# A Clustering-based Approach to Web Image Context Extraction

Sadet Alcic and Stefan Conrad
*Institute for Computer Science, Databases and Information Systems*
*Heinrich-Heine University, D-40225 Duesseldorf, Germany*
{*alcic,conrad*}*@cs.uni-duesseldorf.de*

***Abstract*—Images on the Web come along with textual descriptions that are valuable for different applications, such as image annotation, clustering of images, image categorization, etc. But usually Web pages are poorly structured and cluttered with contents of different topics, which hinder the accurate detection of the image context. Existing approaches are based on heuristic rules and thus cannot handle the variety of documents on the Web. In this paper, we introduce a novel approach to image context extraction, building on a Web content distance measure. Utilizing this distance measure, the addressed problem can be reduced to a content clustering problem where an image is associated with the textual contents of the cluster it belongs to. Our evaluation studies confirm the validity and quality of the proposed method and demonstrate its applicability to the Web.**

***Keywords*-Image Context Extraction, Web Content Mining.**

## I. INTRODUCTION

After text, image is the most basic and commonly used content on the Web. But while text semantics can be extracted from text directly, the automatic detection of image semantics is still an open issue. Considering images on the Web, we recognize a valuable advantage over isolated images: these images come in hand with other textual information on hosting Web pages that can be exploited to describe the images.

However, Web documents are usually cluttered with multi-topic contents, while at same time they do not separate these contents by explicit structure. As a consequence, the problem of estimating the image context as a (hidden) subset of the complete Web page arises. As *image context*, we understand the textual contents of a Web page that share the semantics with an image on this page.

Different parts of a Web page can be considered as possible sources for image context, namely, image url, page title, alternative text (ALT attribute), passages of surrounding text, etc. The first three have been utilized in many approaches [1], [2] due to their easy extraction and promising accuracy to describe the embedded image. However, different researchers inferred in independent empirical studies [3], [4] that these context sources do not describe images satisfactorily. The main reasons are: filenames are often generated (such as img1.jpg); the page title is mostly to general (e.g., New York Times - News); alternative text is hardly available. In comparison, the descriptions derived from passage of text surrounding an image were more reliable.

In recent years, three general approaches have been proposed to extract these passages, (i) a fixed-size window of terms [5], [4], [3], (ii) DOM tree wrappers [6], [7], (iii) content blocks derived by Web page segmentation [8], [9]. While efficient in time, the fixed-window approach is prone to precision as well as recall errors, since the extracted passage can include irrelevant content, or respectively, exclude relevant content. Wrappers are based on heuristic rules that only can cover a small subset of the possible design patterns for Web pages. Each time these patterns change, wrappers have to be adapted manually. Web page segmentation is a more principal approach to estimate the image context by associating images with the textual contents of common segments. However, most of the existing page segmentation algorithms are not designed for image context extraction and deliver to broad or to narrow segments, which affect directly the quality of context information. Due to their complexity, it is difficult to adapt existing page segmentation methods to meet the requirements of image context extraction [10], [7].

Recognizing the shortcomings of existing approaches, we present a more general solution to extract web image context by mining the Web contents of a page based on the underlying DOM tree structure. To make our approach applicable to the Web, we abstain from using visual features of contents, which are very time consuming since they need a Web page to be rendered.

**Contribution.** The main contribution of this work is threefold. First, we introduce a novel distance metric for Web contents based on the hierarchical structure of the DOM tree. Using this metric allows to map the contents of a Web page in a one-dimensional (1D) space. As a result, the image context extraction problem is reduced to finding context separators in 1D space. Secondly, we introduce a solution for the reduced problem by proposing a generic threshold-based clustering algorithm, which exploits the distance of adjacent contents. And finally, we evaluate the proposed method and compare its effectiveness to common approaches from the literature.

**Organization.** This paper is organized as follows. Section II gives a brief overview to related work. Section III introduces by example the different structural clues that can be obtained from HTML and afterwards presents our DOM-based dis-

 ISBN: 978-1-61208-129-8 

tance metric, which is based on the preliminary thoughts. In Section IV, the clustering-based method to image content extraction is presented. Finally, the approach is evaluated in Section V and results are discussed.

