Research of the Web Page Cleaning Technology on Tourism Theme*

Qi Shen, Qingming Song, Meng Zhang and Yan Tang
School of Software Engineering Beijing University of Technology, Beijing, China
shenq@bjtu.edu.cn, qingfengmingxiu@gmail.com

Abstract - With the development of web technology, the use of dynamic web pages and the personalization of page contents become more and more popular. Currently, the information of page is protean and the structures of different pages are vastly different, the traditional thinking of page cleaning technology has been difficult to adapt to the situation.In this paper, proposes a web cleaning method based on regex extraction strategy through the analysis of structural features of web pages on tourist theme.This algorithm avoides the defects of traditional page cleaning technology, it is simple, practical, high cleaning efficiency, accuracy, and saving the overhead of the system.

Index Terms - tourist theme pages; html; page cleaning; regex.

1. Introduction

Resource information on the Internet today is still concentrated in a variety of web pages, the content is rich and the structure is complex. But sometimes we are concerned may only page which a small fraction of the content block[1].How to remove a lot of useless information from different page structure, access to the data only needed and provide better retrieval services, which becames an important work in the process of web resources. Page cleaning technology is to achieve this goal. Most of the web cleaning systems are very similar, which are based on the analysis of document structure tree model DOM[2]. This mechanism is highly dependent on the similar structure of the pages, but for the heterogeneity and complexity of today's Internet pages, this mechanism of processing results is not ideal, and the implementation of the algorithm is complexity and conducive to update.Therefore, it is necessary to consider studying new handling mechanism to solve this important problem of the work of web cleaning.

2. A brief introduction of page cleaning technology

A. The concept of page cleaning

Page Cleaning is a process which is simply said to filter out the useless information and retain the useful data from the web pages. It is a very important part in the process of web information collection, and it is the basis of the treatment process after the acquisition system.The level of efficiency of the cleaning of the page directly affect the performance and efficiency of the entire acquisition system. Page cleaning is divided into three main steps as follows:

1) Remove irrelevant information which the page style sheets, scripts and comment.

2) The page sub-blocks, including the image blocks, the text block and the link block,etc.

3) Further screening of each block in accordance with the specified rules, for example, isolated advertising links, navigation links, useless information, announcements from the link block and advertising,other non-critical information from the text block.

After the processing of the above steps, the page on the structure and semantics to be divided into fine-grained information blocks, so that the subsequent information processing work can be smoothly carried out [3].

B. Problems

From the source of the web page, most of them are concentrated in some portals and various communities. These sites generally have their own unique style, the template is relatively fixed, a common feature of such a web is clear and simple structure.Because of a large number of repeat structure within the page, thus, facilitating the web structure analysis, while the larger the amount of information contained within the page. These similar structural unit not only exist in a single page, in different pages can be kept substantially consistent. For these pages, as long as they can make good use of characteristics on the web page structure, cleansed the useless part of these repeated structures, you can achieve the effect of the cleaning of the page.

Different web page contains information about the ever-changing structure is varied, not only contain text, images, and other traditional information, like animation, video, audio and other complex forms of information has also been introduced into the page. The same time, the demand of the widespread use of the dynamic web pages and web personalization, has been very difficult to find a similar structure in different website page[4]. The page cleaning thinking has been difficult to adapt to the current page cleaning needs. The page content is constantly updated, so page cleaning algorithm is complex and not easy to achieve, the actual effect of the application is also often not satisfactory.

Therefore, it is necessary to design a new page cleaning mode based on page structure characteristics, this page cleaning algorithm not only have universality, and also be able to structure and content for specific pages targeted cleaning and filtering. This personalized, targeted design thinking can not only reduce the difficulty of the current page cleaning

* This work is supported by Scientific Research Project of Beijing Municipal Commission of Education Grant #KM201210005030.

algorithm, while improving the efficiency of the cleaning of the page, and can adapt to the current web information collection technology trends based on the theme, directed, personalized.

3. Analysis and Research

A. Solution ideas

The traditional page cleaning algorithm design are the structural features of the web page based on the page's DOM tree model.Remove a series of "noise" information after analysis and learning algorithm. In this way the algorithm is too complicated, high the page extraction efficiency for a specific structure, but for heterogeneous web treatment effect bad, and often can not meet the requirements[5]. The algorithm is based on DOM tree model, there are certain requirements for memory, wasting memory space, it is clearly not applicable for large information collection system.

Analysis page for theme-based web information directed acquisition system, the following conclusions can be given: A page of useful information is often only a small part of the entire page content.

