Title - Based Extraction of News Contents For Text Mining

Mahananda Tidke
Dept of Computer Engineering
[1, 2] Siddhant College Of Engineering, Pune

Abstract- *As a vital measure to obtain valuable information and intelligence, web news are flooding all corners of the Internet anytime, anywhere. Traditionally, templates or hand-designed features are utilized to extract the content from web pages, but these models have higher time cost and lower extensibility. Recently, many scholars leverage DOM-tree-based or text-ensity-based models to extract the contents which have better extensibility and lower time cost, but most of them are hard to extract the content accurately and completely and are easy to introduce the noises. In this paper, we propose a title-based web content extracting model TWCEM to extract the contents of each web page, which leverage the title information to extract the web content. Compared with other extraction model, TWCEM can _lter the noises effectively and locate the content positions more accurately. In this experiment, we evaluate the proposed model on real-life websites, and TWCEM achieves state-of-the -art results and outperforms its competitors on both extraction performance and time cost.*

Keywords- Title-based extraction model, web news, data mining, information extraction.

I. INTRODUCTION

With the development of the Internet, more than ten millions of news are published, uploaded and shared every day, and the news websites have already become a vital measure for people to obtain concerned information. For example, CNN news is visited about 95 million times every month, and the top 5 most popular news websites and their estimated unique monthly visitors are shown in Table 1.1

Many researchers want to collect a large amount of news to analyze the contents and then obtain valuable information and intelligence, for example, the high-cleaned input information will improve the quality of news summarization effectively [1], [2]. But for each news page, the corresponding contents are embedded in a HTML document which includes lots of noises [3], e.g., navigation panels, advertisements, related news links and etc.2 Initially, many scholars proposed various methods which leverage templates and hand-designed features to extract the news pages [4], e.g., Crescenzi and Mecca *et al.* [5] and Arocena and Mendelzon *et al.* [6] leverage the hand-designed

In experiment, we evaluate the proposed model on real-life websites, and TWCEM achieves state-of-the-art results and outperforms its competitors on both extraction performance.

TABLE 1. The top 5 most popular news websites and their estimated unique monthly visitors.

Popular News Website	Unique Monthly Visitors
Yahoo! News	175 million
Google News	150 million
HuffingtonPost	110 million
CNN	95 million
New York Times	70 million

Features to extract web information which will cost a large amount of time to design features and are hard to be applied in other domains. Soderland [7] and Laender *et al.* [8] trained latent web features with hand-labeled documents, however labeling documents requires a lot of manpower, all of which are hard to be applied on large-scale information extraction. Road Runner [9] and NET [10], leverage template to extract the features which have strong constraints on web page struc- tures.

Recently, DOM-tree-based extraction models [11]_[13] are proposed which leverage the structure of HTML documents to locate the position of correct information. For example, Cai *et al.* [14] introduce the vision page segmentation with DOM-tree to extract the content, and Zheng . To address the issues, content-based models are proposed which omit the structure of web page and utilize the text features to extract contents. CETR [16] and CEPR [17] utilize the *HTML tag ratio* and *path features* to extract news content, respectively. MCSTD [18] lever- ages maximum continuous sum of text density methods to extract web contents which can collect the correct content with high-density property.

FIGURE 1. The framework of information extraction.

II. FEATURE-BASED WEB INFORMATION EXTRACTION

Recently, apart from rule-based web information traction models, there are several models which utilize feature based algorithms to extract the information in each web age

1) DOM-BASED EXTRACTION MODELS

DOM-based extraction models [11]_[13] leverage the structure of HTML documents to locate the position of correct nformation. These models can remove the noises, e.g,advertising and hyper-link groups, effectively. Nevertheless, the establishment of DOM-tree has a high requirement for the HTML. Especially, the construction and traversal of the tree have high time complexity, besides the tree traversal method also differs depending on the HTML tags.

2) VISION-BASED EXTRACTION MODELS

The basic idea of vision-based extraction models [14], [15], [33] is that the core information in each HTML is displayed with signi_cant position or style which can be used to extract news content. Cai *et al.* [14], _rst introduce the vision page segmentation with DOM-tree to extract news pages. Zheng *et al.* [15] propose an improved method which con- siders each item in the DOM-tree as a visual block. A series of visual features in each block is composited to evaluate the importance of each item. Wang *et al.* [33] utilize machine learning methods to combine visual features which have good generalization performance, but have high time and memory- space complexities.

