# A Novel Approach for Extracting Pertinent Keywords for Web Image Annotation Using Semantic Distance and Euclidean Distance

**Payal Gulati and Manisha Yadav**

**Abstract** The World Wide Web today comprises of billions of Web documents with information on varied topics presented by different types of media such as text, images, audio, and video. Therefore along with textual information, the number of images over WWW is exponentially growing. As compared to text, the annotation of images by its semantics is more complicated as there is a lack of correlation between user's semantics and computer system's low-level features. Moreover, the Web pages are generally composed of contents containing multiple topics and the context relevant to the image on the Web page makes only a small portion of the full text, leading to the challenge for image search engines to annotate and index Web images. Existing image annotation systems use contextual information from page title, image src tag, alt tag, meta tag, image surrounding text for annotating Web image. Nowadays, some intelligent approaches perform a page segmentation as a preprocessing step. This paper proposes a novel approach for annotating Web images. In this work, Web pages are divided into Web content blocks based on the visual structure of page and thereafter the textual data of Web content blocks which are semantically closer to the blocks containing Web images are extracted. The relevant keywords from textual information along with contextual information of images are used for annotation.

**Keywords** Image search engine · Annotation · Semantic correlation
Semantic distance measure

P. Gulati
YMCA UST, Faridabad, Haryana, India
e-mail: gulatipayal@yahoo.co.in

M. Yadav (✉)
RPSGOI, Mahendergarh, Haryana, India
e-mail: manishayadav17@gmail.com

M. N. Hoda et al. (eds.), *Software Engineering*, Advances in Intelligent Systems
and Computing 731, https://doi.org/10.1007/978-981-10-8848-3_17

# 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. However, as compared to text, the annotation of images by means of the semantics they depict is much more complicated. Humans can recognize objects depicted in images, but in computer vision, the automatic understanding the semantics of the images is still the perplexing task. Image annotation can be done either through content-based or text-based approaches. In **text-based** approach, different parts of a Web page are 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** approach, image processing techniques such as texture, shape, and color are considered to describe the content of a Web image.

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 for extracting pertinent keywords for Web image annotation using semantic distance and Euclidean distance is proposed. Further, this work proposes an algorithm that automatically crawls the Web pages and extracts the contextual information from the pages containing valid images. The Web pages are segmented into Web content blocks and thereafter semantic correlation is calculated between Web image and Web content block using semantic distance measure. The pertinent keywords from contextual information along with semantic similar content are then used for annotating Web images. Thereafter, the images are indexed with the associated text it refers to.

This paper is organized as follows: Sect. 2 discusses the related work done in this domain. Section 3 presents the architecture of the proposed system. Section 4 describes the algorithm for this approach. Finally, Sect. 5 comprises of the conclusion.

# 2 Related Work

A number of text-based approaches for Web image annotation have been proposed in recent years [1]. There are numerous systems [2–6] that use contextual information for annotating Web images. Methods for exacting contextual information are (i) window-based extraction [7, 8], (ii) structure-based wrappers [9, 10], (iii) Web page segmentation [11–13].

**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 is 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 [12]. Thereafter, semantic similarity is calculated between Web image and Web content block using semantic distance measure. *Semantic distance* is the inverse of *semantic similarity* [14] 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. Generally, text is used for common representation. As per the literature review, there are various similarity metrics for texts [13, 15, 16]. Some simple metrics are based on lexical matching. Prevailing approaches are successful to some extent, as they do not identify the semantic similarity of texts. For instance, terms Plant and Tree have a high semantic correlation which remains unnoticed without background knowledge. To overcome this, WordNet taxonomy as background knowledge is discussed [17, 18].

In this work, the word-to-word similarity metric [19] is used to calculate the similarity between words and text-to-text similarity is calculated using the metric introduced by Corley [20].

## 3 Proposed Architecture

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

### *3.1 Crawl Manager*

Crawl manager is a computer program that takes the seed URL from the URL queue and fetches the Web page from WWW.