With this conclusion, if in a different perspective, the page cleaning from the page to find out “noise” into extracting useful information directly on the page, so that you can greatly reduce the complexity of the page cleaning andThe degree of difficulty. Complex diversity and local fixity tags for web pages, if using regular expressions to directly match and extract the label that we need to retain the useful information and automatically remove out useless information, this approach can be effectivethe page cleaning purposes[6].

Therefore, this paper will present a special tourist theme pages, page cleaning algorithm based on regular expression matching extraction strategy. This algorithm avoided the defects of traditional page cleaning technology, it is simple, practical, high cleaning efficiency and accuracy, while saving the overhead of the system.

B. Travel web features and structural analysis

Travel information web pages has a strong development value, it has features as the stable structure, the large amount of information, practical value, needs extensive,etc. Consider of these characteristics, if using a traditional page cleaning work on these pages apparently can not meet the demand. Such pages close to the functionality of the major sites, web structure is relatively stable, and will not be bound by a specific site, suitable for high-volume integrated acquisition and processing, such as Xie Cheng network, Tu Niu network, Qu Na network,etc. These pages overall structure is essentially the same. Here is to an instance of Qu Na network travel information pages about the hotel information:

```
<html>……<head>……</head>……<body>
……
<!--服务设施:star-->
<div class="b_hotelservice"><h3>服务设施</h3>
<div class="e_hotelservicect">
        <dl class="e_servicelist"><dt>开业时间：1985年</dt>
        <dd>
                <span>最后装修时间：2002年</span>
                <span>房间数：270间</span>
        </dd>
        </dl>
        <dl class="e_servicelist"><dt>服务项目：</dt>
        <dd>
                <span title="商务中心">商务中心</span>
                <span title="停车场">停车场</span>
                <span title="会议室">会议室</span>
        </dd>
        </dl>
        <dl class="e_servicelist"><dt>酒店设施：</dt>
        <dd>
                <span title="健身房">健身房</span>
                <span title="酒吧">酒吧</span>
                <span title="桑拿">桑拿</span>
        </dd>
        </dl>
        <dl class="e_servicelist"><dt>客房设施与服务：</dt>
        <dd>
                <span title="宽带上网">宽带上网</span>
                <span title="洗衣服务">洗衣服务</span>
                <span title="空调">空调</span>
        </dd>
        </dl>
        <dl class="e_servicelist"><dt>酒店周边：</dt>
        <dd>
                <span title="三里屯酒吧街">三里屯酒吧街</span>
                <span title="农业展览馆">农业展览馆</span>
                <span title="朝阳公园">朝阳公园</span>
        </dd>
        </dl>
</div>
</div>
<!--服务设施:end-->
……
</body>……</html>
```

Fig. 1 Instance of tourist hotel information page.

The web page structure expanded on the above is as follows:

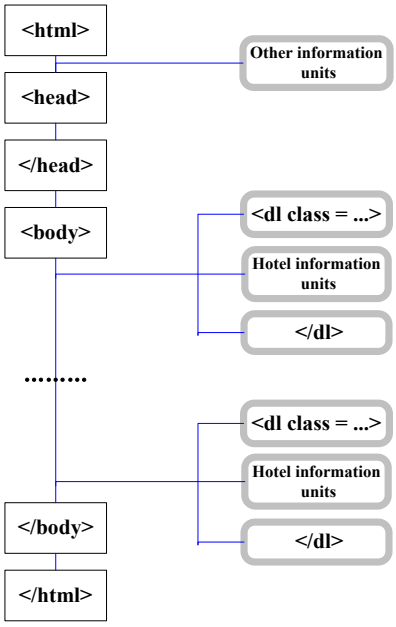

Fig. 2 Web page structure.

As can be seen from the figure, the structure of this site is very clear, we need to travel information that is in the hotel this part of the information unit. Therefore, we do not need to use the traditional cleaning method of the page, a step-by-step analysis of clear out unwanted content blocks up to retain the required contents of the block, just need to <dl> </ dl> between content directly match. This also can achieve the same the page cleaning purposes retain the desired information, filtering the information of the "noise".

Similarly, continue to observe the site diagram can be found between <dl> and </ dl> tag is structured. Screening fragments were analyzed as follows from the above examples:

```
<html>
    <head>
    <title>酒店信息</title>
    </head>
    <body>
        <dl class="e_servicelist">
        <dt>酒店周边：</dt>
            <dd>
                <span title="三里屯酒吧街">三里屯酒吧街</span>

                <span title="农业展览馆">农业展览馆</span>

                <span title="国贸展馆">国贸展馆</span>

                <span title="朝阳公园">朝阳公园</span>

                <span title="工人体育场">工人体育场</span>
            </dd>
        </dl>
    </body>
</html>
```

Fig. 3 dl Label fragments.

