Advances in Intelligent Systems Research, volume 136
4th International Conference on Sensors, Mechatronics and Automation (ICSMA 2016)

A Type of Web Content Extraction Algorithm Based on Adaptive Threshold

Guang Zheng[1, 2, a], Xianghui Hui[1,b], Xin Xu[1,c] and Lei Xi[1,2,d]

[1] College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002, China

[2] HHH Science Observation and Experiment Station of Agricultural Information and Technology, Ministry of Agriculture, Zhengzhou 450002, China

[a]ggzan@163.com, [b]truhui@henau.edu.cn, [c]xuxin468@163.com, [d]ggzan0215@sina.cn

Keywords: new rural community; Web information fetching; text density; adaptive threshold; Otsu threshold algorithm; Web page text extraction algorithm

Abstract. On the basis of the text extraction based on the density of text, the Web page text extraction algorithm based on the adaptive threshold was proposed and applied in the new rural community employment information service system for the employment information fetching from the related government affairs website combined with the Otsu threshold algorithm. Through the web page text extraction contrast experiments to the Webpages including "The ministry of human resources and social security of the People's Republic of China", "The ministry of human resources and social security hall of henan province" and "Sina.com", the text extraction rate of the algorithm reached 90%, 92% and 92% respectively. The results showed that the application of the algorithm in new rural community employment information service system could provide technical support for the directional employment information acquisition and realize accurate employment information retrieval.

Introduction

Advancing the new rural urbanization community construction was the fundamental measure to solve the problem of "agriculture, rural areas and farmers". The core requirements of the new rural community construction were achieving the farmers' local employment with the idea of city and industry integration[2-3].For the employment market based on the new rural community, on the basis of the personalized technology and information fetching technology, the new rural community oriented employment information service system was constructed to provide convenient employment information service for the new rural community residents, small micro enterprises and the new rural production organization.

The government affairs websites represented by functional departments were largely dominated in the development of information service. The service system of such websites possessed characteristics such as strong policy-oriented duties and decree unity and was conducive to the establishment and promotion of rural community employment information service in the blank areas quickly. Such type of web sites was given priority to the one-way information services and provided authoritative labor employment, social security policy and a part of recruitment information, training information and business information. In the process of the new rural community employment information service system construction, the employment information pushing in such type of websites was the key content of the information system service. In this paper, the text extraction algorithm based on adaptive threshold was adopted to capture employment information from related websites, provide directional employment information acquisition and achieve the efficient access to employment information. The algorithm was used in the new rural community employment information service system to realize the practical employment information recommendation, as shown in Figure 1.

Fig. 1 Practical information recommendation for community users

Web page text extraction based on adaptive threshold

As the core content of the new rural community employment information service, the text information extraction of web pages mainly conducted the directional fetching for the employment policy information from the related websites. After information extraction, the information would be classified and saved to structured databases for users' browsing. However, due to the presence of a large amounts of web page noise such as advertising links, navigation information, copyright information, JavaScript information and so on in the web text, the research of how to extract web text information accurately and efficiently had very high application value and practical significance.

The capture goal of this paper were the web pages of "Ministry of Human Resources and Social Security of the People's Republic of China" and " The ministry of human resources and social security hall of Henan province". The target sites belonged to the government news website and had the characteristics of the simple structure, concise web page, outstanding focal point, low noise, neat layout, etc. There was fewer HTML tags and link in the source code structures in such web pages. The text density value of the web page was higher[8].

The webs of the target site were analyzed based on the text density. Randomly selected 50 pages, the text distribution state of such WebPages was shown in Figure 2. In the figure, the abscissa was the number of lines of code. The vertical coordinate was the number of bytes. In the web page text information area, the text bytes was nearly equal to the total number of byte code. However, in the non-text region of web page such as navigation bar, links, sidebar, menu, advertising, header, footer, etc, the number of their code bytes was greater than text bytes number largely.

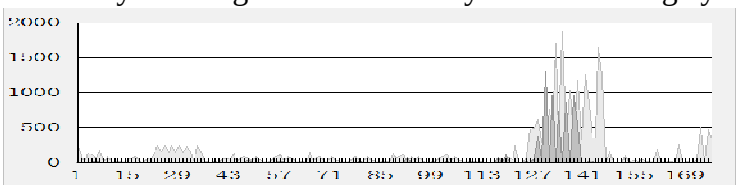

Fig. 2 Text density distribution state diagram

As a basic algorithm for the webpage text extraction, the text extraction based on the text density was based on the calculation of the text density of the each line text and HTML source code in the web page document. After that, a threshold was employed to judge the calculated text density. If the text density was larger than the threshold value, the text in the line was extracted as the body text content. If the text density was less than the threshold value, the extraction was not conducted. However, there were differences in the web page structures of different sites. The threshold values of these web pages were also different. In the case of the threshold value could be determined, the text extraction had obvious effect. However, the threshold value must be inputted manually. For different websites, changeless threshold value could lead to less or more extraction.

