Model-driven Web Page Segmentation for Non Visual Access

Judith Jeyafreeda Andrew, Stéphane Ferrari, Fabrice Maurel, Gaël Dias, Emmanuel Giguet

Model-driven Web Page Segmentation for Non Visual Access

Judith Jeyafreeda Andrew, Stéphane Ferrari, Fabrice Maurel, Gaël Dias and Emmanuel Giguet

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000 Caen, France
{judith-jeyafreeda.andrew, stephane.ferrari, fabrice.maurel, gael.dias,emmanuel.giguet}@unicaen.fr

Abstract Web page segmentation aims to break a large page into smaller blocks, in which contents with coherent semantics are kept together. Within this context, a great deal of approaches have been proposed without any specific end task in mind. In this paper, we study different segmentation strategies for the task of *non visual skimming*. For that purpose, we propose to segment web pages into visually coherent zones so that each zone can be represented by a set of relevant keywords that can be further synthesized into concurrent speech. As a consequence, we consider web page segmentation as a clustering problem of visual elements, where (1) a fixed number of clusters must be discovered, (2) the elements of a cluster should be visually connected and (3) all visual elements must be clustered. Therefore, we study variations of three existing algorithms, that comply to these constraints: K-means, F-K-means, and Guided Expansion. In particular, we evaluate different reading strategies for the positioning of the initial K seeds as well as a pre-clustering methodology for the Guided Expansion algorithm, which goal is to (1) fasten the clustering process and (2) reduce unbalance between clusters. The performed evaluation shows that the Guided Expansion algorithm evidences statistically increased results over the two other algorithms with the variations of the reading strategies. Nevertheless, improvements still need to be proposed to increase separateness.

Keywords: web page segmentation · clustering · reading strategies · processing time · non visual access

1 Introduction

For visually impaired people, accessing the web quickly and efficiently remains a challenge. While research efforts are carried out to design novel interaction models, screen reader is still the dominant technology for non visual web browsing [14].

In the TAGTHUNDER project[1], we introduce the concept of *tag thunder* as a means to produce an interactive and innovative stimulus promoting the emergence of self-adapted strategies for non visual reading [11]. The approach consists in constructing

[1] https://tagthunder.greyc.fr/

oral counterparts to visual concepts (typography, layout, ...) that support the implementation of quick reading strategies such as skimming and scanning.

Skimming and scanning are two well-known reading processes, which are combined to access the document content as quickly and efficiently as possible. Scanning refers to the process of searching for a specific piece of information, and skimming is the action of passing through a document in a first glance to get an overview of its content. Skimming can easily be applied in a visual environment thanks to the visual, logical or textual document structures. Indeed, visual skimming relies on contrasted effects related to layout rendering and typographic styles. However, these effects are not available in a non visual environment. As such, reproducing the document content driven by its structure in a non visual setting is a much harder problem, but essential to be solved to improve web accessibility (e.g. visually impaired people).

In this paper, we focus on the hypothesis that successful non visual skimming strategies can take advantage of the previous identification of relevant zones with coherent semantics, that represent the coarse-grained document structure. This specific task is known as web page segmentation. Within this context, a great deal of approaches have been proposed [15,4,20], which do not focus on any specific end task, and as such are not constrained. Oppositely, we consider that non visual skimming requires three characteristics to be filled.

First, the number of zones has to be fixed in order to foster the emergence of regularities in the output and to comply with the maximum number of concurrent oral stimuli a human-being can cognitively distinguish. Indeed, we assume that each semantically coherent zone can be summarized and simultaneously synthesized into spatialized concurrent speech acts. Within this context, [7,10] have shown that the cognitive load can rise up to five different stimuli, thus limiting the number of zones resulting from the WPS process to 5. *Second*, each zone should be associated to a unique sound source spatially located in accordance with its position in the web page. Thus, each zone should be a single compact block made of contiguous web elements, and the zones should not overlap. *Third*, segmentation must be complete, which means that no web page element should remain outside a given zone, as the objective is to reveal the overall semantics of a document and not just parts of it[2].

In [2], we studied three different algorithms that comply to these constraints: the classical K-means [8], the F-K-means (a variant of K-means, which introduces the notion of force between elements instead of the euclidean distance), and the Guided Expansion algorithm (GE), which follows a propagation strategy including alignment constraints. A manual evaluation of the three algorithms had been performed by three experts measuring two clustering indicators: compactness and separateness, which was followed by a quantitative evaluation introducing different criteria for analysis. From both qualitative and quantitative evaluations, the GE proved to produce the most efficient solution over all criteria. However, as suggested in [2] the clustering process is highly sensitive to the initial seeds positions. By following a diagonal reading strategy, we noted that most algorithms evidence an horizontal segmentation, i.e. vertical cluster are difficult to identify. Thus, in this paper we propose to use different methods to position the seeds based on reading strategies used on the web. As presented in [13] and

[2] Oppositely to advertisement withdrawal for example.