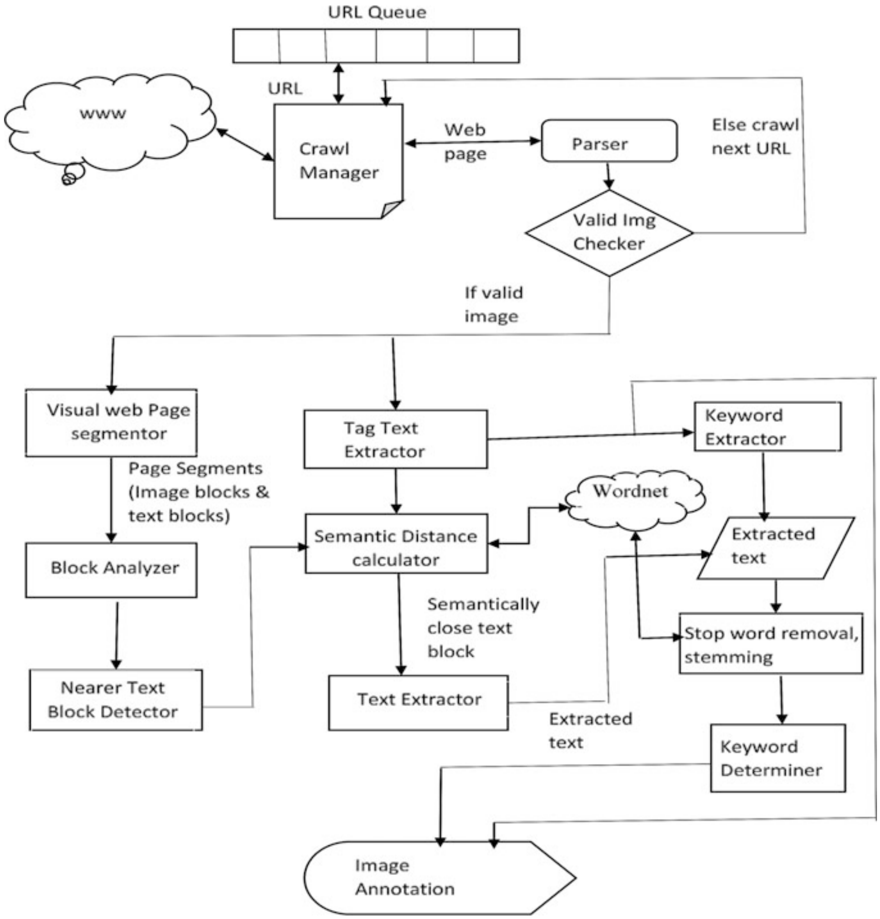

**Fig. 1** Proposed architecture

## *3.2 URL Queue*

URL queue is a type of repository which stores the list of URLs that are discovered and extracted by crawler.

## *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 segmenter; 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.

## *3.4 Visual Web Page Segmenter*

Visual Web page segmenter 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 [12] 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*

Block analyzer analyses the Web content blocks obtained from segmentation. Further, it 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 detects the nearest text blocks to an image block. For checking closeness, Euclidean distance between closest edges of two blocks is calculated. Distance between two line segments is obtained by using Eq. (1):

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

After the distance is calculated between each image block pair 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 does not provide better results, here the words are matched with concepts in a knowledge base and concept to concept matching is computed using WordNet.

Before computing text similarity between image metadata and text blocks, preprocessing of the text blocks is done. After preprocessing process, sentence detection is done. Then tokenization is done, and finally, a part of speech tagging is done for all the words using NLP. At last, stemming of the terms is done and thereafter, terms are mapped to WordNet concepts.