The web page text extraction algorithm based on adaptive threshold

The algorithm based on text density had very good effect for HTML text extraction from these normative web pages. However, some problems still existed in the algorithm application. For example, in some web pages, the text density was low in a few lines of the text area and the text density was relative high in the web link. These problems made the selection of threshold value

difficult commonly. Therefore, combined with OSTU threshold algorithm [11], this paper proposed a type of web page text extraction method based on adaptive threshold to adapt the text information extraction from different sites and improve the accuracy of text information extraction. The algorithm was constructed by the text extraction algorithm based on text density and adaptive threshold calculation method two parts.

The text extraction algorithm based on text density

The extraction algorithm conducted the functional extraction to the HTML text depend on the web line block length and the density of the text area [7]. In the algorithm, the location of the web page content text areas were obtained firstly. After that, the body text was extracted then. The algorithm was defined as follow:

First of all, the HTML documents of the grabbed webs were preprocessed including scripts and format tags removing, special characters transforming, etc. Meanwhile, the original HTML document structures were maintained. All the blank location information was kept after removing labels. That meant the number of lines in the web was invariant. The obtained HTML was the coarse text information F.

1. Row block: any line *Xi* of the *F* was used as the benchmark, the adjacent up and down n lines were taken, such as *{Xi, Xi + n}, {Xi - n, Xi}*. The value of n was set to 5 commonly. The obtained multiple lines *Xi* was recorded as *Blocki* row block.
2. Length of row block: the overall length of $Block_i$ after removing all the blank characters was recorded as the *B-lengthi*.
3. Distribution function of row block: In the every line of X_i, there was a *Blocki* and a *B-lengthi*. Based on this, a discrete distribution function of row piece was obtained and recorded as $f(Block_i, B\text{-}length_i)$.

The discrete distribution function of row piece *f(Blocki， B-lengthi)* was marked in coordinate axis of the experimental figure. From the figure, the B-length of the text area could be seen higher clearly. The obvious surge and dropping points could be found at the beginning and end of the main body area of the web. According to the surge and dropping points shown in the figure, the text area could be identified clearly. For example, a web page was randomly selected from sina.com. The discrete distribution function of row piece for the web page was shown in Figure3.

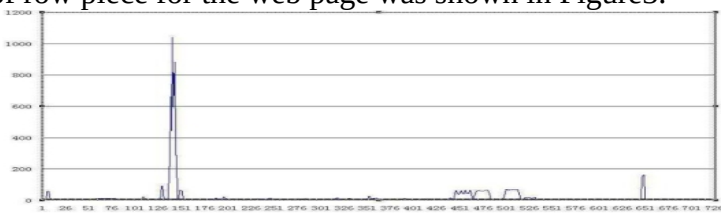

Fig. 3 The correct line number of the text area in the sina news web page was136-145

The purpose of the algorithm was determining the surge and dropping points of the text area. The specific algorithm was as follows:

1. The absolute value of the $B\text{-}length_i$ for each of the surge point $Block_i$ needed more than a threshold *T*.
2. The *B-length* value of the *n* Blocks followed the surge point must be greater than 0 to avoid noise.
3. The $B\text{-}length_{i+n}$ values of the several $Block_{i+n}$ after dropping point $Block_i$ must be zero or close to zero to ensure the end of text.
4. The $B\text{-}length_i$ of HTML must to ensure to take maximum value. Moreover, the $Block_i$ was must between the surge point and plunging point.

