A New Semantic Approach to Improve Webpage Segmentation

Saeedeh Sadat Sajjadi Ghaemmaghami and James Miller*

University of Alberta, Canada
E-mail: sajjadig@ualberta.ca; jimm@ualberta.ca
**Corresponding Author*

Received 14 November 2020; Accepted 31 January 2021;
Publication 10 June 2021

Abstract

Webpage analysis is carried out for various purposes such as webpage segmentation. The goal of webpage segmentation is to divide a page into blocks that have similar elements. A fusion approach that combines different analyses is required in order to obtain high segmentation accuracy. In this paper, we propose a new fusion model for webpage segmentation, where we (1) merge webpage content into basic-blocks by simulating human perception; and, (2) identify similar blocks using semantic text similarity and regroup these similar blocks as fusion blocks. This approach is applied to three public datasets and evaluated by comparing with state-of-the-art algorithms. The results characterize that our proposed approach outperforms other existing webpage segmentation methods, in terms of accuracy.

Keywords: Webpage analysis, webpage segmentation, semantic text similarity, Gestalt Law of grouping.

1 Introduction

The World Wide Web has become a massive repository of information. The content and layout of webpages are getting more complex. Thus, identifying,

Journal of Web Engineering, Vol. 20_4, 963–992.
doi: 10.13052/jwe1540-9589.2042

and categorizing distinct informational elements from webpages has become increasingly difficult. Webpage segmentation provides a solution to this problem. Webpage segmentation is the process of partitioning a webpage into blocks (visually and semantically coherent segments of a webpage), in a manner, where each block contains distinctive content [1]. Also, humans tend to segment a webpage based on their understanding, thus, it is important to generate a segmentation model to segment a page by simulating human perception. Webpage segmentation presents one of the substantial preprocessing steps of webpage analysis and contributes to several applications' domains such as information retrieval, content classification, and content change detection [1]. For instance, in the context of information retrieval, segmentation can be used to extract specific parts of a webpage.

There are two major factors in segmenting a webpage into different blocks, (1) how the content of a webpage is extracted and (2) how the extracted content is processed to retrieve distinct information [6]. Most research extracts and organizes content relying on the DOM (Document Object Model) structure of an HTML page using heuristic or machine learning-based approaches [2–5]. The DOM structure considers a webpage document as a tree structure, where each fragment of the document is related to a particular node of the tree. Therefore, DOM represents the structure of webpage design. However, the information available in the DOM is very limited; the content of the DOM tree nodes are not included in the DOM structure. Additionally, full text content is not extracted from the complex DOM structure which results in lacking of the impressive feature extraction process [6]. To overcome these difficulties, some researchers prefer to segment webpages using the visual information in a webpage.

This vision-based segmentation method focuses on the analysis of visual features of the document content as they are perceived by a human reader. It exploits visual clues such as font size, font color, background color, spaces between paragraphs, etc. [7]. Considering the visual information allows these techniques to achieve a higher segmentation accuracy in comparison to DOM-based approaches [1], webpages are structured more flexible now.

Some research presents segmentation methods based on both the DOM-based and text-based approaches [8–10]. They focus on properties of the text content, such as density, to detect the content segments. However, they do not consider the semantic analysis in order to categorize pages. Semantic analysis includes extracting text from segmented blocks, computing textual similarity and regrouping blocks. A fusion approach that combines different analyses (DOM, vision and text-based segmentation) is required in order

to obtain higher segmentation accuracy. To achieve this, we propose a new fusion model of webpage segmentation, where we (1) merge the content of a webpage into blocks by considering human perception and, (2) identify similar blocks using semantic text similarity and regroup these similar blocks as fusion blocks. Thus, a fusion block includes related context regardless of different text formats.

This paper contributes to current research in webpage analysis in the following ways:

- It provides a new method of webpage segmentation by combining the DOM structure, visual features and text similarity metrics to improve the segmentation performance.
- It is applied to three public datasets and evaluated by comparing with state-of-the-art studies. The results show that our proposed approach outperforms four other existing webpage segmentation methods, in terms of higher precision, recall and accuracy.

The remainder of the paper is organized as follows. Section 2 explains the challenges and our research motivations; while Section 3 reviews related work. A detailed description of our approach is provided in Sections 4. Section 5 discusses an evaluation of the proposed approach. Section 6 concludes the paper and provides some future work directions.

2 Problem Statement and Research Motivation

It is quite easy for humans to recognize related webpage content fast and correctly from complex pages. However, with the huge number of webpages, it is impossible to identify and segment related information manually. Webpage segmentation models generated from different page features provide solutions to several webpage analysis problems. Webpage segmentation is often carried out for various purposes such as to retrieve or cluster information [1]. This paper seeks to improve on these previous attempts in a number of ways which are described below.

Traditional methodologies of webpage segmentation work well on textual pages [7], however due to the rapid development of Web technologies, they cannot efficiently process complex modern pages. Modern webpages present both static and dynamic content. Most of existing works use DOM tree for webpage segmentation. Segmentation using DOM structure of a page can only retrieve very limited information from rich-format webpages. According to Gestalt psychology, humans group objects based on a series of laws –

the Gestalt[1] law of grouping [11–13]. Therefore, generating a segmentation model by utilizing human perception which merges webpage content into segmented blocks can improve the performance of segmentation process [14].

The performance of structural-based segmentation methods can be limited by the shortcomings of the DOM structure. The complex DOM structure leads to the shortening or scattering of the long text of a webpage content. Thus, it is difficult to extract useful features from short text content, which challenges semantic analysis. Also, paragraphs with similar subjects are separated into different blocks because they contain text in different formats. A semantic analysis which computes the textual similarity between the segmented blocks is required to regroup the related blocks and obtain longer text in a merged block. This similarity regrouping model leads to the more stable semantic features. To achieve a high performance segmentation model and avoid biased results, it is required to merge logic, visual and text content of a webpage in a model. With semantic analysis methods such as Natural Language Processing (NLP), it yields better regrouping results. This paper utilizes a fusion webpage segmentation model to address these challenges and improve the semantic analysis of webpages.

3 Related Work

Webpage segmentation as a key step of webpage analysis has been explored in many research papers. The goal is to segment a page into blocks which have similar elements. This has been used for a variety of purposes, including information retrieval, content classification, webpage reconstruction, etc. Attempts of webpage segmentation have been carried out using a variety of methods. It can be divided into three approaches; DOM-based methods, vision-based methods, and text-based methods.

Most of the proposed methods use the DOM tree representation of pages to extract information content and segment pages. For example, a graph-structured webpage segmentation is proposed by Bing et al. [3] to build a training framework based on support vector machine.

Gupta et al. [15] analyzed HTML files to retrieve text. A webpage is passed through an HTML parser that creates a DOM tree representation of the page. The content extractor navigates the DOM tree using filtering techniques in order to modify specific nodes. This method eliminates non-content nodes and performed well on textual webpages, but it does not retrieve vast amount of information from rich format webpages.