3) BLOCKING-BASED EXTRACTION MODELS

CETR [16] utilizes the HTML tag ratio to extract news contents which has good performance and high extensibility. CEPR [17] leverages path features to measure the importance of each node, and then the contents in important nodes are regarded as correct news content. Besides, CEPR utilizes *Gaussian* smoothing method to weight the tag path edit distance which improves the time cost and reduces content extraction performance. However, CEPR will miss critical information when the block distribution function has more than one sudden changing point. MCSTD [18] utilizes maximum continuous sum of text density methods to extract web content which can collect the correct contents with high-density property.

4) OTHER EXTRACTION MODELS

CETD [34] combines DOM-tree and text density to extract news content which achieve good performance. CLG [35] utilizes the DOM tree to train a machine learning model and then uses a grouping technology to further _lter out noisy data. CCB [36] leverages content code vector to judge whether a block is meaningful or not.

Compared with rule-based information extraction models, feature-based models [37] have higher precision and better extensibility which can be applied on on-line information extraction systems effectively. But all the above feature-based models have high time complexity and miss important infor- mation when the text distributions are frequently changed.

In next section, we propose a novel title-based extraction model which can extract the web information ef_ciently, and locate the start and end positions accurately.

III. TITLE-BASED WEB NEWS EXTRACTION MODEL

Given a web news set D in *HTML* format, a document *d*, *d* 2 D. Our model aims to extract the correct news content *c* and removes the noises, e.g., navigation panels, advertisements, related news links etc. We propose a title-based Web content extracting model, namely TWCEM, which can _lter the noises in web pages and collect correct contents effectively. In this section, we _rst process the HTML parsing and then extract the title of each document. Finally, the title features are leveraged to extract news contents.

A. HTML PARSING

Before processing, we _rst need to remove the extra tags andidenti_ers with *Regular Expression* (RE), and the procedure can be seen in Algorithm 1.

Algorithm 1 HTML Parsing

Input: HTML document *d*.

Output: Candidate text contents *c*.

1 *r*1 D *RemoveScriptandCSS*(*d*);

2 *r*2 D *Remove Annotation and Special Char*(*r*1);
3 *c* D *Remove Tags*(*r*2);

First, we remove the contents between (1) tags < *script* > and < =*script* >, (2)tags < *style* > and < =*style* >, which denotes the asynchronous request information and *Cascad- ing Style Sheets* (CSS) in HTML, respectively. In addition, the annotations, blanks and special characters should also be

B. TITLE EXTRACTION

Title is a summarization of news contents and has strong semantic correlation with the news contents. Hence it is pivotal to extract titles from HTML documents which can be leveraged to extract corresponding news contents. In most cases, a title can be easily captured, because the position of the title in HTML is relatively static, e.g., hyperlink tags, meta tags of DOM-tree and so forth. In our model, we utilize both hyperlink and DOM-tree to extract the correct title.

1) HYPERLINK-BASED TITLE EXTRACTION

For each web news page, under most circumstances, it is easy to obtain the URL and corresponding description information from the home page, and the description information can be regarded as the candidate title. If the description information is missing or consists of abnormal characters, we leverage the information included in the web page to extract the title.

2) HTML-BASED TITLE EXTRACTION

In a HTML document page *d*, we _rst convert *d* into a DOM-tree where non-leaf and leaf nodes denote tags and contents, respectively. Hence, the processing of HTML documents can be achieved through the operation of the DOM-tree. Firstly, we leverage DOM-tree to extract the meta title *mt* in tag< *meta* >; secondly, the DOM-tree is utilized to capture can-didate titles *ct* in head tags < *hi* >3, where *i* D 1; 2; _ _ _ ; thirdly, we calculate the similarity between meta title and candidate titles with *Edit Distance* (Levenshtein Distance)[38], and the candidate title with the *minimum* edit distance is chosen as the correct title.