The computing method of adaptive threshold

The adaptive thresholds were calculated to gain the thresholds of the surge points and dropping points in the text area in the text extraction algorithm based on density of text. Combining with the Dajin threshold algorithm, through the binary segmentation for the row block distribution function in the calculating method, the fluctuant range of the row block distribution function was obtained. Then, the

threshold values of the surge points and dropping points were calculated. The calculation method was defined as follows:

1. For a web page F(Block，B-length), assume that segmentation threshold of the text and non-text, the value greater than T was called the main body text. Meanwhile, the value less than T was called the non-text.
2. Sampling for the text areas, a set of discrete data P(Block，B-length) was obtained. The sample size was recorded as X. The variable coefficient of text row block M= (variance of P)/X.
3. Sampling for the non-text areas, a set of discrete data Q(Block，B-length) was obtained. The sample size was recorded as Y. The variable coefficient of non-text row block N= (variance of Q)/Y.
4. The difference value of the variation coefficient g= M-N. For different web pages, the length of the starting line block of the main text was floated between 20 and 140. Therefore, the value range of T should be evaluated between 20 to 140 and increase 10 at a time. As the change of T value, when the g value changed to maximum or tended to be fixed, the T value was the segmentation threshold of the text row block and non-text row block. Then, the threshold was applied in the algorithm for gaining the surge and dropping points in row block distribution function to obtain higher extraction efficiency for different web pages.

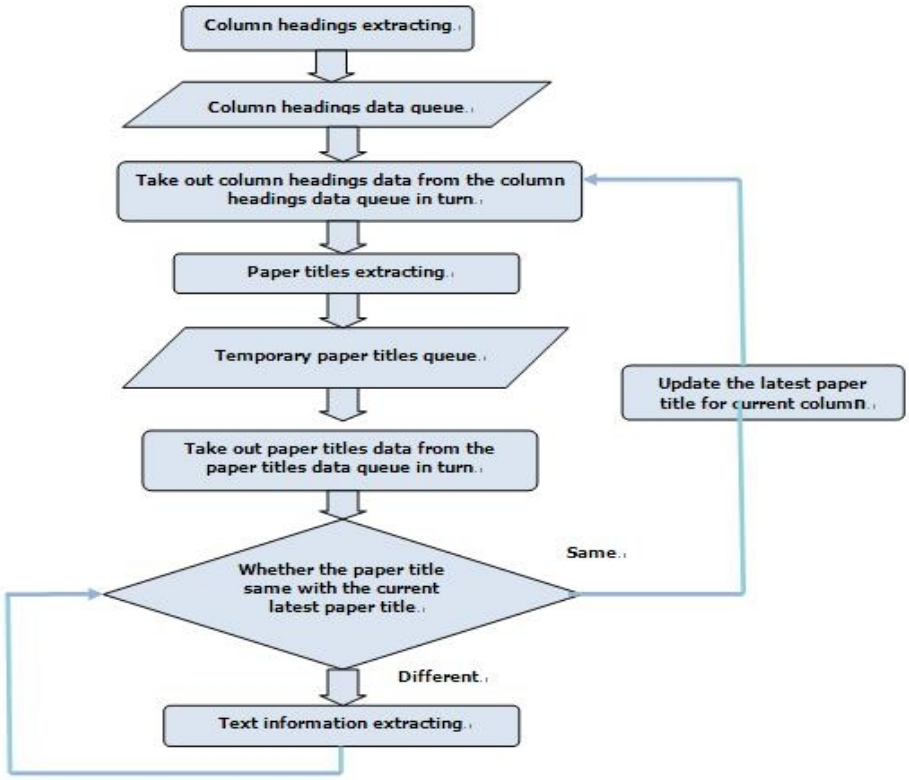

Fig. 4 Flow chart of information collection

Employment policy information collecting

In the new rural community employment information service system, the related employment policy information collecting was directional information acquisition to the employment policy from the websites of “The ministry of human resources and social security of the People's Republic of China” and “The ministry of human resources and social security hall of Henan province”. A complete process of information collecting included primary column header extracting, secondary column header extracting and web page main text extracting, etc. Among them, the column header extracting was utilized to the policy classification. The specific process of the information collecting was shown in Figure 4.

1. Primary content column header extracting: as shown as the each title in left side of Figure 5, utilizing the policies and regulations from the "The ministry of human resources and social security of

the People's Republic of China" webpage as the initial site, each subheading of the title bar part was conducted the URL extracting.

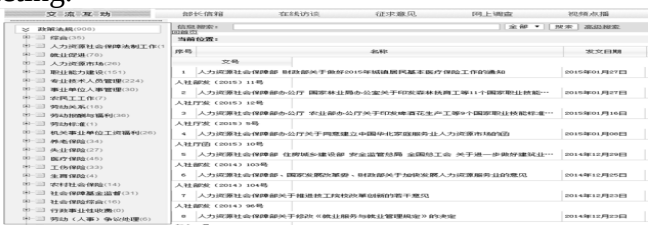

Fig. 5 Primary column header extracting

The DOM tree structure of target page's HTML documents was analyzed to extract the URL. The extracted URL belonged to the relative path. After processing, the URL changed to absolute path and the column header information. The latest articles of current column made up the column heading data structure were stored in the queue. For example, the URL "*http://www.mohrss.gov.cn/gkml/81/83/818/list.htm*" could be transformed to *"comprehensive notice about conscientiously implementing work discipline and regulation for social insurance personnel"*.