The similarity of text is calculated using Corley's approach. In this method, for every noun (verb) that belongs to image metadata, the noun (verb) in the text of text blocks with maximum semantic similarity is identified according to Eq. 2.

$$\mathrm{sim}_{\mathrm{Lin}} = \frac{2.\mathrm{IC}(\mathrm{LCS})}{\mathrm{IC}(\mathrm{Concept}_1) + \mathrm{IC}(\mathrm{Concept}_2)} \tag{2}$$

Here LCS is the least common subsumer of the two concepts in the WordNet taxonomy, and IC is the information content that measures the specificity for a concept as follows:

$$\mathrm{IC}(\mathrm{concept}) = -\log P(\mathrm{concept}) \tag{3}$$

In Eq. 3, $P$(concept) is the probability of occurrence of an instance of concept in a large corpus. For the classes other than noun (verb), a lexical matching is performed. The similarity between two texts $T1$ (text of image metadata), $T2$ (text of text blocks) is calculated as:

$$\mathrm{sim}(T_1, T_2)_{T_1} = \frac{\sum_{w_i \in T_1} \mathrm{maxSim}(w_i, T_2).idf(w_i)}{\sum_{w_i \in T_1} idf(w_i)} \tag{4}$$

where $idf(w_i)$ is the **inverse document frequency** [19] of the word $w_i$ in a large corpus. A directional similarity score is further calculated with respect to $T1$. The score from both directions is combined into a bidirectional similarity as given in Eq. 5:

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

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

$$\mathrm{dist}_{\mathrm{sem}}(T_1, T_2) = 1 - \mathrm{sim}(T_1, T_2) \tag{6}$$

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 the 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

In this work, keyword determiner is used to extract keywords from the text stored in a buffer. 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—all of these describe the image best.

## 4 Algorithm

The algorithm for proposed system is automatic image annotation. This algorithm takes the URL of Webpage as input and provides the description of the Web page as output.

This algorithm is used here for annotating images present on the Web page. Firstly, parsing is done to extract page title, Img_Src, Alt Text of image. Secondly, Web page segmentation is performed using VIPS algorithm. Then validity of image is checked and for valid images find nearest text blocks using the algorithm given below. For closer text block list, semantic distance is calculated using bidirectional similarity between blocks. Then keywords are extracted from the semantically close text block. These keywords are used for image annotation process.

```
Automatic_Image_Annotation (Description of image)
     Begin
   Parse the Web page (URL)
   If contain valid image
        $Text_1$ = Extract Page_Title, Img_Src(valid image),
        alt( valid image)
        Web_Page_Segmentation (URL)
        For each valid image_block
             Text_block_list = Find_Nearest_text_block
             (ImageBlock Cartesian Coordinates, Text Blocks
             Cartesian Coordinates)
   least_distance = some_big_number;
   For each text block in Text_block_list
   Distance = Find_semantic_distance (Text Block, $Text_1$ )
```

```
If (least_distance >  distance)
{
              Least_distance = Distance
   Return $id_{text}$ ;
      }
      Extract keywords from text block ( $id_{text}$ )
      End
End
```

Algorithm for obtaining nearest text blocks is find nearest text blocks. It takes image blocks and text blocks as input and provides a list of nearest blocks as output. This algorithm collects the nearest text blocks to an image block present on the Web page using closest edge Euclidean distance between Web content blocks. It uses the Cartesian coordinates of Web content blocks to calculate Euclidean distance.

```
Find_Nearest_Text_Block (List of Nearest Text Blocks)
    Begin
    For each Text Block
          {
                Distance  =  calculate Euclidean distance
                between image block and text block
     If (distance  <  threshold)
               {
                        Put the id of text block in a list
                 }
          }
   End
```

## 5 Conclusion

This paper presents algorithm for the novel approach for extracting pertinent keywords for Web image annotation using semantics. In this work, Web images are automatically annotated by determining pertinent keywords from contextual information from Web page and semantic similar content from Web content blocks. This approach provides better results than method of image indexing using Web page segmentation and clustering [21], as in existing method context of image, it is not coordinated with the context of surrounding text. This approach will provide good results as closeness between image and Web content blocks is computed using both Euclidean distance and semantic distance.

