Website information extraction based on DOM-model

YaFang Lou[1,a] YiChong Zhang[3,c]
[1,3]Department of Computer Science and Technology
ZhuHai College Of JiLin University , 519041
ZhuHai,China
[a]luckylou@sohu.com

ZhiJun Yuan[2,b]
Pillar Resource Services Inc 4155-84th Ave ,T6B 2Z3
Edmonton, Alberta, Canada
[b]cyy_zjy@yahoo.ca

***Abstract*—With the rapid development of network technology and the promotion of application, web has become the main platform of the issuing and accessing information. It is current research focus, how to obtain the information required by the user from the vast information source. This paper presents an extraction method of website information based on DOM to improve the searching efficiency, which only to preserve the theme information and to filter out the noise information that the users are not interested in.**

Keywords- DOM; Information Extraction; Web Introduction

I. INTRODUCTION

With the rapid development of network technology and the promotion of application, web has become the main platform of the issuing and accessing information. The user will generally find two sections when browsing the webs: the one is the theme page such as the news in a news page; another is the navigation bar, advertising information, copyright information unrelated to the subject content and so on, which we call them as noise information. It has been focused in current, how to find quickly and accurately the theme information from the enormous information resources.

Web information mainly exists in semi-structured HTML documents. It increases the difficulty to extract the information which users are interested in, because there often exist the omitted and non-standardization wording in HTML. Most of the traditional methods is to increase the size of the information being interested, directly to delete the information not-being- interested and to disable JavaScript and so on. These ways have web pages lost their original appearance. The method in this paper is the first to modify the syntax error page with the HTML parser to change the web page into the well-formed HTML document and then to parse the HTML documents into the DOM Tree. DOM (Document Object Model)[1], is the application programming interface of HTML and XML. The web page is parsed into a DOM tree and each node of the tree is an object. DOM model not only describes the structure of the document but also defines the behavior of the node object. With the methods and properties of objects, you can easily access, modify, add, and delete the nodes and contents of DOM Tree. Page being parsed into a DOM tree, it is great convenience to extract website information.

II. THE RELATED TECHNOLOGIES

There are many kinds of descriptions about the concept of information extraction. In 1997,Grishman, the creator of Proteus project[2], described the concept of information extraction as: the information extraction involving to creating a structured representation for the information selected from the text. Much of the information in current internet is taken from the back-end databases and then generated web pages according to a certain pattern. The multiple records are contained in a page. These semi-structured pages can provide a lot of information. But the data only is changed into local one after being extracted and then convenience to be analyzed deeply by computer. But these pages that is including a lot of noise information, flexible structure, hidden data mode information, and no strict constraints to the data are difficult to process generally with the ordinary way. Information extraction can not only help people easily to find the information, but also to access effectively the information interested after the content of the information being analyzed and organized reasonably. People can further data mining, text generation, and the follow-up information processing based-on it.

A. Web information extraction

Web information extraction is to analyze the content and structure of the original document information and to extract the information that the users cared. Its core is to extract the hidden information from the semi-structured HTML pages dispersed in internet and to form a more structured and clearer semantics the representation. The user facilitates to directly inquire and apply to date in web.

Three characteristics of web page information independent extraction:

1) The user can customize the information is that the user can independently extract information according to their requirement.

2) The path expressions in DOM Tree structure is used to locate the information to be extracted in HTML

3) Take a self-learning method to adapt to the dynamic changes of the web page information

B. DOM technology

DOM (Document Object Model) is that W3C provides a standard to establish the tree structure of the XML

Published by Atlantis Press, Paris, France.

document in memory[3]. The element in XML documents can be expressed as a node in the DOM tree structure and each node of the tree is a object. The DOM model is not only to describe the structure of the document but also to find the specific page elements from the web page. It can change the content and attributes of the elements and control the behavior of the elements. With defining the object behavior, you can easily access, modify, add, and delete the nodes and content of the DOM tree by the methods and properties of the objects. DOM has the characteristics to cross platform and to adapt to the different programming languages. HTML documents can also be described with the DOM.

DOM is processing based on the single page. You can remove the noise and extract the topic information in the page with some heuristic rules, depending on the DOM structure of the processing page and the visual information.

Treated with DOM model has several advantages:

- Being the tree long lasting in memory, you can modify any node so that the application programmer can change the data and structure.
- You can navigate in the trees up and down at any time. It is very simple to use. in the tree navigation and simple to use; Programmers can easily create documents and navigate their structures.
- DOM standard appearing, it greatly simplifies the processing of the structured document in the programming environment.

