DOM-Based Clustering Approach for Web Page Segmentation: A Comparative Study

Adrian Sterca[a], Oana Nourescu, Adriana Guran[b] and Camelia Serban[c]
Department of Computer Science, "Babeş-Bolyai" University, Cluj-Napoca, Romania

Keywords: Clustering, Web Page Segmentation, DOM.

Abstract: Web page segmentation plays a crucial role in analyzing and understanding the content of web pages, enabling various web-related tasks. The approaches based on computer vision and machine learning have limitations determined by the need of large datasets for training and validation. In this paper, we propose a Document Object Model (DOM) based approach that uses clustering algorithms for web page segmentation. By leveraging the hierarchical structure of the DOM, our approach aims to achieve accurate and reliable segmentation results. We conduct an empirical study, using a custom built dataset to compare the performance of different clustering algorithms for web segmentation. Our research objectives focus on dataset creation, features identification, distance metrics definition, and appropriate clustering algorithms selection. The findings provide insights into the effectiveness and limitations of our approach, enabling informed decision-making in real-world applications.

1 INTRODUCTION

The increasing in size and complexity of the WWW have led to a growing demand for effective techniques to analyze and understand web page content. Web page segmentation, the process of partitioning a web page into meaningful and semantically coherent regions, plays a crucial role in various web-related tasks such as information extraction, content classification, and layout analysis. Accurate segmentation can enhance the efficiency and effectiveness of these tasks, and improve the user experience.

The web page segmentation approaches based on computer vision have the drawback of needing large datasets of web pages widgets in order to accurately identify them on other web pages. In recent years, researchers have explored alternative approaches [Zhang et al., 2020] based on the Document Object Model (DOM), which represents the hierarchical structure of a web page. DOM-based segmentation offers several advantages, including improved robustness and interpretability.

In this paper, we evaluate several clustering techniques for partitioning a web page into cohesive regions based on the DOM tree structure. We selected a set of DOM-based features and proposed (dis)similarity metrics by considering the inherent structure and relationships between DOM elements, with the aim of achieving accurate and reliable segmentation results.

The main goal of this study is to evaluate the effectiveness of clustering algorithms on DOM features for web page segmentation. We conduct an empirical investigation and report our findings based on a comprehensive set of experiments using a set of diverse web pages. We focus our research objectives on three main directions: create the dataset containing 10 web pages and the ground truth corresponding to them, identify features and define appropriate distance metrics, and design the experiments by creating different combinations of clustering algorithms, features and metrics in order to determine an optimal partition of objects.

Our findings provide insights into the effectiveness, robustness, and limitations of our approach, allowing researchers and practitioners to make informed decisions when applying web page segmentation techniques in real-world applications.

The remainder of this paper is organized as follows: Section 2 provides the related work in web page segmentation based on clustering approaches. Section 3 details our proposed DOM-based clustering ap-

[a] https://orcid.org/0000-0002-5911-0269
[b] https://orcid.org/0000-0002-6172-8156
[c] https://orcid.org/0000-0002-5741-2597

Sterca, A., Nourescu, O., Guran, A. and Serban, C.
DOM-Based Clustering Approach for Web Page Segmentation: A Comparative Study.
DOI: 10.5220/0012184000003584
In *Proceedings of the 19th International Conference on Web Information Systems and Technologies (WEBIST 2023)*, pages 302-309
ISBN: 978-989-758-672-9; ISSN: 2184-3252

proach. Section 4 presents our experimental setup and methodology and discusses the results and analysis of our empirical study. Section 5 discusses the threats to validity, while Section 6 concludes the paper and highlights future research directions.

2 RELATED WORK

Web page segmentation aims to break a page into smaller blocks, in which contents with coherent semantics are kept together. Web segmentation can be useful for advertisement identification, information retrieval, archiving, adaptive web content delivery, topic distillation, focused crawling, and improving the accessibility of web content in non-visual environments.

