Web Information Extraction Algorithm based on Ontology and DOM Tree

Li LIU
School of Information Engineering
University of Science and Technology Beijing
Beijing,100083,China
liuli@ustb.edu.cn

Junfang SHI
School of Information Engineering
University of Science and Technology Beijing
Beijing,100083,China
Shijunfang111@126.com

Xinrui LIU
School of Applied Science
University of Science and Technology Beijing
Beijing,100083,China

Abstract*—Due to the information on the Web being tremendous, dynamic and irregular, it is difficult to search and integrate information from the Web. This paper proposes a Web information extraction algorithm based on Ontology and DOM tree. The areas are accurately found out and the interested information is extracted exactly by information extraction rules generated by ontology. Furthermore this algorithm implements information extraction through traveling DOM tree. Finally, we implement information extraction system and test its performance on news site. Testing result shows that this algorithm doesn't rely on the page structure and it can increase the recall and precision of information extraction.

Keywords-Information Extraction Algorithm; Ontology; DOM tree; Extraction rules

Ⅰ. INTRODUCTION

Due to the information on the Web being tremendous dynamic and irregular, it is difficult to search and integrate information from the Web. Web Information Extraction[1] translates the information expressed by various forms from the Web into united mode of information expression. It provides powerful tool for people to find information needed from tremendous information quickly and accurately. The main problems of current extraction method are as follows: (1) Accuracy and robustness of Information Extraction System need to be improved. (2) The programs of information extraction rely on the structure of Web pages, which makes programs can't be reused. In order to improve the accuracy of information extraction and to present information extraction method suitable for a variety of different pages, the paper proposes a web information extraction method based on ontology and DOM tree[2]. The process of information extraction is as follows: Firstly, establish the page structure ontology. Secondly, extraction rules are established in the guidance of Ontology[3] and are organized into DOM tree. Thirdly, the WebBrowser[4] and MSHTML[5]change HTML Web pages into XML DOM tree. Finally, combining ontology and DOM tree, information extraction algorithm is implemented through analyzing extraction rules and traveling DOM tree. Testing shows that the approach proposed doesn't rely on the page structure and it can increase the recall and precision of information extraction.

Ⅱ. THE EXTRACTION RULES BASED ON ONTOLOGY AND DOM TREE

A. Page Structure Ontology

The information extraction technology only analyses the parts containing relevant information in documents. As to what information is focused, it will be determined by the scope of the domain. Ontology is just the standard description to the sharable and universal conception in a certain domain.

The forms of formal definition of Ontology are as follows[6]:

Concept :PageA;
Super:{Column};
Type :String;
Value:{ContentA_XPath};
End ContentA

Content means information items need to be extracted in web pages. Page A means the pages types of content and ontology of page information will be established base on content. After analyzing HTML forms of a large number of web pages of each website, the Page Structure Ontology

978-1-4244-5392-4/10/$26.00 ©2010 IEEE

shown in Figure 1 was established under the guidance of domain experts.

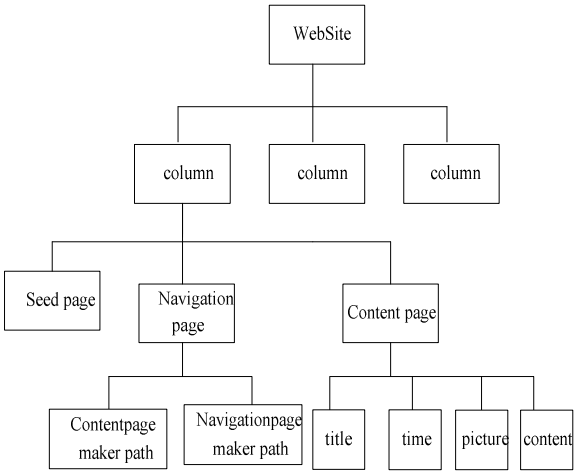

Figure1 Page Structure Ontology

B. Information Extraction Rules

The extraction rules are not the part of Ontology, but they should be corresponding with Ontology objects. Namely objects in Ontology may not appear in extraction rules, it means that there is no need to extract the sample data from the object, but the extraction objects involved in extraction rules should exist in Ontology, otherwise, the output results are not consistent with data structure defined in Ontology. The information extraction rules generated by Ontology are organized into a DOM tree. Figure2 shows the DOM tree of extraction rules.