[3], the users tend to read the web page in a "F" or "Z" strategy. As a consequence, we use this insight to position the inital $K = 5$ seeds of the tested clustering algorithms. Moreover, we study a new methodology to decrease the time complexity of the GE by introducing a simple pre-clustering technique, following the ideas of the QT algorithm [17]. As such, processing time is reduced without major performance loss, and an interesting side effect evidences the fact that more balanced clusters are obtained.

2 Related Works

Web Page Segmentation. Efforts on web page segmentation (WPS) have focused on removing noisy contents from web pages [18,1]. Later, [19] were the first to propose a structural viewpoint of web page segmentation. For that purpose, layout and Document Object Model (DOM) features were used, as well as some hand-crafted heuristics. Although this methodology shows an original research direction, it relies on a fixed structural semantics that does not correspond to the creativity on the Web. More recently, [15] proposed Block-O-Matic, a pipeline strategy, which combines content, geometric and logical structures. Also, [9] developed a method called HEPS (HEading-based Page Segmentation) to extract logical hierarchical structures of HTML documents. In particular, they exploit the properties of headings as headings (1) appear at the beginning of the corresponding blocks, (2) are given prominent visual styles, (3) of the same level share the same visual style, and (4) of higher levels are given more prominent visual styles. One of the main drawbacks of these approaches is the fact that they heavily rely on the DOM, which can be prone to errors due to uncontrolled page creation. Moreover, the number of clusters is automatically determined and thus can greatly vary from page to page. Also, some elements can remain unclustered. In order to overcome some of these limitations, visual-based strategies have been proposed, which mainly focus on the analysis of the visual features of the document contents as they are perceived by human readers. Notable works that follow this paradigm are VIPS [4] and the Box Clustering Segmentation (BCS) algorithm [20]. While VIPS still uses the DOM as a logical view of the document in combination with visual features, BCS exclusively relies on a flat visual representation of the document, that allows great adaptability to new web contents. In particular, BCS follows a sort of hierarchical agglomerative clustering algorithm that includes a threshold, which controls the gathering of visual elements into clusters. As a consequence, the number of coherent zones is automatically determined by the threshold and can vary, and some elements may remain unclustered, similarly to [15]. In [2], we followed the same strategy as the BCS algorithm as we exclusively rely on visual elements to segment web pages, and thus rely on a flat structure. But, we proposed three different clustering techniques (classical K-means, the F-K-means (a variant of K-means, which introduces the notion of force between elements instead of the euclidean distance), and the Guided Expansion algorithm (GE) that comply to the constraints imposed by the non visual skimming task: (1) segmentation into exactly 5 coherent zones, (2) completeness, where all visual elements belong to a given cluster and (3) connectivity of all the elements inside a cluster. In [2], we showed that the initial position of the seeds plays a crucial role in the clustering of web elements. Thus, in this paper, we propose to study variations of the algorithms used in [2] by changing the

position of the initial seeds depending on different reading strategies used on the web. In order to decrease processing time and get more balanced clusters, we also introduce a modified version of the GE algorithm based on an initial pre-clustering step, which follows the ideas presented in [20] and relies on the QT algorithm [17]. Within this context, a quality area around some seeds is used to control the expansion process.

Reading Strategies. [13] propose a study on the "F" reading strategy that users use while reading the Web. The observations of [13] can be summarized as follows: (1) users first read in an horizontal movement, usually across the upper part of the content area. This initial element forms the F top bar; (2) next, users move down the page a bit and then read across in a second horizontal movement that typically covers a shorter area than the previous movement, which forms the F lower bar; (3) finally, users scan the left side of the content in a vertical movement, thus forming the F stem. In particular, the authors [13] show heat maps, which evidence the F pattern of reading on the Web. Another strategy is studied by [3]. They propose a study, which shows that users read the Web in a "Z" shape fashion when the web pages are not centered around its text content. The summary of [3] is as follows: (1) first, users scan from the top left to the top right, forming an horizontal line; (2) next, down and to the left side of the page, creating a diagonal line; (3) last, back across to the right again, forming a second horizontal line. Note that [3] and [13] also suggest other methods used by readers on the Web, but we will skip to both these ones in this study.