There are multiple approaches to the segmentation of web pages. Two main approaches have been identified in addressing this problem: a computer-vision based approach and a DOM-based approach, with different results in different contexts of use. Hybrid methods have also been developed to take advantage of the results obtained by each of them.

If we dive deeper into the scientific literature, we find that each of the two aforementioned approaches can be further divided into other sub directions. The structure based approach refers to the use of the HTML tags, the DOM elements and their hierarchical relationships to detect blocks. The layout based approach focuses on the repetitive elements found in web sites to lead page partitioning. The hybrid approach uses both the structure and the layout of the page to create a hierarchy of blocks through block extraction and recursive refinement. The text based approach retrieves segments from web pages based on the properties of text such as paragraph similarity, clustering, among others. All the above directions are part of the large DOM-based approach to web segmentation.

The computer vision based approach takes an image of a web page (also called snapshot) and applies image processing techniques and usually machine learning to detect blocks.

The paper [Cai et al., 2003] presents the VIPS (VIsion-based Page Segmentation) algorithm to extract the content structure from a web page. To obtain the vision-based content structure of a page, the DOM structure and the visual cues (determined in a heuristic manner) are used.

The authors in [Chen et al., 2001] present an automatic approach to detect the functional properties and categories of objects in websites for FOM (Function-based Object Model) generation. FOM includes two complementary parts: Basic FOM and Specific FOM. Basic FOM represents an Object by its basic functional properties and Specific FOM represents an Object by its category. The function of an object can be of presentation, navigation, interaction or decoration.

In [Mantratzis and Cassidy, 2005] an approach to identify important structures ("table-like" or "list-like" structures of hyperlinks) in a web page by operating at various levels within the DOM tree is presented.

Authors in [Andrew et al., 2019] consider web page segmentation as a clustering problem of visual elements, where all visual elements must be clustered, a fixed number of clusters must be discovered, and the elements of a cluster should be visually connected. For the web segmentation process, the authors rely strictly on the visual elements of a web page. DOM elements are enriched with calculated CSS features, and the basic visual elements correspond to the last block elements in each branch of the DOM tree. Three clustering algorithms are considered: K-means, F-K-means and Guided Expansion (GE).

Block-O-Matic [Sanoja and Gançarski, 2014] is a framework for web page segmentation consisting in three steps: segmentation analysis, understanding and reconstruction. Three corresponding structures are involved: DOM tree, content structure and logical structure. The result is a tree, which contains the structures enriched with the flow.

In [Jayashree et al., 2022b] a multi-objective clustering technique called MCS that relies on K-means is presented. The method uses visual, logical, and text cues and an evolutionary process automatically discovers the optimal number of clusters (segments) as well as the correct positioning of seeds.

In relation to existing approaches that use clustering methods for web page segmentation, our approach offers a comparison of clustering algorithms and various DOM-based features in order to determine the optimal configuration of HTML objects features, distance metrics and clustering algorithms to gain efficiency and effectiveness in the web page segmentation process.

3 THEORETICAL BACKGROUND

In our approach, in order to provide a semantic meaning to the web page blocks, the clustering method is used to group similar objects based on three feature categories: *visual, positional and DOM hierarchy related*.

3.1 DOM-Based Features

For an HTML document, let $X = \{O_1, O_2, ..., O_n\}$ be the set of objects representing all DOM objects representing HTML tags from this document that are visible in the browser, are leaf nodes in the DOM structure and also have a textual inner content (but have no node-type children). The set X does not include tags like $< script >$ and $< style >$ which are not visible in the browser or tags like $< img >$ and $< input >$ which have no textual inner content or tags that do not have textual inner content, but have direct children which contain textual content. We ignored tags that do not contain inner textual content like $< img >$ and $< input >$, because we specifically wanted to segment only the textual content of the HTML document. For each object in the set X, we define a set of features $F_1, F_2, ..., F_m$ that characterize all objects in X.

