ATLANTIS PRESS
Advances in Computer Science Research, volume 44
3rd International Conference on Wireless Communication and Sensor Network (WCSN 2016)

A Survey of Web Page Preprocessing Research

Qi Qi
Department of Information Engineering, Minzu University of China
Beijing, China
Email: 774287402@qq.com

Gui-Xian Xu
Department of Information Engineering, Minzu University of China
Beijing, China
Email: 10658115@qq.com.

***Abstract*-After obtaining the required information through the crawler technology on Web, it also includes a lot of advertisement and navigation bar. So we should take the basic method to remove the noise content on Web page, which is independent of topic, it is necessary to sum up the Web denoising and do a further study. Firstly, we should explain why the page denosing is necessary, define the page denoising, and summarize the method of Web page denosing, Secondly, we should the improve the algorithm on the Web page denoising, Finally we should discuss the webpage denoising problems and the future research direction.**

Keywords-Web page cleaning; data mining; Web mining; information retrieval.

I INTRODUCTION

By Web crawler technology, we can get some information, which contains in a large amount of content and relevant link of Web page, and it refers to as the "noise"[1]. But as for today's development, the bulk of the Web pages on the Web is only composed of HTML code [2], it does not contain the required metadata. So you need a way to get rid of the noise which is not useful with the Web page subject content, the Web page of preprocessing denoising method is there such a function, it can not only to get rid of useless information, to extract useful information, also can make a Web page specification, and improve the efficiency of search engine[3].The template crawler technology is mainly to solve the Internet data , which could generate a collection of original pages, it not only makes use of a set of advertising, but also uses a lot of similar templates. Most of these templates are a navigation bar, article content, and hyperlinks contact information; On one hand , Web page is to facilitate the use of template, beautify the page. But on the other hand that it makes noise exist in the Web page, if it doesn't remove the noise of cleaning, storage space can be gradually reduce or even not enough to use, and the effect of redundant information can also affect the user experience, so it is necessary to use Web denoising technology. Despite technology is good or bad, it will directly affect the efficiency of search engine.

II DENOISING THE DEFINITION OF WEB PAGES

Generally speaking, the Web page of denoising is defined as: 1) we should find and remove duplicate pages (such as mirror sites, to copy the article [4]. 2)we should remove the noise in the Web link, the noise of the so-called link producers, which can improve the Web page in the search engine rankings, deliberately references link; 3) we should remove the noise inside the content page. Noise content is in the process of research and application, solution and the demand is not in conformity with the contents ,these contents includes advertising, templates, and the corresponding Web links,which can improve the quality and the speed of the index. Generally ,it refers to remove the page content denoising, and it does not involve in the first two classes. Nowadays, Web page denoising method is divided into three categories: methods based on the structure of Web pages, the method based on template, the method based on visual information and Web denoising method based on the theme, or classification method is based on the principle of the heuristic method. These two kinds of methods are complementary to each other, they rarely are used alone heuristic method, when they uses in machine learning method ,they could also need to some heuristic rules.

III WEB PAGES DENOISING METHOD

Three kinds of methods have different characteristics and emphasis. In order to more clearly know the property of the method, they are expounded.

A. *Method Based on the Structure of Web Page*

So-called method, which is based on the structure of Web page. It is expressed as a DOM tree or variants of the DOM tree [5], and on the basis of some heuristic rules, we pick up the Web and information relevant to the subject matter, as we know it is used the method of literature. Firstly, we put forward the Content Block, using the < table > tag in the page into pieces. Literature continues to deepen the idea, and puts forward a set of heuristic rules. By using the method of information retrieval, we extract the theme about Web pages, as well as relevant to the subject matter and content. Literature with tag tree representation page file, uses the bottom-up algorithm, extracts with different semantic content of the page.

In the face of the DOM tree and heuristic rules are introduced: Document corresponding to a DOM tree, a Web page within the document is composed of tree layers, it includes the root node of the tree, and the correspond leaf node. Each tag page inside Document represents a node, which is divided into four node types: Document, Element, Comment, Type, and tree node.