Evaluation. With respect to the evaluation of WPS, two strategies have been predominantly proposed. On the one hand, qualitative evaluations can be performed, where human assessors are asked to validate the proposed segmentation against a human ground truth [5]. On the other hand, studies propose quantitative evaluations relying on cluster correlation metrics [20]. In particular, [20] use metrics of a general clustering problem, such as Rand Index or F-measure. However, WPS can not strictly be compared to a general clustering problem. For example, if just one visual element does not belong to its correct cluster, it may break the logical structure of the segmentation, but the quantitative metric will still remain high. Similarly, [16] create a ground truth database by segmenting web pages using the MoB tool, and calculate specifically-tuned metrics. But, as they mostly rely on the DOM structure, they are limited to DOM-based methodologies. In order to overcome the difficulties of quantitative evaluations based on cluster correlation metrics in non-DOM solutions, we proposed in [2] a quantitative evaluation method for the analysis of clusters based on different criteria for non-visual skimming: (1) number of cuts between zones, (2) coefficient of unbalance in terms of surface, text area and number of elements, and (3) number of nested areas . We propose the very same metrics to compare our algorithms in this paper.

3 Clustering Strategies

In this section, we briefly summarize our previous work on clustering strategies for WPS as presented in [2]. WPS for the specific task of non visual skimming can be defined as a clustering problem, where basic visual elements must be gathered into a K

fixed number of clusters, where K is equal to 5. In particular, basic visual elements are first retrieved from a web page after rendering on the user's browser. DOM elements are then enriched with calculated CSS features, and each basic visual element corresponds to the last block element in each branch of the DOM tree[3]. In order to cluster the basic visual elements, we proposed three different strategies in [2]: K-means, F-K-means, and Guided Expansion. In this paper, we also propose the F-Guided Expansion algorithm, an adaptation of the Guided Expansion algorithm based on F-K-means.

K-means. The K-means algorithm [8] is a well-established strategy when the number of clusters must be fixed a priori. Within the context of WPS, some adaptations are required. In particular, the assignment phase is based on the shortest euclidean distance between two visual elements (or between a virtual visual element in the case of the centroid), noted $dist(.,.)$. For our task, the elements to cluster are not data points in an N-dimensional space, but blocks, i.e. rectangle shapes. Thus, we use a border-to-border distance between the rectangles instead of a center-to-center distance.[4] Moreover, in order to calculate the centroid of a cluster, a virtual visual element is computed, instead of relying on the medoid, i.e. a visual element closer to a virtual center.

F-K-means. In the previous proposal, the assignment phase is exclusively based on the geometric distance between visual objects. For this second algorithm, we propose a small variant, which takes into account the area covered by each visual basic element, the rationale being that visually bigger elements are more likely to "absorb" smaller elements than the contrary. So, if two visual elements are close to each other, their assignment function $force(b_1, b_2)$ will also depend on their differences of covered area as defined in equation 1, where a_{b1} (resp. a_{b2}) is the area of the visual element b_1 (resp. b_2) and $dist(.,.)$ is the shortest border-to-border euclidean distance between the basic elements. Thus, the F-K-means algorithm follows the exact same procedure as K-means, to the exception of the function used for the assignment step, which is the $force(.,.)$, i.e. the elements, which show the highest force to their centroid (a virtual visual element) are selected.[5]

$$force(b_1, b_2) = \frac{(a_{b1} * a_{b2})}{dist(b_1, b_2)} \tag{1}$$

Guided Expansion. With the Guided Expansion (GE) algorithm, instead of assigning all visual elements to their closest centroid in a single step, only one visual element is assigned at a time to its centroid, controlled by a set of conditions that include the shortest border-to-border euclidean distance of two visual elements, the alignment between elements, and their visual similarity. The GE algorithm and its illustration is detailed in [2]. In particular, visual similarity $vsim(.,.)$ between two elements b_1 and b_2 is computed as in equation 2 over their respective feature vectors $\overrightarrow{b_1}$ and $\overrightarrow{b_2}$ formed by the following CSS properties of each bounding box: font-color, font-weight, font-family and background-color.

[3] This is our unique use of the DOM structure.

[4] Illustration of this algorithm is presented in [2].

[5] Illustration of this algorithm is presented in [2].

$$vsim(\overrightarrow{b_1}, \overrightarrow{b_2}) = \sum_{i=1}^{|\overrightarrow{b_1}|} \mathbb{1}_{\overrightarrow{b_1^i} = \overrightarrow{b_2^i}} \tag{2}$$

It is important to notice that a cluster is a set of visual elements, except for the first step of the algorithm. So, when the distance and the visual similarity are computed between an element and its cluster candidate, this refers to the computation of each metric between the element and all the elements in the cluster. This situation is formalized in equations 3 and 4, where c_1 is the cluster candidate for b_1. However, the complexity of this algorithm is O(n^2), where n is the number of visual elements in the web page. This is because until there are no unclustered elements, the element under consideration is compared with every other element to form the candidate set of elements. This will be the reason for the definition of a new algorithm detailed in section 5.

$$dist(b_1, c_1) = argmin_{b_i \in c_1} dist(b_1, b_i) \tag{3}$$

$$vsim(\overrightarrow{b_1}, c_1) = argmax_{b_i \in c_1} vsim(\overrightarrow{b_1}, \overrightarrow{b_i}) \tag{4}$$

F-Guided Expansion. The F-Guided Expansion (F-GE) is a variation of the Guided Expansion algorithm presented in [2], which takes into account the area covered by each visual element. Thus, the first criterion to check between elements is the force of attraction, $force(b_1, b_2)$, between them as directed by equation 1, instead of the border-to-border geometric distance. Of course, this is followed by the alignment and visual similarities (equation 2) between elements as in the original GE algorithm.