This is the the hotel information unit module, this module includes a hotel and some neighboring location information, data information is exactly what we need to collect. Matching module information on this unit again, will be able to more precise positioning acquisition of data, leaving the page cleaning work more in-depth and thorough help provide the greatest degree of subsequent information extraction work. Above each one of hotel information, metadata is information reorganization and reproducing the best material.

As can be seen, the cleaning of the core ideas of the entire page is based on matching extraction strategy. Regex is fully equipped with the functionality required by the above matching extraction strategy, therefore, consider the use of regular expressions to assist in the completion of the task of the entire page matching extraction work.

C. Based on regular expression matching extraction strategy to achieve page cleaning work

Regular expressions are the standard package “ava.util.regex” of JDK1.4 added. It can pinpoint any character or string pattern matching, and can be extracted to determine replacement function[7]. Pattern class and Matcher class are two of the most centra. Pattern class main role is to build a regular expression to match the regular expression syntax for. Matcher class is built according to the Pattern class regular expression matching objects for further processing, and matching content.

For travel information pages previously discussed, the module contains tourist information the instance contains <dl> label at the same time there may be other labels also contain similar information we need to collect. Due to the simple structure of the travel information page, type mode is relatively fixed, according to the statistical analysis can be drawn: General tourism information data exist in the body of the page paragraphs, lists and tables, Taking into account the extraction of other necessary information, such as web hyperlink, the page's Meta information,etc. Finally, came to the conclusion that the page matches the need to extract the common label roughly the following categories: <Meta>,<a> <p>,<dl>, <ul>, <table>,etc.

The match abstract tags and Solutions here in order to facilitate presentation, select <ul> label as the example shows:

```
<html>
……
<!－同地区一口价酒店推荐：-->
<ul>
        <li>
            <a href =“http://discount.qunar.com/city/
                    beijing_city/10245.html”>
            燕莎国贸商圈近永安里地铁站四星级品牌酒店</a>
        </li>
        <li>
            <a href =“http://discount.qunar.com/city/
                    beijing_city/10040.html”>
            北京站、建国门地区三星级商务型酒店</a>
        </li>
        <li>
            <a href =“http://discount.qunar.com/city/
                    beijing_city/10332.html”>
            北京农展馆黄金地段五星级国际知名品牌酒店</a>
        </li>
</ul>
……
</html>
```

Fig. 4 ul Label fragments.

Matching extraction work is divided into two steps, and every step of the need to define a Pattern object:

1) The matching outer layer <ul> tag.

2) For the first step of matching results, matching the inner layer <a> tags and extract its contents.

The first pattern object of the regular expressions are as follows:

Pattern pattern1 = Pattern.compile(“<ul([^>]*)>(.*)?”,Pattern.DOTALL|Pattern.MULTILINE)

The above Pattern can match <ul> label from the page. The “<ul>” said this at the beginning of the matching tags, “([^>]*)” Indicates match in addition to"> "outside all characters can appear any number of times, and grouping them together, actually represent <ul> tag attributes, “(*)?” Said content of the <ul> tag, is that we need to collect information block, “”is the end of <ul> tag. The “Pattern.DOTALL | Pattern.MULTILINE” compile method of optional parameters, indicating that this regular expression can be a multi-line match, and is case sensitive.

The second Pattern object of the regular expressions are as follows:

Pattern pattern2 = Pattern.compile(“(.*)?,Pattern.DOTALL|Pattern.MULTILINE)

The above Pattern can match <a> label from the <ul> label. "<A href =" indicates matching tags beginning of this, “\ \ “([^

\ \ “] *) \ \” ” Said “href” attribute value, which is the address of a hyperlink, “(*)?>”Represent other attributes of <a> label, “(.*)?” Said content and the end of <a> label, The "Pattern.DOTALL | Pattern.MULTILINE" compile method of optional parameters, indicating that this regular expression can be a multi-line match, and is case sensitive.

For each Pattern object rules, there will be an the Matcher object corresponding match, and subsequent processing to match the content[8], for example, in accordance with the grouping extracted, replace, delete, here not as the focus of presentation.