The Document represents the root node of the tree, and the root node usually gives several number of child nodes, each tag elements on the Web page Document has its own

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

location in the DOM tree. 2002 Yossef and Rajagopalan, a DOM tree segmentation based method was proposed to detect the template, its principle was the concept that split a Web page, the number of links in your Web page number determined the pagelet, they were divided by the HTML < TABLE > tag page; Lin and Ho, adopted in 2002 as Ma in 2003, the way of dividing the difference was selected the keyword [6], each piece of content by keyword calculated the entropy value of these blocks, the template was usually entropy, which was relatively small .The formula is following:

$$T = Mx\sqrt{\min\{a,b\}}\ xN \tag{1}$$

The formula explains that T is a threshold, M is a constant, a, b is for the length and width of the node. when the sub block is located in the web around, that is equal with 1=>N>0,when the block is located in the middle of the page, N is equal with 1.

This method is the selection of features, which is in line with the requirements and the extracting characteristics of large amounts of time. After Vieira et al. who proposed a idea which tested template by the DOM tree nodes and the subtree Facts, it showed that only the DOM tree did not solve these problems, the Web page denoising technology played a key step. Heuristic Web page denoising method is based on the location, size, font, and color ,the greedy strategy will be used by heuristic rules, which have been defined. For example Kushmerick [7] and Gupta et al. used the DOM structure of the page to confirm elements in Web clip, according to the basis of some rules[8].Deficiency is that there are a lot of links to the HUB page, that will make useful links on this page, and it can not be used. In 2004, cheung chi-kong waited to modify heuristic rules based on predecessors, which was a new information retrieval functions, they were convenient to extract the topics and the related pages. In 2005, the health care article, etc. When people thought of template tags, they should pay more attention to the anchor text of Web page link relations, the generated extraction rules were more targeted. Jian-dong wang in 2008, they put forward a kind of content-based Web purification algorithm rules, this was a comparison of table tree iteration algorithm, the iteration process for Web content filtering noise, and it based on modified editor, which was used to calculate the anchor text similarity algorithm, this kind of method has higher accuracy, semantic relation is that it takes into account a Web page.

B. *Method Based on Template*

Method based on template is drawn using a set of Web pages of the same element[9], the component template will have the effect of extracting information. It extracted from a set of Web pages out of the same template, and then used these templates to extract useful information from the Web Literature which is called DSE arithmetic (data - rich section extraction) algorithm, the algorithm by comparing the top-down two tree with the template page, remove the same subtree, put the rest as the theme of the Web page content and literature to the template type, Web page set presents a direct positioning theme information block, with rapid extraction method of Web page content.

C. *Method Based on Visual Information*

The method based on visual information is the layout of the page element, which contains information. Partition page reserved the middle area, other area classified as noise, in order to improve the accuracy of the content of denoising, Cai VIPS algorithm was put forward in 2003, the core of the algorithm was a collection of Web pages, that will be considered a visual block, visualization of extraction ,separation of the detection and content structure of a series of work, the end tag is cohesion value of each block in comparison with predefined termination conditions. Liu and Meng used the VIPS in 2006, the target was to define a set of rules and to extract the Web data record, this set of rules was carried out on the basis of the four characteristics, including location features (PF), layout (LF), performance features (AF), content (CF), the denoising effects began to gradually appear, the same year Yuan Mingxuan and others put forward the theory based on peer Web similarity matching algorithm, this algorithm was to filter the noise to a Web page, as was shown in figure 1 and figure 2.

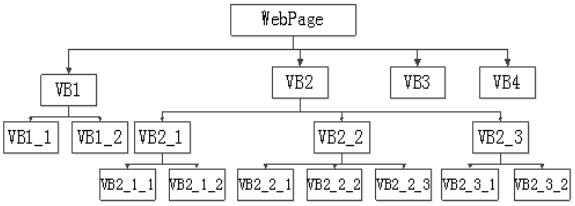

Figure 1. Content structure

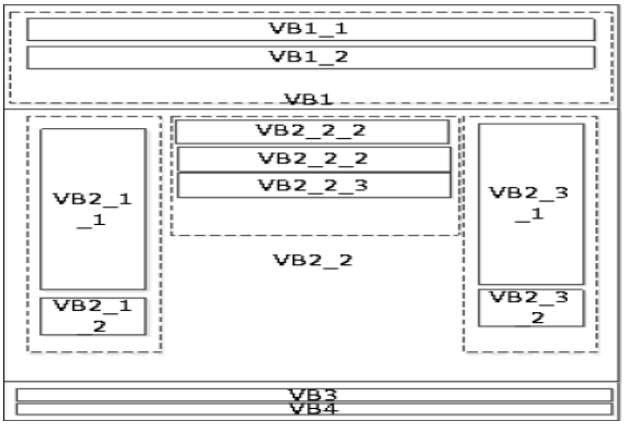

Figure 2. Visual tree

Three methods are applicable to the general situation,[10] but sometimes the effect is not very obvious. Web page denoising based on themes is vital to solve more accurate to remove noise.

