Information Processing and Management 53 (2017) 735–750

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Box clustering segmentation: A new method for vision-based web page preprocessing

Jan Zeleny*, Radek Burget, Jaroslav Zendulka

Brno University of Technology, Faculty of Information Technology, Centre of Excellence IT4Innovations, Bozetechova 2, 61266 Brno, Czech Republic

ARTICLE INFO

Article history:
Received 5 May 2016
Revised 1 December 2016
Accepted 2 February 2017

Keywords:
Clustering
Segmentation
Vision-based page segmentation
VIPS

ABSTRACT

This paper presents a novel approach to web page segmentation, which is one of substantial preprocessing steps when mining data from web documents. Most of the current segmentation methods are based on algorithms that work on a tree representation of web pages (DOM tree or a hierarchical rendering model) and produce another tree structure as an output.

In contrast, our method uses a rendering engine to get an image of the web page, takes the smallest rendered elements of that image, performs clustering using a custom algorithm and produces a flat set of segments of a given granularity. For the clustering metrics, we use purely visual properties only: the distance of elements and their visual similarity.

We experimentally evaluate the properties of our algorithm by processing 2400 web pages. On this set of web pages, we prove that our algorithm is almost 90% faster than the reference algorithm. We also show that our algorithm accuracy is between 47% and 133% of the reference algorithm accuracy with indirect correlation of our algorithm's accuracy to the depth of inspected page structure. In our experiments, we also demonstrate the advantages of producing a flat segmentation structure instead of an hierarchy.

1. Introduction

Web page segmentation presents one of substantial preprocessing steps for data mining from web documents. There has been a lot of development in the area of web page partitioning. While some of the designed methods are targeted at specific problems like cleaning the noise from the web page, others, including page segmentation, are more generic in terms of possible utilization of their results. The problem with most of the segmentation algorithms is that they are quite slow, they depend on implementation details of the documents they process and they produce a hierarchical output that is difficult to process.

The objective this paper pursues is the development of a new web page segmentation method that is purely vision-based, independent of any HTML-related heuristics and implementation details of the processed documents. This requirement is present to make out method resilient to potential future changes in technologies used on the web. Moreover, the method should produce a flat model of the segmented page consisting of a list of visual segments with a consistent granularity level.

* Corresponding author.
E-mail addresses: izeleny@fit.vutbr.cz (J. Zeleny), burgetr@fit.vutbr.cz (R. Burget), zendulka@fit.vutbr.cz (J. Zendulka).

http://dx.doi.org/10.1016/j.ipm.2017.02.002
0306-4573/© 2017 Published by Elsevier Ltd.

And finally, the method must be unsupervised. The quality criteria against which this new method is evaluated include both speed and precision of the algorithm.

1.1. Background

Even though from the technological point of view, the web pages are considered as atomic carriers of information in the World Wide Web, in some research areas, it has been clear for some time (Cai, Yu, Wen, & Ma, 2003) that this granularity is too coarse for processing the contained information. Most web pages are logically split in smaller pieces. From the data mining point of view, some of these pieces can be thrown away as their informational value is negligible. Others can be then used for various purposes by different data mining techniques.

Page segmentation usually presents a preprocessing step in a more complex document processing task. From this point of view, we may find several application domains of page segmentation. Information retrieval and content classification techniques use page segmentation to improve both precision and performance by eliminating those parts of web pages that don't contain useful content Win and Thwin (2014). Separation of multiple topics in one web page is used for example in content classification. This important process can also use segmentation to gain precision Yu, Cai, Wen, and Ma (2003). In the adaptive view transformation (Aguado, 2015; Coondu, Chattopadhyay, Chattopadhyay, & Chowdhury, 2014), segmentation is used to identify coherent parts of the web page that should be kept undivided. Finally, in the information extraction area, page segmentation may be used for the identification of the data-intensive document sections Weng, Hong, and Bell (2011) or even the individual data fields Milička and Burget (2015).

Depending on the target application, different segmentation granularity may be required. The granularity corresponds to visual consistency of segments identified in the page. For the typical applications mentioned above, the following granularity levels may be considered:

- Informative content blocks level – for the page cleaning tasks in the information retrieval and document cleaning areas, the page segmentation is required to discover the basic blocks in the page such as the main content area, header, footer, etc. (Alassi & Alhajj, 2013; Uzun, Agun, & Yerlikaya, 2013; Win & Thwin, 2014; Wu, 2016).
- Paragraph level – for some applications such as vision-based classification of logical parts of the published information Burget (2010); Weng, Hong, and Bell (2014), a finer granularity is required that corresponds to the individual logical parts of the content such as headings, paragraphs, list items, etc.
- Data field level – the finest granularity level is required usually in the information extraction area when the individual data fields have to be identified and extracted Milička and Burget (2015).

Current page segmentation methods such as VIPS and its successors (described in detail in Section 2) produce a hierarchical model of the segmented page that is created by a recursive division (in case of the top-down approaches) or grouping (for the bottom-up approaches) of the detected visual blocks. The required granularity level then corresponds to the size of the leaf nodes of the produced hierarchy and for most segmentation methods, it can be adjusted by setting different parameters of the particular segmentation method such as the *degree of coherence* parameter in VIPS. However, for most of the above mentioned applications, the leaf nodes of the hierarchy are actually the most important ones. The content classification or information extraction methods examine the visual segments of the required granularity and actually do not use the complete hierarchy produced. Therefore, for several applications we have investigated recently (Burget, 2010; Milička & Burget, 2015), we found it more efficient to directly obtain a list of visual segments of the required granularity instead a hierarchical model.

In this paper, we propose the Box Clustering Segmentation (BCS) method that meets the requirements presented in the beginning of this section. Our method is built from ground up and it has virtually nothing in common with existing tree-based algorithms. We embrace a different, so far very marginally explored approach to the page segmentation problem. It is based on processing the rendered page using only very general visual cues. In contrast to the most of current methods, our algorithm does not produce a hierarchy of areas; instead, it aims to put together tiles on the same level of hierarchy. If detected correctly, the tile representation is a much more accurate representation of a web page in terms of user perception and it is more suitable for many application as discussed above. In contrast to most of the existing methods, we don't use any tree-processing approach. Instead, we rely on clustering techniques with a proper distance model in place. The simplified tile representation also allows to achieve a significantly faster segmentation which is traditionally an important issue in case of the vision-based methods.

The rest of our paper is organized as follows: Section 2 introduces the state of the art in the area of web page segmentation. Section 3 then introduces the main concept of the Box Clustering Segmentation. Sections 4, 5 and 7 explain the individual parts of the algorithm in detail. Section 6 covers the metrics we use for the clustering algorithm. Section 8 presents the results of our algorithm and compares them to the reference algorithm and finally, Sections 9 and 10 sum up the achieved results.

2. Related work

Our research deals with the issue of splitting up a web page into smaller segments. There has been a lot of research in this area in recent years and several types of algorithms exist to address this problem. Note that we don't compare