The features $F_1, F_2, ..., F_m$ defined for the objects in the set X are the following:

- visual features: background-color, font-color, font-size, font-family, font-width, border-width, and border-color;
- positional features: left, top, width, height;
- DOM feature: based on the XPATH distance (see Section 3.3) to the closest common ancestor from the DOM hierarchy.

3.2 Clustering Based Web Page Segmentation

Clustering is the division of a data set into groups (clusters) such that, similar objects belong to the same cluster and dis-similar objects to different clusters. One well-known example of a dissimilarity measure is the Euclidean distance.

In what follows we formalize the problem of clustering based web page segmentation. Let $X = \{O_1, O_2, ..., O_n\}$ be the set of n objects to be clustered. The objects considered to be clustered are the elements from the DOM model extracted as described in Section 3.1. Using the vector space model, each object is measured with respect to a set of m initial features $F_1, F_2, ..., F_m$ (a set of relevant attributes of the analyzed objects) and is therefore described by an m-dimensional vector $O_i = (O_{i1}, O_{i2}, ..., O_{im})$, $O_{il} \in \Re$, $1 \leq i \leq n$; $1 \leq l \leq m$.

Our aim is to find a partition that best represents the cluster substructure of the data set X. Objects of the same class should be as similar as possible, and objects of different classes should be as dissimilar as possible.

Definition 1. ***Partition of the set X corresponding to a DOM object d.***
The set $C = \{C_1, C_2, ...C_k\}$ is called a partition of X if and only if the following conditions are satisfied:

- $C_i \subseteq X$, $C_i \neq \emptyset$,$1 \leq i \leq k$
- $C_i \cap C_j = \emptyset$, $1 \leq i < j \leq k$
- $\bigcup_{i=1}^{k} C_i = X$.

In the following we will refer to C_i as the i-th cluster of C and to C as a set of clusters. A typical clustering algorithm uses the number of clusters as input parameter of the algorithm. A possible solution to this drawback is the use of hierarchical clustering algorithms, which produces not only the optimal number of classes (based on the needed granularity), but also a binary hierarchy that shows the existing relationships between the classes. The result of a hierarchical clustering algorithm can be graphically displayed as a tree, called a *dendogram*. This tree graphically displays the clustering process.

There are two basic approaches to generating a hierarchical clustering:

- *Agglomerative*: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters.
- *Divisive*: Start with one, all-inclusive cluster and, at each step, splits a cluster until only singleton clusters of individual points remain.

3.3 Proposed Distance Metrics

We propose the use of three types of distance functions for each of the three types of features (visual, positioning, xpath). The clustering algorithms use the distance metrics individually and then combine all three distance values into one single final distance metric.

The dissimilarity distance function, **visualDissim**, between visual features vectors v_1 and v_2 is:

$$visualDissim(v_1, v_2) = \sum_{i=1}^{|v_1|} \mathbb{1}_{v_1^i \neq v_2^i} \quad (1)$$

where v_1 and v_2 are feature vectors and each of these vectors has 7 visual features: *background-color, font-size, font-color, font-family, font-weight, border-width, border-color*; v_1^i represents the *i-th* element of the vector v_1; the function $\mathbb{1}_{v_1^i = v_2^i}$ returns 1 if $v_1^i = v_2^i$ and 0 otherwise.

The dissimilarity distance function, **positionalDissim**, between positional features vectors is:

$$positionalDissim(v_1, v_2) = bbd \quad (2)$$

where v_1 and v_2 are feature vectors and each of these vectors has 4 feature values defining the bounding box of a DOM element: *left, top, width, height* and *bbd* represents *Border-to-border distance* defined in [Jayashree et al., 2022a].

The dissimilarity distance function is defined by the **XPath** distance. The XPath distance is computed as the sum of the distances from each element to their last common ancestor. We have defined a function that extracts the common ancestor index by comparing the reversed lists of XPath nodes and finding the last index where they differ.

4 RESEARCH DESIGN

4.1 Research Questions

Considering the proposed investigated problem our aim is to find a partition that best represents the cluster substructure of the set of HTML objects from the web page.

