Web page classification based on content extraction

Hu Wanling (2819**), Zheng Xuan (2819****)**
Submission date: 2017-05-29
Southeast University-Monash University Joint Graduate School, Suzhou, China

Abstract

Nowadays, network with diverse messages has been penetrated into all aspects of society and people lives. And along with the evolution of the network technology, the information overload of network has been more and more noteworthy as it become difficult for users to conveniently and accurately locate the useful information among the network. Web classification could classify the cluttered web information to quickly retrieve the desired objective and relevant information. However, the classification method could not balance the accuracy and cost, no matter classified manually or automatically. This paper proposes a new classification method which combines the text characteristic and of the web page structure, with focus on the topic specific web pages. This method can effectively strengthen the result of this process. Moreover, we will give more details about the novelty of this paper in the content.

1. Introduction

With the coming of information era, the information resources of the internet have covered every aspect of our lives, and the number of web page is also experiencing explosive growth. By 2016 December, the number of Chinese web pages is 236 billion, up 11.2 per cent from 2015. And the number of Chinese web sites is 4.82 million, up 11.2 per cent from 2015 [1]. More and more people manage their daily business relying on the Internet, and benefit from the plentiful information contented in these pages.

An important way to deal with those massive data in this information age is to categorize them. The page classification can be categorizing pages based on the information which is carried by the pages. The method of web page classification is intended to help users quickly and accurately find their aim pages by saving it in the corresponding database according to the type. the strength of web

classification includes improve the recall and precision, and enhance the efficiency of search engines. This is also a key technology of network security management [2]. There are two types of classification strategies. One is the manual classification, which is classified those web page manually by experts. Although the accuracy of this method is high, but the efficiency is low. It cannot meet the needs today. The other is automatic classification [3]. It builds a classifier and train it based on the input training set, the classifier will obtain relevant knowledge to prepare the classification.

2. Objectives

Inside web pages, there usually exist content information and irrelevant message. The irrelevant information could bring negative influence to the retrieve and storage of websites. Web classification is significant in many application fields, for example, it could help to extract content, avoid the disturbance of uncorrelated text, scrub web data and generate the article summaries automatically.

The ultimate goal of our research is to propose a method that could filter the useless data to complete web classification. We hope this research is established in previous studies and make a breakthrough. After reviewing many documents in the field of web classification, we find that most of them put forward a single algorithm or advanced algorithms to enhance the efficiency. However, it becomes a bottleneck or them to find a tradeoff among the efficiency, accuracy and cost. Therefore, we consider to combine textual features and structural characteristics to process the classification. In our research, we segment web pages into blocks follow the structural rule. After that, we choose the particular blocks according to the textual feature of Chinese topic specific web pages.

3. Methodology

Web could be classified into three varieties according to the web content [4]. The first kind called topic-specific web, it can be defined as a page that use paragraphs of words to describe one or more themes. Rarely pictures, videos and links will be displayed in this type of web. The second kind of web page called Hub web, which always provides hyperlinks for relevant websites and seldom does these web pages describe a specific thing. The third kind web page named multimedia web page, where pictures and videos occupy the most spaces of the web page and

words are displayed only as a statement for multimedia. Our research mainly focus on the classification of Chinese topic specific web pages.

Firstly, we will preprocess the web pages before extracting the content. Most of the web pages are make up by HTML (Hyper Text Markup Language). In this section, we filter those useless label and code which has no relation with content information by utilizing the HTML labels.

Secondly, we will further distinguish the web pages according to the structural features of the websites. These structural features of the web can be divided into "blocks" visually [5]. In order to coordinate with the page layout rules and provide convenience to the writing task, the designers of the page will usually adopt container labels (<table>,) in the process of building the web page. In this section, we segment the source code into blocks in the form of string based on the container labels. After we obtain the string named “webstr” from preprocessing, we replace the container labels with string “#text” and split webstr into sub-string according to “#text”. We will filter those empty strings and process these remain strings as blocks. At last, we add blocks into BlockList. Due to the linear segment, we wipe out the nested structure and realize the partitioning rapidly and accurately without affecting the content of web page.