2. The secondary content column header extracting: taking out the data from the column header queue in turn and conducting the body text title extraction aimed at URL of each item. As shown in Figure 6, the extracted {article title URL, article title, issuing date, document number, affiliated columns, etc} was used as the column header data structure and stored in the temporary column header queue.

序号	名称	发文日期	文号
1	关于认真贯彻落实社会保险工作人员纪律规定的通知	2013年04月16日	人社险中心函（2013）49号
2	关于印发进一步做好房行节约工作的实施意见的通知	2011年07月23日	人社厅发（2011）60号
3	关于印发人力资源社会保障部政府信息公开实施办法的通知	2011年07月22日	人社厅发（2010）111号
4	关于印发《人力资源和社会保障部工作规则》的通知	2011年07月22日	人社部发（2008）5号

Fig. 6 Secondary column header extracting

3. Web page content information extracting: the article title data was taken out from the temporary article title queue in turn. Then, the article title which needs to be extracted would be contrasted with the latest article title of the current column in the current column headings data structure. If the result was different, the webpage content extraction aimed at article URL would be conducted employing the web content extraction algorithm based on adaptive threshold. If the result was same, the latest article title of the current column in the current column headings data structure would be updated by the first article title of the temporary article title queue. Then, the process would go back to the secondary content column header extracting. The webpage content extracting flow chart was shown in Figure 7.

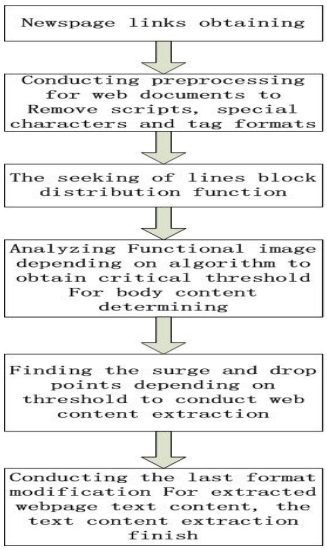

Fig. 7 Flow chart of webpage text content extracting

Firstly, after webpage HTML document resolving, script filtering, special characters replacing and so on, 181 lines' HTML document could be obtained. After that, the rows block distribution function

was calculated to obtain function image. The threshold value for extracting the webpage text content was also obtained through the algorithm, as shown in Figure 8:

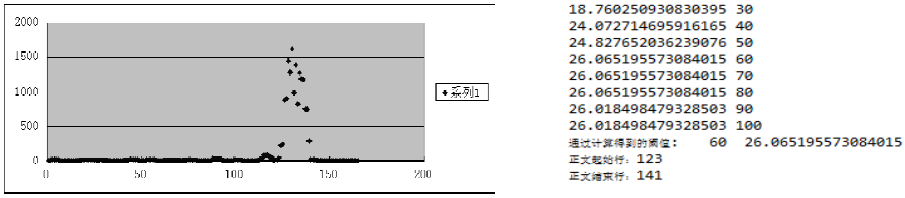

Fig. 8 Function image and threshold of the web text content

Through the calculation, the threshold was obtained as 60. Moreover, the location of the surge and drop points was also get as row 123 and row 141. Then, the extraction for the HTML documents could be conducted to obtain the webpage text content, as shown in Figure 9:

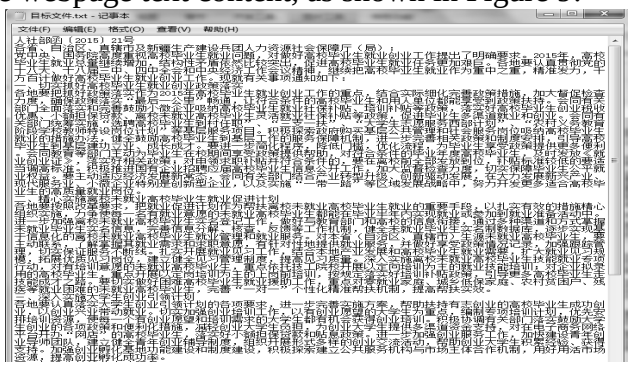

Fig.9 Result of the extracted webpage text content

Design and results of experiment

The design of the experiment was directional capturing to the webpage text content from *"People's Republic of China Department of human resources and social security"*, *"The ministry of human resources and social security hall of Henan province"* and *"Sina.com"* three kinds of mainstream websites. Conducting 50 times web scraping to each site, the actual effects of the adaptive threshold algorithm after optimization were judged by the number of the webs' text content which was completely captured and the grab time. The results were shown in Table 1 and 2.

