Technology of Web Information Extraction

Muqing Zhan

Modern Education Technology Center, Jingdezhen Ceramic Institute, Jingdezhen, 333403, China

294056587@qq.com

Abstract. How to fast, effective and economic retrieve the required information from internet has become a current hot topic. The technology of web information extraction is an effective technology to solve this hot topic. This article analyzes the concept and principle of information extraction technology, and sets forth the working process and its classification of existing information extraction technology and the evaluation index of web information extraction.

Keywords: information extraction; wrapper; abstract regulation; accuracy rate.

1. The Concept of Web Information extraction

The Web information extraction is to study how to find the useful data for the user from the data of semi-structured HTML web which spread on the Internet of Web page, and transform it into a more structured and clear semantic expression form (XML, relation data, object-oriented data, etc.).It provides convenience for user to search information and the application program directly use the data in Web.

Extracting information from Web is mainly completed by wrapper. The so-called wrapper in fact is a software program which consists of a series of information extraction regulations which is already formulated and the program for using these regulations. As to the query request of specific information source of users, it is necessary to find and extract relative and useful data from information sources of Web page, and transform it to a data described in prescribed format and then return to users.

2. The principle of Web information extraction

2.1 Additional Semantics

Additional semantics is connecting the defined semantics and the extracted information. It means that according to users' demand, marking the information block which is corresponded with their own demands by dragging the mouse and then defining the semantics on checking internet. It because that the extracted information can be understood well by machine during processing other data.

2.2 The generation of extraction regulation

The core of Web information extraction is the generating of extraction regulation. It describes various characters and regulations of the information which is up for extraction in order to use it to identify and locate the information which is up for extraction on Web to determine what information need to be extracted. Firstly, locate the information block on internet. Secondly, understand the semantic information reflected from Web pages through the study of sample. Finally, build a reciprocal relationship between the semantic items and the contents of the Web page. This article named this kind of relationship as "extraction regulation". This is the generating of extraction regulation.

2.3 Information Extraction

After the generating of extraction regulation, extract relative information with this regulation.

3. The classification of Web information extraction technology

There are many ways for classifying web information extraction technology. It can be classified into three categories like manual way, semi-automatic and full automation according to the different degrees of automation. According to the different principles adopted by all kinds of information extraction tools, the existing information extraction can be classified into six categories: the way basis on regular expression, the way basis on natural language processing, the way of wrapper induction, the way basis on ontology method, the way basis on HTML's result and the way basis on Web query.

3.1 Information extraction which is basis on regular expression.

The regular expression is actually a formula which matches the string with certain pattern. In the process of all the strings handling, regular expression is a strong and powerful tools for pattern matching of complex string. The handling process of this technology is firstly to treat the Web documents as character stream and handle, then to match the information which is extracted through creating reasonable regular expression, afterwards, to extract information.

The advantage and disadvantage of extracting information by use of regular expression is more obvious. If the information up for extract is with known feature, the accuracy rate of extraction with this technology is so high. But if the information need to extract is with unknown feature, there will be no function to extract with regular expression. Because the information is without feature, it is unable to compile regular expression. Another advantage is that current mainstream programming languages possess the engine of this technology, provide strong support to the regular expression, and it is contributed to establishment of Web extraction system and realize the programming of wrapper. The corresponding disadvantage is that it requires a high technological level of author. Because the complexity degree of regular expression is directly proportional to the quality, complicated regular expression requires the author with superb skills. If the regular expression is in good design, the accuracy rate of character string pattern matching is high. At the same time, the engine running time of regular expression is long, otherwise is opposite.

3.2 Information extraction basis on natural language understating method

This kind of extraction method is to build through the word and sentence structure, paragraph, and the relationship between words and sentences. This method is usually used for the information extraction of free text. The method does not distinguish the Web document and common text, but to treat the Web document as text to handle. The disadvantage is that it needs lot of sample studying. WHISK, RAPLER and SRV system is typical example which is basis on this principle. The distinction among the three systems is that WHISK system can extract more records, but RAPIER and SRV only can extract single record.

3.3 Information extraction basis on the machine learning method

The extraction regulation of this method is basis on delimiter, and the generating of regulation is realized by automatic learning sample examples marked by the users. The main principle of this extraction technology is through delimiter to treat the extracted information and locate. Delimiter means to locate semantic items according to the left and right boundaries of semantic items, namely to describe the context of semantic items which is expected to get by users. WIEN, SOFTMEALY and STALKER system is typical example is basis on this principle.

