A Method Based on Statistics for Extracting Text from Web Pages

Wang Nan and Xiao Chun

***Abstract*—This paper designed a method which is based on statistics for extracting text from news web pages. First it translated a web document into a DOM tree by HTML tag. Then it structure leaf node’s feature vector according to features of punctuations statistics. Last calculate the similarity between two leaf nodes and weights by feature vector. And it extracted text from those leaf nodes whose weights greater than the threshold value. The results show that the method has property accuracy in text extraction and better universal.**

***Index Terms*—Information extraction, DOM tree, statistical characteristics, similarity.**

I. Introduction

With the rapid growth of Internet data, how to effectively from the information ocean mining the data of interest to a user, information search field has become an important research topic in the. Web information extraction refers to from the Web page contains no structure, a structure or to identify the user interest data, and transform it to the structural and semantic clarity to the format of the Web page information extraction process [1]. According to the realization of the principle is different, the existing Web information extraction tool is divided into the following categories：based on natural language processing（NLP）information extraction, based on structure of HTML information extraction, based on inductive learning of information extraction, based on the web query information extraction, based on the characteristics of pattern matching information extraction [2,3]. This paper presents a method for extracting from the Chinese news type webpage: build a DOM tree after web pretreatment, select the sample node to find similar leaf nodes and other treatment, to obtain text information.

II. Punctuation Statistical Features in Web Webpage

Punctuation is not words, but they are as important as text, is the organic part of the language. According to statistics, in the Chinese modern works, words and punctuation at the rate of about ten to one, each of the ten Chinese characters will appear a punctuation mark.

How to filter advertising, news and other unwanted content is the biggest obstacle to a web page content extraction, but this part of the content is often represented by pictures, a short text. In order to give the impression that the clear visual impression and aesthetic feeling, therefore this part's content won't include punctuation. Whether the use of punctuation appearance judging a paragraph of text is useless content? We randomly selected 431 body type webpage from Sina, Netease, Sohu website, Statistics of 16 kinds of common Chinese dots and labels appear number in the text of the page A_1. The number of occurrences on the page A_2

TABLE I: The Statistical Results

Dots	A_1	A_2	Labels	A_1	A_2
Full stop	8137	8509	Quotation marks	229	411
Comma	22938	23907	Brackets	192	574
Interrogation	742	1136	Dash	81	374
Caesura sign	364	583	Emphasis	0	0
Semicolon	327	506	Connect sign	0	15
Colon	511	770	Ellipsis	5	33
Exclamation mark	213	541	Interval sign	162	495
			Book signing	28	63

This shows: punctuation especially the dots appearance text segment is a high probability of web page text, we can accord the characteristic of a web page content whether the web page text.

A. The Information Extraction Method Based on the Statistical Regularities of Punctuation

So far, most of the web webpage is the use of HTML language, because HTML allows for the existence of nonstandard writing style, it is difficult to extract. Therefore, you must first of webpage specification, requirements are as follows:

B. Web Webpage Standardization

1) “<” “>” Can only be used to contain webpage marker (tag) When the emerge in other locations of the two symbols should be “andlt” and “andgt” instead of [5].
2) All start and end tags must match. Each start tag < * * * > must correspond to an end tag < / * * * >.
3) All tag attribute value must be put in quotation marks, such as < table height =“300” >.
4) All tags must be properly nested. For example, < a >... < b >. < / a >... < / b > is not properly nested, properly nested form should be < a >... < b >... < / b >. < / a >.

Manuscript received February 24, 2012; revised April 27, 2012.
Wang Nan (e-mail: 582952853@qq.com, tel.: +8615200321822).

5) All marker size write must be consistent, the default all lowercase letters.
6) Because the text may be text modification markers such as < font > < strong > partition, in order to keep the text content and modification of markers between sequential, increase the custom tag < text > < / text > to nested text content. Through standard web document can be easily according to which HTML marker puts it into an HTML file hierarchy reflects the DOM tree, the tree of each node contains a pair of markers between characters, node name for the corresponding tag name.