## II. Related Work

Many researchers were attracted by the benefits of Web image context in the past. As a result, a variety of context extraction methods, ranging from simple heuristics-based approaches to complex DOM and vision-based extractors have been proposed. Approaches like [5], [4], [3] extract a paragraphs of $n$-terms ($n$ is chosen different, e.g., 10, 20, 32) surrounding the image as context. While this approach is fast and simple, it is prone to errors, i.e., when the image context is placed only under the image.

Tian et. al [6] propose a DOM-based method where the image context is selected by extracting the textual contents of the sibling nodes. Starting at the image node, the DOM tree is traversed upward until a parent node has text nodes. These are then assigned as image context. Fauzi et al. [7] distinguish three different use cases for images in HTML documents: listed images, semi-listed images, and unlisted images. To each case, context extraction rules are defined based on DOM tree.

Cai et al. [9], [8] use Vision based Page Segmentation (VIPS) [11] to partition Web documents into visual blocks. Images are assigned the text of the common visual block. VIPS is an hierarchical top-down approach, which starts with the whole page as initial block. For each block, a Degree of Coherence (DoC) is computed using heuristic rules based on the DOM Tree structure and visual cues obtained from the browser representation. The DoC value determines to what degree the contents within a block correlate to each other. It ranges from 1 to 10, while 10 represents the highest correlation. At the beginning a Permitted Degree of Coherence (PDoC) value is specified (set to 5 in [8]), which controls the segmentation granularity. If a particular block has a DoC value smaller than PDoC, this block has to be subdivided and this rule is repeated until all blocks on the bottom fulfill the mentioned condition. Hattori et al. [12] define a distance function that computes a distance between contents based on structural depth of HTML tags and performs top-down segmentation applying the proposed content distance function. However, their content distance does not satisfy our needs by two reasons: (i) the distance measure actually does not reflect the distance values that correspond to the HTML structure; (ii) the triangle inequality is not met, which is very important for our clustering-based approach. There is a variety of other approaches to page segmentation, i.e. [13], [14], but since they were developed for other applications, their adaptation to image context extraction is of high complexity.

## III. Structural Information in Web pages

In a DOM tree of a Web document, we distinguish two kinds of elements: inner nodes and leaf nodes. The leaf nodes represent basic content units of a Web page, thus image as well as text nodes are elements of this kind. They are arranged from left to right in the order as they appear in the document source. On the other hand, the inner nodes correspond to tags that define the structural as well as functional properties of the contents in their subtree. They further group the underlying contents to DOM blocks. All these hints can be utilized to estimate a structural distance of the content units, which will be motivated by an example.

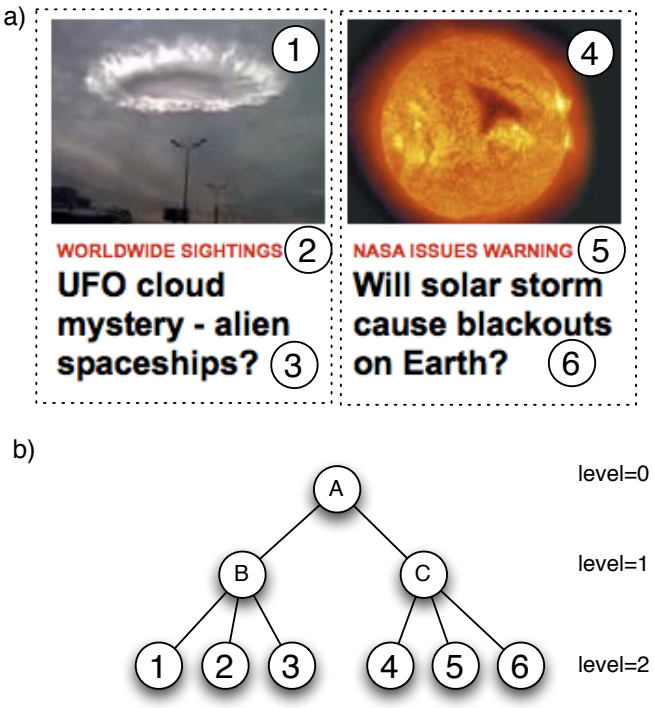


Figure 1. Example snipped of a Web page: a) shows the visual representation and b) the simplified DOM tree.