[1] Gestalt Laws explain the mechanisms of how humans perceive and understand things.

Kang et al. [7] research the HTML tag repeat patterns in a webpage's DOM tree. According to these patterns this method breaks the webpage content into blocks. Most of the modern pages have a non-complicated layout design and these repetition patterns are not included considerably in their design. Although the repetition-based webpage segmentation algorithm can identify these patterns in textual pages, but it cannot perform well on the modern pages.

Chen et al. [16] identify the content blocks from the structure of a webpage. The whole page is regarded as a single content block. The page analysis algorithm partitions a content block into smaller ones. Finally, blocks are categorized into header, footer, left sidebar, right sidebar, and main content.

Tabular tags was used to maintain content in the design of a webpage and content blocks were identified using these tags. For example, Fan et al. [17] propose an approach to discover content from a set of pages within a single Website. The actual contents and the common presentation styles of the pages are captured using a Site Style Tree (SST) [18, 19]. The tree structure is built by aligning pages of the site. According to entropy and threshold method, informative content of each node of SST are discovered.

Webpages have more complex design, properties and format due to the development of browsers. Some research have been carried out to represent the design structure of a webpage using visual properties. Cai et al. [20, 21] propose a well-known Vision-based Page Segmentation algorithm (VIPS) which considers visual properties of a page. However, this method does not have as high performance in modern webpages as traditional pages [14].

Sanoja et al. [22] propose an approach inspired by automated document processing methods and visual-based content segmentation techniques. The segmentation process is split into three phases: analysis, understanding, and reconstruction of a webpage. Logical, visual and structural features of webpages are combined to analyze and understand the content.

Zeleny et al. [1] present a box clustering segmentation model which uses visual properties, the distance of elements and their visual similarity. The Segmentation model consists of three steps, box extraction, computing distances between the boxes and clustering boxes.

Mehta et al. [23] propose a segmentation algorithm which uses both visual and content information. The visual information is utilized (1) the VIPS algorithm to divide pages in to small elements, and (2) the content information using a pre-trained Naive Bayes classifier to create bigger blocks.

Jiang et al. [6] propose a webpage segmentation method that merges visual and logic features of content. It also employs text density (the number of words within a particular document), as its segmentation algorithm.

Blocks of webpages containing mostly textual content are well identified using traditional methodologies. However, these methodologies are not able to identify blocks of modern complex webpages with high performance. A method is required to segment blocks accurately based on a specific laws (the Gestalt laws of grouping) [11–13]. These laws simulate the process of understanding human perceptions in identifying related contents. "Gestalt Layer Merging" (GLM) model is proposed by Xu and Miller [14] and utilized the Gestalt laws to identify blocks in the complex modern webpages [14]. We used this method in this paper to generate basic-blocks.

Text-based segmentation approaches [8–10] focus on the properties of the text content, such as density, to detect the content segments. For example, Kohlschütter et al. [5] present a densitometric approach for identifying segments of a page using the text density metric. This method limits some text density metrics and cannot perform well on the modern technologies. These approaches do not use semantic analysis in order to webpage segmentation. They merely focused on a page structure and vision features without considering the text similarity metrics of blocks.

Our approach generates semantic blocks using Gestalt laws of grouping, and compares the text similarity of blocks to regroup these similar blocks as fusion blocks. Thus, a fusion block is composed of related blocks in terms of similar text contents using Natural Language Processing (NLP) algorithms. It is believed that these enhanced semantically-based, visually initiated blocks will deliver superior performance across a wide array of tasks on modern webpages. It is believed that this is the first time a model uses basic-blocks and fusion blocks together to semantically segment a webpage using a NLP algorithm.

4 The Proposed Fusion Webpage Segmentation Approach

Webpage segmentation is the process of partitioning a page into segments that represent the related information and be consistent with people [3]. To facilitate the description of our approach, an overview of segmentation procedure is given in the following paragraph.

4.1 Overview

The DOM elements represent the structure of webpages [22]. So far, most of the studies utilized the DOM elements to segment webpage content [22].

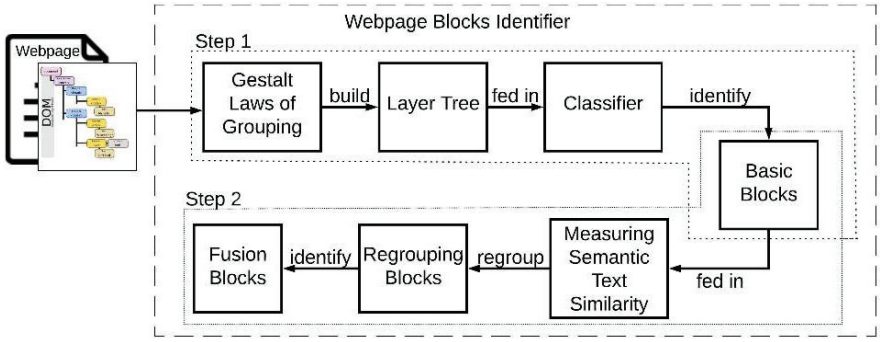

Figure 1 The framework of the webpage segmentation approach.

Relevant content groups together as blocks; each block contains distinctive content. The goal of webpage segmentation is to construct a content structure from webpage features which groups the elements of a webpage using metrics such as distances, locations and semantic context. A fusion model which includes DOM, vision and text-based segmentation approach is required to achieve a superior segmentation result.

4.2 Proposed approach

Our framework is designed and implemented to generate a webpage segmentation model for different webpages and to overcome the limitations of the former approaches. The main steps of our proposed framework are shown in Figure 1.

As shown in Figure 1, our model is mainly categorized in two steps; (1) it identifies basic-blocks in webpages in order to segment a webpage using the Gestalt laws of grouping technique, and (2) the model compares text similarity of basic-blocks identified in step1 and regroups the semantically related blocks as fusion blocks using a semantic analysis approach. These steps are explained in the following paragraphs.

4.2.1 Step 1: Basic-block

For the first step, our model segments webpage content into basic-blocks inspired by human understanding. This paper uses Gestalt laws to simulate human understanding and perception. In order to present a webpage, a "layer tree" is designed [14]. Two major ways, explored by researchers, to represent a webpage for visual similarity evaluation are: screen shots (images) and DOM trees. A webpage representation model (layer tree) is generated using

these two techniques according to the DOM elements of a page [24]. Details of constructing a layer tree can be found in [14]. The Gestalt laws are transformed in to six rules usable by machine which are expressed in the following paragraph. Further details can be found in [14].

1. The Gestalt law of simplicity indicates that basic format of shapes is preferred by people. In a webpage, the simplest representation of the content is DOM tree elements. Figure 2 shows one of the news' pages of "cbc.ca/news/". In this figure, the middle image between the texts contains multiple elements (i.e., the text "FOR BREAKING NEWS", an image, and video). However, they have various styles, they are grouped as a single image.