4 Reading Strategies and Seeds Positioning

As mentioned in section 2, users tend to scan/skim the Web using several reading strategies. In [2], we showed that for the algorithms mentioned in section 3, the positioning of the initial seeds plays a crucial role in the clustering process. Indeed, by following a classical diagonal reading strategy, we noted that most algorithms evidence an horizontal segmentation, i.e. vertical clusters are difficult to identify. Another related issue concerns the F-K-means. If some seed is associated to a small element, this cluster will hardly expand as the $force(.,.)$ metric tends to benefit larger visual elements, thus clearly disadvantaging this algorithm compared to the other ones.

Thus, we intend to study the other reading strategies mentioned in section 2. The diagonal method places the seeds on a diagonal virtually drawn on the web page from top-left to bottom-right. In particular, two seeds are positioned on each extremities, another one in the center and the two other ones between the extremities and the center of the diagonal. In this paper, we propose to place the seeds in a "F" and "Z" fashion motivated by the studies of [12,13] and [3]. The strategies are shown in figure 1. In figure 1, the blocks represent the visual blocks of the web page, the blue lines through the blocks represent the reading strategies and the red blocks indicate the chosen seeds.

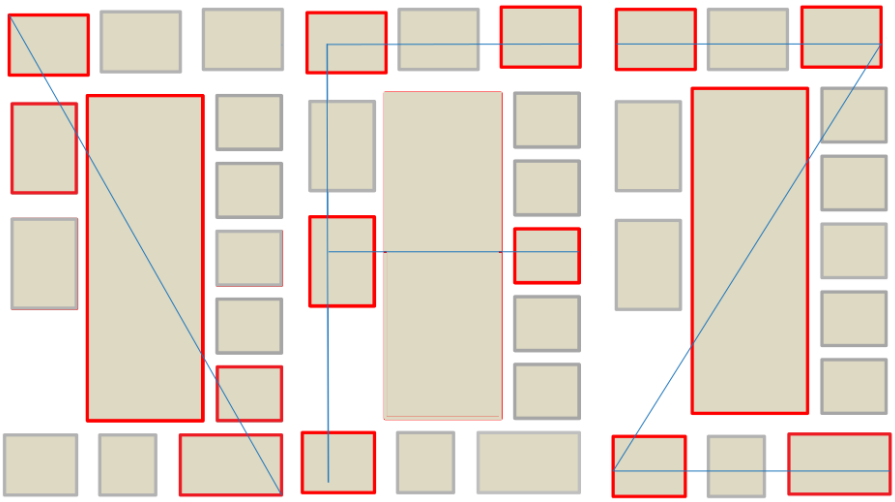

Figure 1. Diagonal (left), F (center) and Z (right) strategies to position the seeds.

5 Pre-clustering Guided Expansion

As noticed in section 3, the complexity of the Guided Expansion algorithm is O(n^2). Thus, in order to decrease the time complexity and as a consequence the running time, a simple pre-clustering of the visual elements is performed. To perform this pre-clustering, we rely on the Quality Threshold algorithm [17], which clusters elements within a confidence area defined by a distance threshold. This can be viewed as a coarse-grained clustering that gathers all visual elements reliably within a small area as suggested in [20]. As such, five clusters can easily be formed using this simple pre-clustering method. The visual blocks are selected in the ordered list of visual elements[6]. For the first element in the list, the QT is applied and assigns its visual elements depending on if the border-to-border distance is within the given threshold. Then, the assigned elements are withdrawn from the list, and the same process is iterated four times based on the updated list.

This clustering obviously leaves some visual elements unclustered. So, for the visual elements that are unclustered, the Guided Expansion algorithm is used to finalize the clustering process. within this context, the five pre-clusters serve as basis for the final assignment. Note that with this simple pre-clustering step, we can reduce the complexity to O($\alpha \times n^2$), where $\alpha < 1$, where α depends on the size of the web page and the maximum distance between two visual elements. The Guided Expansion along with the QT pre-clustering step is given in algorithm 1.