Figure 1 contains a small excerpt of a Web page and a simplified version of its corresponding DOM subtree. From both representations, we can simply infer that contents 1 - 3 form a structural block, and respectively contents 4 - 6 do the same.

This example can now be used to introduce the basic ideas for a DOM-based distance measure. Starting with the contents 1 and 2, we may set their distance to

$$d(①,②) = c,$$

where $c > 0$ is an arbitrary constant. Because contents 2 and 3 are on the same level and both under the same parent, we set the distance

$$d(②,③) = c.$$

However, between content 1 and content 3 there is content 2 and thus the distance between content 1 and 3 can be computed transitively

$$d(①,③) = d(①,②) + d(②,③) = 2c.$$

The same rules can be applied in the right subtree resulting in following three equations:

$$d(④,⑤) = c; d(⑤,⑥) = c; d(④,⑥) = 2c.$$

The only missing distance is that one between contents 3 and 4. As these elements belong to different blocks, we must ensure that their distance is greater than the maximum distance of contents in the left or right block. In this example, this maximum distance of siblings on level 2 is $2c$ and therefore the distance between content 3 and 4 might be set to

$$d(③,④) = 4c,$$

which is a high distance, that separates the blocks at this position. The described distance measure can be further used to map content units of a Web page in a 1D space, simply by setting the point of origin at content 1 as depicted in Figure 2.

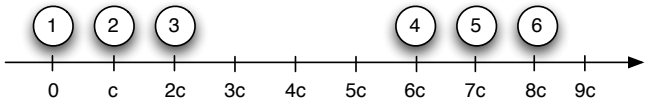


Figure 2. Six content elements spread over 1D space by proposed distance metric.

Following the ideas in this example, we formulate a general DOM distance metric, which is applicable to any DOM tree.

**Definition: DOM Distance.** Let $d$ be a Web document and $D_d(N, E)$ the corresponding simplified DOM tree, with $N$ as a set of nodes and $E$ as a set of edges. Further let $P = (p_1, ..., p_n)$ be the sequence of all nodes in $N$ ordered by traversing $D_d$ in preorder traversal. The index $i \in \{1, ..., n = |N|\}$ gives the order of $p_i \in N$ in the sequence $P$. Based on this formulation, we define the *DOM Distance Metric* $d : N \times N \rightarrow \mathbb{R}$ for two nodes $p_a, p_b \in N, 1 < a < b < n$ (if $b < a$, wlog. switch $p_a, p_b$) as follows:

$$d(p_a, p_b) = \sum_{i=a+1}^{b} w_{p_i}, \tag{1}$$

where $w_{p_i}$ is the *block weight* of element $p_i$, which has to be further specified with the knowledge from Section III. In general, the metric $d$ is a sum of block weights.

The block weights $w_{p_i}$ can be explained as the cost needed to reach the content block of $p_i$ from its predecessor $p_{i-1}$. Therefore $w_{p_i}$ corresponds to the distance $d(p_{i-1}, p_i)$ of two neighbored contents $p_{i-1}, p_i$. For nodes on the same tree level $l$, the block weights are all equal. On a lower level $l-1$ the block weight has to be at least greater than the maximal distance of two sibling nodes at level $l$. The maximal distance of two sibling nodes on level $l$ corresponds to the degree of the nodes on level $l-1$, which is the maximal count of children of the nodes at level $l$.

With these considerations, the block weight function $w : \mathbb{N} \rightarrow \mathbb{R}$ can be formulated by following recurrence:

$$w(l) = \begin{cases} c & : \quad d_l = 0 \\ d_l \cdot w(l+1) & : \quad d_l > 0, \end{cases} \tag{2}$$

where $d_l$ refers to the maximal degree of nodes at level $l$ and $c$ is an arbitrary constant value. To apply $w$ in Equation 1, we define function $l : N \rightarrow \mathbb{N}$ that delivers the tree level of a node. Thus the block weight $w_p$ for a node $p \in N$ is $w_p = w(l(p))$.