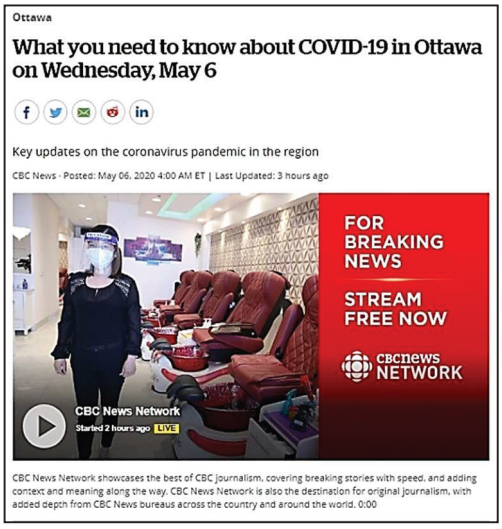

Figure 2 Gestalt law of simplicity ("cbc.ca/news/").

2. The Gestalt law of closure states that a complete format of shapes is preferred to be in a single category by human. For example, in Figure 3 ("ualberta.ca/admissions-programs"), the middle part of the background image is covered by a search box but according to this Gestalt law, the image is considered as a complete one.
3. The Gestalt law of proximity illustrates that close objects (in terms of distance) are preferred to be in a category by human. Based on this law, objects are categorized into different blocks by distance. The Normalized Hausdorff Distance (NHD) is used to calculate the distance of layers [26]. The elements with the same distances with adjacent

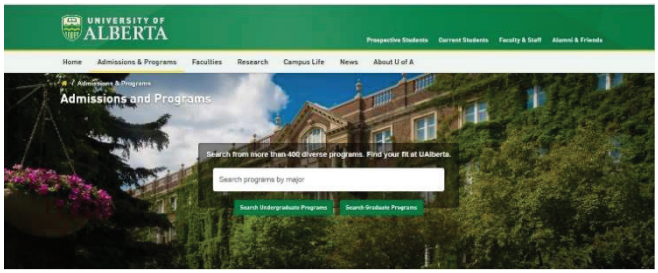

Figure 3 Gestalt law of closure (Webpage of "ualberta.ca/admissions-programs/).

elements are merged. Using the sign up page of Instagram (https://www.instagram.com/accounts/emailsignup/) as an example shown in Figure 4, the four boxes regarding sign up ("Mobile Number or Email", "Full Name", "Username", and "Password") are related and regarded as a group.

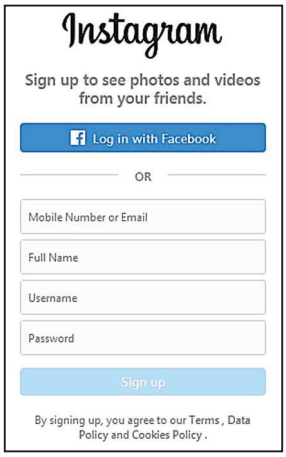

Figure 4 Gestalt law of proximity.

4. The Gestalt law of similarity describes that similar objects are preferred to be considered in a category by human. Background, foreground and size similarity of objects are considered in this law. Image and Color are included in background similarity; style and format of text are considered in foreground similarity; and the size of the blocks are included in the size similarity. See [25] for a more precise set of definitions. This paper uses CIE-Lab color space in the same manner as [26] in order to simulate human vision. We select $\triangle E_{00}^{12}$ as the color difference metric

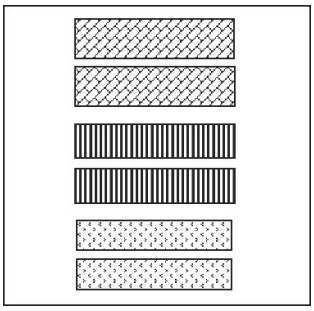

Figure 5 Gestalt law of similarity.

same as [26], calculated by Equation (1). The parameter list can be found in [26], and it is omitted in this paper for brevity.

$$\triangle E_{00}^{12} = \sqrt{\begin{array}{c}\left(\frac{\triangle L'}{k_L s_L}\right)^2 + \left(\frac{\triangle C'}{k_C s_C}\right)^2 + \left(\frac{\triangle H'}{k_H s_H}\right)^2 \\ +R_T \left(\frac{\triangle C'}{k_C s_C}\right)\left(\frac{\triangle H'}{k_H s_H}\right)\end{array}} \tag{1}$$

Structural Similarity Index (SSIM) [11] is employed to calculate digital image comparison in order to imitate human understandings. This paper uses SSIM in the same manner as [26]. As shown in Figure 5, the six objects are categorized into three groups in terms of styles. The two objects in the bottom, the two objects in the middle and the two objects at the top are categorized in three different groups.

5. The Gestalt law of continuity states that aligned objects are categorized in one group. The left-aligned lines in the orange rectangle shown in Figure 6 (the example of University of Alberta's homepage), includes "Student Information", "Register", "Student Union", etc., are categorized as a single group.
6. The Gestalt law of common fate categorized the elements with similar motion tendency in a group. For example, the lower ribbon with the red background color in Figure 7 (the homepage of global health in the Amazon, "amizade.org") hangs at the bottom and does not move with scrolling the page, but other content moves accordingly.

According to these six laws, a model can allow elements to be categorized whether in a group or not. This group of similar elements includes the results of six laws merged. Our approach uses naïve Bayes classifier [27] same

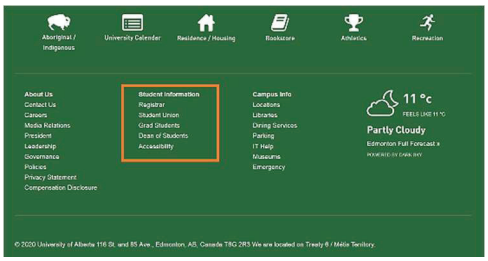

Figure 6 Gestalt law of continuity.

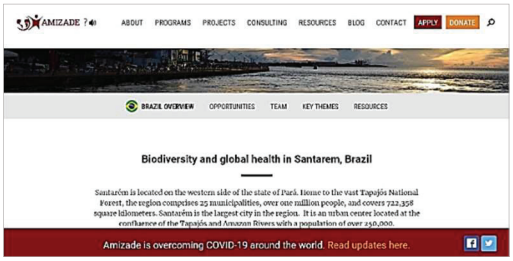

Figure 7 Homepage of "amizade.org".