The proposed approach aims to determine the clustering algorithms that increase the accuracy in the web page segmentation problem, and to explore the most relevant attributes associated with the DOM elements. Another important aspect in this investigation is the definition of a distance metric that quantify objects' dissimilarity. Therefore, we formulated the following research questions:

RQ1: What DOM elements features would be most suitable for applying clustering in the problem of web page segmentation?

RQ2: What kind of clustering algorithms give better results for web page segmentation?

4.2 Data Set Creation

To verify our approach we have used a custom data set created for this purpose, similar to other studies like [Cai et al., 2003, Chen et al., 2001, Sanoja and Gançarski, 2014]. Initially, we have used simple web pages that are easy to be verified by a human expert. Then, we have increased the complexity of the web pages used for our study. Our custom data set contains 10 web pages from different domains to ensure a variety of information layout and structure. The web pages have been downloaded together with their external resources such that in the future they can be used by other authors without difficulties (images and other resources that become unavailable). We tried initially to use a publicly available dataset for web segmentation [Kiesel et al., 2020], but we encountered difficulties related to establishing the ground truth (i.e. the ground truth labeling depended on the screen resolution which was not specified in the dataset) and various resources of the HTML pages from the dataset were missing (i.e. they were not included in the dataset, but were referenced remotely and were not available anymore at test time). From the selected pages, we have extracted the non-empty, textual HTML leaf nodes. The data extraction is performed after opening the webpage in the browser, by running a JavaScript script in the developer tools console. The JavaScript identifies the DOM elements within the body of the HTML DOM, then iterates through each element and applies certain conditions to filter out non-textual elements (e.g., ignoring script tags etc.) and checks if the element is visible on the page.

Each selected HTML element from a web page is described with a total of twenty three attributes (features) which are in a CSV file. These attributes refer to the visual, positional and hierarchical features of each element as they are described in Section 3.1. The CSV files will serve as input for the clustering algorithm.

The selected web pages used for clustering are described in Table 1.

To validate the proposed approach regarding web segmentation using clustering, we also built the ground truth for the selected pages. This is a challenge often encountered in the field of web segmentation regarding the decision of what can be considered the ground truth in the clustering problem. For smaller web pages (eg: fitting in the viewport without requiring scrolling, fewer HTML elements, etc.), it may be easier to agree on the result of manual web segmentation, but this process is still a subjective task and may vary depending on the person performing the manual grouping due to one's own understanding of UI and previous experiences. One of the team members who participated in building the ground truth has Human Computer Interaction as main research field. Unlike some image segmentation tasks, where the ground truth can be easily obtained by manual annotation, web page segmentation has no definitive standards or universally accepted segmentation rules and guidelines. To avoid subjectivity as much as possible, three authors of this paper have independently labeled the ground truth clusters on the selected web pages and after that they reached a consensus which formed the final ground truth.

4.3 Research Method

To answer the research questions we have designed 17 experiments as described in Table 3. Each exper-

Table 1: The data set used for clustering.

Page Id	Web page	URL	No. of HTML elements	No. of ground-truth clusters
1	Azure	https://azure.microsoft.com/en-us/	997	23
2	Baeldung	https://www.baeldung.com/	122	15
3	CSUBB	https://www.cs.ubbcluj.ro/	119	22
4	Jira	installed locally	264	13
5	Maven	https://maven.apache.org/	114	6
6	CityHall	https://primariaclujnapoca.ro/	554	13
7	ScienceDirect	https://www.sciencedirect.com/	177	11
8	Smurd	https://smurd-cluj.ro/	68	9
9	UAIC	https://www.uaic.ro/	65	11
10	W3Schools	https://www.w3schools.com/	918	26

iment is characterized by three aspects: the features used for the DOM elements, the distance metric, and the algorithm used for clustering. Table 2 describes the mapping of the attributes/features that characterize the DOM objects with the distance metrics selected for the clustering algorithms.