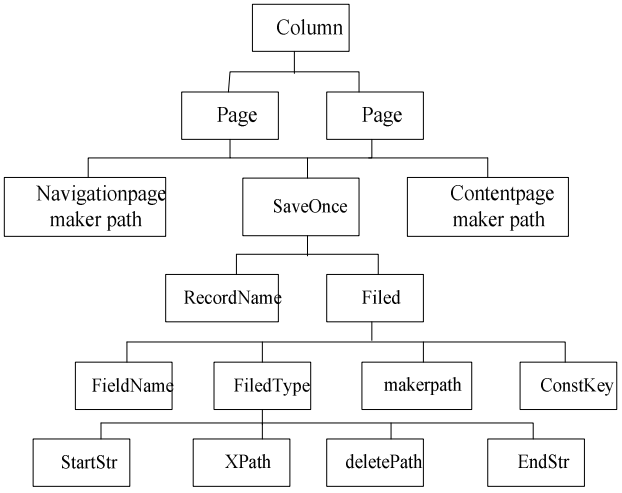

Figure2 DOM tree of Extraction Rules

The rules file use as the current node in DOM tree, and then we can use the attribute of selected nodes in the DOM to obtain the dealing page type and the Xpath[7] of extracted information. According to the Xpath value to travel the XML DOM,and then we can extract the information interested.

As a example of Sina-- http://roll.news.sina.com.cn,We define extraction rules as follows:

```
<xml>
<column> domestic news </column>
<pagetype="contentpage"
Schema="^http://news.sohu.com/\d+/n\d+\.shtml ">
<saveonce>
<field>
<fieldname>title </fieldname>
<fieldtype>
<Xpath></Xpath>
…
</fieldtype>
<makerpath>//DIV[@ id='contentA']//H1</makerpath>
</field>
<field>…</field>
</saveonce>
</page>
<page>…</page>
</xml>
```

Wherein, "//DIV[@ id='contentA']//H1" stands the XPath of the title, other fields like picture, time, content, etc, we can use the same method to express.

The attribute of content page schema is formed with regular expressions (regex), which match the URL of different page types. The regex in the column is: "^http://news.sina.com.cn/c/\d+-\d+-\d+/\d+\.shtml",where in:"\d" means matching number, "+ "stands for matching one or more times, the regex above match the type of URL like"http://news.sina.com.cn/c/2010-08-24/123120964132.shtml".

As the regular expression, it can match the processing pages accurately. If the pages can't match the regex, they will not be handled, so the precision of information extraction is improved.

Ⅲ. INFORMATION EXTRACTION ALGORITHM

The HTML documents analyzed can be converted to XML DOM. The extracted nodes in the DOM tree can be located by XPath, and information can be extracted through using XPathAPI. Figure 3 describes the flow chart of information extraction algorithm.