HTML source code as the follows:

```
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<title></title>
</head>
<body>
<table width="100%">
<tr><p>paragraph</p>
</tr>
<tr>
<a href="1..."> </a>
<a href="2..."> </a>
<a href="3..."> </a>
</tr>
</table>
</body>
</html>
```

HTML source code corresponding to DOM tree as shown in Figure 1[4]:

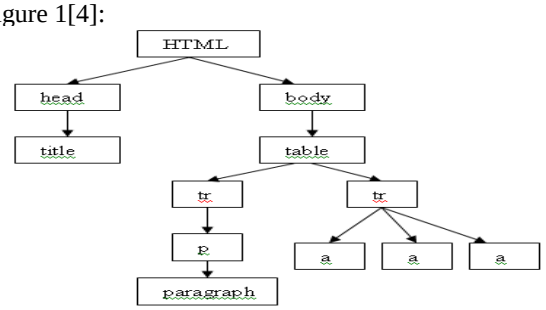

Figure 1. the HTML source code corresponding to DOM tree

C. *The ideology and algorithm of DOM information extraction*

1) *The ideology of information extraction*

The basic idea of the DOM information extraction: the irregular HTML documents are organized into well-formatted XHTML documents and then XHTML document are parsed into the tree model - DOM tree. Then you can extract the information and search the similar structure around the DOM tree. The results extracted are expressed with XML document form and are stored structured. The four steps can be described as the follows:

- Accessing to the Web and organizing:

Two cases will be encountered to search for web pages by site link. The one is the pages containing the data required. Another is the hyperlinks pages to the target page that contains the data required. The navigation rules of web site could be edited, combined the careful analysis of the target site with the characteristics of the target site.

Organizing is that the data source is mapped into XHTML. Organizing mainly contains three aspects as the following: ①put the terminator "/" for the unpaired mark ② mark the quotation for all property ③ change all "\" into "/." in the URL

- The data source parsing

The XHTML document converted is constructed DOM tree and all the elements of the document are mapped into nodes of DOM tree. The parsing process is as follows: the first is to find the start tag of the web page, and its name is stored in the mark table. Identify each mark successively and check if its start tag is corresponding to end or comment tag. Remove the tag if there is no corresponding end or comment tags; Otherwise, the content between the end mark and start end will be stored in the mark table. This content is a leaf node. Repeat the operation until to complete processing each mark in pages. The marks and their contents are constituted a table. The whole tree is stored in the table, decomposed by n sub-trees.

- Web clustering depended upon the similarity

In this paper, the DOM judge if the web pages collected are similar to the sample structure and determine whether to extract the information from pages collected by the existing model. In a collection of web pages, the web page with the same and similarity degree can be looked as the pages generated by the same template. That is that this set of pages has the similar DOM tree structure. This collection of pages can be divided into k classes. The template of each class will be extracted in the next step.

- The extraction of the same class template

The template is the common DOM tree of a type of pages. That is the intersection of all the DOM tree. The page template is got by comparing the HTML parsing tree of two pages with similar structure[5].

The extraction process of website information based on DOM is shown in Figure 2:

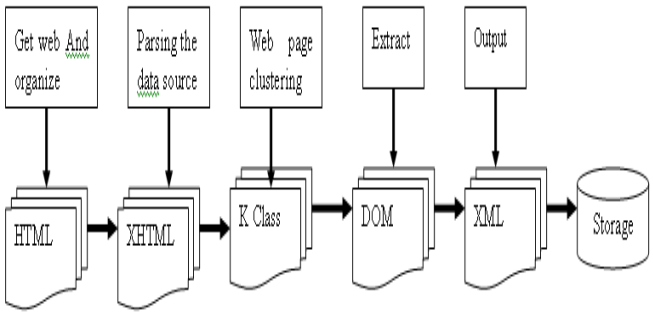

Figure 2 . flow chart of DOM-based website information extraction

2) Information extraction algorithm

Extract information in two steps: Generate extraction mode, and then extract information with it. Generating extraction mode is divided into three steps: Summarize in a single sample web information block to position path; Sum up the information block of sample pages collection to position path. Position the information point path within the information block.

