Heuristic Method for Automatic Image Annotation in HTML Documents

Jorge Luis Betancourt González[1], Adisleydis Rodríguez Alvarez[2]
[1]*University of Informatics Sciences, Cuba, jlbetancourt@uci.cu*
[2]*Quality Software National Centre, Cuba, aralvarez@uci.cu*

Abstract– An automatic heuristic method for embedded image annotation in HTML documents is exposed. This method exploits the tree structure present in HTML documents trying to identify nodes that contain relevant information about the embedded image, and then using the text in these nearest nodes to expand the information collected about the image, increasing the recall of a Web Search Engine. The proposed heuristic was evaluated using the Agreement Index: the text contained in the identified nodes and the corresponding image was assessed and assigned a category of how well the text was related (i.e. described) with the image. In our test cases the calculated Agreement Index was over 85%, validating the proposed method.

Keywords— image annotation, HTML, information retrieval, search engine, web

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2015.1.1.057
ISBN: 13 978-0-9822896-8-6
ISSN: 2414-6668

Heuristic method for automatic image annotation in HTML documents

Jorge Luis Betancourt González[1], Adisleydis Rodríguez Alvarez[2]
[1]University of Informatics Sciences, Cuba, jlbetancourt@uci.cu
[2]Quality Software National Centre, Cuba, aralvarez@uci.cu

***Abstract*– An automatic heuristic method for embedded image annotation in HTML documents is exposed. This method exploits the tree structure present in HTML documents trying to identify nodes that contain relevant information about the embedded image, and then using the text in these nearest nodes to expand the information collected about the image, increasing the recall of a Web Search Engine. The proposed heuristic was evaluated using the Agreement Index: the text contained in the identified nodes and the corresponding image was assessed and assigned a category of how well the text was related (i.e. described) with the image. In our test cases the calculated Agreement Index was over 85%, validating the proposed method.**

***Keywords*-- image annotation, HTML, information retrieval, search engine, web**

I. Introduction

The Internet has no dearth of content. The challenge is in finding the right content for yourself: something that will answer your current information needs or something that you would love to read, listen or watch. Search engines help solve the former problem; particularly if you are looking for something specific that can be formulated as a keyword query [1].

The web is growing at an increasingly rapid pace. More importantly, faster computers and network connections are allowing creators of web content more freedom to add, with fewer constraints, larger quantities of images, graphics, and video. At the same time, people's interest in using images from the web has also increased [2].

Web Search Engines allow the use of some keywords as a query which is perfectly logic considering that most of the Web pages in the Internet are fill mainly with textual content; Nevertheless image search is a different case, an image is a binary stream which means no textual content that identifies it; is true that the existence of metadata allows to associate some extra information about a particular image, but in must cases this metadata is insufficient to characterize a particular resource. Also, the Web heterogeneous character doesn't allow assuming that all the published images has any of the required metadata.

In a recent study [3], for instance, it was found that in 2001 the keyword "fotos" (photos) was the second most searched keyword in the Chilean search engine TodoCL.

Characterizing the multimedia contents of the web, however, is a challenging technical problem. First, one must deal with huge amounts of distributed data. Second, it is necessary to use media-specific content-based analysis tools to be able to determine the content of the multimedia elements. With images and video, this means developing tools to automatically determine their visual characteristics: color, texture, shape, etc. More interestingly, it implies using algorithms to automatically detect objects of interest (e.g., faces). Obviously, given the large amounts of data, manual classification is not an option [2].

Image retrieval has been a very active research area since the 1970s, with the thrust from two major research communities, database management and computer vision. These two research communities study image retrieval from different angles, one being text-based and the other visual-based [4]. Even in present days taking into account the amount of images and the ever-growing volume of the Web the processing of images to identify the visual attributes can be expensive. In this paper we intent to explode the textual nature of the Web and the dispositional layout of the embedded images, basically there is a tendency to put the text that describes o mention the image near the image object itself.

II. Related work

The HTML language provides semantics mechanisms to describe the information contained in the embedded image. Images in the web are inserted into web pages using the IMG html tag. The attribute alt of the img tag allows us to specify a text alternative to the image, which is automatically displayed when the browser cannot display the image. Some images are included within a hypertext anchor: in this case an image may behave as a button linked to other documents or resources. The text in the alt attribute, along with the text inside the hypertext provides additional information about the image. However a study conducted revealed that only a small fraction of the crawled images contained such attributes [5]. And in many cases this attributes when present are auto generated by the CMS used in the site (if any).

