XML-based Web Information Extraction System Design and Implementation

Ma Jun, Li Tihong
Information Engineering College, Jiaozuo University
2230600@sina.com.cn

Abstract-based on the research of the existing Web information extraction techniques, this paper proposes a XML-based Web information extraction system design. This design can extract the interested information points from HTML pages and express the extracted results by using structured XML with strong scalability. The system has certain versatility and flexibility, users can quickly customize for the Web information extraction wrapper to be used in different fields.

Keywords-Web information extraction; XML; extraction rules; XSLT

I. INTRODUCTION

With the rapid development of Internet in recent years, Web has become a huge platform to distribute and share information resources. But how to fast and effectively access to information from the Web is still a problem troubled to Web users. Although at present there is a lot of research on the Web information extraction technology, but existing technology does not meet user requirements. XML provides a consistent data model and description language for the Web, has become the facto standard of the diversity of the Web.

II. XML AND WEB INFORMATION EXTRACTION TECHNOLOGY

A. XML technology

XML is a meta-markup language, can provide the format of describing structured data to allow different applications to read these structured data, to separate the data and the user interface. This will help document content statement, and in line with cross-platform search operation. In addition, XML will also be the key technology of new generation of network data display and running.

XML has a good format, authentication mechanism, rich display style, convenient data processing, and object-oriented features, is a universal electronic data exchange format.

B. Web information extraction technology

Information extraction is to carry out structural processing for the information the text contains, to form the organization with same form. The input of the information extraction system is the original text, and the output is information point with fixed format. Information points is extracted from a variety of documents and then integrated with a unified form. This is the main task of information extraction.

Web information extraction is to carry out extraction with the Web as a class of information source, there is such a document library in the Internet. Online, the information with the same topics is usually stored in different sites and has different performance forms. If this information can be collected together and stored with the structured form, it would be useful. As the major information carrier in the Internet is text, so, information extraction technology is essential for the people who take the Internet as a source of knowledge. Information extraction system can be seen as a system to transfer the information from different documents into database records. Therefore, the successful information extraction systems will regard the Internet as a huge database.

Extraction technology is continuously enriched with the increasing of demand, in recent years there were a variety of information extraction methods have emerged, and according to the difference of the extraction principle and extraction methods, which are divided into the following five types: natural language-based approach, wrapper-based induction method, ontology-based approach, method based on HTML structure and the Web-based query. These methods have focused solution to Web information extraction problems, in general achieved good results.

C. XML application in Web information extraction

With the development of XML technology and the need for information extraction, XML-based Web information extraction technology is increasingly subject to attention from information retrieval, data mining, database and other field researchers, the specific research and application is mainly reflected in the following areas:

1) XML-based data acquisition, mainly to study how to get the data source users needed.
2) XML-based data conversion, mainly to study how to convert the document with HTML format into the corresponding document with XML format.
3) XML-based data extraction, mainly to study the data extraction techniques based on XML family.
4) XML-based data storage, mainly to study the methods of sorting and storing the extracted results obtained. XML can be viewed as a semi structured data model, the document description of which can easily be related to the relational database attributes for storage to implement the precise query and model extraction.

XML-based next-generation WWW environment is directly facing Web data, which can not only be well compatible with existing Web applications, but also can better realize the Web information sharing and exchange.
III. XML-BASED WEB INFORMATION EXTRACTION SYSTEM

A. System whole structure

The main task of Web information extraction is to verify the information users interested from the unstructured or semi-structured information the Web contained and convert it into the data with strong structure and clear semantics. The input of information extraction system is original text and the output is information with fixed format. Finally, the data extracted will be stored in the relational database after cleaning and sorting for further accurate data inquiries and pattern extraction.

The whole approach of Web information extraction includes the following five stages:
1) Data acquisition: is to download the HTML page sample of a specific area from the Internet, to store in the domain knowledge base after classification.
2) Page optimization: is to optimize the HTML pages in the domain knowledge base, to convert the non-standard HTML documents into structured XHTML documents according with XML standards, and use XML parser to parse the XHTML document obtained into XML DOM tree.
3) Rule learning: take the page after page optimization as the training sample, and the user will mark the interested information point to obtain the XPath path expression of the node to be extracted, and combined with the XSLT technology to compile the information extraction rules and store the different extraction rules in the extraction rule base, and the rule learning process can be combined with optimization methods to optimize the rules.
4) Information extraction: when users request page information extraction, they will use the extraction rules in the library to extract the HTML page, if there is no corresponding extraction rule in the rule base, then the structure of such HTML page should be re-learned to get new extraction rules and store in the rule base. The extracted results are expressed with XML format.
5) Data storage: to store the extracted data in XML format into a relational database.

The overall structure of the Web information extraction system is shown in figure 1.

The work flow of this Web information extraction system us shown in figure 2.

B. XML-based Web information extraction system implementation

1) Page optimization

In the extraction rule learning phase JTree tree is needed to be used to display the source document DOM tree structure, and the document loaded into the JTree tree must conform to XML standards, therefore, it is necessary to optimize HTML sample page. Page optimization consists of two steps: page cleaning and page analysis. Page cleaning is mainly to clean up the HTML page illegal characters, non-standard and error nested tags, remove the non-subject elements to achieve the conversion from HTML documents to XHTML documents, in this process HTML parser is used, first of all the HTML documents will be parsed to HTMLDOM tree, and then through traverse of the DOM tree the HTML document is cleaned up, finally building XHTML documents met the XML standard; page analysis is mainly to use XML parser to parse the XHTML document obtained after page washing into XMLDOM tree structure, so that we can carry out rule extraction learning to this tree.