To demonstrate the applicability of the defined distance, we consider again the example from Figure 1. First, we determine the preorder sequence of the nodes:

$$Ⓐ - Ⓑ - ① - ② - ③ - Ⓒ - ④ - ⑤ - ⑥$$

Secondly, we compute the level weights $w_l$ according to Equation 2. For level 2 we get $w_2 = c$ since the level degree $d_l$ at level 2 equates to 0. At level 1 we have $d_1 = 3$ and consequently $w_1 = w_2 \cdot d_1 = 3c$. For level 0 the level weight equates $w_0 = w_1 \cdot d_0 = 3c \cdot 2 = 6c$ . After this initialization steps, the distance between contents can be computed by summing up appropriate weights as defined in Equation 1. We can verify, that the distances correspond to that assumed in our preliminary thoughts.

**Time complexity.** The proposed distance measure consists of two steps, an initialization step, which has to be executed only once for a complete Web page, and the distance computation step. The initialization step includes a pre-order traversing of the DOM tree consuming linear time depending on the *total number of nodes* $n$, and further the weights computation for $l$ levels while $l$ equates in average $log(n)$. Thus the time complexity for initialization is in $O(n + log(n)) = O(n)$. The distance computation includes a sum over $n$ node weights at maximum and thus can also be computed in $O(n)$. By an additional step in initialization, where we map all nodes in the metrical space derived by the proposed metric, we can minimize the effort for distance computation to one subtraction computable in O(1). The additional effort costs $O(n)$ and thus does not affect the time complexity of initialization.

## IV. Image Context Extraction

In this section, we will present our proposed method to Web context extraction by clustering contents in 1D space supplied by the proposed distance measure.

 ISBN: 978-1-61208-129-8 

**Problem Formulation.** Given an image $I$ in a Web document $d$, Web Image Context Extraction (WICE) denotes the process of determining the textual contents $t_i$ of document $d$ that are associated with the image $I$. The proposed DOM distance maps the basic content units of a Web page to a 1D space. By setting cuts at appropriate positions in this space, contents are partitioned (or clustered) to content blocks. The image $I$ can now be associated with the textual contents $t_i$ of the block, $I$ belongs to. Thus the WICE problem is reduced to clustering in 1D space, or in other words, to estimating suitable positions in this space to separate contents.

**1D-Clustering based on Distance Threshold.** The idea to a clustering method for 1D data points will be motivated by the example page excerpt from Section III. Consider the simple one-dimensional optimization problem in Figure 2: we want to find a good clustering for the six data points so that the variance of the distances between each pair of adjacent points in same cluster is minimized. It is not hard to understand that by cutting at the largest distance between a pair of adjacent points, we will find a good solution to the clustering problem. Actually, if we are able to define threshold that the distance of adjacent contents should not exceed, the clustering can be done as described in Algorithm 1.

**Input**: Sequence of Web contents $S = (s_1, ..., s_n)$, threshold $t$
**Result**: Set of computed clusters $C$
$c$ = newCluster($s_1$);
**for** $i = 2$ *to* $n$ **do**
  **if** $d(s_{i-1}, s_i) > t$ **then**
    $C$.add($c$);
    $c$ = newCluster($s_i$);
  **else**
    $c$.add($s_i$); ;
  **end**
**end**

**Algorithm 1**: 1D-clustering in by thresholding

The algorithm starts by initializing a new cluster $c$ with the first element $s_1$. Then a loop iterates over the sequence $S$, in which the elements are ordered by their appearing in the document, and computes the distance between every pair of adjacent contents. If the computed distance is greater than a predefined threshold $t$, the actual cluster $c$ is put in the set of clusters $C$ and $c$ is initialized again with content $s_i$. Otherwise, content $s_i$ expands the actual cluster $c$.

**Threshold estimation.** A static threshold value $t$ could be computed by averaging over all distances of adjacent content pairs:

$$t = \frac{1}{n-1} \sum_{k=2}^{n} d(s_{k-1}, s_k).$$

This baseline threshold might work well for Web documents that consist of content clusters with similar density. However, since Web contents are usually distributed over different levels, the cluster density can significantly differ among different clusters. Thus the threshold should be more adaptive to distances in the environment.

