ATLANTIS PRESS
Advances in Computer Science Research (ACRS), volume 54
International Conference on Computer Networks and Communication Technology (CNCT2016)

An Application Based on Regular Expression

Jian-Ping DU[1] and Ya-Shu LIU [2,*]

[1] Library System Department, Beijing Union University, Beijing, China

[2] Department of Computer Science, Beijing University of Civil Engineering and Architecture, Beijing, China

*Corresponding author

Keywords: Hypertext, Regular Expression, Information extraction, Chinese Information Processing.

Abstract. Hypertext is the main format on the Internet, which has the characteristic of the non-continuity and simple formal standard. This paper researches how to extract information from hypertext by regular expression, gives the example on extracting weather information base on the web of Sohu, Sina and Tencent to forecast weather. The paper gives the feasible extraction approach of hyper text, which is a new method of Chinese Information Processing. It is very important practice for them who want to research the research engine

Introduction

With the popularity of computers and the rapid development of the Internet (WWW), a lot of information appears in the form of electronic documents. In order to deal with the serious challenges caused by information explosion, it is urgent to need some automated tools to help people quickly find the real needs in the mass of information sources. So, Information extraction (Extraction Information, IE) is produced.

In this paper, we use the regular expression as a tool to study how to extract specific information from the weather forecast Web. As the same time we research the relevant technology about the Web pages, finally give specific examples of extraction.

Information Extraction Technology

Information extraction technology is a new field that has developed in the past ten years, and it has met many new challenges. The original goal of information extraction is to find specific information from natural language documents, which is a very useful research field in Natural Language Processing. The key component of the information extraction system is a series of extraction rules or patterns. The patterns can determine what information needs to be extracted. With the great increase of the online text information, the research has been highly improved in this field.

Text information extraction and processing can be divided into three categories: unstructured text, semi-structured text and structured documents. The original purpose of information extraction is to extract the main information from the unstructured text. The unstructured text information extraction system usually uses the Natural Language Processing method, which mainly through the extraction rules based on syntactic relations between words and lexical category. In order to get the syntactic relations, we need to train and learn on a large number of texts combined with machine learning and artificial intelligence technology. Structured text is a text that is generated in a certain format. To extract specific information from such a text, which is necessary to specify

the rules in accordance with the prescribed format. Semi structured text is a form of text between unstructured and structured text, such as WEB pages. Another, such as the text format of the legal treaty, patent documents, etc., which seems to be unstructured, but its content structure is followed by a certain pattern structure, so it can also be seen as semi structured.

In this paper, we will research some of the weather forecast websites, analysis of extraction rules, extract and store weather information in the local system, in order to facilitate the different sites of weather information query and display on the same platform.

The Regular Expression

The regular expression is a string describing the structure of formal language, which is a method to search the specific set of characters from the string, proposed by American mathematician Stephen Kleene in 1956. Mainly it is used to describe the regular set on algebra. Then people use this expression to calculate the search algorithm in some early researches, the first practical application of regular expression is QED Unix editor. At present, with the successful practice of regular expressions in a high-level language makes it has greatly exceeded the traditional mathematical limitations. It can be understood as a special string, a common character (literal) and special characters (meta characters) which can be interpreted in accordance with the rules of particular grammar. The regular expression is in the form of "/pattern/", which is going between "/" delimiters in the target object matching mode. As long as the user finds the object content matching with the pattern of "/" delimiter, the content can be extracted. In order to be more flexible to customize the content of the model, the regular expression provides a special "meta character". The so-called meta character refers to those special characters that have special meaning in regular expressions, which can be used to specify the appearance patterns of their leading characters (i.e. characters in front of meta characters) in the target object. For example, Eq. 1 is a regular expression that judges the Email address.

$$\[\w-\]+(\.\[\w-\]+)*@\[\w-\]+(\.\[\w-\]+)+\ \quad (1)$$

Special Application Example

In this paper, Sohu, Tencent and Sina weather websites are as the research objects. We research their Web structures, design regular expression and realize the weather information extraction platform.

Web Page Structure

At present, many web sites have provided a special weather information prediction pages, that includs text, hyperlinks, images and other information, as shown in Fig. 1. If you want to extract specific weather information from such a Web page, you need to clean the page information. Web page cleaning is mainly divided into three steps: the first step is to remove the page notes, scripts, style sheets, and other unrelated information. The second step the page is divided into a number of blocks, including the text block, the link block, the image block, etc.. Finally, according to the semantics of each block for further distinction, such as the text block ads and other non key information blocks; distinguish between the link blocks, navigation links, advertising

links, such as different content. After the above processing, the Web page is divided into fine granularity information blocks in the structure and semantics, so that the subsequent information processing can be carried out smoothly.