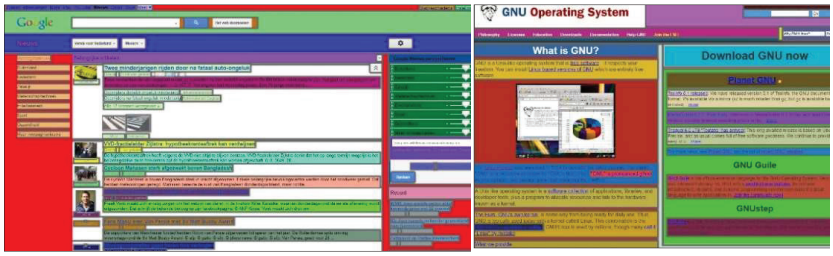

Figure 8 Two examples of segmented webpages using semantic-block algorithm [14].

as [14] to merge these laws. Figure 8 shows two webpages segmented by [14], where for each block, a different background color is assigned.

4.2.2 Step 2: Fusion block

Human understandings are simulated using Gestalt laws of grouping to identify basic-blocks. According to the Gestalt law of similarity, the similarity between blocks includes three features, (1) background, (2) foreground, and (3) size similarity. Block features such as color, textual styles, width and height are compared regardless of considering the semantic analysis. Hence,

objects may be segmented in different blocks, even though they have semantically related text. It is required to utilize semantic analysis to address this problem. This paper utilizes semantic text similarity to identify semantically related blocks and regroup them as a fusion block. A description of semantic analysis is explained in the following paragraphs.

Natural language processing (NLP) is a series of techniques simulated human technique of processing and analyzing a language. NLP system inventions are associated with the semantic analysis of linguistic structures [28]. Semantic analysis within the framework of NLP evaluates and represents human language and analyzes texts with an interpretation similar to those of human beings. Recently, text analysis is of one the popular topic of research, therefore, a wide range of text analysis applications exists including information retrieval, text representation, text similarity, etc. [29].

Text representation plays an important role in NLP. Tasks in this field rely on representations that can express the similarity and dissimilarity between textual elements [30]. Such representations and their associated similarity metrics have many applications. For example, one of the most common fixed-length methods of text representation is a bag-of-words [31]. Another text representation is TD-IDF (Term Frequency-Inverse Document Frequency) [32], based on the bag-of-words philosophy, which involves the assumption that a document is simply a collection of words, and thus, the document can be vectorized by computing the relative importance of each word, i.e., by considering the word's frequency in the document and its popularity in the corpus. TF-IDF and bag-of-words do not consider semantics of the words and also the ordering of the words are lost, which are the major weaknesses of these methods. For example, assume that the three words of "Country", "City" and "Flower" are equally distant in a document despite the fact that the first two words are semantically more related than "Flower".

There are some methods to determine the semantic similarity of texts. For example, Word2vec proposed by Google [33, 34] contains continuous bag-of-words models (CBOW) and continuous skip-gram models (skip-gram). Word2vec can learn word vectors in a short period of time from a large scale document and has been applied in different aspects of text processing such as text representation [35]. Some research utilizes word embedding [36–38] to understand the semantic logic's of text. Doc2vec was proposed by Mikolov et al. [39], inspired by Word2vec. The neural-network-based document embedding known as Doc2vec extends Word2vec from the word level to the document level [39]. Each document has its own vector values in the same space as that for words. Thus, the distributed representation for

both words and documents are learned simultaneously. Some of the major weaknesses of the bag-of-words models are addressed by Doc2vec. First advantage of Doc2vec is that it includes the semantic information of text. The second advantage is that it does not ignore the word order. In addition, this method can be applied to variable-length pieces of texts regardless of its length. Furthermore, it does not rely on the parse trees and does not require task-specific tuning of the word weighting function [39].

Doc2vec starts with training the model, subsequently, a vector space is built based on the word vectors, therefore, semantically similar words show similar vector representations (e.g., "City" is more related to "Country" than to "Flower") [39]. By training the word vectors, the document vector is trained, and the document's numeric representation is held in the end of training. The document vector represents the concept of a document as the concept of a word is represented by the word vector. Some studies have represented that Doc2vec results in better classification accuracy than other representation methods in different domains [39–41].

In this paper we compare the semantic similarity of nearby blocks using the Doc2vec technique. Doc2vec captures both semantic and syntactic information of words, and can be used to measure text similarity. We used genism's [42] implementation of Doc2vec to determine the semantic similarity of the blocks' text content. Therefore, related blocks are regrouped into a fusion block, which not only forms based on visual features by simulating human perception, but also utilizes semantic analysis of blocks to improve webpage segmentation.

As it is mentioned earlier, the shortcomings of the DOM structure can limit the performance of structural-based segmentation methods. Full text content is not extracted from the complex DOM structure which results in lacking of the impressive feature extraction process and challenges semantic analysis. Also, texts with similar subjects are categorized into different blocks since they contain text with different formats such as different font size, font color, etc. Our proposed approach addresses these problems using semantic analysis and regrouping the related scattered blocks in to a fusion block that contains longer text. Our semantic regrouping method is presented in Algorithm 1. In this algorithm, there are two inputs, basic-blocks and text difference limit t. (Text difference limit threshold can be set as $t = 0.5$, which is based on the empirical results presented in Section 5.) This algorithm semantically regroups blocks based on the text_similarity and Gestalt Laws of grouping (proximity and continuity). We used Doc2vec to evaluate the text similarity of blocks.

Algorithm 1 Semantic regrouping algorithm

Input: Two Basic Blocks (B_1, B_2). Text Difference Limit t
Output: Fusion Block
Begin
  if *Text_Similarity* * $(B_1, B_2) > t$
    Fusion_Block $\leftarrow$ *Regroup* ** (B_1, B_2)
  end
  return *Fusion_Block*
End
  * We used Doc2vec algorithm
  ** The blocks are regrouped according to Gestalt laws

According to the Gestalt law of proximity, close objects (in terms of distance) are preferred to be in a category by human. To measure distances between blocks, NHD is employed similar to [14] calculated as follow:

$$NHD(L_1, L_2) = \max\left\{\left(\frac{hd_{1,2}}{Re_{L_1}}, \frac{hd_{2,1}}{Re_{L_2}}\right)\right\} \tag{2}$$

where, Re_{L_1} and Re_{L_2} are the relevant lengths of layers L_1 and L_2, and $hd_{1,2}$ and $hd_{2,1}$ are the Hausdorff distance from L_1 to L_2 and L_2 to L_1, respectively. According to Gestalt law of continuity, humans tend to categorize aligned objects in one group. This law evaluates the positions of layers (left, top, right and bottom coordinates of layers (blocks)); an object belongs to a different group if it is not aligned with its siblings. The two blocks are continuous if the any of the four edges between these blocks have the same value [14]. We assume that the two blocks are different if the text difference limit of them is less than t. Thus, this model regroups nearby blocks with semantically related text together as a fusion block. The regrouping procedure considers the semantic similarity of text; related text may have different formats but they regroup as a fusion block to aggregate related contents into the same block. Having Text_Similarity calculated, the semantically related blocks can be grouped according to Gestalt laws of proximity and continuity [14]. If a series of sibling layers have the same proximity and blocks have the text difference limit greater than t, then they belong to one single group; the two layers will be placed into a different groups if any pairs of them share a different proximity than other pairs [14]. We utilized the Gestalt laws in addition to the text similarity of blocks in order to segment a webpage according to human perception. After regrouping, the blocks are transformed into bigger fusion blocks that contain much more stable semantic features than before. Due to the bigger blocks and longer text sentences, these features

can be extracted more accurately and can result in better performance on webpage segmentation.