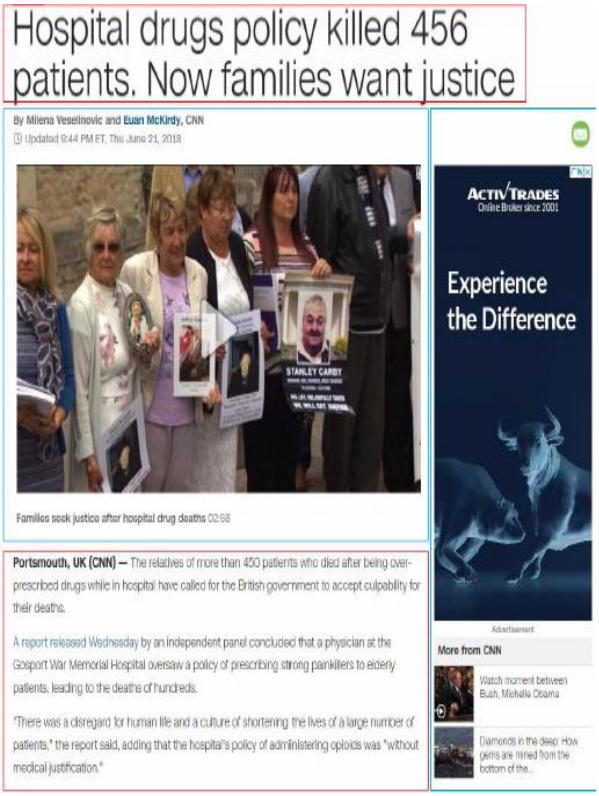

FIGURE 2. A sample of news page on CNN.

IV. CONCLUSION

In this paper, we propose a novel title-based information extraction model TWCEM to extract the contents of each webpage which can _lter the noises effectively and locate the content positions more accurately. We verify the proposed model on various real-life web pages, and TWCEM achieves state-of-the-art results on both extraction effectiveness and time cost.

We will explore the following future works:

1) Besides the title information, the news content also can be leveraged to calculate the correlation between different *C block*s which can improve the precision of location.
2) For the websites with different languages, we will set adaptive preprocessing measure to extract the news pages which can also improve the performance of the extraction.

REFERENCES

[1] O. Shapira, H. Ronen, M. Adler, Y. Amsterdamer, J. Bar-Ilan, and I. Dagan, ``Interactive abstractive summarization for event news tweets," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Copenhagen, Denmark, Sep. 2017, pp. 109_114. [Online]. Available:

https://aclanthology.info/papers/D17-2019/d17-2019

[2] R. A. Chongtay, M. Last, and B. Berendt, ``Responsive news summariza - tion for ubiquitous consumption on multiple mobile devices," in *Proc. 23rd Int. Conf. Intell. User Interfaces (IUI)*, Tokyo, Japan, Mar. 2018, pp. 433_437, doi: 10.1145/3172944.3172992.

[3] D. Gibson, K. Punera, and A. Tomkins, ``The volume and evolution of Web page templates," in *Proc. 14th Int. Conf. World Wide Web (WWW)*, Chiba, Japan, May 2005, pp. 830_839, doi: 10.1145/1062745.1062763.

[4] M. I.Varlamov and D.Y. Turdakov, ``Asurvey of methods for the extraction of information from Web resources," *Program. Comput. Softw.*, vol. 42, no. 5, pp. 279_291, 2016, doi: 10.1134/S0361768816050078.

[5] V. Crescenzi and G. Mecca, ``Grammars have exceptions," *Inf. Syst.*, vol. 23, no. 8, pp. 539_565, 1998, doi: 10.1016/S0306-4379(98)00028-3.

[6] G. O. Arocena and A. O. Mendelzon, ``WebOQL: Restructuring docu-ments, databases, and webs," *Theory Pract. Object Syst.*, vol. 5, no. 3, pp. 127_141, 1999.

[7] S. Soderland, ``Learning information extraction rules for semi-structured and free text," *Mach. Learn.*, vol. 34, nos. 1_3, pp. 233_272, 1999, doi: 10.1023/A:1007562322031.

[8] A. H. F. Laender, B. A. Ribeiro-Neto, and A. S. da Silva, ``DEByE_Data extraction by example," *Data Knowl. Eng.*, vol. 40, no. 2, pp. 121_154, 2002, doi: 10.1016/S0169-023X(01)00047