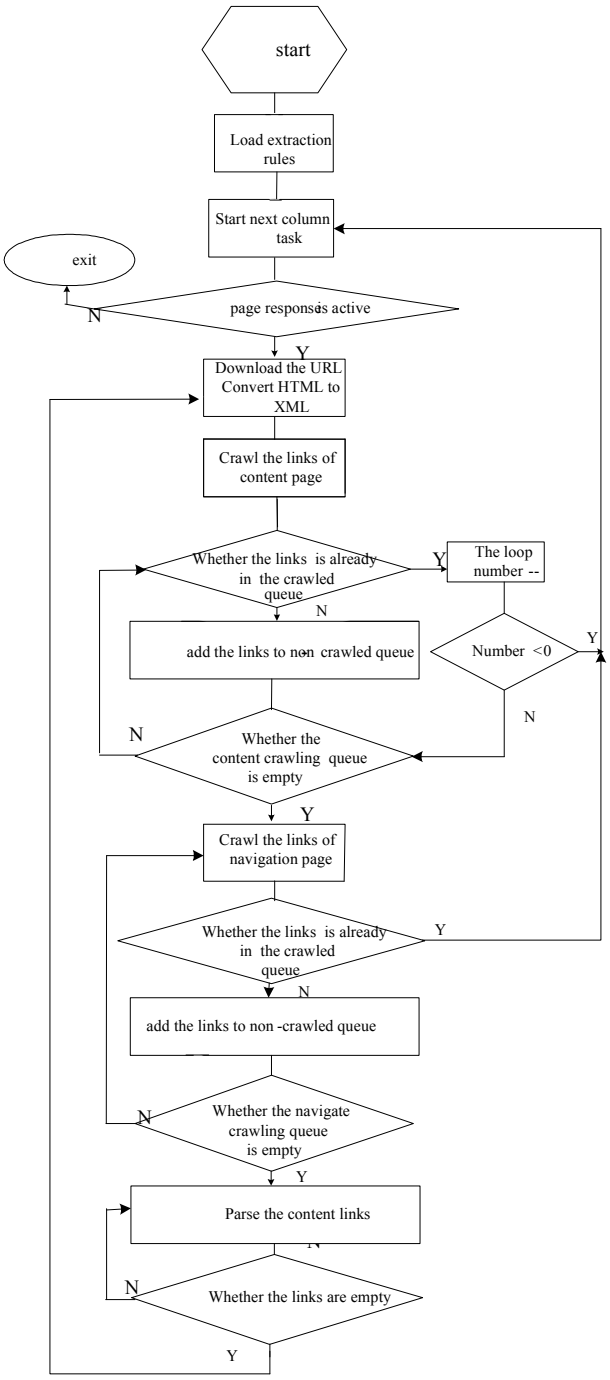

Figure3 the Flow chart of Information Extraction Algorithm

Our key algorithm for information extraction is as follows:

- The information extraction algorithm of navigation pages.

Download and convert the navigation page,and then extract the links of content page and add the links to non- crawled queue. Extract the links of navigation page and add the links to non-crawled queue.If the navigate crawling queue is empty ,return to content page algorithm, else continue to crawl the navigate page links.

- The information extraction algorithm of content pages.

The information extraction algorithm of content pages.

Download and convert the content page then extract the XPath value（such as content, time ,title and picture etc）. from the configuration rules file. Then use XML DOM to extract and store the information into XML Document. Judge whether the queue of content page being crawled is empty or not, if it is empty then goto navigate page Algorithm, else continue to parse the content inks.

Usually, in order to judge whether the news has been crawled or not we need to know whether its headlines, release time, contents and other fields have been extracted. This paper presents a method which only needs to compare the page URL to infer whether the page has been processed or not. The illation algorithm eliminates the crawled content page and the influence of page noise on the page information processing effectively , which also improves the efficiency of the program greatly.

Ⅳ. ARCHITECTURE OF INFORMATION EXTRACTION SYSTEM

Information extraction converts a group of HTML pages into structured form. Figure 4 shows the entire extraction process.

