Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 10; April-June, 2015 pp. 83-86
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

A Novel Approach for Extracting Relevant Keywords for Web Image Annotation using Semantics

Manisha Yadav[1] and Payal Gulati[2]

[1,2]YMCA UST, Haryana, India
E-mail: [1]manishayadav17@gmail.com, [2]gulati.payal@yahoo.co.in

Abstract—*Due to improved digital technologies, facility of sharing of images on the web through social networking sites like facebook, instagram, google+, etc, the number of images over www is exponentially growing and will continue to increase in future; leading to the challenge for image search engines to search relevant images from the exploding web image repositories. Therefore there is a need for efficient image annotation, indexing and retrieval. Since manually annotating images is difficult and time consuming process, automatic image annotation came into account. Most of the existing image search engines use contextual information from page title, image src tag, alt tag, meta tag, image surrounding text for annotating web images. In addition to this, there exist some image search engines that annotate web images based on content such texture, shape and color histogram, they all have problem of scalability. Moreover images are not annotated with their semantic descriptors therefore it acts as a challenge for general users to find specific image from the web. This paper proposes a novel approach for extracting relevant keywords for web image annotation using semantics. In this work, the relevant keywords from contextual information along with semantic similar content are then used for annotating web images.*

Keywords: *Image Search Engine, Annotation, Semantic Correlation, Semantic Distance Measure.*

1. INTRODUCTION

WWW is the largest repository of digital images in the world. The number of images available over the web is exponentially growing and will continue to increase in future. For better image retrieval and enhancing relevance of the searched image, images must be properly annotated. Image annotation can be done either through content-based approach or by text based approach. In **Text based** approaches, different parts of a web page is considered as possible sources for contextual information of images, namely, image file names (ImgSrc), page title, anchor texts, alternative text (ALT attribute), image surrounding text. In the **Content-based** approaches, image processing techniques are considered to describe the content of a web image. Texture, shape and color histogram are the common features that are used in image processing techniques. Most of the image search engines index images using text information associated with images i.e. on the basis of alt tags, image caption. Alternative tags or alt tag provides a textual alternative to non-textual content in web pages such as image, video, media. It basically provides a semantic meaning and description to the embedded images. However, the web is still replete with images that have missing, incorrect, or poor text. In fact in many cases, images are given only empty or null alt attribute (alt =“ ”). Thereby such images remain inaccessible. Image search engines that annotate web images based on content based annotation have problem of scalability.

In this work a novel approach is proposed that automatically crawls the web pages and extract the contextual information from the pages containing valid images. The web pages is segmented into web content blocks and thereafter semantic correlation is calculated between web image and web content block using semantic distance measure. The relevant keywords from contextual information along with semantic similar content are then used for annotating web images. Further image is indexed with the associated text it refers to.

The paper is organized as follows: Section 2 discusses the related work done in this domain. Section 3 presents the architecture of the proposed system. Finally Section 4 comprises of the conclusion.

2. RELATED WORK

A variety of text based approaches for web image annotation has been proposed in recent years [1]. There are several systems [2,3,4,5] that use contextual information for annotating web images. Methods for exacting contextual information are: (i) window based extraction [6,7], (ii) Structure based wrappers[8,9] (iii) web page segmentation[10,11,12].

Window based extraction is a heuristic approach which extracts image surrounding text; it yields poor results as at times irrelevant data is extracted and relevant data is discarded. **Structure based wrappers** use the structural information of web page to decide the borders of the image context but these are not adaptive as they are designed for specific design patterns of web page. **Web page segmentation** method are adaptable to different web page styles and divides the web page into segments of common topics and then each image is associated with the textual contents of the segment which it belongs to. Moreover it is difficult to determine the semantics of text with the image.

In this work web page is segmented into web content blocks using Vision based Page Segmentation algorithm [11]. Thereafter semantic similarity is calculated between web image and web content block using semantic distance measure. *Semantic distance* is the inverse of *semantic similarity [13]* that is the less distance of the two concepts, the more they are similar. So, *semantic similarity* and *semantic distance* are used interchangeably in this work.