D. *Method Based on Visual Information*

Its main process is divided into three parts[11], the first part is training data, the second is training generated classifier, the third is application classifier to remove noise. In the second step, the typical text classification algorithms are: Naive Bayes algorithm, KNN (K - on his Neighbor) classification algorithm and the SVM (Support Vector Machine) classification algorithm. Due to the effect of the classifiers directly, it determines the stand or fall of noise removal algorithm, which requires the choice of classifier has a faster speed and higher accuracy.

Through experiment contrast, we should choose to have a higher accuracy of the SVM classifier, using Sogou Labs to provide Chinese corpus, after Chinese word segmentation and feature extraction, training for a precision up to 97.8% (accuracy) of the SVM classifier. The third step is the key step of Debnath et al. put forward a method named L - Extractor, this method need to define a label set in advance, within the set to the DOM node began training, training result is to generate a classifier, the classifier to classify the contents of the Web, so as to achieve the Web denoising effect, but the qualification must be specified after the tag. Kao et al. after a paradigm based on information entropy is put forward, its emphasis is links and rich information of Web pages, the occupation ratio which can reduce the content of the template and Web denoising effect. For certain types of modules, the researchers also took time. For example arcandor et al. proposed an algorithm: by constructing content block (a DOM tree varieties), with classifier to determine page noise block, formula is as follows:

$$Noise(E)=\begin{cases}\mathrm{sgn}((1-r^{l})C(t)+r^{l}\sum_{i=1}^{k}\frac{isNoise(Ei)len(t_i)}{len(t)} & \text{E is not leaf node}\\ C(t) & \text{E is leaf node}\end{cases} \quad (2)$$

The formula explains that is r is the damping coefficient, r belongs to 0 and 1, which generally set to 0.8. L is the depth of the contents of the block tree, which is set to 1. Len (t) is the length of the content block text, and sgn() is the symbol function, isNoise (E) = - 1 shows that content block node E where the noise is, isNoise (E) = 1 explains that nodes E where the content block is not noise.

Davison's idea was to learn by decision tree [12], he detected and removed a part of the page there are many links; Kushmerick developed a learning and deleted pages of advertising ADEATER, it used the principle which was very simple, that was the predefined certain rules, his idea was to let it learn from some offline data and these rules, according to combined with the actual ads automatically, they can remove a Web page. Chakrabarti et al. summarized the experience of these people, designed a need human intervention, fully automated. He extended the good at the same time, but there was a problem: the template offered monotonicity properties, a DOM node in the DOM tree was template only if all of its child nodes were template.

IV WEB DENOISING EXPERIMENTS

For Web denoising, its accuracy and efficiency of assessment is very important. This several Web denoising methods used by the experiment data, which sets experiment method, and it is necessary to summarize.

A. *Date Set*

In previous studies, researchers from basic network selection on certain pages set LYRICS data developing slowly, after they used artificial marking on the collected data, but as a result of these data sets which were selected by the researchers, it did not have representative. Later in the study, network laboratory of Peking University and Peking University institute of computer establish and maintain information retrieval study BBS for test set, which includes a massive Chinese Web information.It aims at promoting information retrieval technologies and adopting in the study of relevant CWT100G and CWT200G .So it concludes that Web denoising is not a collection of data standards, but also it cannot be effected among different algorithms, they establishes a standard of test data sets or experimental platform is very necessary.

B. *Experimental Method*

Evaluation method of denoising methods can be divided into two categories, one is the direct evaluation of the denoising method, including precision and efficiency; a denoising method and data mining or search engines combine into using the denoising method .Performance measure is one of the important criteria, including F-measure, recall, precision and signal to noise ratio.