To meet these requirements, we propose to use a gaussian weighted threshold function $t : \mathbb{N} \rightarrow \mathbb{R}$:

$$t(k) = \frac{1}{\sum_{i=2}^{n} G(i, k, \sigma)} \sum_{i=2}^{n} G(i, k, \sigma) d(s_{k-1}, s_k),$$

with the gaussian function $G(x, \mu, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$. The remaining parameter $\sigma^2$ is the variance (the measure of the width of the gaussian peak) and has to be considered empirically.

**Visual Representation.** To give a more intuitive description to the proposed algorithm, we visualize its main components. The solid-line curve in Figure 3 refers to the distances of adjacent Web contents, while the values on the $x$-axis correspond to the index of the contents in the contents sequence (e.g., $x$-value 280 means the distance of adjacent contents 280 and 281). The dashed-line curve in the same plot is the gaussian smoothed version of the red function, and corresponds to the threshold $t$. For each peak of red (distances) function exceeding the blue (threshold) function, we have drawn a circle at this function value and pointed with an arrow to the corresponding position in the browser representation of the document. This example shows empirically the quality of our method, since all blocks were properly recognized.

## V. EVALUATION

The accuracy of the proposed method was evaluated using the evaluation framework proposed in [10]. The ground truth data consists of different test collection gathered from real Web servers. Table I comprises the main information about the test collections. The *diverse* collection consists of 79 documents for which the context extraction was performed manually. The other collections were created by recalling the main page of a Web site and storing the gathered document whenever a significant change of the content to the previously stored was detected. In this way, we collected a large amount of documents based on the same template. A rule based extractor was then implemented for each collection, that extracted the image context. This resulted in a large amount of real testdata, consisting of 12,907 documents is and 155,565 context to image pairs.

**Quality measure.** In order to compare the extracted image context with the ground truth data we understand both texts as sequences of words and measure their overlap by computing the longest common subsequence. By treating

 ISBN: 978-1-61208-129-8

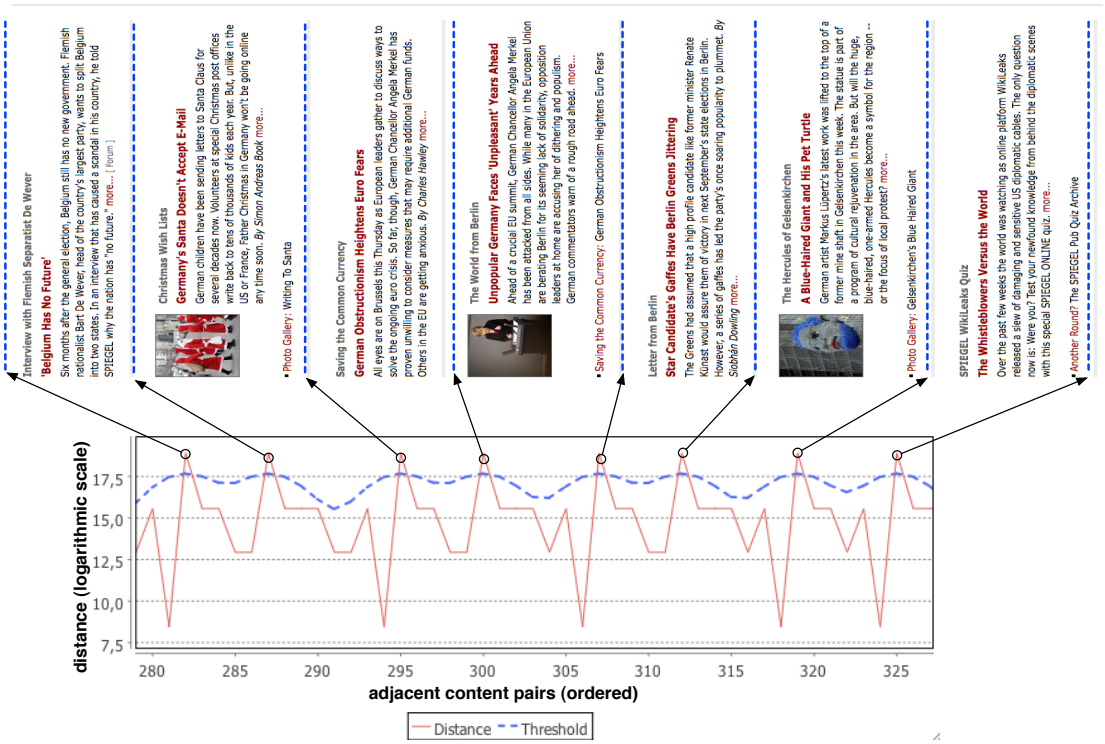


Figure 3. Distances of adjacent contents and the threshold values computed by weighted averaging with gaussian. The peaks at which the distance exceeds the threshold point to the corresponding separators in the browser output of the document.