Semantic distance between web content blocks is calculated by determining a common representation among them. As text is most prevalent media type in web pages so text is selected for common representation. There is a variety of similarity metrics for texts are present in the literature [14,15,16]. Some simple metrics are based on lexical matching between units of one text to the others. These approaches are successful to a certain degree, but they fail to identify the semantic similarity of texts. For example synonyms Plant and Tree have a high semantic correlation which cannot be detected without background knowledge. To overcome these limitations a variety of more sophistic metrices which uses the WordNet taxonomy [17] as background knowledge are discussed in [18]. In this work, the word to word similarity metric introduced by Lin [19] is used to calculate the similarity between words and text to text similarity is calculated using the metric introduced by Coley [16].

3. PROPOSED ARCHITECTURE

The architecture of proposed system is given below in Fig. 1. Components of proposed system are discussed in following subsequent subsections:

3.1 Crawl Manager

Crawl Manager is a computer program that browses the www in an automated manner to gather information from webpage. It takes the seed URL from the URL Queue and fetches the web page from www.

3.2 URL Queue

URL Queue is a type of data structure which is used to store the list of URLs that are discovered and extracted by Crawler. It doesn't contain duplicate URLs.

3.3 Parser

Parser is used to extract information present on web pages. Parser downloads the web page and extracts the XML file of the same. Thereafter, it convert XML file into DOM object models. It then checks whether valid images are present on the web page or not. If valid image is present on the web page then the page is segmented using visual web page segmentor otherwise next URL is crawled. The DOM object models which contain Page Title of web page, Image Source and Alternative Text of valid images present on the web page are extracted from the set of object models of the web page.

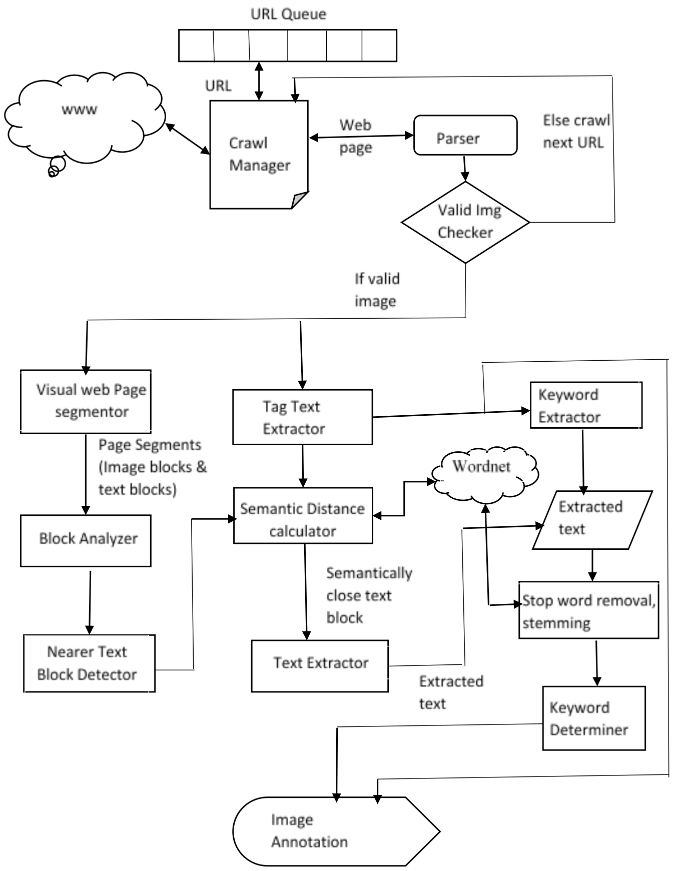

Fig. 1: Proposed Architecture.

3.4 Visual Web Page Segmentor

Visual Web Page Segmentor is used for the segmentation of web pages into web content blocks. By the term segmentation of web pages, means dividing the page by certain rules or procedures to obtain multiple semantically different web content blocks whose content can be investigated further.

In the proposed approach, VIPS algorithm [11] is used for the segmentation of web page into web content blocks. It extracts the semantic structure of a web page based on its visual representation. The segmentation process has basically three steps: *block extraction, separator detection and content structure construction*. Blocks are extracted from DOM tree structure of the web page by using the page layout structure and then separators are located among these blocks. The vision-based content structure of a page is obtained by combining the DOM structure and the visual cues. Therefore, a web page is a collection of web content blocks that have similar DOC. With the permitted DOC (pDOC) set to its maximum value, a set of web content blocks that consist of visually indivisible contents is obtained. This algorithm also provides the two dimensional Cartesian coordinates of each visual block present on the web page based on their locations on the web page.