The three clustering algorithms have been used with 3 types of features and 3 distance metrics. Out of the seventeen designed experiments, three of them have been designed using the K-Means algorithm, six experiments using the Agglomerative clustering algorithm, and six experiments using the Optics algorithm [Ankerst et al., 1999, Breunig et al., 1999]. For the KMeans algorithm, we have used only the Euclidean distance with visual features, positional features, and all features (visual and positional); because we could use only the Euclidean metric with KMeans, we also could not use the XPath features with this clustering algorithm. The Python library we used for running the KMeans algorithm does not provide the option of having the distance metric as input parameter. This limitation reduced the number of experiments for the KMeans algorithm.

The Agglomerative algorithm has been used with the custom metrics (visual, positional, xpath and their combination) and has benefited of the number of clusters as input parameter. The Optics algorithm automatically determines the optimal number of clusters. In the case of the 6 experiments designed by us, the Optics algorithm determined 5 clusters for 3 of the experiments, using the Euclidean metric, and 12 clusters for the other 3 experiments using the proposed positional and visual metrics.

4.4 Results Analysis

In order to validate our proposed approach we selected two clustering performance evaluation metrics named: Random Index (RI) and Adjusted Random Index (ARI).

4.4.1 Clustering Performance Evaluation

Given the knowledge of the ground truth class assignments *labels_true* and our clustering algorithm assignments of the same samples *labels_pred*, the (adjusted or unadjusted) Rand index [Saaty, 1980] is a function that measures the similarity of the two assignments, ignoring permutations. If C is a ground truth class assignment and K the clustering one, let us define as: a, the number of pairs of elements that are in the same set in C and in the same set in K and b, the number of pairs of elements that are in different sets in C and in different sets in K.

Random Index. The Unadjusted Rand Index is then given by:

$$RI = \frac{a+b}{\binom{n}{2}}$$

where $\binom{n}{2}$ is the total number of possible pairs in the dataset.

Adjusted Random Index. To better understand the results of our research we have computed the Adjusted Random Index [Hubert and Arabie, 1985, Chacón and Rastrojo, 2023], which adjusts for a chance the raw RI. The formula for ARI is:

$$ARI = \frac{RI - Expected\_RI}{max(RI) - Expected\_RI}$$

where $max(RI) = 1$ and $Expected\_RI$ is the expected value of the RI index when the partitions are made at random.

Table 2: Selected distance metrics.

Features/ Distance Metric	Euclidean (EM)	Visual Similarity (VSM)	Positional Dissimilarity (PDM)	XPath (XM)
Visual (VF)	X	X		
Positional (PF)	X		X	
XPath (XP)	X			X
All (AF)	X			

Table 3: Performed Experiments.

Exp	Clustering Alg.	Features	Metric
1	K-Means	VF	EM
2	K-Means	PF	EM
3	K-Means	AF	EM
4	Agglomerative	VF	EM
5	Agglomerative	VF	VSM
6	Agglomerative	PF	EM
7	Agglomerative	PF	PDM
8	Agglomerative	XF	XM
9	Agglomerative	AF	EM
10	Agglomerative	AF	VSM + PDM
11	Optics	VF	EM
12	Optics	VF	VSM
13	Optics	PF	EM
14	Optics	PF	PDM
15	Optics	XF	XM
16	Optics	AF	EM
17	Optics	AF	VSM + PDM

The obtained results for our experiments are depicted in Table 4.