- Position path, summarized in a single sample web information block

Sample pages are blocked with similar structure, according to the structure characteristics of the sample pages provided by user

Single sample algorithm as the follows:

IBPATHi = NULL;
Pre-order traversal parse tree DOMi;
The resulting path expressions written in treePath;
Sequentially scan treePath;
while (treePath not ending) {
Comparing the corresponding path nodes in two paths;
if (the index value of the two paths and the child node are the same) {
Write the path in IBPATHi;
Compare the next set of path expressions;
}
else (the index value of nodes are same, but the index value of the child node is different) {
Intercept the node in the path expression and the path before the node, write the path in IBPATHi[6];
Enter the next set path to compare;
}
}
return IBPATHi;

- Positioning path, summarized the sample pages collection information block

The algorithms are described as follows:

LocationIBs = null;
for (i = 1; i <= m; i + +) {
Path [i] = null;
LocationIB [i] = null;
}
for (i = 1; i <= m; i + +)
for (j = 1; j <= n; j + +) {
Scan the DOM tree of the j-th sample pages;
Write the path expressions of i-th content in j-th sample page in path [i], path [i] = path [i] + {path [i] [j]};
}
for (i = 1; i <= m; i + +) {
while (path [i]! = null) {
Randomly extract a path [i] [j]=apath;
Compare apath and other path expressions in path [i] and obtain the positive examples collection S covered by apath
path [i] = path [i]-S ;/ / delete the positive examples covered
LocationIB [i] = LocationIB [i] + apath;
}
}
LocationIBs = {LocationIB [1], LocationIB [2], ….., LocationIB [m]};
return LocationIBs;
}

- The information point positing path in the information block

After determine the positioning path of the sample collection information block, we can position the specific path of information point by preorder traversing within a information block, expressing it with Xpath.

With the induction- learning to obtain XPath, we edit XSLT document. It can generate an XML document according to the document conversion DOM node. Only nodes specified by XPath are retained in the XML document, thus completing the information extraction.

III. TECHNOLOGY IMPLEMENTING

The people mainly uses the following three indexes to evaluate the information extraction technology: Recall, precision and F value. Recall rate is to measure the proportion of information extracted correctly, while the precision is to measure the proportion of correct information in the extracted. Calculated as follows (P precision, R is the recall rate)[7]:

$$P = \frac{\text{the number of correct information extracted}}{\text{The number of information extracted}}$$

$$R = \frac{\text{The number of correct information extracted}}{\text{The number of information in the samples}}$$

Both values are between 0 and 1, the closer to 1 the value is, the higher the recall or precision is.

Here is the weighted geometric mean of Recall rate and Precision rate. F value evaluation method:

$$F = \frac{(b^2 + 1)PR}{b^2P + R}$$

b, the relative weighting of the P and R, is a preset value. That b is greater than 1 indicates that P is more important. That b is less than 1 means that r is more important. Normally, it is preset to be 1, indicating that both are equally important. By F value, we can know that the system is good or bad. The closer to 1 the F value is, the better it is [8].

We randomly selected 20 web page samples from Sohu http://www.sohu.com/and 15 pages from NetEase http://www.163.com/ for testing, website 15 pages sample we randomly selected for testing. The test results are shown in Table 1:

Table 1 System Test Results

Website Address	Number of sample pages	R%	P%	F%
www.sohu.com	20	96.1	91.1	93.5
www.163.com	15	93.7	88.7	91.1

Seen from Table 1, the information extraction methods based on DOM can obtain higher recall and precision.

CONCLUSION

The subject content extraction of pages has been essential for Web information pretreatment link. It can reduce the browsing time, promote the speed of user accessing to information, improve efficiency and enhance the usability of the web, with extracting the topic information by DOM model. Therefore, the research about the theme information extraction based on DOM model in website is certain practical significance.

REFERENCES

[1] A. Arasu , H . Garcia-Molina, Extracting structured data from web pages, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2003: 480- 499.

[2] Suhit Gupta, Gail E. Kaiser,Peter Grimm, Michael F. Chiang,Justin Starren. Automating Content Extraction of HTML Documents[J]. Kluwer Academic Publishers, 2004 : 12.

[3] CaiD eng, Yu Sh ipeng, W en J irong, M aW eiy ing. VIPS: a V ision-based Pages Segm entation Algorithm [R]. M icrosoft T echn ical Report MSR-TR-2003-79, Novem ber, 2003.

[4] H an W, Buttler D, Pu C. W rapping W eb Data into XML [J]. S IGMOD Record, 2001, 30 (3): 35-38.

[5] The Open Group. TOGAF Version 9: TheOpen Group Arch itecture Frame-w ork [M]. Ap r, 2009.

[6] IBM Corporation. Bu siness Systems Plann ing- In form ation System s P lanning [M]. New York: IBM Press, 1975.

[7] Dawn G.Gregg.Steven Walczak Adaptive web information extraction 2006(05)

[8]Valter Crescenzi.Giansalvatore Mecca Automatic Informarion Extraction from Large Websites 2004(05)

Published by Atlantis Press, Paris, France.