3.5 Block Analyzer

In this work, Block Analyser analyses the web content blocks obtained from previous step. It analyses the contents of web content blocks by using the HTML source code that correspond to each of these blocks. This analyser divides the web content blocks into two categories: image blocks and text blocks. Web blocks which contain images are considered as image blocks and rest are considered as text blocks.

3.6 Nearest Text Block Detector

Nearest Text Block Detector assign nearest text blocks to an image block. To achieve this objective distance calculator is used to calculate the distance between each image/ text block pair using cartesian coordinates. For each web content block, obtained in previous step, the VIPS algorithm returns the two dimensional Cartesian coordinates of its location in the web page. In this work, closest edge between image block and text block is determined by calculating the Euclidean Distance between each pair of edge of these blocks. Distance between two line segments is obtained by using equation (1):

$$Euclidean\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

After the distance is calculated between each image and text block pair, the text blocks whose distance from image block is below the threshold are assigned to that image block. In this way each image block is assigned with a group of text blocks which are closer in distance with that image block.

3.7 Tag Text Extractor

In the proposed approach, Tag Text Extractor is used for extracting text from the html tags. Parser provides the DOM object models by parsing a web page. If the image present on this web page is valid i.e. it is not a button or an icon, which is checked by valid image checker, are extracted from metadata of image like Image Source (Imgsrc), Alternative Text (Alt). Page Title of the web page which contains this image is also extracted.

3.8 Keyword Extractor

In this work, Keyword Extractor is used to extract keywords from the metadata of images and Page Title. Keywords are stored into a text file which is further used for obtaining semantically close text blocks by calculating semantic distance.

3.9 Semantic Distance Calculator

Semantic Distance Calculator is used to determine the semantic correlation among the web content blocks. As lexical matching between words doesn’t provide better results, words are mapped to corresponding concepts in a knowledge base and a concept to-concept accordance is computed using Wordnet.

Before computing text similarity between image metadata and text blocks, some pre-processing is needed to bring the text blocks in proper format. First a sentence detector is applied, then a word level tokenizer and finally part of speech (POS) tagging is done to identify the word class for each word using the Open NLP Library. The terms are stemmed using the WordNet stemmer and mapped to suitable WordNet concepts if possible.

The text similarity is computed by using Coley [16] approach which is used as follows: for each noun (verb) belongs to image metadata, identify the noun (verb) in text of text blocks with the highest semantic similarity (maxSim) according to the concept similarity Sim $_{\text{Lin}}$.

$$sim_{Lin} = \frac{2.IC(LCS)}{IC(Concept_1) + IC(Concept_2)} \quad (2)$$

Where LCS is the least common subsumer of the two concepts in the WordNet taxonomy, and IC returns the Information Content [41] which is defined as:

$$IC(c) = -\log P(c) \quad (3)$$

and P(c) is the probability of encountering an instance of concept c in a large corpus. For the classes other than noun (verb), a lexical matching to their counterparts in text of text block is performed. The similarity function between two texts T1 (text of image metadata), T2 (text of text blocks) is defined as follows:

$$sim(T_1,T_2)_{T_1} = \frac{\sum_{w_i \epsilon T_1} maxSim(w_i,T_2).idf(w_i)}{\sum_{w_i \epsilon T_1} idf(w_i)} \quad (4)$$

With idf (w_i) identifying the **inverse document frequency** [20] of the word w_i in a large corpus. With the defined similarity metric in Equation (3), a directional similarity score is computed with respect to T1. The score from both directions can be combined into a bi-directional similarity as shown in [4]:

$$sim(T_1,T_2) = sim(T_1,T_2)_{T_1}.sim(T_1,T_2)_{T_2} \quad (5)$$

This similarity score has a value between 0 and 1. From this similarity score, semantic distance is calculated as follows:

$$dist_{sem}(T_1,T_2) = 1 - sim(T_1,T_2) \qquad (6)$$

$$dist_{sem}(T_1,T_2) = 1 - sim(T_1,T_2)$$

In this way, semantic distance is calculated among image block and its nearest text blocks. The text block whose semantic distance is less, is the semantically correlated text block to that image block.

3.10 Text Extractor

Text extractor is used to extract text from text blocks present on the web page. Text of semantically close text block obtained in previous step, is extracted and buffered. This text along with the text extracted from image metadata and page title of web page is used to extract frequent keywords.