Table I
TEST COLLECTIONS WITH TOTAL NUMBER OF DOCS AND IMAGES.

| Collection | #Documents | #Images |
|---|---|---|
| BBC | 1077 | 7878 |
| CNN | 874 | 11612 |
| Golem | 789 | 3061 |
| Heise | 79 | 1403 |
| MSN | 375 | 9264 |
| New-York Times | 556 | 10927 |
| Spiegel | 1076 | 36310 |
| Telegraph | 530 | 10503 |
| The Globe and Mail | 735 | 15808 |
| Wikipedia | 3000 | 6728 |
| Yahoo! (english) | 3737 | 41170 |
| diverse (manual) | 79 | 901 |
| total | 12907 | 155565 |

Table II
EVALUATION RESULTS SHOWING THE AVERAGE F-SCORES OF CONTEXT EXTRACTION METHODS ON DIFFERENT COLLECTIONS.

| | 10 terms | 20 terms | monash | siblings | full text | vips 5 | vips 6 | vips7 | DOM-dist |
|---|---|---|---|---|---|---|---|---|---|
| cnn | 0,40 | 0,27 | 0,75 | 0,91 | 0,02 | 0,16 | 0,28 | 0,29 | **0,92** |
| golem | 0,51 | 0,62 | **0,96** | 0,95 | 0,10 | 0,15 | 0,47 | 0,47 | 0,95 |
| heise | 0,43 | 0,52 | 0,93 | 0,95 | 0,03 | 0,31 | 0,77 | 0,77 | **0,96** |
| msn | 0,40 | 0,47 | **0,95** | 0,93 | 0,04 | 0,16 | 0,23 | 0,23 | 0,89 |
| nytimes | 0,41 | 0,45 | 0,86 | 0,76 | 0,03 | 0,16 | 0,46 | 0,63 | **0,89** |
| spiegel | 0,45 | 0,40 | 0,90 | 0,84 | 0,03 | 0,10 | 0,25 | 0,19 | **0,98** |
| telegr. | 0,63 | 0,61 | 0,92 | 0,23 | 0,03 | 0,12 | 0,62 | 0,80 | **0,93** |
| gl.&m. | 0,55 | 0,50 | 0,94 | 0,97 | 0,03 | 0,22 | 0,35 | 0,49 | **0,99** |
| wiki | 0,42 | 0,33 | 0,92 | **0,94** | 0,02 | 0,07 | 0,66 | 0,32 | 0,89 |
| yahoo | 0,54 | 0,59 | 0,91 | 0,49 | 0,04 | 0,05 | 0,14 | 0,26 | **0,89** |
| diverse | 0,41 | 0,41 | 0,81 | 0,80 | 0,05 | 0,21 | 0,36 | 0,36 | **0,85** |
| overall | 0,47 | 0,48 | 0,90 | 0,80 | 0,04 | 0,15 | 0,42 | 0,45 | **0,93** |

the extracted context as retrieved data and the ground truth as relevant data we can apply standard information retrieval concepts of *precision* $P$, *recall* $R$ and *F-score* 1 as performance measures.

**Parameter estimation.** Our proposed method to context extraction has one open parameter $\sigma^2$ that has to be specified. $\sigma^2$ is part of the gaussian smoothing kernel and considers its variance. In order to avoid overfitting, the parameter was trained iteratively on a smaller subset of our testdata consisting of five documents of each collection. The maximal average F-score was reached when $\sigma^2$ was set to 2.25.

**Results.** Table II contains the average F-scores computed on different test collections. Besides the proposed algorithm, we have extracted image context with two DOM-based methods – Monash [7] and [6] Siblings; a heuristics based method –

N-terms [5], [4]; and a vision-based method – VIPS [11]. To show the benefit of extraction methods, we have further included a full text extractor in the evaluation as baseline.

