A New Approach for Web Information Extraction

R.Gunasundari  
Research Scholar  
Karpagam University  
Coimbatore, India  
E-mail: gunasoundar@rediff.com

Dr.S.Karthikeyan  
Director,School of Computer Science  
Karpagam University  
Coimbatore, India  
E-mail: skaarthi@gmail.com

Abstract

With the exponentially growing amount of information available on the Internet, an effective technique for users to discern the useful information from the unnecessary information is urgently required. Cleaning web pages for web data extraction becomes critical for improving performance of information retrieval and information extraction. So, we investigate to remove various noise patterns in Web pages instead of extracting relevant content from Web pages to get main content information. To solve this problem, we put forward an extracting main content method which firstly removes the usual noise and the candidate nodes without any main content information from web pages, and makes use of the relation of content text length, the length of anchor text and the number of punctuation marks to extract the main content. In this paper, we focus on removing noise and utilization of all kinds of content-characteristics, experiments show that this approach can enhance the universality and accuracy in extracting the body text of web pages.

Keywords- information extraction; web page content extraction; removing noise content.

I. Introduction

The rapid expansion of the Internet has made the WWW a popular place for disseminating and collecting information. Extracting useful information from Web pages thus becomes an important task. Usually, apart from the main content blocks, web pages usually have such blocks as navigation bars, copyright and privacy notices, relevant hyperlinks, and advertisements, which are called noisy blocks. Although such information items are functionally useful for human viewers and necessary for the Web site owners, they often hamper Web page clustering, classification, information retrieval and information extraction [8]. Therefore how to identify and separate main content blocks from noisy blocks is a crucial issue.

Whenever a user query the web using the search engine like Google, Yahoo, AltaVista etc, and the search engine returns thousands of links related to the keyword searched. Now if the first link given by the user has only two lines related to the user query & rest all is uncluttered material then one needs to extract only those two lines and not rest of the things.

The current study focuses only on the core content of the web page i.e. the content related to query asked by the user. The title of the web page, Pop up ads, Flashy advertisements, menus, unnecessary images and links are not relevant for a user querying the system for educational purposes.

The traditional web page text extractions generally have three steps: web pretreatment such as tag correction, removing useless labels etc, then the web pages content extraction according to various text extraction algorithms, and the result correction. In the preprocessing stage, most methods just simply remove invisible code which is irrelevant to content extraction such as style, comment, script etc. In fact, the web pages purified by this method still have a lot of visible noise, such as navigation links, sidebar, website copyright statement, etc., It will directly affect the accuracy of following text extraction algorithm [9]. Meanwhile, the traditional methods can’t work well on body text
which contains many links, because they assume that lack of hyperlink is a feature of body text, while a lot of links exist in financial news. In a word, it is very necessary to put forward a more universal and accurate method to extract body text from web pages.

In order to solve these problems, we propose a new approach to extract body text. In preprocessing stage, it is designed not only to remove irrelevant invisible code, and also to some extent remove the visible noises mentioned earlier. After this step, we obtain more pure web pages to reduce noises interference for content extraction from web pages. We make full use of relations among text length, punctuation marks and links to extract body text. Our work is focused on the combination of eliminating noise and comprehensive utilization of characteristics of body text [10]. In the following section, we will give related work and detail steps of content extraction. Finally, the progress in the accuracy and applicability of our method is proved by a series of experiments on random news pages from different sources.

II. Related work

Currently, there are already a lot of content extraction algorithms based on different features of main text.

1) Lin and Ho [1] proposed an extraction method based on information entropy, the web page is divided into content block according table tag, each block has entropy, and then information blocks are obtained by comparing with threshold value. But this method just applies to web pages which contain table tags, while increases the complexity of the algorithm.

2) Yi and Liu [2] put forward an extraction approach based template, this method assumes that the same part of two web pages having same format, so it is simple and effective to remove noises by comparison of two web pages coming from one source. But it is difficult to identify so many templates for a variety of web pages.

3) The extraction approach based on framework web pages and rules [3], supposes it is reasonable the noise blocks generally locate in the secondary positions in the page. This method compares the ratio of width and height attributes of every table tag, and removes the tags of bigger ratio. It is hard to work well on table tags with less height and width attributes. The table tag is the only Processing Objects to this approach.

4) Web content information extraction method based on tag window [4], which could cope with some special circumstances that web pages content text locate in table and div tags, all page content information is put into one td or several tds, and the length of body text is short as that of the other information such as navigation bars, and the copyright, etc. But during the process of judging body text, it involves word segmentation and computing similarity of string, which has enhanced the complexity of information content extraction.

5) Pan and Qiu [5] put forward a web page content extraction method based on link density and statistic, which recognizes main content according to the different properties between content nodes and non content nodes of web page represented as a tree. But the threshold values in this paper do not always adapt to some news pages such as stock news, so it is still hard to find a set of universal parameters.

These methods, based on removing noise is suitable to delete a lot of nodes without any web content information, they can contribute to filter the unrelated part using the layout properties of noises in the web pages. The most current ways lack of enough considering on removing noise in the preprocessing. Furthermore, large numbers of financial news have so many hyperlinks in the information text that those methods which over-reliance on links have poor results. After analyzing these methods about removing noise and information content extraction, and moreover, the relation among punctuation mark density, length of information text and anchor text is considered enough in the extraction stage. We put forward a method based on removing noise and characteristics of body text.

