Web Page Segmentation for Non Visual Skimming

Judith Jeyafreeda Andrew, Stephane Ferrari, Fabrice Maurel, Gaël Dias and Emmanuel Giguet
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000 Caen, France
Email: {judith-jeyafreeda.andrew, stephane.ferrari, fabrice.maurel, gael.dias,emmanuel.giguet}@unicaen.fr

Abstract

Web page segmentation aims to break a page into smaller blocks, in which contents with coherent semantics are kept together. Examples of tasks targeted by such a technique are advertisement detection or main content extraction. In this paper, we study different segmentation strategies for the task of *non visual skimming*. For that purpose, we consider web page segmentation as a clustering problem of visual elements, where (1) all visual elements must be clustered, (2) a fixed number of clusters must be discovered, and (3) the elements of a cluster should be visually connected. Therefore, we study three different algorithms that comply to these constraints: K-means, F-K-means, and Guided Expansion. Evaluation shows that Guided Expansion evidences statistically-relevant results in terms of compactness and separateness, and satisfies more logical constraints when compared to the other strategies.

1 Introduction

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 structure. 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, for the visually impaired, for instance.

In this paper, we focus on the hypothesis that successful non visual skimming strategies can take advantage of a prior identification of the coarse-grained document structure. This specific task is known as Web Page Segmentation (WPS). WPS aims to break a page into zones that appear semantically coherent. A large number of approaches have been proposed to automate this process (Sanoja and Gançarski, 2014; Cai et al., 2003a; Zeleny et al., 2017). However, they deal with tasks that imply constraints far from ours. In our TAG THUNDER project[1], 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, (Guerreiro and Gonçalves, 2015; Manishina et al., 2016) have shown that the cognitive load can rise up to five different stimuli, thus limiting the number of zones resulting from the WPS process. The 5-zone WPS should also ease the association of a particular sound position to the logical function of the zone in a given

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

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 423-431, Hakodate, Japan, September 13-15, 2019

web page. As a consequence, it may enable the advent of new non visual reading strategies. *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, opposite to advertisement withdrawal for example.

In this paper, we study three different algorithms that comply to these constraints: the classical k-means (MacQueen, 1967), 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 is performed by three experts measuring two clustering indicators: compactness and separateness. However, human evaluation may be subject to bias as each expert evaluates the WPS process with his/her own subjectivity. As a consequence, we propose a quantitative evaluation that introduces different criteria of analysis.

The paper is structured as follows. Section 2 provides a brief overview of WPS and its evaluation policies. Section 3 introduces the three clustering algorithms. Sections 4 and 5 present the manual and automatic evaluations. Finally, 6 concludes the paper with a discussion and outlines future works.

2 Related Work

Web Page Segmentation. Efforts on WPS have focused on removing noisy content from web pages (Yi et al., 2003; Chen et al., 2003; Alassi and Alhajj, 2013; Barua et al., 2014). Later, (Yin and Lee, 2005) were the first to propose a structural viewpoint of web page segmentation, by developing a graph-based strategy to classify elements into categories. 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, (Sanoja and Gançarski, 2014) proposed Block-O-Matic, a pipeline strategy, which combines content, geometric and logical structures. One of the main drawback of this approach is the fact that it heavily relies on the DOM, which can be prone to errors due to uncontrolled page creation (Zeleny et al., 2017). 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 (Cai et al., 2003a) and the Box Clustering Segmentation (BCS) algorithm (Zeleny et al., 2017). 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 (Sanoja and Gançarski, 2014).

In this paper, we follow 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 propose three different clustering techniques 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.

Evaluation. With respect to 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 (Cai et al., 2003b).

On the other hand, studies propose quantitative evaluations relying on cluster correlation metrics.

Within this context, (Zeleny et al., 2017) compare BCS to VIPS using classical clustering evaluation metrics, the F-score and the Adjusted Rand Index. In particular, they create pairs of automatically detected areas and manually annotated areas, which share at least one rendered box. For each such pair, they calculate Precision and Recall. If there are any manually selected areas that do not share boxes with any automatically detected areas, the recall value for each of them is set to 0. The resulting F-score is calculated using average values of Precision and Recall for the entire web page. So, (Zeleny et al., 2017) use the techniques of a general clustering problem. 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, (Sanoja and Gançarski, 2015) create a ground truth database by segmenting web pages using the MoB tool Then, a block in the automatic segmentation is said to be correctly segmented if its geometry and location are equal to only one block in the ground truth database; thus proposing specifically-tuned metrics. But as they mostly rely on the DOM structure, they are limited to DOM-based methodologies.

