, predefined threshold, m_d the mean value of the distances found in the cluster c and s_d the standard deviation of these distances.

3.3 Keyword extraction based on Natural Language Processing

In this step, our goal is to extract keywords from text blocks created after the visual page segmentation. To achieve this, the Maui tool [10] for keyword extraction is used. Maui automatically identifies main topics in text documents. Depending on the task, topics may be tags, keywords, key phrases, vocabulary terms, descriptors, index terms or even titles of Wikipedia articles. The keywords and/or key phrase extraction from the text blocks is realized in a two step procedure that consists of the model building and the tagging steps.

3.3.1 Model Building

Maui uses the machine learning toolkit Weka [1] for creating a topic indexing model. This is achieved by using documents to which annotators have already assigned topics, and applying the newly created model to new documents. Since Maui includes the complete Weka library it offers the opportunity to tailor the indexing model to specific data sets.

3.3.2 Tagging

Maui utilizes a two-step process of automatic indexing and keyword extraction: (i) candidate selection and (ii) filtering. Both of these steps are inherited by Kea [2], a tool that extracts n-grams of a predefined length that do not start or end with a stopword for the candidate selection. Several features are then computed for each candidate (TFxIDF, first occurrence, length, node degree) in order to realize the filtering stage and output the most significant candidates as the index of the input documents. In the current implementation the described system uses the Maui toolbox with its default settings, and no important changes take place either on the model building or the tagging step.

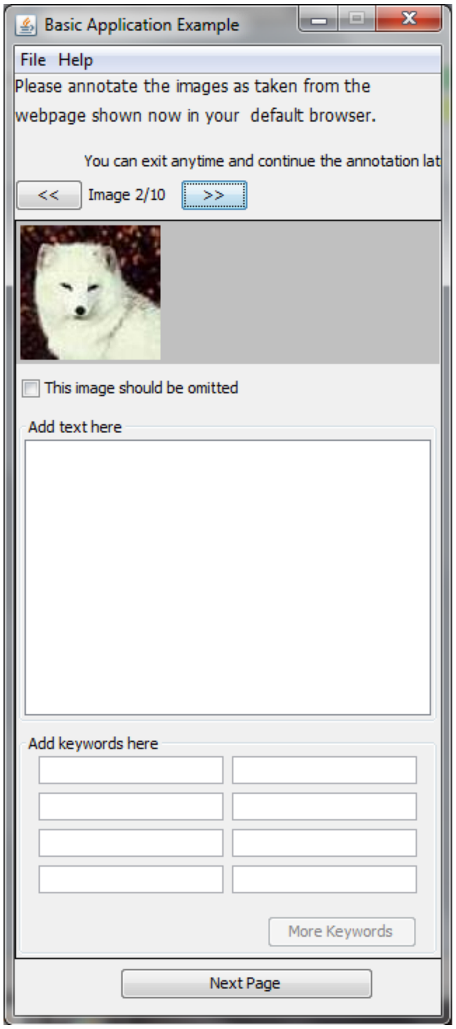
4 Evaluation

4.1 Evaluation Corpus

In order to evaluate the clustering of text blocks and the keyword extraction from these blocks, it was necessary to collect a set of manually labelled images. To achieve this, an annotation tool that facilitates web image labelling was developed. The user interface of this tool is shown in Figure 3. The annotator is able to add related text and keywords to a web image. The annotation tool starts both the default browser of the system and the illustrated annotation window. The user has to assign text and keywords to each one of the images that exist in the web page shown in the default browser.

The assigned text forms the annotation which is used for the clustering evaluation while the assigned keywords are used for the evaluation of the keyword extraction module.

A dataset that consists of 15 web pages was created in order to evaluate our method. One user was asked to annotate a total number of 44 images by providing their related text and one to six representative keywords. The web page clustering obtained from the described system is evaluated using the measures described in the next paragraphs. In the current experiment the main concern is to evaluate the success of the system to correctly retrieve the visual blocks that refer to a certain image, rather than the keyword extraction from these blocks.



Please annotate the images as taken from the webpage shown now in your default browser.

You can exit anytime and continue the annotation lat

Figure 3: The designed annotation tool.

4.2 Defined measures

In order to evaluate the page segmentation and clustering results we use measures inspired from the traditional information retrieval measures *Precision*, *Recall* and *F-measure*.

Let A be the manually labelled region that the annotator determined to be connected with an image and $b_1, b_2, \ldots, b_N$ the N regions that the proposed system assigned to this image. As shown in Figure 4, there are three possible cases: (i) only a small part (sub-block) of b_i is correctly identified (for instance in block b_1 and block b_{N-1}), (ii) the whole block b_i is correctly

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.nzdl.org/Kea/index.html

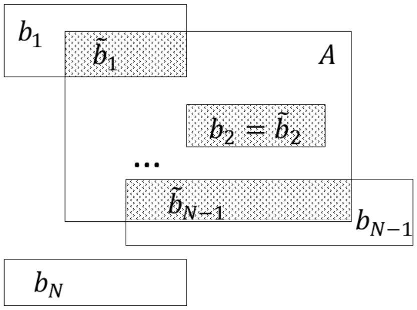


Figure 4: The possible cases of textual blocks assigned to an image with annotation A.

identified (*e.g.* block b_2) and (iii) no part of the block b_i is related to the image (block b_N).