3.11 Keyword Determiner

Keyword Determiner is used to extract keywords from the text. Here, keywords are extracted from the text stored in a buffer. Keywords are extracted after applying stemming and removing stop words from the text. Frequent keywords are determined by applying a threshold on the frequency count of keywords. Keywords whose frequency is above the threshold are extracted and used for annotating images.

3.12 Image Annotation

Page Title of web page, Image source of image, Alternative text of image and frequent keywords extracted in the previous step are put together into a text file as all of these describes the image best. Index the text file to the regarding image as this is the annotation of the image.

4. CONCLUSION

This paper proposes a novel approach for extracting relevant keywords for web image annotation using semantics. In this work web images are automatically annotated by determining relevant keywords from contextual information from web page and semantic similar content from web content blocks. This approach will provide good results as closeness between image and web content blocks are computed using both Euclidean distance and semantic distance

REFERENCES

[1] T. Sumathi1, C.Lakshmi Devasena2, and M.Hemalatha, “An Overview of Automated Image Annotation Approaches”, *International Journal of Research and Reviews in Information Sciences* Vol. 1, No. 1, March 2011 Copyright © Science Academy Publisher, United Kingdom.

[2] M. Swain, C. Frankel, and V. Athitsos. Webseer: "An image search engine for the World Wide Web". *In CVPR, 1997.*

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

[4] 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.* Academy Publisher, United Kingdom.

[5] 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.* Publisher, United Kingdom.

[6] Tatiana Almeida Souza Coelho, P´avel Pereira Calado, Lamarque Vieira Souza, Berthier Ribeiro-Neto, and Richard Muntz. Image Retrieval Using Multiple Evidence Ranking. *IEEE Transactions on Knowledge and Data Engineering, 16(4):408–417, 2004.*

[7] Lexin Pan. Image8: an image search engine for the internet. *Honours year project report, School of computing, National University of Singapore, April. 2003.*

[8] Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. *Data-Centric Systems and Applications. Springer, 2007. 16(4):408–417, 2004.*

[9] Fariza Fauzi, Jer-Lang Hong, and Mohammed Belkhatir. *Webpage segmentation for extracting images and their surrounding contextual information*. In ACM Multimedia, pages 649–652, 2009.

[10] Deepayan Chakrabarti, Ravi Kumar, and Kunal Punera. “*A graphtheoretic approach to webpage segmentation*”. In Proceeding of the 17th international conference on World Wide Web, WWW ’08, pages 377– 386, New York, USA, 2008.

[11] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. *VIPS: a Vision based Page Segmentation Algorithm.* Technical report, Microsoft Research (MSR-TR-2003-79), 2003.

[12] Gen Hattori, Keiichiro Hoashi, Kazunori Matsumoto, and Fumiaki Sugaya. “*Robust web page segmentation for mobile terminal using contentdistances and page layout information*”. In Proceedings of the 16th international conference on World Wide Web, WWW ’07, pages 361–370, New York, NY, USA, 2007. ACM.

[13] Hoa A. Nguyen, B.Eng. New semantic similarity techniques of concepts applied in the Biomedical domain and wordnet. *Master thesis .The university of houston-clear lake 2006.*

[14] Voorhees, E. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference*, 1993.

[15] Landauer, T. K.; Foltz, P.; and Laham, D. Introduction to latent semantic analysis. *Discourse Processes* 25, 1998.

[16] C. Corley and R. Mihalcea. *Measuring the semantic similarity of texts.* In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, EMSEE '05, pages 13-18, Morristown, NJ, USA, 2005. Association for Computational Linguistics. 1998.

[17] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. *WordNet: An on-line lexical database*. International Journal of Lexicography, 3: 235-244, 1990.

[18] S. Patwardhan, S. Banerjee, and T. Pedersen. *Using measures of semantic relatedness for word sense disambiguation.* In Proceedings of the 4th international conference on Computational linguistics and intelligent text processing, CICLing'03, pages 241-257, Berlin, Heidelberg, 2003. Springer-Verlag.

[19] D. Lin. *Automatic retrieval and clustering of similar words.* In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2, ACL-36, pages 768{774, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[20] K. Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*, pages 132-142. Taylor Graham Publishing, London, UK, UK, 1988.