4.4.2 Response to Research Questions

Analyzing the above results for the 10 previously described web pages, we can observe the following:

- For the Azure page, experiments 5, 7, 8, 10 gave the best results, respectively for the Agglomerative clustering algorithm, the RI and ARI values being 0.863 and 0.682. The features implied in these experiments are: VF, PF, XP, All with EM, XM, VSM+PDM distance metrics.
- Considering the Baeldung page, the same experiments as in the Azure page, 5, 7, 8, 10 gave the best results, the RI and ARI values being 0.499 and 0.499.
- The Experiments 5,7,8,10 gave the best results in case of CSUBB page also, the RI and ARI values being 0.86 and 0.639.
- For the JIRA page, the best results are obtaind with the experiments 4,6,9 respectively for the Agglomerative clustering algorithm, the RI and ARI values being 0.772 and 0.506 for the VF, PF and AF features, and the distance metric being the Euclidean metric.
- For the Maven page, the results are very similar to JIRA. Experiments 4,6,9 gave the best results, respectively for the Agglomerative clustering algorithm.
- For the CityHall, ScienceDirect, SMURD, UAIC and W3Schools pages, the best results are obtained for experiments 5,7,8,10, respectively for the Agglomerative clustering algorithm. Analyzing these pages in comparison Azure, Baeldung and CSUBB the experiments with the best results are the same, only the AI and ARI values differ.

Concluding the analysis of the results, we can answer the two research questions stated previously, namely:

Response to RQ1: What DOM elements features would be most suitable for applying clustering in the problem of web page segmentation?

All selected features conducted to the best clustering results, both individual or combined together.

Response to RQ2: What kind of clustering algorithms give better results for web page segmentation?

The best results are obtained for the Agglomerative clustering algorithm for all analyzed pages. Still, the KMeans and Optics algorithms give similar results to the Agglomerative algorithm, excepting two of the analyzed pages: Azure and CityHall. Additionally, analyzing the distance metrics used during the experiments, we observe that for Jira and Maven pages the best results are obtained with the Euclidean metric, while for the rest of the pages the custom metrics provide the best results. As a conclusion, for 8 of 10 pages the best results are obtained using our proposed distance metrics.

5 THREATS TO VALIDITY

Our reported results are based on an empirical study, thus being subject to certain threats to validity. In

Table 4: IR and AIR results for experiments.

Ex	Azure		Baeldung		CS		Jira		Maven		CityHall		ScienceDirect		SMURD		UAIC		W3schools	
	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI	RI	ARI
1	0.55	0.245	0.886	0.303	0.952	0.532	0.767	0.495	0.783	0.484	0.627	0.343	0.852	0.188	0.8	0.268	0.946	0.78	0.679	0.294
2	0.55	0.245	0.886	0.303	0.952	0.532	0.767	0.495	0.783	0.484	0.627	0.343	0.852	0.188	0.8	0.268	0.946	0.78	0.679	0.294
3	0.55	0.245	0.886	0.303	0.952	0.532	0.767	0.495	0.783	0.484	0.627	0.343	0.852	0.188	0.8	0.268	0.946	0.78	0.679	0.294
4	0.55	0.245	0.884	0.298	0.949	0.506	**0.772**	**0.506**	**0.808**	**0.531**	0.622	0.336	0.853	0.198	0.8	0.268	0.945	0.779	0.68	0.297
5	**0.863**	**0.682**	**0.905**	**0.499**	**0.96**	**0.639**	0.72	0.393	0.742	0.333	**0.925**	**0.833**	**0.854**	**0.339**	**0.841**	**0.415**	**0.954**	**0.812**	**0.701**	**0.349**
6	0.55	0.245	0.884	0.298	0.949	0.506	**0.772**	**0.506**	**0.808**	**0.531**	0.622	0.336	0.853	0.198	0.8	0.268	0.945	0.779	0.68	0.297
7	**0.863**	**0.682**	**0.905**	**0.499**	**0.96**	**0.639**	0.72	0.393	0.742	0.333	**0.925**	**0.833**	**0.854**	**0.339**	**0.841**	**0.415**	**0.954**	**0.812**	**0.701**	**0.349**
8	**0.863**	**0.682**	**0.905**	**0.499**	**0.96**	**0.639**	0.72	0.393	0.742	0.333	**0.925**	**0.833**	**0.854**	**0.339**	**0.841**	**0.415**	**0.954**	**0.812**	**0.701**	**0.349**
9	0.55	0.245	0.884	0.298	0.949	0.506	**0.772**	**0.506**	**0.808**	**0.531**	0.622	0.336	0.853	0.198	0.8	0.268	0.945	0.779	0.68	0.297
10	**0.863**	**0.682**	**0.905**	**0.499**	**0.96**	**0.639**	0.72	0.393	0.742	0.333	**0.925**	**0.833**	**0.854**	**0.339**	**0.841**	**0.415**	**0.954**	**0.812**	**0.701**	**0.349**
11	0.454	0.156	0.824	0.187	0.94	0.528	0.702	0.352	0.71	0.209	0.616	0.328	0.761	0.08	0.732	0.209	0.935	0.763	0.656	0.241
12	0.454	0.155	0.891	0.355	0.934	0.51	0.728	0.405	0.684	0.135	0.615	0.327	0.835	0.2	0.769	0.287	0.931	0.751	0.661	0.253
13	0.454	0.156	0.824	0.187	0.94	0.528	0.702	0.352	0.71	0.209	0.616	0.328	0.761	0.08	0.732	0.209	0.935	0.763	0.656	0.241
14	0.454	0.155	0.891	0.355	0.934	0.51	0.728	0.405	0.684	0.135	0.615	0.327	0.835	0.2	0.769	0.287	0.931	0.751	0.661	0.253
15	0.454	0.155	0.891	0.355	0.934	0.51	0.728	0.405	0.684	0.135	0.615	0.327	0.835	0.2	0.769	0.287	0.931	0.751	0.661	0.253
16	0.454	0.156	0.824	0.187	0.94	0.528	0.702	0.352	0.71	0.209	0.616	0.328	0.761	0.08	0.732	0.209	0.935	0.763	0.656	0.241
17	0.454	0.155	0.891	0.355	0.934	0.51	0.728	0.405	0.684	0.135	0.615	0.327	0.835	0.2	0.769	0.287	0.931	0.751	0.661	0.253