The authors of [6] propose a system to automatic index images crawled from the WWW. A category is assigned to each image based on the text surrounding the image and several extracted visual attributes. In [7] a similar system was built which also incorporate face detection. Our approach combine several ideas found in literature and the adaptation into our particular environment, including several techniques to effectively detect the surrounding text of an image based on

current Web development techniques, but without taking in consideration the visual features of the images. In [8] a joint model is proposed to generate a sequence of words $S = \{S_1, S_2 \ldots S_n\}$ that describe the image. A Recurrent Neural Network is used and some techniques from automatic translation are applied. Although the results reported are promising no information about the performance of the overall system in terms of memory or CPU usage are found. Other inconvenient about this approach is the need of a dictionary with a large collection of terms that will be used to generate the desired description.

Finally in a similar approach using Recurrent Neural Networks is used, the main difference with [9] is that in this approach two sub-networks are used: a deep recurrent neural network for sentences and a deep convolutional network for images, in this particular case the power of moderns GPUs are used.

III. Our Approach

In this paper the terminology presented in [3] is used. The crawler defines a page as an indexed document. A subdomain, intranet.uci.cu (that belongs to the uci.cu domain by instance), identifies a logical Web server. The crawling of the Web is the process, in which web pages and link structure are recollected for later use as the main information source in the search process [10].

The first step in any Web search engine is the crawling process, in this step the crawler extract relevant information out of the Web page content and indexes the extracted data to be later used in the visualization interface.

The crawler need a set of initial URLs to start from, this set of "seed" URLs are then fetched, parsed and outlinks of the pages are extracted and added into the LinkDB; finally the extracted content is stored (indexed) into the indexing system. In our case we use Solr as the indexing backend, inside Solr we have several cores[1], which allow us to have a separated concerns depending on what we are indexing, as shown in Fig. 1.

In the particular case of images we not only store the metadata extracted but also the surrounding text of the image, using the technique proposed in this paper, also the thumbnail of the image encoded as Base64[2] is stored, this is done in order to guarantee a rapid response of the interface when the user does a query. Some of the metadata extracted from the image itself we can find: dimensions, URL, domain, fetch timestamp, inlinks, etc.

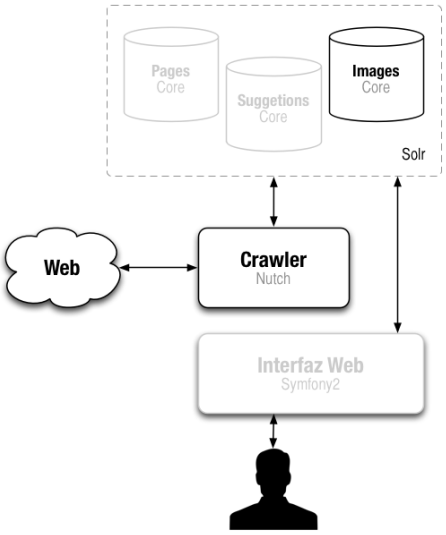

Fig. 1 Architectural overview of a Web Information Retrieval System. Additional components used in the present investigation are shown.

A. Document Object Model

The Document Object Model is a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents [9]. Following this standard the attributes and the text are embedded inside the nodes [9].

The nodes in each HTML document are arranged in a tree structure, known as DOM (DOM Tree). This structure shown in Fig. 2 allows movement in two directions: between nodes in different levels (vertical) and between nodes on the same level (horizontal). This representation provides the opportunity of making changes through certain methods executed in each node [9].

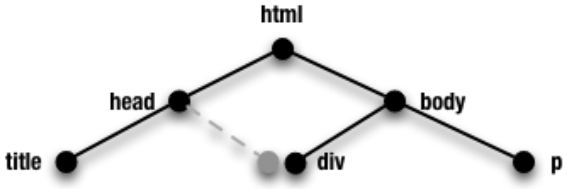

Fig. 2 Tree structure provided by the DOM representation of a HTML document.