3 Clustering Strategies

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 retrieved from a web page after rendering on the user's browser. DOM elements are then enriched with calculated CSS features, and the basic visual elements correspond to the last block elements in each branch of the DOM tree[2]. In order to cluster the basic visual elements, we propose three different strategies: K-means, F-K-means, and Guided Expansion.

3.1 K-means

K-means (MacQueen, 1967) is a well-established algorithm, when the number of clusters must be fixed in advance. 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, noted $dist(.,.)$. For our task, the elements to cluster are not data points in an N-dimensional space, but blocks, i.e. rectangle shapes. In particular, we use a border-to border distance instead of a center-to-center distance. Indeed, as shown in Figure 1, a border-to-border distance is more appropriate in a visual context than a center-to-center distance. In our example, the center-to-center strategy selects the two visual bounding boxes positioned on the right while border-to-border strategy selects the two positioned on the left.

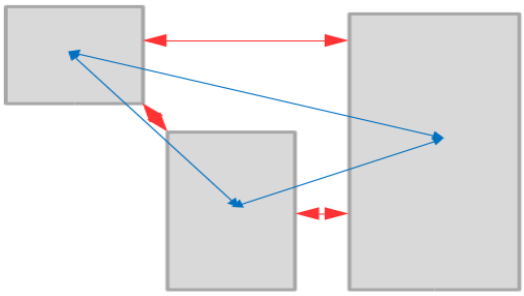

Figure 1: Blue lines showing center to center and red lines showing border to border distances.

Moreover, the classical K-means relies on the random selection of initial seeds. However, this strategy does not adapt to our approach because we need comparable algorithms. As a consequence, we propose to fix the 5 initial seeds following the diagonal reading strategy, i.e. if a diagonal is drawn on the web page from top-left to bottom-right, two seeds are positioned on each extremities, one in the center and the two other ones between the extremities and the center of the diagonal. The underlying idea is that within a skimming process, readers adopt a fast reading strategy, that focuses on particular areas of the web page. In this paper, we propose to test the diagonal strategy but other reading approaches exist (Fitzsimmons et al., 2014; Pernice et al., 2014), which study remains as future work. The K-means clustering process is detailed in Algorithm 1. An illustration of the K-means on a real web page is given in Figure 2.

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

Input: The set of basic visual elements; K
Output: K clusters
Initialization: Select K centroid elements;
while *true* **do**
  Assign each visual element to its closest centroid based on $dist(.,.)$;
  Compute K new centroids as the average virtual visual element of each cluster;
  if *centroids do not change* **then**
    break;
  end
end

Algorithm 1: K-means algorithm.

3.2 F-K-means

In the first 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.

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

So, the F-K-means algorithm follows the exact same procedures as algorithm 1, 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 centroids are selected. An illustration of the F-K-means on a real web page is given in Figure 3.

3.3 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 euclidean distance between the borders of two elements, the alignment between elements, and their visual similarity. The GE is defined in algorithm 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.

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

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. An illustration of the GE algorithm on a real web page is given in Figure 4.

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

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

4 Qualitative Evaluation

In this section, we propose to perform a qualitative evaluation, where 3 human experts are asked to evaluate two common indices in clustering, i.e. compactness and separateness (Acharya et al., 2014). Each expert must produce his/her own segmentation and evaluate both indicators on his ground truth. Compactness is defined at the cluster level and evaluates how many of the elements within a cluster belong to a same cluster in the (individual) ground truth. Separateness is defined at the web page level and evaluates how much the proposed segmentation guarantees the separability between clusters when compared to the expert ground truth segmentation. In this case, the expert must evaluate how much, on average, elements that should belong to the same cluster following the (individual) ground truth are separated in different clusters.

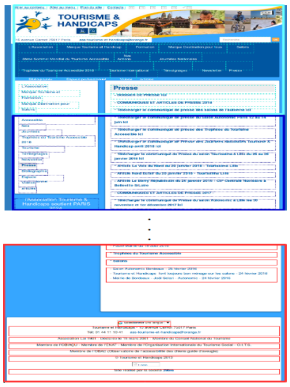

Figure 2: K-means

Figure 3: F-K-means

Figure 4: Guided Expan.

Figure 5: Manual Segm.