5 Evaluation

This section presents experiments we have performed to verify the effectiveness of our proposed approach. To verify the effectiveness of our approach, we apply it on open-source public datasets. Also, we compare our method with four existing well-designed algorithms. We use four evaluation metrics for evaluating the performance of our method: precision, recall, F-1 score, and the Adjusted Rand Index (ARI). Our approach segments a webpage into fusion blocks. The experiments and their results are presented and discussed below.

5.1 Research Goal

The goal of this paper is to propose a new fusion model of webpage segmentation by combining the DOM structure, visual features and semantic text similarity metrics together to achieve better segmentation performance. Our method merges webpage content into basic-blocks by simulating human perception. It identifies similar blocks using semantic text similarity and regroups these basic-blocks as fusion blocks.

5.2 Dataset

The effectiveness of our proposed approach is evaluated against the following three datasets. These datasets are utilized to segment the content of webpages according to human judges (by using the semantic analysis approach).

1. DSpopular, a public dataset of 70 homepages of popular Websites with manually labeled ground truths for segmentation collected in 2014 [43]. This dataset contains three versions of each page including (1) the basic HTML, (2) a serialized version of the DOM after all external resources are loaded, and (3) a DOM page with manually labeled semantic blocks.
2. DSrandom, a public dataset of 82 homepages of popular Websites with manually labeled ground truths for segmentation collected in 2014 [44]. Each page includes three versions the same as DSpopular.
3. DSnew, a dataset of 50 homepages of popular Websites from Alexa Topsites [45] collected in 2017. These pages are viewed and labeled according to human judges.

Table 1 The statistics of the datasets

Dataset	Number of Webpages	Average Number of Blocks
$DS_{popular}$	70	12.59
DS_{random}	82	8.46
DS_{new}	50	18.33

The number of webpages and the average number of blocks in each dataset are shown in Table 1. It is crucial to have a ground truth, validated by human assessors to check algorithm correctness. Thus, the accuracy of our proposed method is evaluated using the manually labeled ground truths provided for each dataset. These datasets are collected from a real world environment and include type-rich content; therefore, they are suitable for evaluating our method of webpage segmentation.

5.3 Comparison Methods

We compare our proposed method (Fusion-Block) with the following four well-known existing webpage segmentation algorithms. The results show that our method is superior to all these algorithms in terms of semantic webpage segmentation based on human judgement (ground truth).

1. VIPS [20], a well-known approach of segmenting a webpage content structure based on its visual representation. This paper is used an open source implementation of VIPS [46] which was utilized in other papers [1, 47]. As the default setting of the tool, permitted Degree of Coherence parameter in VIPS is set to 8 the same as [6].
2. BoM [22], a hybrid webpage segmentation method which combines structural, visual and logical features of webpages. This method consists of three phases; analysis, understanding, and reconstruction of a webpage.
3. A webpage segmentation method [6] which combines visual, logic and features of the content on a webpage. For simplicity, we name this segmentation method as SegBlock in this paper.
4. Semantic-Block [14], a webpage block identification algorithm utilizing the Gestalt laws of grouping in order to simulate human perception.

5.4 Segmentation Accuracy

In order to verify the accuracy of the segmentation method computed by our approach and the other four comparison algorithms, we employ the following

four evaluation metrics (precision, recall, F-1 score and Adjusted Rand Index (ARI)). The segmentation result generated by our approach groups the elements of a webpage into cohesive regions visually and semantically. Similar to the previous works [4, 5], we regard each generated segment (block) as a cluster and employ cluster correlation metrics to conduct the evaluation.

1. Precision represents the ratio of correctly segmented blocks over the blocks segmented by the algorithm as Equation (3).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

   TP denotes two similar blocks are identified as similar, correctly; while FP indicates that two different blocks are identified as similar, incorrectly.
2. Recall represents the ratio of correctly segmented blocks over the ideal blocks that are manually obtained by humans (ground truth) as Equation (4).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

   FN indicates that two similar blocks are identified as different, incorrectly.
3. F-1 score which combines precision and recall computed as follow.

$$\text{F} - 1 \text{ score} = 2.\frac{\text{Precision.Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

4. Adjusted Rand Index (ARI) [48], which identifies the agreement between two clusters (segmented blocks and ground truth clustering) on a particular dataset shown in Equation (6). Value of the Rand Index is between 0 and 1; clusters' agreement on any pair of elements leads to value 1 shows these clusters are the same, and 0 states that the clusters do not agree on any elements. A version of the Rand Index is called ARI which has a value between 0 and 1; 1 shows that two blocks are identical and for random blocks the value is 0 on average. ARI can be calculated as follow.

Consider a set of n objects $S = \{O_1, O_2, \ldots, O_n\}$, and suppose that $X = \{x_1, x_2, \ldots, x_r\}$ and $Y = \{y_1, y_2, \ldots, y_s\}$ represent two different partitions (blocks) of the objects in S. Given two partitions, X and Y, with r and s subsets, respectively, the contingency Table 2 can be formed to indicate group

Table 2 Contingency table for comparing partitions X and Y

		Y				
Partition	Group	y_1	y_2	...	y_s	Sums
X	x_1	n_{11}	n_{12}	...	n_{1s}	a_1
	x_2	n_{21}	n_{22}	...	n_{2s}	a_2
	...	...	...	...	...	...
	x_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
	Sums	b_1	b_2	...	b_s	

overlap between X and Y as n_{ij}, where $n_{ij} = |x_i \cap y_j|$. In Table 2, a generic entry, n_{rs}, represents the number of objects that were partitioned in the rth subset of partition r and in the sth subset of partition s [50]. Thus, the ARI can be expressed as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{6}$$

5.5 Results

The evaluation of the segmentation algorithm is an important challenge. In order to compare one algorithm with another, it is crucial to have a ground truth, validated by human assessors to check algorithm correctness. Thus, in this section, we evaluate the performance of our algorithm against a ground truth. We also apply the same method for all the four comparison methods (BoM, VIPS, SegBlock and Semantic-Block) using the four evaluation metrics (precision, recall, F-1 score and ARI) and explain the results as follows.