Within the boundaries of this framework in the tree structure the different tags of the HTML language are positioned. Some of the tags provided by the HTML language are containers, these containers can hold within other tags (this is the case of the div, span and p tags, for instance) becoming effectively in the root of a sub tree. The img tags, on the other hand, will always be positioned as leafs within any DOM document, because the img tag cannot act as a container.

Taking into account the structure previously described we can think that the sibling nodes of an img node, i.e. those nodes in the same level (with a certain degree of "closeness")

[1] A Solr *core* is a logical separation inside the same Solr instance that has a unique structure. In our case we use one core to store only images, one to store HTML, PDF, and other more general formats.

[2] Binary to text codification scheme that represents the information as an ASCII string.

that holds textual information can describe or contain information related to the image itself (Fig. 3). These sibling nodes are very important to identify and to keep related with the image, increasing the probability of a match between the queries introduced by user and the indexed text.

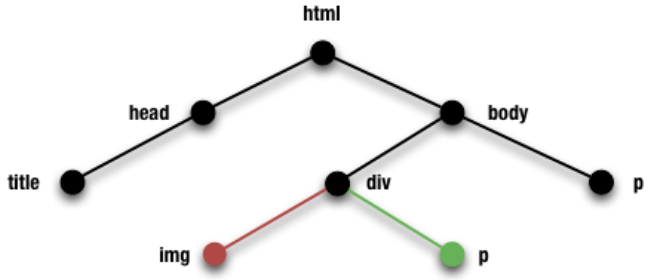

Fig. 3 The p (paragraph) node has a very high probability of containing text related to the img (image) node.

It's very important to highlight that due to the freedom provided by the HTML language, the previous statement is not always true.

A tree is defined as a graph $T = (V, E)$ where elements of V are the vertices and elements of E are the edges; is T is minimally connected: $(T - e)$ provides a non connected graph for each edge $e \in T$, not posses any cyclic and is maximally acyclic: $(T + xy)$ provides a graph with cycles for any pair of non adjacent vertices $x, y \in T$; if this conditions are met, then we can conclude that the graph T is a tree. In our case the vertices correspond to the HTML tags and the edges represent the hierarchical relation between the tags.

As consequence a node I of image can be associated to a set of nodes H in the same level that can be found at a maximum distance of N_h. Extrapolating this into the vertical dimension, also a set of nodes V of superior levels that can be found at a maximum distance of N_v can be related to the node I, in short: $I \leftarrow H \cup V$.

B. Node selection

Taking into consideration the previous section an important question arises: How many nodes (n) can be associated with an image? As shown in [11] where a similar approach is applied the absorption of attributes for an ever-increasing number of close nodes doesn't improve the quality of the classification, no relevant information can be extracted from these nodes.

In our case can be convenient to use two separated variables to control the number of nodes to be related with the image in both directions (horizontal and vertical). This differentiation probed to be useful in our tests, starting from the tree representation of the DOM is logical to assume that when you include the textual information of a node in an upper level, the textual information of the entire sub-tree gets added to the metadata extracted from the image itself.

As shown in Fig. 4 even the election of a small number for the maximum level of vertical nodes (N_v) to relate with the img node can have a big impact. In this case ($N_v = 2$) the highlighted nodes will be related to the img node, which are far less probable to be related to the image. Which can cause that the same image matches a very diverse criteria set hurting the precision of the system.

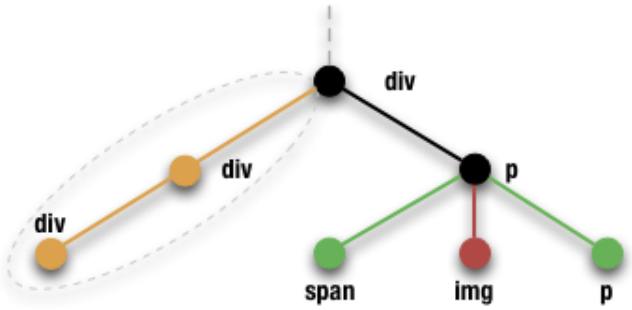

Fig. 4 Example of vertical node selection associated with an image.