what follows, we present the major threats to the validity of our research and we also explain our actions to mitigate them.

Internal Validity refers to factors that could have influenced the obtained results.

Features: The selection of features used for web page segmentation could impact the results. We carefully identified relevant features based on established practices in the field.

Distance Metrics: The choice of evaluation metrics for assessing the quality of segmentation plays a crucial role. We selected metrics commonly used in the literature and we defined new custom distance metrics, which we believe are appropriate for evaluating the effectiveness of our approach.

Algorithms Selected: We employed the KMeans clustering algorithm as it is a well-known and widely used algorithm in the literature for clustering tasks. Additionally, we explored the OPTICS algorithm as an alternative. However, we encountered difficulties in determining the appropriate epsilon parameter for DBSCAN, leading us to opt for OPTICS instead. Although these algorithms are commonly employed, there could be other clustering algorithms that might yield different segmentation results.

To mitigate these threats, we conducted extensive literature reviews, consulted with experts in the field, and performed preliminary experiments to ensure the robustness of our choices regarding features, metrics, and algorithms.

External Validity concerns the generalization of our findings.

The following factors could potentially affect the external validity of our research:

HTML Code Structure: The manner in which HTML code is structured, including the usage of div tags and spaces for alignment, may impact the web page segmentation process.

Dataset Size: The size of the dataset used for training and evaluation might affect the generalization of our findings. Although we collected a diverse and representative dataset, variations in dataset size or composition in other contexts could potentially lead to different segmentation outcomes.

Content of the Dataset: The content of the dataset we used for experimentation might not capture the entire range of web page structures and layouts. Different domains or types of websites may exhibit distinct characteristics that could impact the segmentation results. Consequently, generalizing our findings to web pages with dissimilar structures should be done with caution.

To address these concerns, we made efforts to collect a varied dataset with a sufficient number of web pages from diverse sources. However, we acknowledge that there might still be variations and complexities in real-world web pages that could influence the generalizability of our findings.