C. Webpage Text Information Extraction

We know the information present in the leaf nodes, so we are concerned only with the leaf nodes, A leaf node with a feature vector V representation, $V = \{w_1, w_2, \ldots, w_n\}$, In which w_i express the weight of punctuation first i . n is the total number of punctuation. w_i use the following formula:

$$F_i = \left(\sum_{j=1}^{m} b_j f_{ij} \right) \log(H / n_i) \tag{1}$$

$$w_i = F_i / \left(\sum_{j}^{n} \left(F_j \right)^2 \right)^{1/2} \tag{2}$$

Equation (1): f_{ij} is the punctuation mark p_i frequency of occurrence in the first j leaf nodes; b_j is the weights of first j leaf node; H is the total of all leaf nodes in Leaf node set S ; n_i is the number of nodes contain first i punctuation in Leaf node set S ; m is the number of leaf nodes in the Dom tree corresponding to a webpage document; As can be seen, a punctuation in a leaf node in the higher the frequency, the greater the weights w_i of punctuation, But if it is in all the leaf nodes of the frequency is higher, so the punctuation is not very well put the leaf node distinguish leaf nodes to distinguish, so its weight is smaller. Equation (2): It puts the weight normalization, guarantee value w_i of from 0 to 1. Definition 1: similarity between leaf nodes, hypothesis, $V_j = \left\{ v_{j1}, v_{j2}, \ldots, v_{jn} \right\}$ they are feature vectors of two leaf node, Definition similarity (v_i , v_j) $= \cos(v_i, v_j)$,it's the similarity between leaf nodes. When similarity (v_i , v_j) is greater than a certain threshold ε , we say the two leaf node is similar. Definition 2: the weight of the leaf node, hypothesis S is leaf node set, the number of nodes which are similar to V in set S . Definitely $I = \log\left[1/(N+1)\right]$, it's the weight of leaf node V . In the test, we gave a there should θ node greater eight of leaf node V greater than θ , we output it's content as the content of the text. The steps of the algorithm are as follows:

1) Traverse the DOM tree, to extract all the < text > point, generates a leaf node set.
2) For each leaf node V in the set S we count the frequency of occurrence of 7 points in Table 1,
3) For each leaf node V_i ,we count the similarity between it and the other leaf nodes V_j , If the similarity is greater than a certain threshold, then re-calculated the weight value I of leaf node V_i
4) We select those leaf nodes value I is greater than a given threshold θ , output its content as the content of text.

We randomly sampled 200 webpage from 10 sites, Calculation ε , θ with different values, corresponding to the correct rate, After a lot of experiments, we found that when $\varepsilon = 0.8$, $\theta = -2.5$ best. As shown in figure 1:

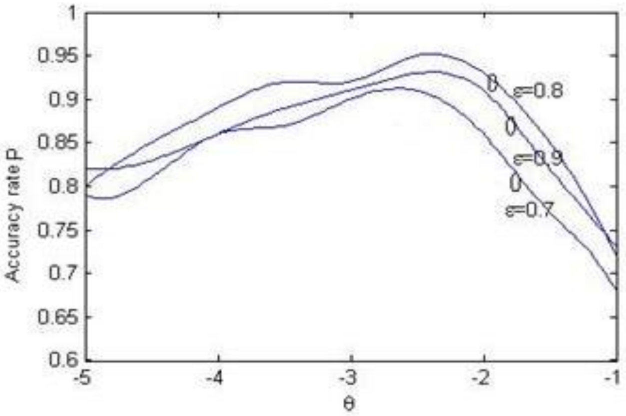

Fig. 1. Threshold ε , θ

III. The Experimental Results

In order to verify this method of actual effect, we listed in Table 2 randomly selected parts of webpage website, In order to meet the methodological assumptions, try to select the text information in WebPages. The extraction result is acceptable (on a web page text extraction accuracy is above 85%), not acceptable (on a web page text extraction accuracy below 85%), With the artificial extraction of text comparison, the experimental results are as follows: In this experiment the threshold $\varepsilon = 0.8, \theta = -2.5$.