A similar problem can be found in the variable that control the number of nodes in the same level (N_h) although the noise introduced in this tends to be lower than in the vertical direction, even if the same value is used for both variables. If we increase in k the amount of vertical nodes to consider we are adding actually the information of S sub-trees: $S = \sum_{i=0}^{k} d(n_k)$ where $d(n_k)$ is known as the node valence or node degree.

Fig. 5 shows the selected nodes to be associated with the img node in a case where both variables N_h and N_v has a value of 2.

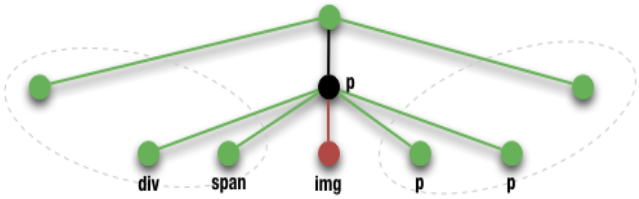

Fig. 5 Nodes related to the img node using the proposed algorithm.

In our particular case only the textual value of the p nodes, div nodes and span nodes were associated, as a technique of reducing the noise introduced by the variety of HTML tags.

IV. Results

The evaluation of the implemented system presents its own set of challenges. First of all the absence of a controlled collection of the Cuban Web the use of the precision and recall measures are ineffective which brings the need to use some alternative measures. For this investigation a discrepancy metric was selected, this kind of metrics are designed to measure the different points of view of a set of

judges that will evaluate how well the extracted text using the proposed algorithm describes the associated image. The answer to the previous questions should lead the validation of the proposed algorithm.

A. *Kappa statistic.*

A first option to measure the agreement level between judges is the Kappa statistic [12], a statistical measure between judges:

$$k = \frac{P - P_e}{1 - P_e} \tag{1}$$

The kappa statistic is defined as the difference between how consensus is actually present $(P - P_e)$ and the random expected cohesion value $(1 - P_e)$. P is the consensus among judges and P_e is the probability of random cohesion. In particular the Kappa de Fleiss statistic [13] [14] is used, which is a variant to the Cohen proposal applied to a fixed number of judges and a fixed number of documents. The scale presented in Table 1 was used, similar to that presented by [15] for the interpretation of the obtained value.

TABLE I
INTERPRETATION SCALE FOR THE KAPPA STATISTIC

Value of Kappa	Consensus among judges
< 0	No possibility of agreement
0.01 – 0.20	Light
0.21 – 0.40	Considerable
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect

Considering our three possible classifications (good, regular, bad) we observed a kappa value of 0.64 that can be interpreted as a substantial agreement between judges. A recurrent difficulty to assign the "regular" category was observed on the feedback received from the volunteers. Taking this into account, and restricting the category set to only good and bad the kappa value rises until 0.82. This is due to the fuzzy nature of the "regular" category, and the evaluation made by each judge. The obtained values reflect a good consensus among the judges, meaning that the extracted text for each image describes quite well.

B. *Agreement Index*

Another form of evaluating the disagreement among judges is using the Agreement Index. To use this metric a cost matrix needs to be defined:

TABLE II
COST MATRIX FOR THE AGREEMENT INDEX METRIC

	Bien	Regular	Mal
Bien	0	0.5	1
Regular	0.5	0	0.5
Mal	1	0.5	0

This matrix reflects the cost of the difference in the evaluations given by the judges, the goal is to provide a higher cost for those differences in evaluation that are less likely i.e. the more "different" the evaluation between the judges is, higher the cost should be. In this particular case if a judge provides an evaluation of "good" and other provides an evaluation of "bad" for the same test case, then the disagreement level is higher.

Within this framework given a pair of judges i y l the agreement index $A_j(i,l)$ in the subset of images $S_i \cap S_l$ for each $j \in S_i \cap S_j$ is defined as:

$$A_j(i,l) = 1 - \text{cost}(a,b) \tag{2}$$

In the previous equation a is the label (classification) provided by the judge i and b is the label provided by the judge j.

Using (2) the Agreement Index of the pair of judges (i,l) can be defined as:

$$AI(i,l) = \sum_{j \in S_i \cap S_j} A_j(i,l) \Big/ |S_i \cap S_j| = \sum_{j \in S_i \cap S_j} A_j(i,l) \Big/ O(i,l) \tag{3}$$

$O(i,l)$, is defined as the number of overlapping images classified by both judges. The average, maximum and minimum values of the Agreement Index for increasing values of $O(i,l)$, basically for each value x of overlap the subset of pair of judges (i,l) were restricted so that $O(i,l) \geq x$.