After these two steps, the rest of the content of <ul> label structure contains is the information block to be extracted, while structures within <a> labels have been extracted, will not cause interference on subsequent information extraction. Like the examples, the structure of the web is relatively simple, the label is not complicated, could directly match the need to collect information block data during the cleaning of the page and combine with subsequent information extraction work[9]. This will be more help to improve the efficiency of the entire acquisition system.

4. Algorithm evaluation

Selecting from Xie Cheng network, Tu Niu network and Qu Na network to each of a random sample of 276, 324, 372 (total of 972 pages, the page size of the sum is 21546KB) travel pages,for evaluation of the analytical results of the regular expression method algorithm. The evaluation work is divided into two: speed and accuracy.

Using regular expressions on the 972 page batch parsing shared about 19 seconds, the average 54 parsing web pages per second, an average speed of 1197KB/S. But web cleaning systems use the traditional method of parsing speed to 889KB/S. Obviously, regex algorithm is a direct match exactly ideas, stronger than the web cleaning parse page algorithm running speed.

Randomly selecting 254 pages from the 972 pages above,artificial rated by their algorithm to parse the regex quality. In total, divided into three levels:

1) Good(Page records of all tourism related topics are accurately extract and the important information of each travel record is accurately extracted).

2) Fair(Overall acceptable, but there is a small amount of errors of general errors).

3) Poor (Serious error or errors are more difficult to accept).

The result, a score of "good", "ordinary", "poor" pages proportions were 77%, 21% and 2%. That is, the accuracy of the analytical site of this algorithm on the whole 98% is acceptable and unacceptable is only 2%. The reason why there is such a high accuracy because the algorithm on the handle of the object under the premise of the master page structure, high matching success rate. But still 2% unacceptable results, it is because the web structure or different, for example, the travel web pages records of Tu Niu network sometimes are pictures, sometimes are no pictures, this brought a certain degree of difficulty to the algorithm written, so the algorithm may further improve.

5. Conclusion

This paper analyzes the lack of traditional page cleaning algorithm, and proposes a cleaning method of page extraction strategy based on regular expressions to match. This approach cleverly avoides some defects of the traditional page cleaning algorithm. It is universal, could targeted clean and filter the pages with special structures and contents. This personalized, targeted design thinking can not only reduce the difficulty of the current page cleaning algorithm, while improving the efficiency of the cleaning of the page, and can adapt to the current web information collection technology trends based on the theme, direction and personalization.

In today's society, with the rapid development of information technology and mature, people’s psychological needs for access to information has become more sophisticated. It will be a huge challenge that how to make web information collection technology adapt to the situation, to be better and more powerful, to meet the individual needs of people access to information, to provide greater convenience to people's lives and help for researchers in the field of information. Therefore, there is a range of issues to be further in-depth study.

6. Acknowledgment

This work is supported by Scientific Research Project of Beijing Municipal Commission of Education (Grant No. KM201210005030), the support is gratefully acknowledged.

7. References

[1] Bergman M K. The Deep web:surfacing Hidden Value[J]. Journal of Electronic Publishing, 2001, 7(1): 1174-1175.

[2] Menczer F, Pant G, SriniVasan P. Topic Web crawlers: Evaluation Adapti Vealgorithms[J]. ACM Trans On Internet Technologies, 2004, 4(4): 378-419.

[3] M. YuVarani, N. Ch. S. N. lyengar, A. Kannan. Lscrawlef: A Framework for an Enhanced Focused Web Crawler Based on Link Semantics[A]. 2006 IEEE/ WIC/ ACM International Conference on Web Intelligence, 2006, 794-800.

[4] Yuanyuan Zhou. Research and Implementation of a Web Page Cleaning Technology[J]. Computer Engineering, 2002, (9): 48-50.

[5] Bin Xia, Jun Gao, Tengjiao Wang, Dongqing Yang. An Efficient Dynamic Script Website Page for the Method Effectively[J]. Journal of Software, 2009, (20): 176-183.

[6] Chen Cheng, Kaiyue Qi, Jianbo Chen. Web2.0-Based Search Engine[J]. Computer Applications and Software, 2010, 27(1): 180-182.

[7] Chong Cheng. SDI Service System Research Report Based on the Java Platform Network Information Retrieval[Z]. Nanjing Agricultural University, 2004.

[8] Shasha Li. Research and Implementation of Incremental Web Information Retrieval and Information Extraction System[J]. Computer Science and Technology, 2011, (9).

[9] Zhijin Wang, Zhengbiao Han, Peng Zhou. Framework and Mechanism of Network Information Mobile Search[J]. China Index, 2011, (1).