[6] Different strategies can be used to order the visual elements. In this paper, we use the order in which the visual elements appear in the DOM, using a depth-first search.

```
Input: The ordered list of basic visual elements; K
Output: K clusters
Threshold ← max(distance between two visual elements)/10;
K ← 1;
while K ≤ 5 do
    Choose the first visual element, as the parent element, from the ordered list;
    Remove the first element from the ordered list;
    for each visual element in the ordered list do
        Calculate dist(.,.) between the visual element and the parent element;
        if dist(.,.) < Threshold then
            Add the visual element in the cluster of the parent element;
            Remove the visual element from the ordered list;
        end
    end
    K ← K + 1;
end
while the ordered list is not empty do
    Select each closest element to every cluster using dist(.,.);
    Order these elements by the minimum distance to their candidate cluster;
    Remove all elements that do not evidence the smallest distance for possible
      assignment;
    if there are no ties then
        Assign the closest element overall to its cluster;
    end
    else if there are ties then
        Check whether the elements are vertically or horizontally aligned with at least
          one element of their cluster;
        Order elements by alignment;
        if there are no ties AND one aligned element then
            Assign the aligned element to its cluster;
        end
        else if there are ties OR no aligned element then
            Order elements by the maximum visual similarity to their cluster;
            Remove all elements that do not evidence the highest visual similarity for
              possible assignment;
            if there are no ties then
                Assign the most visually similar element to its cluster;
            end
            else if there are ties then
                Assign all elements to their cluster;
            end
        end
    end
end
```

Algorithm 1: Guided Expansion with QT pre-clustering.

6 Quantitative Evaluation

In [2], we presented five criteria for a quantitative evaluation that were derived from the conclusions of a qualitative evaluation performed by 3 experts on 25 web pages. The criteria that emerged are as follows:

- Logical constraints embodied by specific HTML tag sequences such as <li> items, <title> and the following paragraph <p>, <header>, <footer> or <nav> elements should not be broken. Indeed, such breaks are likely to lead to odd clustering. As such, we propose to count one cut for each broken logical rule. This value is shown in column 1 of Table 1.
- An efficient algorithm should produce zones neither completely balanced nor too much unbalanced. To evaluate such a criterion, we test three different balance properties of the clusters: standard deviation of the surface area of the clusters, standard deviation of the number of characters within the clusters, and standard deviation of the number of visual elements within the clusters. The higher the standard deviation, the more unbalance the clusters are. The results of this property is shown in columns 2,3 and 4 of Table 1.
- Zones should not be nested, i.e. the clustering should avoid non-rectangular clusters. To evaluate this phenomenon, the number of overlaps between the outer rectangles of all clusters is calculated, i.e. the smallest rectangle including all the elements of each cluster. So, if two clusters overlap in terms of outer rectangle, this stands for the presence of a non rectangular zone, and it is counted as a nested situation. The results of this property are shown in column 5 of table 1.

Table 1 shows the results of the automatic evaluation for the three main criteria for a set of 150 web pages (47 tourist domain, 58 e-Commerce domain and 45 news domain[7]) segmented using all the versions of the three algorithms (K-means, F-K-means and Guided Expansion). In particular, each criterion receives the average value and the standard deviation $\pm\sigma$ for the set of 150 pages.

First, results show the superiority of the Guided Expansion algorithm over the other two algorithms in terms of number of cuts. In particular, it evidences a minimum average value of 1.34 with the GE with Z reading strategy and a maximum of 1.83 with the F-GE with the F reading strategy, while K-means shows a minimum 2.12 score and F-K-means shows worst results with a minimum score of 2.63. In the case of K-means, using the F and Z reading strategies does not seem to improve the results over the diagonal strategy. But, in the case of F-Kmeans, the F and Z reading strategies give better results in terms of cuts. Thus, the three algorithms, irrespective of the reading strategies, can be sorted according to their ability to minimize the cut criterion with statistically significant values, i.e. GE is superior to K-means, which is in turn superior to F-K-means, however, the reading strategies do not seem to play a great role in this criteria. We will confirm these results in section 7, where we present a complete statistical analysis of the results.

Second, balance results show similar observations whether we compare surface area, text area or number of elements between clusters. In all cases, the F-GE with

[7] All part of our project corpus.