In particular, each expert must give a mark ranging from 0 (unacceptable), 1 (bad), 2 (passable), 3 (good) and 4 (perfect). Based on this protocol, the three algorithms presented in section 3 have been tested on a total of 53 web pages from 3 domains: Tourism (23 web pages), E-Commerce (12 web pages) and News (18 web pages), that are part of our TAG THUNDER project corpus[3]. To avoid bias, the experts are unaware of the algorithms strategies he/she is evaluating. Overall results are presented in table 1 and an example of the expert manual segmentation is illustrated in Figure 5.

It is clear that the GE algorithm shows the best figures both in terms of compactness and separateness for the 3 human experts. However, while compactness receives average values between passable and good, separateness receives much lower values, between passable and bad. This finding is transverse to all three algorithms, clearly evidencing that finding coherent zones that match human expectations is a hard task, while building internally semantically coherent zones is easier. Also, figures show differences between K-means and F-K-means. In particular, both algorithms show similar compactness, but the F-K-means evidences worst results for separateness. This result can easily be explained as the F-K-means tends to create unbalanced clusters, that are either very small or rather big. This is confirmed by the higher standard deviation in terms of compactness for F-K-means than for K-means, signifying that F-K-means tends to create very compact clusters (but small) and uncondensed big ones, thus penalizing separateness.

To statistically confirm these results, we computed a global segmentation score (GSS) taking into account both compactness and separateness (equation 5) and performed a Wilcoxon signed-rank test between all algorithms for each human expert. In equation 5, the evaluation scale refers to the scoring scale of separateness ($separat$) and compactness ($compact$), i.e. in our case 5 (0 to 4 grade). Results in table 2 show that GE evidences statistically superior results to both K-means and F-K-means, and that K-means provides statistically higher results than F-K-means, for all three experts in all tested situations, to exception for Expert 3 when comparing K-means and F-K-means.

$$GSS = \frac{(1 + separat) \times (1 + \overline{compact})}{|\text{evaluation scale}|^2} \quad (5)$$

5 Quantitative Evaluation

As seen from the manual evaluation, each expert evaluates the segmentation in a different way depending on his/her perception of coherency of the visual elements. In order to reduce human bias in evaluation, quantitative metrics should be used. However, as stated in section 2, clustering metrics are not adapted to our task. As a consequence, we propose to compute a set of metrics that characterize clustering results based on three different criteria: (1) number of broken logical constraints, (2) cluster balance and (3) cluster geometrical overlap.

[3]This dataset is freely available for research purposes.

Input: The set of basic visual elements; K
Output: K clusters
Initialization: Select K centroid elements (clusters) based on the reading strategy;
while *there are unclustered elements* **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 2: Guided expansion algorithm.

		Compactness		Separateness		GSS	
		Avg.	$\pm\sigma$	Avg.	$\pm\sigma$	Avg.	$\pm\sigma$
K-M	E1	2.42	1.16	1.15	0.64	0.30	0.12
	E2	1.90	0.87	1.20	0.60	0.26	0.11
	E3	3.10	0.74	0.70	0.80	0.29	0.15
F-K-M	E1	2.43	1.46	0.62	0.57	0.23	0.09
	E2	1.83	1.15	0.40	0.50	0.16	0.07
	E3	3.05	1.22	0.30	0.50	0.21	0.095
GE	E1	**2.89**	1.24	**1.62**	0.93	**0.42**	0.19
	E2	**2.41**	0.81	**1.90**	0.90	**0.41**	0.16
	E3	**3.40**	0.68	**1.50**	0.90	**0.44**	0.18

Table 1: Overall results for K-means (K-ME.), F-K-means (F-K-ME.) and Guided expansion (GE).

H_1	F-K-ME. $<$ K-ME.		F-K-ME. $<$ GE		K-ME. $<$ GE	
	z-score	S/NS	z-score	S/NS	z-score	S/NS
E_1	4.365	S	5.392	S	3.726	S
E_2	5.291	S	4.997	S	3.548	S
E_3	2.169	NS	4.304	S	3.021	S

Table 2: Wilcoxon signed-rank test for the GSS. S stands for significant statistical difference and NS for non significant. E_i is i-th expert. Tests are computed for $p < 0.05$.

Three criteria emerged from the manual evaluations conducted by all three experts. First, experts evaluated negatively when logical constraints were broken, i.e elements embodied by specific HTML tag sequences such as <li> items, <title> and the following paragraph <p>, <header>, <footer> or <nav> elements. So, each time one of these logical constraints is broken, this counts for one cut, and each web page is evaluated based on its overall number of cuts. The higher the number of cuts, the worst the clustering result must be evaluated. Overall results are given in table 3 (column 1). 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 1.47, while K-means shows a 2.12 score and F-K-means shows worst results with a score of 2.80. Moreover, the three algorithms 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. This criterion seems all the more important that there seems to be a correlation between manual and automatic results. Indeed, as illustrated by figure 4 for GE and figure 5 for manual segmentation, similar behavior seems to stand. However, this situation does not stand for the other two algorithms, where for instance menu sec-

tions are cut as illustrated in figures 2 and 3.