V EXISTING PROBLEMS AND RESEARCH DIRECTION

In information retrieval, information extraction bases on Web and knowledge discovery. In these areas, Web page denoising is regarded as a basic and important part, if it supports and falls of its effect and its efficiency, what directly affects the subsequent research. From the point of the previous sections, the best Web page is that the performance and effect of denoising method, when we have the very big enhancement compared to the original method, it applies to the study that also can have very good effect. But it is relative to the big data processing, Web denoising method is not perfect enough. As each Web page of denoising efficiency is increased by one percent, we realize the overall performance is relatively easy. As far as data collection concerned, there is not a set of complete standard test set and unified criteria, it makes the Web great difficulties in the further research of the de-noising.

Further research for Web denoising is mainly concentrated in the following aspects: 1) we should improve the performance of the algorithm, performance is Web denoising criteria. 2) we should improve the result of algorithm, the denoising effect of ascension can only prove that denoising effect obviously has the possibility to succeed. 3) we should improve degree of unsupervised algorithm, when we realize fully automated, this will

become the direction of future efforts. 4) general Web page denoising algorithms, and specific types of denoising ,which will have double parallel development. Only in this way, can Web denoising play a role in every aspect. 5) we should build a standard test data set or experimental platform, a perfect set of test data is the foundation of the Web page denoising.

VI CONCLUSION

This paper analyzes and summarizes the necessity of Web page denoising, the main methods and framework of Web page denoising, the experimental method, the problem and the research direction of the Web page denoising. The original data set by Web crawler technology, the Web page pretreatment can be directly applied to the research, and the important step in the preprocessing is based on Web page denoising, denoising now faces many problems, for instance denoising efficiency is not high enough, the effect is not obviously, there is no standard test data set and test platform. Meanwhile, test is a variety of methods, there are some problems in each method of Web denoising. So it is necessary to research a set of good results, at the same time the accuracy and efficiency is a denoising method which measures good or bad, all of the above issues still need to be a step by step in-depth study, so as to draw more accurate conclusions.

ACKNOWLEDGMENT

This work is supported by “the National Natural Science Foundation of China (No.61331033)” and “Beijing Social Science Foundation (No.14WYB040)”.

REFERENCES

[1] A. Broder. On the resemblance and containment of documents. In Compression and Complexity of Sequences, 10(3): 21-30, 1998.

[2] Zhang Zhigang, Chen Jin, Li Xiaoming. An approach toreducing noise in HTML pages[J]. Journal of the ChinaSociety for Scientific and Technical Information, 2004, 23(4): 387- 393

[3] Wang Jiandong, Wang Jimin, Tian Feijia. An algorithm fornoise reduction in Web pages based on a group of content-related rules [J]. New Technology of Library andInformation Service, 2008, 22(3): 51- 54

[4] Wan Le, Zuo Wanli, Gao Jin. Web pages noise removalbased on focused topics[J]. Computer Engineering andDesign, 2008, 29(8): 2072- 2076

[5] Fetterly D, Manasse M, Najork M, et al. A large-scalestudy of the evolution of Web pages[J]. Software: Practiceand Experience, 2004, 34(2): 213- 237

[6] Gibson D, Punera K, Tomkins A. The volume and evolutionof Web page templates[C] Proc of the 14th Int Conf onWorld Wide Web. New York: ACM, 2005: 830- 839

[7] Bar-Yossef Z, Rajagopalan S. Template detection via datamining and its applications(2002: 580- 591)

[8] Manku G S, Jain A, Sarma A D. Detecting near-duplicatesfor Web crawling(2007: 141- 150)

[9] Coughlan J, Yuille A, English C, et al. Efficient deformabletemplate detection and localization without user initialization[J]. Computer Vision Image Understanding, 2000, 22(78):303- 319

[10] Ma L, Goharian N, Chowdhury A, et al. Extractingunstructured data from template generated Web documents(2003: 512-518)

[11] Jing Tao,Zuo Wanli. An algorithm for the elimination of thenoise in Web pages based on visual layout information[J]. Journal of South China University of Technology: NaturalScience Edition, 2004, 32(Suppl 1): 84- 88

[12] Ou Jianwen, Dong Shoubin, Cai Bin. Topic informationextraction from template Web pages[J]. Journal of TsinghuaUniversity: Science and Technology, 2005, 45(9): 1743-1747