Algorithm	Nb. of Cuts	Surface Area	Text Area	Nb. of Elements	Exterior Rectangle
	Avg. $\pm\sigma$	Avg. $\pm\sigma$	Avg. $\pm\sigma$	Avg. $\pm\sigma$	Avg. $\pm\sigma$
K-means D	2.12 $\pm$2.05	11.80 $\pm$6.46	11.40 $\pm$5.52	10.95 $\pm$8.01	5.21 $\pm$2.54
K-means F	2.59 $\pm$2.50	12.57 $\pm$6.54	12.52 $\pm$5.64	12.85 $\pm$9.63	4.13 $\pm$2.29
K-means Z	2.50 $\pm$2.40	13.20 $\pm$6.14	13.46 $\pm$6.02	14.85 $\pm$10.45	4.04 $\pm$2.21
F-K-means D	2.80 $\pm$2.76	21.14 $\pm$8.18	18.55 $\pm$7.74	22.79 $\pm$16.73	4.54 $\pm$2.20
F-K-means F	2.66 $\pm$2.40	20.58 $\pm$8.61	19.18 $\pm$8.63	23.87 $\pm$18.12	3.54 $\pm$1.94
F-K-means Z	2.63 $\pm$2.36	21.14 $\pm$7.82	19.40 $\pm$7.57	25.32 $\pm$18.33	3.53 $\pm$1.95
GE D	1.47 $\pm$1.85	17.34 $\pm$6.95	16.78 $\pm$6.37	19.67 $\pm$13.47	5.39 $\pm$2.22
GE F	1.43 $\pm$1.85	22.64 $\pm$7.23	22.37 $\pm$6.70	30.42 $\pm$19.93	4.91 $\pm$2.01
GE Z	1.34 $\pm$1.66	23.69 $\pm$7.10	22.77 $\pm$6.70	32.45 $\pm$21.82	5.26 $\pm$2.03
GE P	1.57 $\pm$1.98	12.55 $\pm$6.76	12.24 $\pm$6.35	15.04 $\pm$11.12	6.72 $\pm$2.11
F-GE D	1.75 $\pm$1.94	28.50 $\pm$8.27	27.41 $\pm$7.74	38.80 $\pm$24.62	3.46 $\pm$1.89
F-GE F	1.83 $\pm$2.08	31.12 $\pm$7.29	29.65 $\pm$7.52	43.85 $\pm$25.21	3.53 $\pm$1.89
F-GE Z	1.77 $\pm$1.97	31.35 $\pm$6.88	30.26 $\pm$6.75	44.90 $\pm$25.96	4.18 $\pm$2.12
F-GE P	1.80 $\pm$2.15	13.70 $\pm$7.12	12.12 $\pm$6.70	14.64 $\pm$11.07	5.92 $\pm$2.36

Table 1. Automatic evaluation results for K-means, F-K-means and Guided Expansion (GE) for all reading strategies plus the pre-clustering GE. The evaluation is performed over 150 web pages. Note that D stands for Diagonal, F for F reading strategy, Z for Z reading strategy, GE P means the pre-clustering version of GE. Note also that $\pm\sigma$ stands for the standard deviation value over the 150 web pages.

diagonal, F and Z reading strategies show highest unbalance, while K-means shows the lowest unbalance. The Guided Expansion algorithm evidences some tendency to unbalanced clustering, which seems to better approximate human segmentation as explained in [2]. However, it is important to note that using a pre-clustering step with GE increases the balance between the zones in a huge way. This is because the pre-clustering step uses a threshold on distance thus restricting the number of elements in a zone and in turn producing clusters with similar sizes.

Finally, the "Exterior Rectangle" criterion, that aims to measure the number of non-rectangular shapes evidences similar results between all algorithms with around five overlaps per web page on average. Nevertheless, there is a clear statistical tendency for the F reading strategy to produce less non-rectangular zones amongst the other strategies. This is because when the seeds are placed near the border, the zones propagate only in one direction. Thus, this tends to produce rectangular zones. However, it is important to notice that the exterior rectangle criterion goes down to almost 0 for human annotators as shown in [2], who rarely proposed non-rectangular zones in their manual segmentation. As such, one might think that all algorithms are far from achieving human-like behavior. Although this is a strict reality from the figures, this difference against the manual evaluation observation may also indicate a lack of possible solutions by human annotators. Indeed, we think that acceptable segmentation can be proposed by some algorithms, although human annotators may not have thought about[8]. Thus further discussion should clearly be about the way to refine this criterion in order to distinguish between good and bad overlaps automatically.

[8] This situation is explained in detail in [2]

7 Statistical Evaluation

Box Plots. In order to verify the significant difference between all tested algorithms, we first show a box plot analysis for the five criteria mentioned here before. From Figure 2, it is witnessed that Guided Expansion a evidences a minimum of zero cuts irrespective of the strategy used to position the seeds. However, the minimum number of cuts for the other algorithms reaches the levels of the GE. From figures 3, 4 and 5, we can notice that the algorithms which are pre-processed with a simple clustering method tend to have more balance between the clusters. This is due to the fact that the first clustering produces initial clusters using a threshold and thus ensuring balance in the first stage of the process. From figure 6, it can be noticed that the more the seeds are placed near the border of the web page, the more they tend to make zones that are rectangle. Indeed, the F strategy places 5 seeds along the borders of the web page and thus witnesses less exterior rectangles i.e. more rectangular zones. Instead, the Z strategy places 4 seeds near the borders of the web page, while the diagonal strategy places 2 seeds on the borders of the web page. Thus, the order of algorithms with less nested zones can be summarized as follows: F $>$ Z $>$ D $>$ PC.