On the Web page files, they also contain a lot of redundant information except the type information contained in the variety, such as Fig.1 shows the weather information on www.weather-forecast.com. In the Fig.1 the left column shows the weather map, but the middle column is ad which is no use for weather information, and is required to be removed. Therefore, it is necessary to carry out the clean processing of this page in order to extract the weather information. Fig.2 shows the weather information about Beijing.

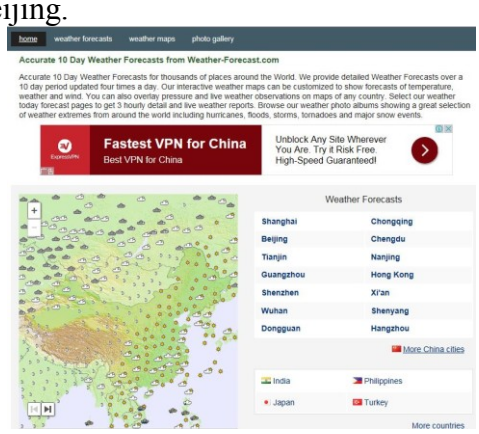

Fig. 1 Weather Forecasting Website Example

Beijing 1 – 3 Day Weather Forecast Summary: Mostly dry. Freeze-thaw conditions (max 9°C on Mon afternoon, min -3°C on Sun night). Wind will be generally light.

	Sun 18		Monday 19			Tuesday 20			Wed 21	
	Night	AM	PM	Night	AM	PM	Night	AM	PM	
Weather Map See all maps										
Wind (km/h)	5	5	5	5	5	10	10	5	10	
Summary	clear	clear	clear	clear	some clouds	some clouds	some clouds	clear	clear	
Rain (mm)	-	-	-	-	-	-	-	-	-	
Snow (cm)	-	-	-	-	-	-	-	-	-	
Max. Temp (C)	-2	6	9	-1	5	6	0	5	7	
Min. Temp (C)	-3	-3	0	-2	-2	1	-3	-3	1	
Wind Chill (C)	-3	-3	-2	-4	-5	-3	-6	-3	-2	
Humid. %	35	29	24	27	28	29	52	64	46	

Fig.2 Beijing Weather Information

We analyze the structure of the web page, find the required weather information, then design the corresponding regular expression to extract the information. Finally we will get the required information, and remove redundant and noise information.

Get the Regular Expression

Regular expressions provide a mechanism for searching for a particular string from a character set. It allows users to use special characters through a series of construction, mode, to match mode and data files. On our platform, we input object, according to whether the target object is included in the matching model, and get the weather information.

This paper mainly introduces the design process of the regular expression of Sohu, Tencent and Sina weather website. For weather.news.sohu.com, in the last section we have introduced that how the weather forecast in the region on the page, we selected 'China map Flashmap plugin in the Web page', the 'hypertext Flashmap plugin' corresponding to the following:

```
<div  id="flashMap">
<embed  height="380"  width="654"allowfullscreen="false"  quality="high"
bgcolor="#8bd9ff"src="http://news.sohu.com/upload/weather_10/index091230.swf"
type="application/x-shockwave-flash">
</div>
```

Thus regular expressions can be obtained, as shown in the Eq. 2.

```
/<div id="flashMap">(.*)</div>/                                    (2)
```

The Sina weather website in the weather part of the Web page is as follows:

```
<!-- weather begin -->
<div class="weather">
```

<iframe width="150" height="20" frameborder="0" marginheight="0" marginwidth="0" scrolling="no" src="http://news.sina.com.cn/iframe
/2009/1230/408.html"></iframe>

<!-- weather end -->

In the code of weather information is included in the '<! -- weather and begin -- > <! -- weather end -- >'. The code displays specific weather information which is written between the and the tag, the regular expression as shown in Eq. 3.

/<div class=\"weather\">(.*) / (3)

Realization of Weather Information Extraction System

Design of Weather Information Extraction System

In order to browse the weather information of each website more easily, this paper presents an example of weather information extraction system based on regular expression. On the platform, we can wirte the URL address of the website to be extracted and input the corresponding regular expression. Then we can extract the weather information of the website. The page design of the system is shown in Fig. 3.