All the pages in the datasets have been segmented with our proposed approach and the other four methods. Table 3 represents the results of the evaluation metrics in each dataset. "Total" includes the results obtained by using the three datasets (DSpopular, DSrandom and DSnew) together. The average number of correctly segmented blocks of webpages is represented in the "correct" column for each dataset; a block is said correctly segmented if its geometry and location are equal to a labeled block in the ground truth. According to this table, BoM, VIPS, SegBlock, Semantic-Block and Fusion-Block method achieve 25.74%, 24.14%, 38.16%, 41.67% and 48.06% of the average number of correctly labeled blocks in the Total dataset, respectively. Thus, it indicates that our approach (Fusion-Block) obtains an improvement in terms of average number of correctly labeled blocks.

Table 3 Evaluation results

	$DS_{popular}$					DS_{random}				
	Correct	Precision	Recall	F-1 Score	ARI	Correct	Precision	Recall	F-1 Score	ARI
BoM	2.78	30.5%	26.1%	28.1%	0.452	2.55	30.8%	33.0%	31.8%	0.473
VIPS	2.78	23.7%	26.2%	24.9%	0.420	1.97	27.8%	26.4%	27.1%	0.371
SegBlock	5.62	38.1%	40.2%	39.1%	0.530	3.74	41.9%	44.8%	43.3%	0.531
Semantic-Block	6.75	40.3%	43.4%	41.8%	0.532	3.82	43.3%	53.6%	48.0%	0.549
Fusion-Block	8.70	44.7%	54.1%	48.9%	0.598	4.54	49.3%	61.2%	54.6%	0.610
	DS_{new}					Total				
	Correct	Precision	Recall	F-1 Score	ARI	Correct	Precision	Recall	F-1 Score	ARI
BoM	5.54	30.5%	21.7%	25.3%	0.426	3.38	31.4%	27.9%	29.5%	0.450
VIPS	6.38	20.7%	24.4%	22.4%	0.415	3.17	24.7%	25.8%	25.2%	0.405
SegBlock	5.59	39.2%	38.9%	39.0%	0.464	5.01	39.6%	42.4%	40.9%	0.514
Semantic-Block	5.89	40.6 %	41.7%	41.1%	0.483	5.47	41.2 %	46.7%	43.8%	0.526
Fusion-Block	6.02	45.8%	48.2%	47.0%	0.533	6.31	46.3%	54.9%	50.2%	0.583

Table 4 ARI values of different threshold

Threshold	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ARI (Total Dataset)	0	0.085	0.092	0.210	0.584	**0.601**	0.562	0.525	0.528	0.523	0.514

The second higher value of the average number of correctly segmented blocks belongs to Semantic-Block algorithm. The Semantic-Block approach simulates human perception using the Gestalt laws of grouping (Gestalt law of simplicity, closure, etc.) in order to segment webpages. Simulating human perception allows this method to achieve the second highest average number of correctly labeled blocks. SegBlock reaches the third highest amount of average number of correctly labeled blocks. The SegBlock method segments webpages using visual and logic features of content. Also it utilizes the number of words within a particular document. Since SegBlock employs more features than BoM and VIPS, we can see that it reaches the third maximum average number of blocks after our approach and Semantic-Block. The BoM algorithm segments webpages using visual and logical features of the webpages. The VIPS algorithm relies mainly on visual separators of webpages to identify blocks. Modern webpages' layouts are more complicated than before and the visual separators are much less obvious [51]. Thus, BoM has a slightly better performance than VIPS in the amount of correctly segmented blocks found over the whole datasets.

Our approach groups semantically similar blocks using semantic text similarity method (we used Doc2vec algorithm shown in Algorithm 1). Text difference limit threshold (t in Algorithm 1) is set to $t = 0.5$, which is based on the empirical results shown in Table 4. This table shows different thresholds from 0 to 1. It represents that $t = 0.5$ results in the highest amount of ARI (comparing to the ground truth) in the Total dataset which is 0.601. The ARI distribution over the different text difference limit is shown in Figure 9. Thus, according to this result, we set $t = 0.5$.

As shown in Table 3, our approach (Fusion-Block) outperforms all the four comparison methods in terms of precision, recall, F-1 score and ARI. It obtains 46.3%, 54.9%, 50.2% and 0.583 in precision, recall, F-1 score, and ARI, respectively which shows a noticeable improvement on the segmentation's quality. Comparing our approach (Fusion-Block) with the other comparison methods, the following highlights are provided:

- On precision, Fusion-Block reaches 47.4%, 87.4%, 16.9% and 12.4% improvements against BoM, VIPS, SegBloack and Semantic-Block respectively.

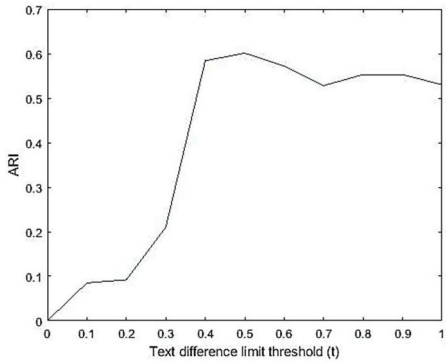

Figure 9 ARI distribution of different t.

- On recall, Fusion-Block reaches 96.8%, 112.8%, 29.5% and 17.5% improvements against BoM, VIPS, SegBloack and Semantic-Block, respectively.
- On F-1 score, Fusion-Block reaches 70.2%, 99.2%, 22.7% and 14.6% improvements against BoM, VIPS, SegBloack and Semantic-Block, respectively.
- On ARI, Fusion-Block reaches 29.5%, 43.9%, 13.4% and 10.84% improvements against BoM, VIPS, SegBloack and Semantic-Block, respectively.

These results demonstrate that our approach achieved the best performance and out-performed the other methods significantly. According to Table 3, the maximum amount of precision and recall values are belong to the DSrandom dataset compared to the DSpopular, and the DSnew datasets. It shows that webpages in DSrandom are tend to be less complicated (with less content) rather than DSpopular and DSnew. As shown in Table 3, the recall value of all the methods has its highest amount in DSrandom, and the second and third place belong to DSpopular and DSnew, respectively.

Our approach segments webpages according to human perception by combining the logic, visual and text content of a webpage in the segmentation model. It merges webpage content into blocks by utilizing the Gestalt laws of grouping to simulate human understandings. In addition, it compares the text similarity of blocks to regroup these similar blocks as fusion blocks. The other approaches do not use semantic analysis in order to segment webpages. They only focused on the page structure and the visual features without considering the semantic text similarity metrics of blocks. Utilizing the semantic

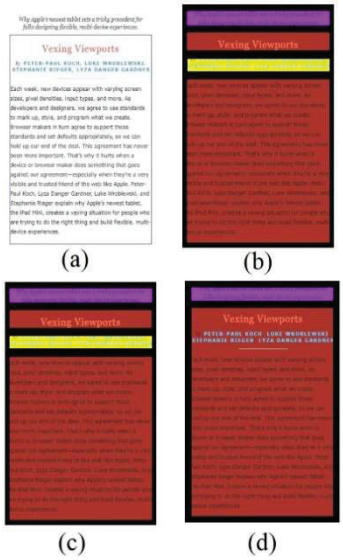

Figure 10 Homepage of "www.alistapart.com" from DSpopular.

text similarity, our approach reaches superior performance against all other four methods.