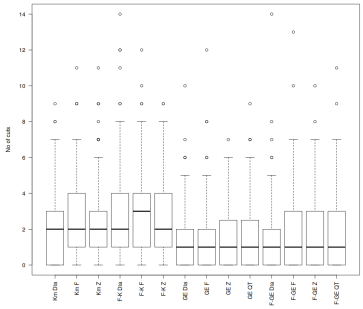

Figure 2. Box plot for the number of cuts (column 1 in table 1).

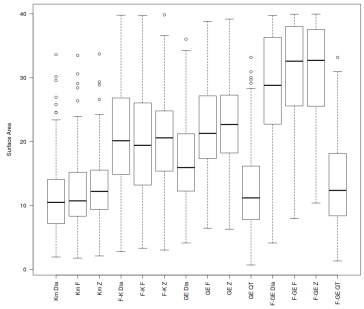

Figure 3. Box plot for balance in surface area (column 2 in table 1).

Dunn Test. Once the initial ANOVA has found a significant difference in more than means, the Dunns Test [6] can be used to pinpoint which specific means are significant from the others. Thus, the Dunns Multiple Comparison Test is a post hoc (i.e. its run after an ANOVA) non parametric test, which is done to determine which groups are different from others. In order to verify the differences between algorithms in terms of statistical significance, we propose to use this test. The results of the analysis with the Dunn test are shown in Table 2. Note that the algorithms within each group are not significantly different from each other. However, algorithms in different groups are significantly different from each other. Moreover, the rank of each group shows how well a given group performs for a given criterion. From this analysis, it seems that the Guided Expansion algorithm with pre-clustering (GE P) is globally the more suitable

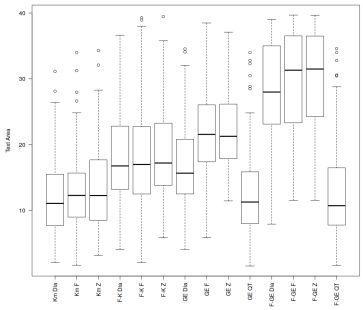

Figure 4. Box plot for balance in text area (column 3 in table 1).

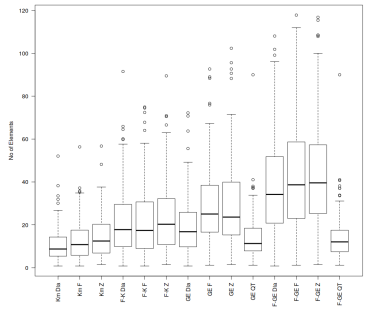

Figure 5. Box plot for balance in visual elements (column 4 in table 1).

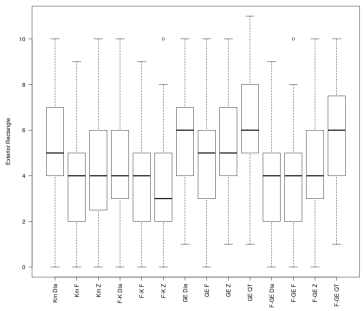

Figure 6. Box plot for the criteria Exterior Rectangle (column 5 in table 1)

solution for WPS in the specific context of non visual information access. However, this needs to be confirmed by a qualitative analysis[9] as this algorithm shows the worst results in terms of nested clusters.

8 Conclusions

In this paper, we presented Web Page Segmentation as a clustering problem driven by the task of non visual information access. In particular, we tested three well-known algorithms, namely K-means, F-K-means and the Guided Expansion, with three reading strategies used on the web, namely diagonal, "F" and "Z". We also presented a new methodology to reduce the time complexity of the Guided Expansion algorithm by introducing a pre-clustering step based on the QT algorithm. We also tested the Guided Expansion algorithm combined with the force measure. Quantitative and statistical evaluations showed that the Guided Expansion algorithm is a good baseline, in particular

[9] For lack of space, we do not present this study in this paper.

Criterion	Groups	
Cuts	1	{GE F, GE Z, GE P}
	2	{GE D}
	3	{F-GE D, F-GE P, F-GE Z}
	4	{F-GE F}
	5	{K-means D}
	6	{K-means F, F-K-means D, F-K-means Z}
	7	{F-K-means F}
	8	{K-means Z}
Surface Area	1	{K-means D}
	2	{GE P, K-means F, K-means Z}
	3	{F-GE P}
	4	{GE D}
	5	{F-K-means F}
	6	{F-K-means Z}
	7	{F-K-means D}
	8	{GE F}
	9	{GE Z}
	10	{F-GE D, F-GE F, F-GE Z}
Text Area	1	{GE P, F-GE P, K-means D, K-means F, K-means Z}
	2	{GE D, F-K-means D, F-K-means F, F-K-means Z}
	3	{GE F, GE Z}
	4	{F-GE D, F-GE F, F-GE Z}
Number of Elements	1	{K-means D}
	2	{K-means F}
	3	{GE P, F-GE P, K-means Z}
	4	{GE D, F-K-means D, F-K-means F, F-K-means Z
	5	{GE F, GE Z}
	6	{F-GE D, F-GE F, F-GE Z}
Exterior Rectangle	1	{F-GE D, K-means Z}
	2	{F-GE F}
	3	{F-K-means F, F-K-means Z}
	4	{K-means F}
	5	{F-GE Z}
	6	{F-K-means D}
	7	{GE F}
	8	{GE Z, K-means D}
	9	{GE D, F-GE P}
	10	{GE P}