6 CONCLUSIONS

In this paper, we presented a comparative study on web page segmentation using DOM-based clustering. This approach leverages the hierarchical structure of the DOM to partition web pages into cohesive and semantically meaningful regions. We have considered various features categories: visual, positional, and XPath and predefined (Euclidean) and custom metrics to identify the most appropriate approach. From the experiments developed, we have discovered that the Agglomerative clustering algorithm applied with custom metrics like positional similarity and XPath metric return the best results for the set of web pages

we have explored.

In conclusion, our research contributes to the field of web page segmentation by comparing various DOM-based clustering algorithms. The empirical study evaluates the effectiveness of the studied clustering algorithms for web segmentation and provides valuable insights for its application in real-world scenarios. Future research directions include refining the approach to handle more complex web page layouts, investigating the scalability of the approach for large-scale web page datasets, and exploring additional features and clustering techniques for further improvements.

ACKNOWLEDGEMENT

The present work has received financial support through the project: *Integrated system for automating business processes using artificial intelligence*, POC/163/1/3/121075 - a Project Cofinanced by the European Regional Development Fund (ERDF) through the Competitiveness Operational Programme 2014-2020.

REFERENCES

Andrew, J. J., Ferrari, S., Maurel, F., Dias, G., and Giguet, E. (2019). Web page segmentation for non visual skimming. In *The 33rd Pacific Asia Conference on Language*, Japan. hal-02309625. Information and Computation (PACLIC 33) Hakodate.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (1999). Optics-of: Identifying local outliers. In *Principles of Data Mining and Knowledge Discovery: Third European Conference, PKDD'99, Prague, Czech Republic, September 15-18, 1999. Proceedings 3*, pages 262–270. Springer.

Cai, D., Yu, S., Wen, J.-R., and Ma, W.-Y. (2003). Extracting content structure for web pages based on visual representation. In *Web Technologies and Applications: 5th Asia-Pacific Web Conference, APWeb 2003, Xian, China, April 23–25, 2003 Proceedings 5*, pages 406–417. Springer.

Chacón, J. E. and Rastrojo, A. I. (2023). Minimum adjusted rand index for two clusterings of a given size. *Adv Data Anal Classif*, 17:125–133.

Chen, J., Zhou, B., Shi, J., Zhang, H., and Fengwu, Q. (2001). Function-based object model towards website adaptation. In *01). Association for Computing Machinery, New York, NY, USA*, pages 587–596. Proceedings of the 10th international conference on World Wide Web (WWW.

Hubert, L. and Arabie, P. C. p. (1985). Journal of classification 2. pages 193–218.

Jayashree, S. R., Dias, G., Andrew, J. J., Saha, S., Maurel, F., and Ferrari, S. (2022a). Multimodal web page segmentation using self-organized multi-objective clustering. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–49.

Jayashree, S. R., Dias, G., Andrew, J. J., Saha, S., Maurel, F., and Ferrari, S. (2022b). Multimodal web page segmentation using self-organized multi-objective clustering. acm trans. *Inf*, 40:3.

Kiesel, J., Kneist, F., Meyer, L., Komlossy, K., Stein, B., and Potthast, M. (2020). Web page segmentation revisited: Evaluation framework and dataset. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3047–3054, New York, NY, USA. Association for Computing Machinery.

Mantratzis, C. and Cassidy, S. (2005). Dom-based xhtml document structure analysis separating content from navigation elements. In *International Conference on Computational Intelligence for Modelling*, pages 632–637, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), Vienna, Austria. Control and Automation and International Conference on Intelligent Agents.

Saaty, T. L. (1980). *The analytic hierarchy process*. McGraw-Hill, New York, NY.

Sanoja, A. and Gançarski, S. (2014). Block-o-matic: A web page segmentation framework. In *2014 international conference on multimedia computing and systems (ICMCS)*, pages 595–600. IEEE.

Zhang, S., Wu, J., and Yang, K. (2020). A webpage segmentation method based on node information entropy of dom tree. In *Journal of Physics: Conference Series*, volume 1624, page 032023. IOP Publishing.