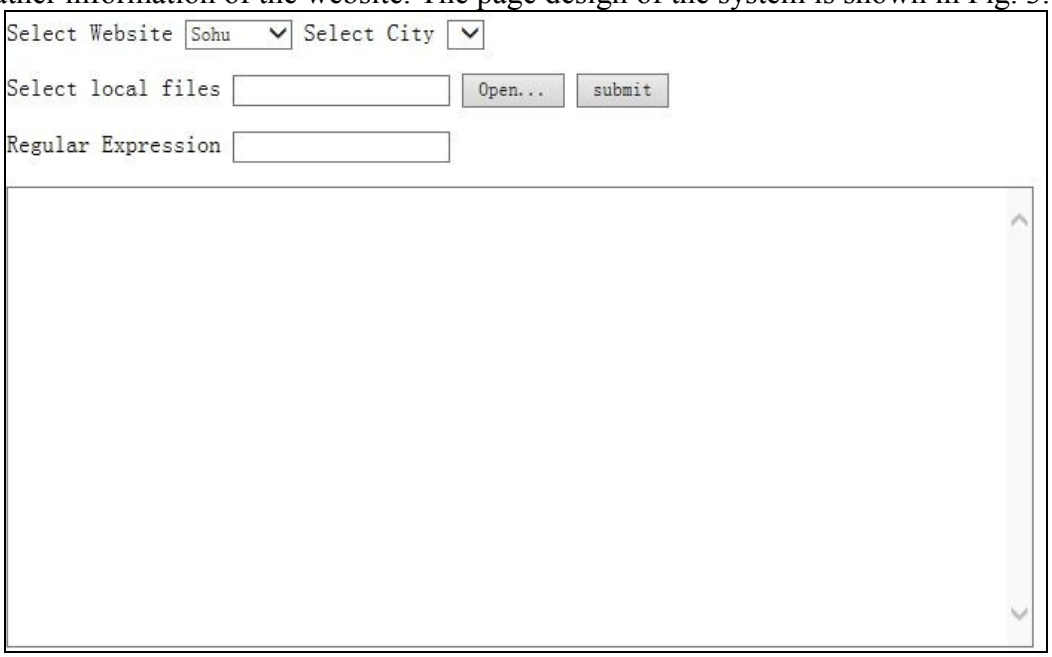

Fig. 3 Weather Information Extraction Platform

Key Technology of Weather Information Extraction Platform

In order to parse the page and obtain weather information, we use SUN's java.util.regex package, which provides comprehensive support for regular expressions, the Java.lang.String class replaceAll, split and related function of weather information and matching regular expressions. There are three classes in the java.util.regex package: Matcher, patternSyntaxException, and Pattern. Pattern class is mainly in accordance with the syntax of regular expressions to match the structure of the object.

According to the structure of the regular expressions, it needs to complete two tasks, the pattern matching and finding the string. For example, sohu weather website, the platform needs to find and match the ' ... '.

The core algorithm is used in the system to match the regular expression which is shown as follows:

```
public String getSelCities(String link,String regularExpression){
   String data="";
XMLReader xr=new XMLReader(link);
String rst =XMLReader.parse(null).asString();
Perl5Compiler compiler = new Perl5Compiler();
Pattern pattern = compiler.compile(regularExpression);
PatternMatcher pm = new Perl5Matcher();
PatternMatcherInput matcherInput = new PatternMatcherInput(rst);
if (pm.contains( matcherInput, pattern)) {
MatchResult result = pm.getMatch();
data+=result.group(1));
      }
      return data;
      }
```

Summary

In this paper, we research the information extraction method in regular expressions, using the Tencent, Sina and Sohu weather website as examples. It is also a new way to Chinese information processing and information extraction and other related research work, but also it provides a technical basis for the development of its own search engine, which is worth further study. However, this approach still has problems, such as it needs to analyze the structure of the web page, the structure of regular expressions manually, which is not flexible enough. However, due to the strong ability of regular expressions, this method still has a broad research and application prospects.

Acknowledgement

This research was financially supported by the National Science Foundation.

References

[1] Jeffrey E.F.Friedll . Mastering Regular Expression. Publisher:Oreilly.2006:146-166.

[2] Qin B, Wang S, Du X Y, et al. Graph - based Query Rewriting for Knowledge Sharing Between Peer Ontologies. Information Sciences, 2008, 178 (18):3525- 3542.

[3] Marcelo A, Leonid L. XML Data Exchange: Consistency and Query Answering. Journal of the ACM , 2008, 55 (2) : 29 - 60.

[4] Du Dong mei, Xu Cai Xin.The Application of Regular Expression on Web System. Computer Systems and Applications,2007(8):87-90.(In Chinese)

[5] Yaru Sheng, Zhengang Wei, Meng Liu.Data Collection Research and Application based on topic Web Crawler.Electronic Technology & Software Engineering. 2016(7):168-169. (In Chinese)

[6] Huili Tang, Xiaomei Zheng. Regular Expression Research and Application in Web. Computer Technology and Development.2013(2):82-84.. (In Chinese)

[7] Shenyuan Zhang. Regular Expression implemention. Science and Technology Innovation Herald. 2010(1):28-31. (In Chinese)

[8] XU Qian, E YuePeng, GE JingGuo.Efficient Regular Expression Compression Algorithm for Deep Packet Inspection.Journal of Software.2009, 20(8):2214-2226. (In Chinese)

[9] Bin Li. Application in PHP Based on Regular Expression. Computer Development & Applications. 2015(3):54-57. (In Chinese)