Information Sciences 342 (2016) 132–149

Contents lists available at ScienceDirect

Information Sciences

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

Language independent web news extraction system based on text detection framework

Yu-Chieh Wu*

Department of Communication and Management, Ming-Chuan University, 250 Zhong Shan N. Rd., Sec. 5, Taipei 111, Taiwan

ARTICLE INFO

Article history:
Received 23 September 2014
Revised 22 November 2015
Accepted 22 December 2015
Available online 15 January 2016

Keywords:
Content extraction
Information filtering
Web mining
Block segmentation
HTML

ABSTRACT

Web news provides a direct and efficient way to construct large text corpora. The creation of text data requires an understanding of HTML code and the preparation of customized parsing rules to identify text content in a webpage. Typically, parsing rules are written manually and cannot be applied to pages with different layouts. In this study, we present a web news extraction system that is based on a text detection framework. The proposed method scans the input HTML page and creates text statistics as a projection profile. Then, text block identification is applied to determine a set of content candidates. To filter noise, text verification determines whether a given text block can be included with content. We evaluate the proposed approach with the L3S-GN1 corpus and 3506 multilingual news data items randomly sampled from 325 websites (15 geographic regions and 11 distinct languages). We also compare the proposed method to 23 well-known state-of-the-art techniques. The experimental results show that the proposed method outperforms the second best method (NReadability) by 7.30% in the macro F-measure rate and is 16.91 times faster than NReadability. In terms of the perfect rate, the proposed method demonstrates 46.38% accuracy, whereas the Boilerpipe algorithm demonstrates only 21.54% accuracy. The proposed method is very useful for constructing a multilingual corpus because it requires no language-specific processing component.

1. Introduction

Web news media contains a huge amount of text sources that provides wide coverage and rich text information for many text mining applications [7,11]. For example, the creation of language models for optical character recognition, speech recognition, and information retrieval often requires large text corpora. However, the main text in web news is usually surrounded by various unrelated content, such as banners, commercial ads, link navigation bars, and user comments. The ability to filter noisy information while preserving the main text is useful for constructing clean text corpora and potentially improving web-based text retrieval systems. Moura et al. [27] reported improved results by applying the best (content) blocks for BM-25 ranking methods. Miliaraki et al. [23] presented a large-scale *N*-gram mining approach based on cleaned web news pages (derived from [14]). The well-known CleanEval competition, held in 2007 by the ACL-SIGWAC, also addressed the cleaning of web pages with the goal of developing a system that can identify structured text on a given Web page. Rahman et al. [32] also attempted to identify text content in HTML documents (content extraction).

* Tel.: +886 2 2882 4564x2100; fax: +886 2 2881 8675.
E-mail address: wuyc@mail.mcu.edu.tw, bcbb@db.csie.ncu.edu.tw

http://dx.doi.org/10.1016/j.ins.2015.12.025
0020-0255/

The importance of news content extraction systems depends on the fact that large and rapidly growing amounts of information are being produced every day and news content extraction systems allow efficient crawling of such information with limited human effort. Traditionally, constructing a news text corpus has been very time consuming because of the need to manually obtain the news content. By the introduction of information extraction technology [16,18], content extraction systems can generate symbolic wrappers for a web page by parsing the HTML code structure. The wrapper locates patterns in an HTML document and provides a method for extracting the content. Although the wrapper-based approach is well-developed, automatic news content extraction from web pages remains very challenging. Layout styles vary widely for different news pages, and styles change over time. The existence of banners, side bars, and commercial advertisements increases the difficulty of news content extraction. Because it is difficult to determine a uniform guideline for developing online news structures, the investigation of new methods that will allow for an effective deployment of a more structured view of web news remains an open issue.

Adelberg [1] studied writing wrapper rules for web data extraction. Zheng et al. [40] integrated both visual [2] and structural features to derive the V-wrapper for content extraction. However, this method requires a webpage to be rendered to derive visual features, such as position, size, and block. Reis et al. [33] proposed a multiple-page-based approach and applied a similar wrapper for news pages using a tree edit distance measurement. Machine learning methods have also been employed to learn content in a web page ([3,8,14,37]). In the training phase, some annotated data is prepared, and in testing, the learned model is used to identify a content block of an unseen web page. Semantic-based approaches have focused on word or phrase information to identify content areas. NCleaner [5] adopted character *N*-gram language models based on conditional probabilities. JusText [30] was proposed to discover content areas by counting the frequency of stopwords in a content block. More recently, statistically-based approaches have been widely used to solve the content extraction problem. The well-known combinE framework [9] is an integrated implementation of multiple statistical content extraction algorithms, such as content code blurring (CCB) and document slope curve (DSC). Weninger et al. [38] presented a cluster-based content extraction algorithm based on a tag-ratio histogram, known as content extraction via tag ratios (CETR). Sun achieved better accuracy than CETR by exploiting the HTML document object model (DOM) tree with a proposed composite text density scoring function. A hybrid CETR model using machine learning and heuristic method was presented by [29] to determine the optimal token (word)-level subsequence in a webpage.

Previous studies have not addressed the efficiency and effectiveness of large scale news text extraction methods; however, some methods that do not require building a DOM tree have been proposed. Given such complex approaches, it is difficult to specifically identify the main contributions of these studies. Furthermore, few studies have provided evaluations of their news content extraction algorithms for large volumes of Web news. For example, published studies [13,15,29,30,35,36,38] have examined methods using merely several hundreds of web pages. Thus, to date, the determination of their actual performance with respect to other domains and languages has not been performed. Moreover, recently developed algorithms such as Boilerpipe [14] and semantic approaches have been ignored thus far.

Text detection in video has also been developing over recent decades. The goal of video text detection is to identify the text components in complex backgrounds. Typically, a raw input frame is transformed into a domain map (e.g., edge-map or wavelet) for downstream processing. The transformed image map serves as the basis for text detection. Lienhart and Wernicke [17] used an edge projection profile to localize text in videos. Lyu et al. [20] achieved improved results by integrating several elaborate techniques, such as edge map smoothing and coarse-to-fine localization. Although such techniques cannot be applied directly to text content detection in webpages, with considerable modification and transformations the main concept and processing flow are useful.

By introducing a video text detection framework in this paper we present a language independent web news extraction system. Video processing techniques can produce useful results and improve text component detection because they are proven to perform well in identifying closed captions on complex backgrounds. Motivated by edge map transformation, we treat an entire HTML page as a raw input frame and transform the HTML code into a projection profile using compound text-tag difference (CTTD) statistics. To reduce the bolder effect, we also smooth the CTTD statistics using a designed simple linear kernel. Then, we create a set of text-like candidates using a text-box creation algorithm. Finally, text verification determines whether the given candidate belongs with the news content. In addition, we have also designed a method to eliminate user posts or message text, which sometimes includes much more text information than the content text and can mislead the content extraction algorithm. We evaluated our method with 3506 news webpages, derived from 326 distinct websites covering 15 geographic regions and 11 languages. To validate the performance of our proposed method, we compared it with 21 well-developed published content extraction algorithms and two commercial tools using the same testing data and hardware.

2. Related work

There have been many studies on detecting content area from HTML webpages and a number of methods have been proposed. One of the earlier content extraction methods creates hand-crafted web scrapers that directly find content text by looking for known HTML cues containing regular expressions. Examples of tools using this method include [1], Softmealy [28], and XWRAP [19]. The identification of HTML cues depends on specific pages and must be revised when the page is updated. One obvious limitation of this approach is that it is expensive to maintain a comprehensive set of different