Second, experts negatively evaluated strong imbalance between clusters, but also high balance between clusters. This can be motivated by the fact that a great deal of web pages contain a main (rather large) body section, while all other zones show similar sizes. Note that this issue is usually not taken into account by classical clustering metrics such as Adjusted Rand Index or F-score. As a consequence, this notion of balance is tested over three different properties of the clusters: surface area of the cluster, number of characters within the cluster, and number of visual elements within the cluster. In particular, the surface area of the cluster is calculated as the maximum rectangle that embodies all the visual elements contained in it. So, each web page receives an overall score that stands for the standard deviation between all clusters for each of the three balance criteria (i.e. surface, text and visual elements). Overall results are given in table 3 (columns 2-4).

Third, experts evaluated negatively when the zones were intertwined with each other, i.e. they penalized non rectangular clusters. To evaluate this phenomenon, we computed the number of overlaps between the outer rectangles of all clusters, 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. Overall results are given in table 3 (column 5).

Table 3 shows the results of the automatic evaluation for the three main criteria for a set of 150 web pages (47 tourist web pages, 58 e-Commerce web pages and 45 news web pages[4]) segmented using 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. Table 4 completes results of table 3 with statistical significance by including the Wilcoxon signed-rank test.

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 1.47, while K-means shows a 2.12 score and F-K-means shows worst results with a score of 2.80. Moreover, the three algorithms 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. This criterion seems all the more important that there seems to be a correlation between manual and automatic results. Indeed, as illustrated by figure 4 for GE and figure 5 for manual segmentation, similar behavior seems to stand. However, this situation does not stand for the other two algorithms, where for instance menu sections are cut as illustrated in figures 2 and 3.

Second, balance results show similar observations whether we compare surface area, text area or number of elements between clusters. In all cases, the F-K-means shows highest unbalance[5], while K-means shows the lowest unbalance. This situation can be observed in figures 2 and 3, where respectively, K-means tends to create evenly distributed zones and F-K-means usually discovers a large zone and a set of smaller clusters. Oppositely, the Guided Expansion algorithm evidences some tendency to unbalanced clustering, that seems to better approximate human segmentation as shown in figures 4 and 5, where human annotators may allow a disequilibrium between the main body of the web page and the satellite zones such as headers, footers or menus. Indeed, humans tend to prefer little unbalanced zones in order to both respect the task condition (i.e. non visual skimming) and maintain the structural and logical aspects of the web page. Note that with respect to statistical significance, we can conclude that F-K-means is clearly the algorithm that steadily produces more unbalanced results. while this hypothesis is not so strong between K-means and the GE algorithm.

Third, 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 of the F-K-means to produce less non-rectangular zones. This can be explained by the unbalance constraint. Indeed, as the F-K-means produces highly unbalanced clusters, i.e. usually a large big zone and a set of rather small clusters, it is unlikely that overlap between zones exist, and

[4] All part of our project corpus.

[5] This situation has already been evidenced in the qualitative evaluation.

	Nb. of Cuts Avg. $\pm\sigma$	Surface Area Avg. $\pm\sigma$	Text Area Avg. $\pm\sigma$	Nb. of Elements Avg. $\pm\sigma$	Exterior Rectangle Avg. $\pm\sigma$
K-means	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
F-K-means	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
GE	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

Table 3: Automatic evaluation results for K-means, F-K-means and Guided Expansion (GE) for 150 web pages. Note that the column $\pm\sigma$ gives the standard deviation value over the 150 web pages.