2) Learning extraction rules

The core task to structure a wrapper is to compile extraction rules and the process of extraction rule learning is the process of extraction rules generation. The information extraction rules in this paper use XSLT language to describe, and the main component of the XSLT is XPath path expression which is used to locate information points to be extracted in the XML document. In rule learning, JTree is used to display the XMLDOM tree formed after page optimization, and the user is to mark the interested information and through a DOM-based XPath generation algorithm this paper presented to obtain the marked node.
XPath expression, and combined with XSLT technologies to compile the extraction rules.

In which, XPath is a major component of XSLT, XSLT is the extraction rule, and to generate extraction rule is the core task of structuring an information extraction wrapper. Using the standard XSLT as information extraction rules, making extraction rules are easy to understand and modify, at the same time, XSLT has powerful and flexible syntax structure to allow us to more easily manipulate the data in the document. The main component XPath of XSLT is used to locate the document node, and the single information block extraction mainly uses XPath positioning method based on tree path, which not only can accurately locate information, but also can automatically access the XPath expression users have marked nodes, this is on the basis of the DOM-based XPath generation algorithm this paper presented, thus users can easily learn rules and define the wrapper. When extracting multi information block, we can construct extraction rules through merging template after obtaining the XPath expression of the nodes to be extracted. The obtained extraction rules can be optimized by using the above described positioning optimization. This shows that the use of XSLT and XPath powerful features can build a flexible and universal extraction rules.

3) Information extraction

XSLT style file defines the transformation rules from the source XML data to target XML data. XSLT itself is a markup language consistent with XML standard, which does not depend on any programming language, and plays XML meta-language features, is an application of XML. In the implementation of the XSLT transformation process, the first XML source document tree will be traversed to find the node to be dealt with, and then the defined template information will be added in the node, and the nodes in each XMLDOM tree will be compared with a model, if the two matched then to convert according to the rules the template defined, else continue down the match. This cycle until the entire document is completely processed.

Using XSLT processor and according to extraction rules to convert the XHTML document obtained after cleaning the HTML source, then the target XML documents can be obtained.

Through information extraction, we can get the XML document that contains useful data, the final extraction results are expressed with XML document format, because as the main form of data exchange which can facilitate the data to be easily stored in a relational database, for the same class of pages with similar structure, we use semi-automatic wrapper induction method based on sample learning to generate extraction rules and store the extraction rules after optimization in rule base, when you need to extract information from the same page, you can select the appropriate rules from the rule base to match with, if there is no matched rules in it, we can through rule learning and compare with the original rules, and then modify the rules of such class page structure to summarize more adaptable extraction rules.

C. System performance testing

Information extraction system evaluation mainly uses two important indicators of MUC conference: precision P and recall R, however, for comprehensive evaluation of system performance, the weighted geometric means F index of the P and R also often used, the formula is:

\[ F = \frac{(1 + F_1^2)PR}{F_1^2P + R} \] (1)

In which, \( F_1 \) is the relative weight of the precision and recall. If \( F_1 \) is equal to 1 the both is equally important; greater than 1, accuracy is more important; less than 1, the recall is more important. Usually take a value of 1. To take 200 HTML pages to test, in which the 120 pages for single information block extraction, 80 pages for multi information block extraction, the test data as shown in table 1.

<table>
<thead>
<tr>
<th>Information block extraction type</th>
<th>Number of data to be extracted</th>
<th>Number of extracted data</th>
<th>Number of data correctly extracted</th>
<th>P (%)</th>
<th>R (%)</th>
<th>( F(\frac{F_1}{1}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single information block extraction</td>
<td>120</td>
<td>110</td>
<td>100</td>
<td>90.9</td>
<td>83.3</td>
<td>0.869</td>
</tr>
<tr>
<td>Multi information block extraction</td>
<td>340</td>
<td>322</td>
<td>280</td>
<td>87.0</td>
<td>82.4</td>
<td>0.846</td>
</tr>
</tbody>
</table>

That can be seen, for a single information block extraction, the system can achieve better results and implement the automatic generation of extraction rules, to some extent reduce the human involvement. For multi information block extraction, user can easily use Web information extraction platform to help build rule template, and generate multi-block extraction rules through combining with the XPath path expression of the marked nodes. Overall, this system can achieve relatively good extraction results, in addition, this system has better portability and scalability, has some versatility and flexibility, users can quickly customize the Web information extraction wrapper applied in different fields.

IV. Conclusion

With the development of the Internet and the rapid surge in the amount of network information, Web information extraction technology is increasingly important. XML technology provides important support for the Web information extraction and other applications, while its semantic mark can support more precise query. This paper designs and implements a XML-based Web information extraction system, and the system performance tests show that the system can achieve relatively good extraction effect.
In addition, this system has better portability and scalability, has some versatility and flexibility, and thus is a system with practical significance.

REFERENCES