TABLE II: The Experimental Results

Website	Total	Acceptable	Unacceptable	Accuracy rate
sina	200	193	7	0.965
sohu	200	195	5	0.975
infzm	160	154	6	0.963
cctv	100	89	11	0.890
yahoo	150	145	5	0.967
Total	810	776	34	0.958

Table 2 shows, this method for webpage text information extraction accuracy up to 97.5%, the lowest 89%, for an average of 95.8%, that the method has higher accuracy and feasibility. Make a comparison between This algorithm and The method based on DOM from Literature [4]. This algorithm's accuracy has been greatly improved, as shown in table 3.

TABLE III: ACCURACY COMPARISON

Website	Accuracy rate	
	The method form Literature [4]	This algorithm
sina	0.980	0.965
sohu	0.970	0.975
infzm	0.907	0.963
cctv	0.829	0.910
yahoo	0.943	0.967
Total	0.929	0.958

Experimental error refers to the extraction information contains a text or text content is not extracted. For error webpage, our study found, their containing text information is very short, or evens just a simple word. This algorithm is not limited to a specific type of webpage, has very strong versatility, while maintaining high accuracy.

IV. CONCLUSION

This paper presents a method based on punctuation statistical feature extraction method for webpage text information. The method of the type of news webpage has higher accuracy, can effectively remove the webpage of navigation, image, advertising and copyright and other irrelevant information, accurate extraction webpage text content, experimental results show that the method is feasible. This method is applied to most sites, but there are also part of the site is not suitable for. The next step is to further explore the two thresholds on extraction effect, improve the leaf node similarity calculation method, and further improve the universality of the algorithm, in order to adapt to various types of web page text information extraction.

REFERENCE

[1] Q. J. Liu, H. Jia, and Hui-bo,"Research on Approaches of Information Extraction System," *Application Research of Computers*, vol. 24, no. 7, pp. 6-9, 2007.

[2] J. G. Liu, G. S. Liu, L. Y. He, and S. Chen, "Web based information extraction technology status and development," *Fujian computer*, 2007, vol. 7, pp. 48-49.

[3] G.-P. Hu, W. Zhang, and R.-H. Wang, "Precise Content Extraction from News Web Page Based on Decisions of Two Layers," *Journal of Chinese information processing*, 2006, vol. 20, no. 6, pp. 1-10.

[4] M. Song, R. Zhang, X. Wu, and W. Li, "A new approach to content extraction from web pages," *Journal of Dalian University of Technology*, 2009, vol. 49, no. 4, pp. 594-597.

[5] D. Buttler and L. Liu, *et al*., "A Fully Automated Object Extraction System for the World Wide Web," *Proceedings of the 2001 International Conference on Distributed Computing Systems [C]*. 2001, pp. 361-370.

[6] D. Cai, S. Yu, J. Wen, and W. Ma, "VIPS a Vision-based Pages Segmentation Algorithm," Microsoft Technical Report MSR-TR-2003-79, November, 2003.

[7] D. Ikeda and Y. Yamada. "Expressive Power of Tree and String Based Wrapper," In: *online proceedings of IJCAI'03 workshop on Information Integration on the Web [c]*. 2003.

[8] L. Yi, B. Liu, and X. Li, "Eliminating Noisy Information in Web Pages for Data Mining," *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[9] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "DOM-based Content Extraction of HTML Documents," *The Twelfth International World Wide Web Conference*, 2003.

[10] Knobloccak, Mntons, and Ambitejl, *et al.*, "The Ariadne approach to Web-based information integration," *Journal on Cooperative Information Systems*, 2001.