Thirdly, we will clarify the criterion to measure whether the block can be regarded as a feature of the web. Descriptive text usually occupies the most proportion of Chinese topic specific web pages and Chinese punctuations play an important role in locating the main content [6-7], so we set the concepts of “text density” and “punctuation number” based on the characteristics of Chinese web pages. Text density of a block was defined as the ratio of the number of non-linked texts to the number of all texts in the block. Punctuation number of a block was defined as the total number of periods and commas. And we judge the possibility of a block to be the main part through setting the threshold p and q. According to the analysis of the text characteristic of pages, blocks can be divided into two categories: the text block, which mainly exists in the main body of a web page with higher text density, and link block, which mainly exists in the form of hyperlinks and always contained in the “noise” area. We will classify a block into text block if its text density is greater than p and punctuation number is bigger than q. Otherwise, the block is more likely to be a link block.

Finally, we will extract the content of web pages. Not all text blocks will always appear in the main body of the web pages, sometimes link blocks may also exist in

the text part. But continuous link blocks are likely to be the "noise" area of the page. In the experiment, we will use the rule that the "noise" part is consists of continuous link blocks to construct the content extraction algorithm. The extraction of web content could be obtained by the following steps. In the beginning, we find the text block j that contains the most text. After that, we search forward and backward for the first link blocks i and k from j. both i and k should meet the requirement that their neighboring blocks will also be link blocks. At last, we extract the blocks between the link block i and k we search in the former step as the content.

4. Novelty

Our team propose a new text extraction method with high universality for web classification. It analyzes the textual features and structural features of the Chinese topic specific web pages, and combines the textual features and structural features of the web pages. This method linearly segments the structure of web pages and avoids the process of parsing the source code through the DOM tree during the partitioning process [8]. It could remove the nested structure without affecting the content of the page, and the implementation process is simple. The method considers roundly about the possible appearance of both text blocks and link blocks in the content of a page. This method is not for the specific page. Therefore, it has the advantages of versatility that can be developed without considering for particular web pages.

5. Conclusion

In conclusion, eliminating the interface of useless information to improve the efficiency of the web page classification is the ultimate goal of our research. The web page classification can enhance the user experience by utilizing page information maximumly to delivers optimal value of the web. This paper proposes a method of classification that combines the characteristic of text and the structure of web page. This method has taken excellent advantage of that the web pages could be easily divided into blocks, and focus on the topic specific web pages. This method is not for the specific page, it has the advantage of versatility. However, the information of images and videos could affect the final extraction result. Therefore, in order to obtain better result, the future work will pay more attention to handling the information of images and videos.

References

[1] China Internet Network Information Center. The 39th Statistical Report of China Internet Development. Beijing: CNNIC, 2017.

[2] Krishnakumar, L., & Varughese, N. M. (2013). High speed classification of vulnerabilities in cloud computing using collaborative network security management. *International Conference on Advanced Computing and Communication Systems*, 1-6.

[3] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using k-nearest neighbor (knn) classification method. *Applied Computing & Informatics, 12*(1), 90-108.

[4] Zhang Zhigang, Chen Jing, & Li Xiaoming. (2004). An Approach to Reducing Noise in HTML Pages. *Journal of The China Society for Scientific and Technical Information, 23*(4), 387-393.

[5] Nie Hui, & Zhang Jinhua. (2012). Page Content Extraction Based on Web Page Segmentation. *Journal of The China Society for Scientific and Technical Information, 31*(1), 31-39.

[6] Song, M., & Wu, X. (2007). Content Extraction from Web Pages Based on Chinese Punctuation Number. *International Conference on Wireless Communications, NETWORKING and Mobile Computing*, 5573-5575.

[7] Xiong Zhongyang, Lin Xianqiang, Zhang Yufen,& Ya Man.(2013). Content Extraction Method Combining Web Page Structure and Text Feature. *Computer Engineering, 39*(12), 200-203.

[8] Sun, F., Song, D., & Liao, L. (2011). DOM based content extraction via text density. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 245-254.

1780 words