H_1	F-Kmeans $>$ Kmeans		F-Kmeans $>$ GE		Kmeans $>$ GE	
	z score.	S/NS	z score	S/NS	z score	S/NS
Nb. of Cuts	3.64	S	7.08	S	4.85	S
Surface Area	10.29	S	5.65	S	9.12	NS
Text Area	9.59	S	2.36	S	8.83	NS
Nb. of Elements	9.96	S	2.53	S	9.54	NS
Exterior Rectangle	3.35	NS	3.60	NS	1.11	S

Table 4: Wilcoxon signed-rank test for the automatic evaluation for K-means, F-K-means and GE for 150 web pages. S stands for significant statistical difference and NS for non significant. Tests are computed for $p < 0.05$.

as a side-effect less non-rectangular zones are created. However, it is important to notice that the exterior rectangle criterion goes down to almost 0 for human annotators, who rarely proposed non-rectangular zones. 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. For example, the top of figure 4 shows a non-rectangular red zone with an outer rectangle overlapping the yellow one, that might satisfy some logical coherence, as menus are gathered together. Although this situation has not been proposed by any of the three annotators, we agree that such a segmentation is clearly satisfactory. Based on a deeper manual analysis of these results, we found that the Guided Expansion algorithm seems to be best performing algorithm on this criterion by producing better non-rectangular zones. Nevertheless, further discussion should clearly be about the way to refine this criterion in order to distinguish between good and bad overlaps automatically.

6 Conclusions and Research Directions

In this paper, we presented Web Page Segmentation as a clustering problem driven by the task of non visual skimming. In particular, we tuned the well-known K-means algorithm and designed two other algorithms, namely the F-K-means and the Guided Expansion, all dedicated to our objective and respecting the task constraints of a fixed number of zones, completeness of the coverage, and connectivity of visual elements. In particular, we showed that human and automatic evaluations are complementary to rank the algorithms according to several parameters (the number of cuts of HTML elements, the number of overlaps between zones and the balance of created clusters), each parameter performing a specific complementary role for both compactness and separateness criteria. From both qualitative and quantitative evaluations, the Guided Extension algorithm seems to be the most efficient solution over all criteria. The superiority of the GE algorithm is probably due to the introduction of the alignment constraint. Indeed, the alignment constraint is more difficult to encode in a K-means family algorithm as alignment is a local feature. Still, some clear limitations exist. 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. 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. As such, immediate future work must deal with finding optimal reading strategies for all algorithms.

References

Sudipta Acharya, Sriparna Saha, José G. Moreno, and Gaël Dias. 2014. Multi-objective search results clustering. In *25th International Conference on Computational Linguistics (COLING)*, pages 99–108.

Derar Alassi and Reda Alhajj. 2013. Effectiveness of template detection on noise reduction and websites summarization. *Information Sciences*, 219:41–72.

Jayendra Barua, Dhaval Patel, and Ankur Kumar Agrawal. 2014. Removing noise content from online news articles. In *20th International Conference on Management of Data (SIGMOD)*, pages 113–116.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003a. Extracting content structure for web pages based on visual representation. In *5th Asia-Pacific Web Conference on Web Technologies and Applications (ApWeb)*, pages 406–417.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003b. Vips: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft, November.

Yu Chen, Wei-Ying Ma, and Hong-Jiang Zhang. 2003. Detecting web page structure for adaptive viewing on small form factor devices. In *12th International Conference on World Wide Web (WWW)*, pages 225–233.

Gemma Fitzsimmons, Mark J. Weal, and Denis Drieghe. 2014. Skim reading: an adaptive strategy for reading on the web. In *ACM Web Science Conference (WebSci)*, pages 211–219.

João Guerreiro and Daniel Gonçalves. 2015. Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech. In *17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS)*, pages 3–11.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *15th Berkeley Symposium on Mathematical Statistics and Probability (BSMSP)*, volume 1, pages 281–297.

Elena Manishina, Jean-Marc Lecarpentier, Fabrice Maurel, Stéphane Ferrari, and Busson Maxence. 2016. Tag Thunder : Towards Non-Visual Web Page Skimming. In *18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.

K. Pernice, K. Whitenton, and J. Nielsen. 2014. *How People Read on the Web: The Eyetracking Evidence*. Nielsen Norman Group.

Andrés Sanoja and Stéphane Gançarski. 2014. Block-o-matic: A web page segmentation framework. In *International Conference on Multimedia Computing and Systems (ICMCS)*, pages 595–600.

Andrés Sanoja and Stéphane Gançarski. 2015. Web page segmentation evaluation. In *30th Annual ACM Symposium on Applied Computing (SAC)*, pages 753–760.

Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating noisy information in web pages for data mining. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 296–305.

Xinyi Yin and Wee Sun Lee. 2005. Understanding the function of web elements for mobile content delivery using random walk models. In *14th International Conference on World Wide Web (WWW)*, pages 1150–1151.

Jan Zeleny, Radek Burget, and Jaroslav Zendulka. 2017. Box clustering segmentation: A new method for vision-based web page preprocessing. *Information Processing & Management*, 53(3):735–750.