# References

1. Sumathi, T., Devasena, C.L., Hemalatha, M.: An overview of automated image annotation approaches. Int. J. Res. Rev. Inf. Sci. **1**(1) (2011) (Copyright © Science Academy Publisher, United Kingdom)
2. Swain, M., Frankel, C., Athitsos, V.: Webseer: an image search engine for the World Wide Web. In: CVPR (1997)
3. Smith, J., Chang, S.: An image and video search engine for the world-wide web. Storage. Retr. Im. Vid. Datab. 8495 (1997)
4. Ortega-Binderberger, M., Mehrotra, V., Chakrabarti, K., Porkaew, K.: Webmars: a multimedia search engine. In: SPIE An. Symposium on Electronic Imaging, San Jose, California. Academy Publisher, United Kingdom (2000)
5. Alexandre, L., Pereira, M., Madeira, S., Cordeiro, J., Dias, G.: Web image indexing: combining image analysis with text processing. In: Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS04). Publisher, United Kingdom (2004)
6. Yadav, M., Gulati, P.: A novel approach for extracting relevant keywords for web image annotation using semantics. In: 9th International Conference on ASEICT (2015)
7. Coelho, T.A.S., Calado, P.P., Souza, L.V., Ribeiro-Neto, B., Muntz, R.: Image retrieval using multiple evidence ranking. IEEE Trans. Knowl. Data Eng. **16**(4), 408–417 (2004)
8. Pan, L.: Image 8: an image search engine for the internet. Honours Year Project Report, School of Computing, National University of Singapore, April, 2003
9. Liu, B.: Web data mining: exploring hyperlinks, contents, and usage data. Data-Centric Syst. Appl. Springer 2007 **16**(4), 408–417 (2004)
10. Fauzi, F., Hong, J., Belkhatir, M.: Webpage segmentation for extracting images and their surrounding contextual information. In: ACM Multimedia, pp. 649–652 (2009)
11. Chakrabarti, D., Kumar, R., Punera, K.: A graphtheoretic approach to webpage segmentation. In: Proceeding of the 17th International Conference on World Wide Web, WWW'08, pp. 377–386, New York, USA (2008)
12. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: VIPS: a vision based page segmentation algorithm. Technical Report, Microsoft Research (MSR-TR-2003-79) (2003)
13. Hattori, G., Hoashi, K., Matsumoto, K., Sugaya, F.: Robust web page segmentation for mobile terminal using content distances and page layout information. In: Proceedings of the 16th International Conference on World Wide Web, WWW'07, pp. 361–370, New York, NY, USA. ACM (2007)
14. Nguyen, H.A., Eng, B.: New semantic similarity techniques of concepts applied in the Biomedical domain and wordnet. Master thesis, The University of Houston-Clear Lake (2006)
15. Voorhees, E.: Using WordNet to disambiguate word senses for text retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference (1993)
16. Landauer, T.K., Foltz, P., Laham, D.: Introduction to latent semantic analysis. Discourse Processes **25** (1998)
17. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: WordNet: An on-line lexical database. Int. J. Lexicogr. **3**, 235–244 (1990)
18. Patwardhan, S., Banerjee, S., Pedersen, T.: 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, pp. 241–257. Springer, Berlin, Heidelberg (2003)
19. Lin, D.: 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, vol. 2, ACL-36, pp. 768–774, Morristown, NJ, USA. Association for Computational Linguistics (1998); Sparck Jones, K.: A Statistical

Interpretation of Term Specificity and Its Application in Retrieval, pp. 132–142. Taylor Graham Publishing, London, UK (1988)

20. Corley, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, EMSEE'05, pp. 13–18, Morristown, NJ, USA, 2005. Association for Computational Linguistics (1998)
21. Tryfou, G., Tsapatsoulis, N.: Image Indexing Based on Web Page Segmentation and Clustering (2014)