As $N$ in the $N$-terms extractor we chose 10 and 20 since these are the frequently used parameters in the literature. The PDoC value of the VIPS algorithm was set to 5, 6, and 7 during the evaluation. As observable, there are some results missing for VIPS on the bbc collection. The reason for this lies in the implementation of the VIPS library that is based on the Internet Explorer 6 (IE6). However, IE6 was not able to properly display the crawled documents from the BBC Web page due to javascript errors.

As a first results, we find out that the baseline extractor has a significantly low performance compared to other extraction methods. This is because in most Web documents the length of the image context is significantly smaller than the length of the full text. Therefore it is worth investigating image context extraction algorithms.

The heuristics-based methods extracting the text within a frame of $N$ terms surrounding the image achieve both F-scores around $0,5$. A possible reason is the fact that images are often placed next to the borders of articles. As a consequence, half of the associated text does not belong to the image. Both parameters deliver similar results and thus no one can be preferred.

VIPS is traditionally a page segmentation algorithm that was frequently used for context extraction in the past. However, the performance of VIPS ranges over the lower third. While the segments that VIPS extracts with a PDoC value of 5-7 are too broad, higher PDoc values yield to segments that contain only the image and no text.

The DOM-based methods – monash and siblings, as well as our proposed method – perform best on all collections. While the F-score of the siblings method varies for different test collections, the other two reach constantly high values with a small advantage for our method compared to the monash extractor.

## VI. Conclusion and Future Work

This work presents a new method for image context extraction based on the distances of Web contents. Distance is computed using structural clues from the document. Using this distance function, the complex problem of image content extraction is reduced to the familiar 1D clustering. The results of the evaluation task show that the proposed method delivers highest accuracy on almost all test collections. As future work, other traditional clustering approaches could be applied to the 1D clustering problem. Further, we plan to estimate the impact of our method to applications, like image ranking or image classification.

## References

[1] J. R. Smith and S.-F. Chang, "Image and Video Search Engine for the World Wide Web," in *Storage and Retrieval for Image and Video Databases (SPIE)*, 1997, pp. 84–95.

[2] H. T. Shen, B. C. Ooi, and K.-L. Tan, "Giving meanings to www images," in *Proceedings of the eighth ACM international conference on Multimedia*, ser. MULTIMEDIA '00. New York, NY, USA: ACM, 2000, pp. 39–47.

[3] 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*, ser. MULTIMEDIA '04. New York, NY, USA: ACM, 2004, pp. 960–967.

[4] T. A. S. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, and R. Muntz, "Image Retrieval Using Multiple Evidence Ranking," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 408–417, 2004.

[5] S. Sclaroff, M. L. Cascia, and S. Sethi, "Unifying textual and visual cues for content-based image retrieval on the World Wide Web," *Computer Vision and Image Understanding*, vol. 75, no. 1-2, pp. 86–98, 1999.

[6] T. Yong-hong, H. Tie-jun, and G. Wen, "Exploiting multi-context analysis in semantic image classification," *J. Zhejiang Univ. SCI. 6A(11)*, pp. 1268–1283, 2005.

[7] 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.

[8] 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.

[9] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, New York,USA, 2004, pp. 952–959.

[10] S. Alcic and S. Conrad, "Measuring performance of web image context extraction," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, ser. MDMKDD '10. New York, NY, USA: ACM, 2010, pp. 8:1–8:8.

[11] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "VIPS: a Vision-based Page Segmentation Algorithm," Microsoft Research (MSR-TR-2003-79), Tech. Rep., 2003.

[12] G. Hattori, K. Hoashi, K. Matsumoto, and F. Sugaya, "Robust web page segmentation for mobile terminal using content-distances and page layout information," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 361–370.

[13] C. Kohlschütter and W. Nejdl, "A densitometric approach to web page segmentation," in *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 1173–1182.

[14] D. Chakrabarti, R. Kumar, and K. Punera, "A graph-theoretic approach to webpage segmentation," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 377–386.

 ISBN: 978-1-61208-129-8 