As represented in Table 3, VIPS has the minimum amount of precision, recall, F-1 score and ARI comparing to other methods respectively. This is because the VIPS employs only on visual features of webpages, which makes it perform less accurately on the evaluation metrics.

According to the results in Table 3, the three methods (Fusion-Block, Semantic-Block and SegBloack) were achieved ARI value greater than 0.5, which are 0.583, 0.526 and 0.514, in the Total dataset, respectively. It shows that their segmentation is not close to randomness. Additionally, these three methods were achieved F-1 score values more than 40% in the Total dataset, which are 40.9%, 43.8% and 50.2% for SegBloack, Semantic-Block and Fusion-Block, respectively.

In Figures 10 and 11, we offer a perspective of the impact caused by our approach on a webpage segmentation task in comparison with the impact caused by the SegBloack and Semantic-Block methods, which do not segment blocks using semantic text similarity of blocks; this limitation is indicated and can be found in the paper [6]. These figures show the homepages of the two Websites ("www.alistapart.com" and "www.koreanconsulate.on.ca") from DSpopular and DSrandom, respectively. Figures 10(a) and 11(a)

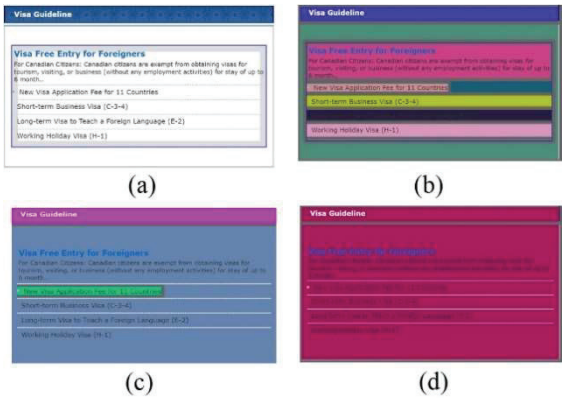

Figure 11 Homepage of "www.koreanconsulate.on.ca" from DSrandom.

represent the manually labeled ground truth of these webpages. Figures 10(b), 11(b) and 10(c), 11(c) show the result of webpage segmentation using SegBlock and Semantic-Block, respectively; and finally, Figures 10(d) and 11(d) demonstrate the pages segmented using our approach. According to Figure 10(b) and (c), we can see that SegBlock and Semantic-Block segmented the blocks identically. However, our method (Figure 10(d)) grouped the whole paragraph as a fusion block. As represented in the Figure 11(b), SegBlock did not group the whole paragraph; this is because the different font styles were used in the paragraph. As shown in Figure 11(c), since Semantic-Block simulates human perception using Gestalt laws of grouping, it segmented the blocks better than SegBlock. However, our approach used semantic analysis method and grouped the whole paragraph as a block regardless of the different font styles. Thus, our approach groups similar semantic text contents as a fusion block and overcome the scattering or shortening of the long text of webpage content mentioned in Sections 1 and 2.

6 Conclusion and Future Work

6.1 Limitation and Future Work

We demonstrate that the fusion model proposed in this paper outperforms the existing methods. However, there are also limitations in this segmentation technique which we plan to address in future work. We use Doc2vec to compare the textual similarity of basic-blocks to regroup them in a fusion block using Gestalt laws of grouping. Figure 12 shows the homepage of

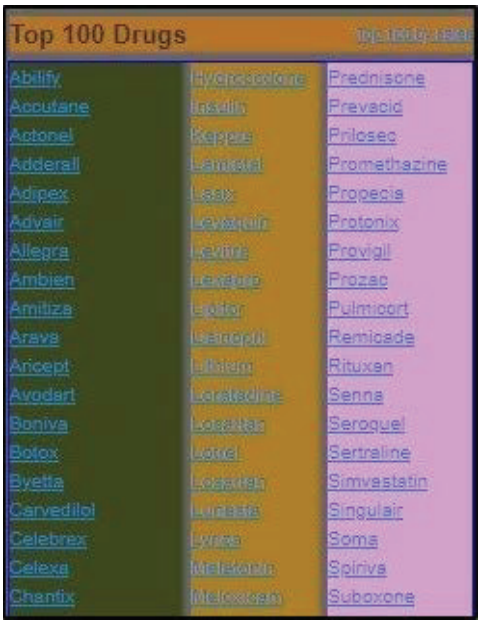

Figure 12 Homepage of "www.drugs.com".

"www.drugs.com" that is segmented in several blocks using our approach. Each block contains paragraphs represented the name of different drugs. Although there are different drug names in each block, they have a similar concept which is the "name of drugs". Using semantic analysis methods such as transformers may yield better results when paragraphs have the same concept during the regrouping stage. In our future work, we intend to utilize other semantic analysis methods to extend our model. Also, we plan to test our model on additional datasets.

6.2 Conclusion

In this paper, we present a new segmentation model to semantically segment webpages into fusion blocks. Our model merges webpage content into basic-blocks by simulating human perception. Additionally, it utilizes semantic text similarity to identify similar blocks and regroups these similar blocks as fusion blocks. To verify the effectiveness of our approach, we (1) applied it to the open-source public datasets, (2) compared it with the four existing state-of-the-art algorithms. The results show that our approach outperforms all the comparison methods in terms of precision, recall, F-1 score, and ARI.

References

[1] J. Zeleny, R. Burget, and J. Zendulka, Box clustering segmentation: A new method for vision-based web page preprocessing, Information Processing & Management, vol. 53, no. 3, pp. 735–750, 2017.
[2] S. Baluja, Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework. pp. 33–42.
[3] L. Bing, R. Guo, W. Lam, Z.-Y. Niu, and H. Wang, Web page segmentation with structured prediction and its application in web page classification. pp. 767–776.
[4] D. Chakrabarti, R. Kumar, and K. Punera, A graph-theoretic approach to webpage segmentation. pp. 377–386.
[5] C. Kohlschütter, and W. Nejdl, A densitometric approach to web page segmentation. pp. 1173–1182.
[6] Z. Jiang, H. Yin, Y. Wu, Y. Lyu, G. Min, and X. Zhang, Constructing Novel Block Layouts for Webpage Analysis, ACM Transactions on Internet Technology (TOIT), vol. 19, no. 3, pp. 1–18, 2019.
[7] J. Kang, J. Yang, and J. Choi, Repetition-based web page segmentation by detecting tag patterns for small-screen devices, IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 980–986, 2010.
[8] Z. Bu, C. Zhang, Z. Xia, and J. Wang, An FAR-SW based approach for webpage information extraction, Information Systems Frontiers, vol. 16, no. 5, pp. 771–785, 2014.
[9] H. F. Eldirdiery, and A. Ahmed, Web document segmentation for better extraction of information: a review, International Journal of Computer Applications, vol. 110, no. 3, 2015.
[10] C. Kohlschütter, P. Fankhauser, and W. Nejdl, Boilerplate detection using shallow text features. pp. 441–450.
[11] K. Koffka, Principles of Gestalt psychology: Routledge, 2013.
[12] S. E. Palmer, "Modern theories of Gestalt perception," Mind Lang. 5(4), 289–323, 1990.
[13] R. J. Sternberg, and K. Sternberg, Cognitive psychology, 3rd edn, Wadsworth, Belmont, 2003.
[14] Z. Xu, and J. Miller, Identifying semantic blocks in Web pages using Gestalt laws of grouping, World Wide Web, vol. 19, no. 5, pp. 957–978, 2016.
[15] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, DOM-based content extraction of HTML documents. pp. 207–214.