3.4 Information extraction basis on Ontology method

This information extraction technology is mainly depending on a complete repository. All kinds of defined element extraction pattern and the relations between them are all stored in the repository. It is realized by the description information of data itself when extracting information. So this method does not affect the web page's structure and form. UIXOTEQ and BYU system is typical example of this principle.

3.5 Information extraction basis on HTML structure

This information extraction firstly uses the parser to parse Web documents into the syntax tree, and create the extracting regulation by use of automatic or semi-automatic method. The characteristic of this technology is to locate information by HTML and realize the information extraction by operating the syntax tree. LIXTO, XWRAP, W4F and ANDES system is the typical example of this principle.

3.6 Information extraction basis on Web query

This information extraction method is to search the Web pages by using standard Web query language to get information data and it is universal. The dominant idea of this method is to manage and search internet data by database technology to get required information. This kind of query is more accurate. PQAGENT and Web-OQL system is the typical example of this principle.

4. The working progress of Web information extraction

It is difficult to extract data of Web pages directly and need to process data because the information on Web is unstructured, semi-structured, and dynamic and easily to be confused. The processing procedure is word processing, feature extraction, structure analysis, text extraction, text categorization, text clustering and correlation analysis, etc. of the content gathered by Web document. The extraction procedure of a mature information extraction system includes the following four steps:

4.1 Data Collection

The data resource currently provided by Web includes Web page data, hyperlinks data, emails, news-group, log data of website and the transaction database formed through Web. According to the principle of related topics, widely collect users' information by various methods, build necessary database and datasheet and prepare for data extraction.

4.2 Data Processing

Automatic filter and pre-handling specific information from obtained web. Information collected for operations such as noise deduction, etc, to provide basic platform for Web information extraction in next step to ensure that the data can truly reflect the extracted object.

4.3 Data Exchange

Certain format conversion of de-noising data and make it adaptable to the processing requirements of the data mining system or data mining software.

4.4 Pattern Analysis

Evaluate the discovered mode, verify and explain the mode created in last step. This work can be done by machine automatically and also can be completed by interact with analysis personnel. Using some methods and tools to analyze the discovered mode and regulation, to find the mode and regulation we are interested in. After the pattern discovered, the number of rules generated can be very large and the expression may be more obscure. Therefore, it is necessary to analyze and evaluate mode, and to show the result up in a way that is easy to understand and accepted.

5. The evaluation indicator of Web information extraction

Up to now, the evaluation indicator of information extraction system is mainly according to the index determined in MUC meeting: P and R, namely precision and recall. The calculation formula is as follows:

P=the points of correct information extracted /all points of information extracted

R= the points of correct information extracted /all correct information points

The value of P and R is between O and 1. Generally speaking, P and R is in reverse relation, that is reducing R will resulting in P's increasing, in opposite, increasing R will cause the reduction of P. When evaluating a system, we can not only see a P or R but to compare both at the same time. So far as now, if the information extraction task is simpler, P and R can reach 90% in highest. But for the complicated extraction task, the highest percent P can reach is 70%; the highest of R is 50%.

In the comprehensive performance comparison of two extraction systems, it is normally to use P and R's synthetically value F to measure. The calculation formula is as follows:

$$\mathrm{F}=\frac{(F1^{2}+1)PR}{F1^{2}P+R} \tag{1}$$

Of which F1 is a preset value and normally it is 1 which is the relative weight of R and P. The accuracy rate is more important when F1is higher than 1. The recall rate is more important if F1 is lower than 1.It is possible to evaluate the quality of a system only with one value of F.

Conclusion

With rapid development of Internet, the requirements of people to obtain information from the Internet gradually upgrade and cause the research boom of Web information extraction technology. But the final aim of research is still to develop practical information extraction system, and then can draw the information which users expected from the Web page, and to analyze and tease information, so that to get more valuable information by spending more less time and energy. In the near future, Web information extraction technology will become more important and will play a more important role in more fields.

References

[1] Chen Yufang, Ge Suihe. Research of web data collection module basis on XML [J]. Computer engineering and application, 2004(10):p.150-152.

[2] Zhang Zhixiong, Sun Tan, Wu Zhenxin. The research on concept, ideas, methods and trends of knowledge extraction[R]. The Documentation and Information Center of the Chinese Academy of Sciences, 2008.

[3] Ding Weiwei. Research based on Web information resource data mining technology [J]. Journal of Changchun University of Science and Technology, 2012(12).

[4] Zheng Junxi. The research based on Web information extraction technology of XML [D]. Dalian Jiao tong University, 2013.