For all cases we define $\tilde{b}_i$ the correctly identified part of a block b_i. Using the above notation we define the *content Precision*, C_p, of the system as follows:

$$C_p = \frac{\tilde{b}_1 + \tilde{b}_2 + \cdots + \tilde{b}_N}{A} = \frac{\left| \bigcup_{i=0}^{N} b_i \bigcap A \right|}{|A|} \qquad (3)$$

and the *content Recall*, C_r, as:

$$C_r = \frac{\tilde{b}_1 + \tilde{b}_2 + \cdots + \tilde{b}_N}{b_1 + b_2 + \cdots + b_N} = \frac{\left| \bigcup_{i=0}^{N} b_i \bigcap A \right|}{\left| \bigcup_{i=0}^{N} b_i \right|} \qquad (4)$$

Using equations 3 and 4 the *content F-measure*, C_f, is defined as:

$$C_f = 2\frac{C_p C_r}{C_p + C_r} \qquad (5)$$

The C_f measure is used for the evaluation of the proposed system. It is mentioned here that since both, the C_r and C_p measures, can have values in the interval $[0, 1]$ the C_f measure can have 0 for the lowest and 1 for the highest success rate.

4.3 Results

The cumulative distribution function of the results obtained from the text to image assignment is presented in Figure 5. 79% of the text blocks are identified with a value of *content Recall* higher than 0.8, while 75% of the blocks are identified with *content Precision* higher than 0.8. The results yield an average content F-measure equal to **0.81** for the total of the 44 annotated images. Finally, the keyword extraction that takes place for each text block is evaluated. From the 44 images, the 39 were randomly chosen and used to built the Maui model while the remaining 5 were used to test it giving an average F-Measure equal to **0.82**.

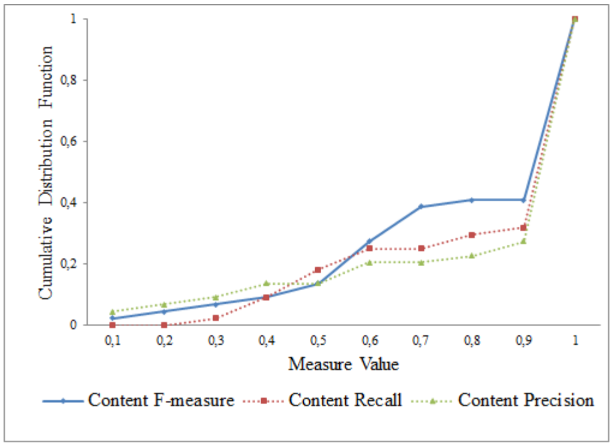


Figure 5: Cumulative distribution function for the results based on the defined evaluation measures.

It is important to mention that the total time needed to segment, cluster and process a web page and then extract keywords from the text that is assigned to each image is less than 1 second (on a platform with an Intel Core Duo 2.53GHz CPU and 4GB RAM).

5 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 2D distances among these segments. It delivers a clustering of the contents of a web page in order to assign textual information to the existing images. The results so far are encouraging and indicate that the combination of visual information for the web page segmentation and natural language processing for the keyword extraction is an appropriate solution for the problem of image indexing.

References:

[1] M. Swain, C. Frankel, and V. Athitsos. Webseer: An image search engine for the world wide web. In CVPR, 1997.

[2] J. Smith and S. Chang. An image and video search engine for the world-wide web. Storage. Retr. Im. Vid. Datab, pp. 8495, 1997.

[3] M. Ortega-Binderberger, V. Mehrotra, K. Chakrabarti, and K. Porkaew. Webmars: A mul-

timedia search engine. In SPIE An. Sym. Elect. Im., San Jose, California, 2000.

[4] L. Alexandre, M. Pereira, S. Madeira, J. Cordeiro, and G. Dias. *Web image indexing: Combining image analysis with text processing.* In Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS04), 2004

[5] H. Feng, R. Shi, and T.S. Chua. A bootstrapping framework for annotating and retrieving www images. In Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004, pp. 960-967.

[6] F. Fauzi, J.L. Hong, and M. Belkhatir. Webpage segmentation for extracting images and their surrounding contextual information. In ACM Multimedia, 2009, pp. 649-652.

[7] X. He, D. Cai, J.R. Wen, W.Y. Ma, and H.J. Zhang. Clustering and searching www images using link and page layout analysis, ACM Trans. Multimedia Comput. Commun. Appl., vol. 3, no. 2, p. 10, 2007.

[8] S. Alcic and S. Conrad. A Clustering-based Approach to Web Image Context Extraction In MMEDIA ’11, 2011.

[9] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma. VIPS: a Visionbased Page Segmentation Algorithm. Technical report, Microsoft Research, 2003.

[10] O. Medelyan. Human-competitive automatic topic indexing. PhD thesis, Department of Computer Science, University of Waikato, New Zealand, 2009.