[16] Y. Chen, W.-Y. Ma, and H.-J. Zhang, Detecting web page structure for adaptive viewing on small form factor devices. pp. 225–233.
[17] Q. Fan, C. Yan, and L. Huang, Discovering Informative Contents of Web Pages. pp. 180–191.
[18] L. Yi, and B. Liu, Web page cleaning for web mining through feature weighting. pp. 43–48.
[19] L. Yi, B. Liu, and X. Li, Eliminating noisy information in Web pages for data mining, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003, pp. 296–305.
[20] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, Vips: a vision-based page segmentation algorithm, 2003.
[21] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, Extracting Content Structure for Web Pages Based on Visual Representation, Web Technologies and Applications. pp. 406–417.
[22] A. Sanoja, and S. Gançarski, Block-o-matic: A web page segmentation framework. pp. 595–600.
[23] R. R. Mehta, P. Mitra, and H. Karnick, Extracting semantic structure of web documents using content and visual information. pp. 928–929.
[24] Z. Xu, and J. Miller, Estimating similarity of rich internet pages using visual information, International Journal of Web Engineering and Technology, vol. 12, no. 2, 2017.
[25] Z. Xu, and J. Miller, A new webpage classification model based on visual information using gestalt laws of grouping. pp. 225–232.
[26] Z. Xu, and J. Miller, Cross-browser differences detection based on an empirical metric for web page visual similarity, ACM Transactions on Internet Technology (TOIT), vol. 18, no. 3, pp. 1–23, 2018.
[27] A. Mccallum, and K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification, Work Learn Text Categ, vol. 752, 2001.
[28] J. Hirschberg, and C. D. Manning, Advances in natural language processing, vol. 349, no. 6245, pp. 261–266, 2015.
[29] W. H. Gomaa, and A. Fahmy, A Survey of Text Similarity Approaches, International Journal of Computer Applications, vol. 68, pp. 13–18, 2013.
[30] P. Neculoiu, M. Versteegh, and M. Rotaru, Learning text similarity with siamese recurrent networks. pp. 148–157.
[31] Z. S. Harris, Distributional structure, Word, vol. 10, no. 2–3, pp. 146–162, 1954.

[32] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, Journal of documentation, 2004.
[33] T. Mikolov, Q. V. Le, and I. Sutskever, Exploiting similarities among languages for machine translation, arXiv preprint arXiv:1309.4168, 2013.
[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.
[35] T. Ming, Z. Lei, and Z. Xianchun, Document vector representation based on Word2vec, Computer Science, vol. 43, no. 6, pp. 214–217, 2016.
[36] Y. Wang, Z. Liu, and M. Sun, Incorporating linguistic knowledge for learning distributed word representations, PloS one, vol. 10, no. 4, 2015.
[37] M. Alsuhaibani, D. Bollegala, T. Maehara, and K.-i. Kawarabayashi, Jointly learning word embeddings using a corpus and a knowledge base, PloS one, vol. 13, no. 3, 2018.
[38] Y. Li, B. Wei, Y. Liu, L. Yao, H. Chen, J. Yu, and W. Zhu, Incorporating knowledge into neural network for text representation, Expert Systems with Applications, vol. 96, pp. 103–114, 2018.
[39] Q. Le, and T. Mikolov, Distributed representations of sentences and documents. pp. 1188–1196.
[40] J. H. Lau, and T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation, arXiv preprint arXiv:1607.05368, 2016.
[41] C. Xing, D. Wang, X. Zhang, and C. Liu, Document classification with distributions of word vectors. pp. 1–5.
[42] gensim, https://radimrehurek.com/gensim/, 2020.
[43] dataset-popular 2014. A dataset of popular pages (taken from dir.yahoo.com) with manually marked up semanticblocks. Retrieved from https://github.com/rkrzr/dataset-popular.
[44] dataset-random 2014. A dataset of random pages with manually marked up semantic blocks. Retrieved from https://github.com/rkrzr/dataset-random.
[45] Alexa. 2016. The top 500 sites on the web. Retrieved from http://www.alexa.com/topsites.
[46] VIPS-JAVA [n.d.]. Implementation of Vision Based Page Segmentation Algorithm in Java. Retrieved from https://github.com/tpopela/vips-java.
[47] A. S. Bozkir, and E. A. Sezer, Layout-based computation of web page similarity ranks, International Journal of Human-Computer Studies, vol. 110, pp. 95–114, 2018.

[48] L. Hubert, and P. Arabie, Comparing partitions, Journal of classification, vol. 2, no. 1, pp. 193–218, 1985.
[49] W. M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical association, vol. 66, no. 336, pp. 846–850, 1971.
[50] K. Y. Yeung, and W. L. Ruzzo, Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data, Bioinformatics, vol. 17, no. 9, pp. 763–774, 2001.
[51] Z. Xu, Visual Similarity Analysis of Web Pages based on Gestalt Theory, Department of Electrical and Computer Engineering, University of Alberta, 2017.

Biographies

Saeedeh Sadat Sajjadi Ghaemmaghami received the BS and MS degrees in computer engineering from QIAU, Iran, in 2007 and 2012, respectively. She is currently working toward the PhD degree in the Department of Electrical and Computer Engineering at the University of Alberta. Her research interests include webpage analysis, machine learning, natural language processing, and image processing.

James Miller, P.Eng (Alberta) has been a full professor with the Dept. Electrical and Computer Engineering at The University of Alberta since 2000. Previously, he was a professor at the University of Strathclyde (U.K.) and a principal research scientist at the National Electronics Research Initiative (U.K.). He has been an active researcher for more than thirty years across a wide range of topics, ranging from Computer Vision, Pattern Recognition, Embedded System Design, Software Engineering, Web Engineering and Proactive Analytics. He has published more than 100 articles in peer-reviewed journals including many IEEE and ACM venues.