III. Approach used

On Internet large amount of information is available in the form of HTML pages. Most sites use the same presentation style for maximum or all web pages. The non-content blocks share some common contents and presentation styles. While main content blocks are different in their content and /or their presentation style. Various algorithms developed to extract content blocks from web pages share this observation.
A. Related Outset

The main text often contains the comparatively many characters, in addition, the amount of punctuation marks is proportional to length of information text. Most of the noise nodes are composed almost exclusively of anchor text, almost no texts, and few punctuation marks. We describe the extraction algorithm by introducing the following parameters.

1) Node-Link Text Density (NLTD): the ratio of the anchor text length to all text length in a given node. To a certain extent, this value plays a good role in identifying main content of general news pages. Comparing noise blocks, the smaller this value, the greater the possibility of main content. But it is not uncertainty in some special circumstances that the greater value always represents the higher possibility of noise blocks. This indicator can work well when cooperates with other indices.

2) Non-Anchor Text Density (NATD): the ratio of the non-link text length in a given node to the total non link text length in all nodes of news pages. This indicator denotes the absolute proportion of useful information in the whole news page. So the possibility of information content blocks is in direct proportion to this value.

3) Punctuation Mark Density (PMD): the ratio of the number of punctuators in a given node to the number of punctuation marks in the news page. Because of the lack of punctuation marks in the noise blocks, it is reasonable to consider punctuation density as an indicator of judging information content blocks.

These three parameters are used to calculate the weight of each information content block. None of them is decisive factor in judging main content, so we obtain a good result when they coordinate with each other according a certain proportion.

B. Steps of main content extraction

Step 1 : Preprocessing web page

a) Standardizing the page tags is the first step of most content extraction algorithms. Tags should be paired and nested correctly, etc.

b) General useless tags which are unrelated to valid information, such as <script> <img> <style> <declaration> <option> <comment> <meta>, etc, should be deleted.

c) The potential useful nodes which are virtually noise nodes without any valid information, such as navigation bar, sidebar, etc. Most of them frequently have such a feature that only anchor text and very few punctuation marks are included in the noise nodes. According to this feature, we can remove the tags which the length of non-anchor text is equal to zone, this step should be done on the premise that the tags are the container tags such as <td> <div> <table> <p>, etc, the aim of this precondition is to prevent the useful hyperlinks from being removed [6]. After carrying out this rule, many noise nodes are deleted from web pages which are expressed as structure tree [7]. This purification measure is simple, but critical step in the preprocessing of web page.

d) Web page tags are still expressed by tag windows [4], this form can deal with the conditions that the main content is put into one or several tags. The tag window contains all the information of the tag, such as original text and code, length of text, amount of punctuation marks, etc.

Step 2 : Extracting the main content

Each tag window has a weight which indicates the possibility of main content. They are calculated according to the following formula.

\[
\text{Weight} = \text{NLTD} + \text{NATD} + 0.7 \times \text{PMD} \quad (1)
\]

The bigger the tag window, the more information content and noise it contains. After this measure, we obtain rough weights of tag windows.

Step 3 Adjusting the weight

It is impossible to completely eliminate the noise from web pages in the front work, most pages still have some noise, such as "ad-serving", "contact", "company profiles", "copyright ","©", "all rights reserved" and other noise words. It is necessary to adjust the weight to make the result more precise. Though they may appear in the main content, the possibility
of appearance together is low. Accordingly, we stipulate that if five or more of these words appear in some tag window, and then subtract 0.6 from weight of tag window. After sorting the weights, it is easy to find the optimum tag window which has the greatest weight.

IV. Flowchart of content extraction

![Flowchart of content extraction]

V. Experimental Results

In this paper propose a new content extraction algorithm. It differentiates noisy blocks and main content blocks. We present here the experimental results to testify the effect of algorithm. Fig. 2 is a news page about university grants commission. In this example, there are so many links in the main content that they can produce enough noise. The block of recommended reading is in the main content, but the block is contained in the div tag, besides, it only has the link text, this noise block is trimmed in the preprocessing.

At the same time, So many links in the text, reduces the weight of the text, but the punctuation density effectively supplement the weight value of tag window.
In this work, the experiments are performed for the applicability and accuracy of method. We choose 1000 web pages which are general news pages from some well-known websites such as yahoo, sum Up, xfocus, tom etc. Besides some web pages selected randomly such as job, comxc etc.

VI. Conclusion

Web page content extraction are more and more essential for mining the main content of pages, especially in such ever-growing and unstructured web. In spite of the many different approaches presented, no universally recognized algorithms are available for extracting exactly the main content from different websites. This approach trims invalid code and useless candidate information nodes, provides relatively pure input for the step of extracting. The algorithm extracts the content by making full use of those characteristics above. We will do a lot of work in applications based on web pages, such as generating ontology instances and event information extraction.

VI. References


R. Gunasundari is presently working as a Assistant Professor in the Departmetnt of MCA, Karpagam University, Coimbatore. She has six years teaching experience. She has participated and presented seven papers in national conferences.Currently she is working in the area of data mining and its application to web mining.

Dr. S. Karthikeyan received the Ph.D. Degree in Computer Science and Engineering from Alagappa University, Karaikudi in 2008. He is working as a Professor and Director in School of Computer Science and Applications, Karpagam University, Coimbatore. At present he is in deputation and working as Assistant Professor in Information Technology, College of Applied Sciences, Sohar, Sultanate of Oman. He has published more than 14 papers in National/International Journals. His research interests include Cryptography and Network Security.