Table 2. Dunn test analysis for the 14 algorithms over the 5 different criteria. Algorithms within a group show no statistical difference between them. Rank evidences the performance order for each criterion.

in its new version including pre-clustering. Pre-clustering not only reduces the time complexity of the Guided Expansion algorithm but also improves the balance between the zones without significantly increasing the number of cuts. We also showed that the position of the initial seeds does change the results of the algorithms in a significant way. However, there are still other reading strategies used on the web that are open to exploration. As a consequence, future work needs to be endeavour to strengthen these first findings. This goes with performing a qualitative analysis and an exhaustive analysis of reading strategies. But, the automatic selection of optimal seeds seems to be the priority research direction.

References

1. Alassi, D., Alhajj, R.: Effectiveness of template detection on noise reduction and websites summarization. Information Sciences **219**, 41–72 (2013)
2. Andrew, J.J., Ferrari, S., Maurel, F., Dias, G., Giguet, E.: Web page segmentation for non visual skimming. In: The 33rd Pacific Asia Conference on Language, Information and Computation (2019)
3. Babich, N.: Z-Shaped Pattern For Reading Web Content (2017), `https://uxplanet.org/z-shaped-pattern-for-reading-web-content-ce1135f92f1c`, last access on September 2019
4. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Extracting content structure for web pages based on visual representation. In: 5th Asia-Pacific Web Conference on Web Technologies and Applications (ApWeb). pp. 406–417 (2003)
5. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Vips: a vision-based page segmentation algorithm. Tech. Rep. MSR-TR-2003-79, Microsoft (November 2003), `https://www.microsoft.com/en-us/research/publication/vips-a-vision-based-page-segmentation-algorithm/`
6. Dunn, O.J.: Multiple comparisons among means. Journal of the American statistical association **56**(293), 52–64 (1961)
7. Guerreiro, J., Gonçalves, D.: Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech. In: 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS). pp. 3–11 (2015)
8. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: 15th Berkeley Symposium on Mathematical Statistics and Probability (BSMSP). pp. 281–297 (1967)
9. Manabe, T., Tajima, K.: Extracting logical hierarchical structure of HTML documents based on headings. PVLDB **8**(12), 1606–1617 (2015)
10. Manishina, E., Lecarpentier, J.M., Maurel, F., Ferrari, S., Maxence, B.: Tag Thunder : Towards Non-Visual Web Page Skimming. In: 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS) (2016)
11. Maurel, F., Dias, G., Ferrari, S., Andrew, J.J., Giguet, E.: Concurrent speech synthesis to improve document first glance for the blind. In: 2nd International Workshop on Human-Document Interaction (HDI 2019) associated to 15th International Conference on Document Analysis and Recognition (ICDAR 2019), September 20-25. Sydney, Australia (2019)
12. Nielsen, J.: F-Shaped Pattern For Reading Web Content, `https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content-discovered/`
13. Pernice, K.: F-Shaped Pattern of Reading on the Web: Misunderstood, But Still Relevant (Even on Mobile) (2017), `https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content`, last access on September 2019

14. Ramakrishnan, I.V., Ashok, V., Billah, S.M.: Non-visual web browsing: Beyond web accessibility. In: 11th International Conference on Universal Access in Human-Computer Interaction (UAHCI). pp. 322–334 (2017)
15. Sanoja, A., Gançarski, S.: Block-o-matic: A web page segmentation framework. In: International Conference on Multimedia Computing and Systems (ICMCS). pp. 595–600 (2014)
16. Sanoja, A., Gançarski, S.: Web page segmentation evaluation. In: 30th Annual ACM Symposium on Applied Computing (SAC). pp. 753–760 (2015)
17. Xin, J., Jiawei, H.: Quality Threshold Clustering, pp. 1–2. Springer US, Boston, MA (2016)
18. Yi, L., Liu, B., Li, X.: Eliminating noisy information in web pages for data mining. In: 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 296–305 (2003)
19. Yin, X., Lee, W.S.: Understanding the function of web elements for mobile content delivery using random walk models. In: 14th International Conference on World Wide Web (WWW). pp. 1150–1151 (2005)
20. Zeleny, J., Burget, R., Zendulka, J.: Box clustering segmentation: A new method for vision-based web page preprocessing. Information Processing & Management **53**(3), 735–750 (2017)