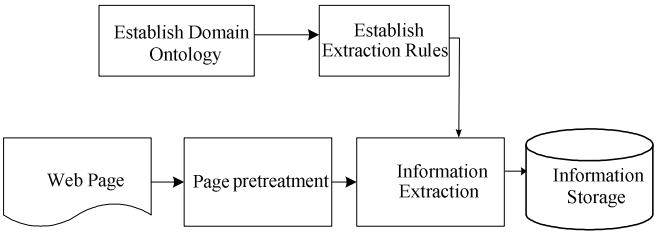

Figure4 Architecture of Information Extraction System

The process of information extraction is as follows:

Establish an page structure Ontology to sample page which users are interested. Then define the web information extraction rules. Download web pages and convert them to XML document. Thirdly, analyse

information extraction rules and implement the information algorithm to finish the process of information extraction. Store the extracted information in XML format.

V. THE ALGORITHM TESTING

Information extraction system is usually evaluated from the three measures of extraction rate, recall, precision. The experimental environment of this system are: Athlon 2.8GHz CPU,2.0Gmemory,Windows7 Operating system and C# Programming language.

This system test the domestic news column of two websites, such as Sina and China in one day, and also test the recall, precision and the average extraction rate of the system, wherein:

Precision is the ratio of the news extracted exactly to the news extracted. Recall is the ratio of the news extracted exactly to the news should be extracted. The average extraction rate(ARE) is the ratio of the total information extraction time(s) to the number of actually extracted news. Figure 5 shows the algorithm testing results. Wherein, the number of actually extracted news, for sina, that is 260,for China ,that is 99, the number of exactly extracted news, for Sina, that is 253,for China that is 94.

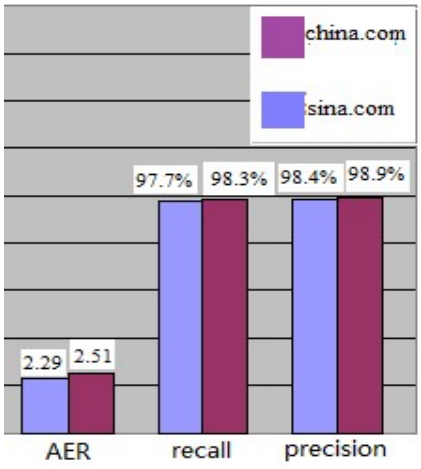

Figure5 Result of Algorithm Testing

The information extraction system has an efficient recall and precision. For the page processing rate, the time of downloading picture is longer than the time of analyzing content, so this will decrease the extraction rate, but in general the extraction rate of this system is satisfactory.

VI. CONCLUSION

Basing on the ontology of web page and the DOM tree technology, the proposed information extraction algorithm has adaptability. It can adapt the page format characteristics changed by reforming the rules files instead of changing programs. Moreover it has popularity, which can deal with web pages of different domain. It also improves the recall and precision of information extraction. But as to the slow response of pages, how to improve the extraction rate is our further research.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under grant N0.60873193 and No.60873192.

REFERENCES

[1] Silvescu A, Reinoso Castillo J, Honavar V. Ontology - driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources[C /OL]. In: Proccedings of the LJCAI2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources, 2001.

[2] Jing Yang, Wenzhu Yang,Yue Gao .Rules Construction and Implementation in DOM-based Web Information Extraction[J]. Journal of Hebei University,2007,2(27):209-212

[3] David W.Embley.Ontology-based extraction and structring of information from Data-rich unstructured document[EB/OL]. http://pages.cs.wisc.edu/~smithr/pubs/cikm98.pdf, 2008-06-12

[4] Tim Anderson. Working with the WebBrowser[J]. Personal Computer World,2006,29(5).

[5] Shujin Lv.Extract Data from a Web Page by using MSHTML Component[J]. Journal of Baoding teachers college,2004, 17(4):15-17

[6] Mingjian Zhou, Ji Gao,Fei Li.Ontology-Based Information Extraction from Web Sources[J].Journal of computer-aided design&computer graphics, 2004 , 16(4) : 535 – 541

[7] Wenfei Fan,Geerts F,Xibei Jia.Rewriting Regular XPath Queries on XML Views.Data Engineering,2007,ICDE 2007,IEEE 23rd International Conference on, Istanbul,2007:666 – 675