Table 1 Contrast of the number of completely captured pages

	People's Republic of China Department of human resources and social security	The ministry of human resources and social security hall of Henan province	Sina.com
Basic algorithm	40	41	29
Adaptive threshold algorithm	45	46	32

Table 2 Contrast of grab time(second)

	People's Republic of China Department of human resources and social security	The ministry of human resources and social security hall of Henan province	Sina.com
Basic algorithm	1.065	1.783	2.134
Adaptive threshold algorithm	5.735	6.232	7.341

The experimental results showed that, for the target WebPages *"People's Republic of China Department of human resources and social security"* and *"The ministry of human resources and social security hall of Henan province "*, because their pages' normative format and no advertising, the extraction rate could reach above 90%. For *"Sina.com "*, due to its business category, relatively more advertisement and picture and complex webpage structure, the extraction rate was only 64%. Through the analysis for the webpage which could not be correctly extracted, most of the error pages were found existing multiple-text. The text contents of these pages commonly shorter. Moreover, the

title link information of these pages also was found longer. Due to the shorter text information, these pages were often filtered out during threshold selection. The overlong title links were often mistaken for text content of pages. In addition, when the text content belonged to the itemized list, these text main be mistaken for a title link and filtered.

Conclusions

To adapt to the adaptivity of the threshold for different website during conducing the web text content extraction, improve the extracting accuracy of webpage text content, this paper proposed a type of webpage text content extraction algorithm based on text density and adaptive threshold. Depending on the length of the webpage's row block and the density of the webpage's text content area, this algorithm carried out the functional extraction for the HTML text to gain the location of the web content areas. Combining with the Otsu threshold algorithm, binary segmentation was conducted on the row block distribution function to obtain the fluctuation area of the row block distribution function. Then, the thresholds of the surge and drop points were calculated to implement the text content extraction. The results of contrast experiment between the algorithm and basic web content extraction algorithm showed that the text content extraction rate of the algorithm could reach more than 90% for the government portal websites with relatively normative format. The capture accuracy was also higher. Due to the introducing of adaptive threshold calculation, the complexity of the algorithm was improved. At the same time, the fetching time was also increase. Therefore, how to improve the efficiency of algorithm effectively as well as guarantee the text content extraction accuracy needed further research.

Acknowledgements

This work was financially supported by the Henan Major Science and Technology projects (131100110400,121100111300-04), National "Twelfth Five-Year" Plan for Science & Technology Support (2014BAD10B06).

References

[1] GE T. The Job Security Problem Analysis for Land-lost Farmers[J]. Journal of Shenyang Agricultural University (Social Science Edition)，2014, 16(2): 144-146.

[2] WEI X, HUANG Q. Analysis of Demand and Supply of Employment Services of Migrants[J]. Forward Position or Economics, 2014, 11(03): 152-160.

[3] LIU Z, LIU L, LIU P. Learning Resource Personalizing Recommendation Algorithm Based on Semantic Web[J]. Journal of Jilin University (Engineering and Technology Edition), 009,39(2): 392- 395.

[4] ZHAO X, SUO H, LIU Y. Web Content Information Extraction Method Based on Tag Window[J]. APPLICATION RESEARCH OF COMPUTERS, 2007, 24(3):144-180.

[5] ZHU Z, LI M, ZHANG J. Web Content Extraction Based on Text Density Model[J]. Pattern Recognition and Artificial Intelligence, 2013, 11(07): 278-279.

[6] HANG Z, LI W, MO Q. Research on Methods For Extracting Text Information From HTML Pages [J]. APPLICATION RESEARCH OF COMPUTERS, 2008, 12(12): 358-359.

[7] LUO Y, ZHAO C. Extracting Method of Emergency News Headline and Text From Webpages[J]. Journal of Computer Applications, 2014, 23(10): 89-91.

[8] Nobuyuki Otsu. The Between-cluster Variance Adaptive Threshold[EB/OL]