In Fig. 6 Agreement Index over number of overlapping classified images., the calculated agreement index for values of $x \leq 70$ was plotted and we can see the evolution of the calculated values.

Fig. 6 Agreement Index over number of overlapping classified images.

The average agreement index is never above 89% and never less than 73%. In addition, the agreement index does not

appear to increase with the number of overlapping images. In fact, begins to decrease to values greater than 50 overlapping images, when the number of pairs of reviewers over which the average is calculated is still relatively low (between 7 and 10). This result should probably be confirmed in subsequent larger studies, but suggests that a non negligible number of disagreement between reviewers can not be the result of statistical noise, but due the ambiguous nature regarding the categories used; that translates into ambiguity in deciding whether or not the selected text describes the associated image and can be solved by selecting more disjoint categories.

V. Conclusions

The analysis of the collected metrics during the experiment showed that most pairs of reviewers had little overlap, so it did not evaluated the same images; this suggests that this experiment can be performed on a larger scale in order to cover a larger amount of collected images, so that the overlap between each pair of judges be greater. Further analysis also showed that despite the above statement, many pairs of reviewers significantly overlapped their ratings; at least enough to reach some conclusions according to their agreement index, demonstrating that they agreed on the assigned evaluation.

The high percentage of the agreement index shows that the selected heuristic increases the system's accuracy without introducing too much ambiguity. The method outlined in this article, focusing on textual attributes for annotation of images can be extensible to other formats present on the Web. Also, do not limit or interfere with the application of image processing methods for identifying objects and other advanced techniques, constituting a cheaper alternative in terms of computing power to other methods.

References

[1] Mayur Datar, Ashutosh Garg, Shyam Rajaram Abhinandan Das, "Google News Personalization: Scalable Online Collaborative Filtering," *WWW 2007 / Track: Industrial Practice and Experience*, 2007.

[2] A. Jaimes et al., "On the Image Content of the Chilean Web," *Proceedings of the First Latin American Web Congress (LA-WEB 2003)*, 2003.

[3] B. J. Poblete, F. Saint-Jean R. Baeza-Yates, "Evolución de la Web Chilena 2001-2002 (Evolution of the Chilean Web 2001 - 2002)," 2003.

[4] Thomas S. Huang, Shih-Fu Chang Yong Rui, "Image Retrieval: Current Techniques, Promising Directions, and Open Issues," *Journal of Visual Communication and Image Representation*, pp. 39-62, 1999.

[5] Ricardo Baeza-Yates, J. Ruiz-del-Solar, R. Verschae, C. Castillo, and C. Hurtado, "Content-based Image Retrieval and Characterization on Specific Web Collections," 2002.

[6] M.J. Swain, V. Athitsos C. Frankel, "“WebSeer: An Image Search Engine for the World Wide Web," 1996.

[7] S.-F. Chang J.R. Smith, "An Image and Video Search Engine for the World-Wide Web," *Proc. of SPIE Storage & Retrieval for Image and Video Databases V*, vol. 3022, pp. 84-95, 1997.

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator," *ArXiv e-prints*, November 2014.

[9] W3C. (2013, Aug.) W3C. [Online]. http://www.w3.org

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *An Introduction to Information Retrieval*.: Cambridge University Press, 2008.

[11] Soumen Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*.: Morgan Kaufmann Publishers, 2003.

[12] J. Cohen, "A coefficient of agreement for nominal scales," *Psychological Bulletin*, vol. 20, pp. 37-46, 1960.

[13] Annette M. Green, "Kappa statistics for multiple raters using categorical classifications," *Proceedings of the Twenty-Second Annual Conference of SAS Users Group, San Diego, USA*, 1997.

[14] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, pp. 378-382, 1971.

[15] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159-174, 1977.

[16] Junhua Mao, Wei Xu, Yi Yang, and Alan L Yuille, "Explain Images with Multimodal Recurrent Neural Networks," *ArXiv e-prints*, October 2014, http://adsabs.harvard.edu/abs/2014arXiv1410.1090M.