The Automatic Extraction of Web Information Based on Regular Expression

Li Ji [1,2] , Jiang Guangyu[1,2], Xu Aijun[1,2*], Wang Yunzhen[3]

[1] School of Information Engineering, Zhejiang A&F University, Lin'an 311300,China.
[2] Zhejiang Provincial Key Laboratory of Forestry Intelligent Monitoring and Information Technology, Zhejiang A&F University, Lin'an 311300,China.
[3] Jiande Xin'anjiang Woodland , Jiande 311600.

* Corresponding author. Tel.: 8618058124658; email:xuaj1976@163.com

Manuscript submitted December 12, 2016; accepted January 20, 2017.
doi: 10.17706/jsw.12.3.180-188

Abstract: Based on search engine , this paper built a Web information retrieval matching and structure extraction model. And realized the algorithm of locating and automatically extracting multi-web Baidu news information. Getting the standard mathematical expression of URLs by analyzing the search results URLs and analyzing the DOM tree structure of web pages, this article designed the key tags regular expression. Finally, the method of multi-page location retrieval and structured extraction based on search engine is realized. The experimental results showed that the average extraction result is 99.60%, and the matching ratio is 99.56%. It can be used for Web information structure and automatic extraction and local preservation.

Key words: Search engine; extraction model of web information; regular expression; web information.

1. Introduction

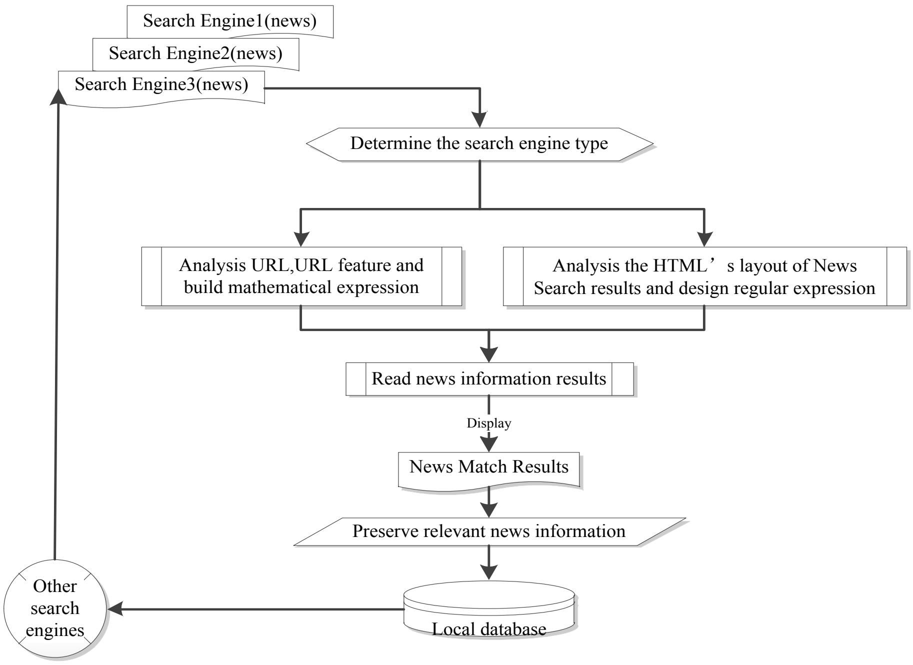
With the development of information technology, web information has been explosive growth in today's society [1]. But as far as on its effective use, there comes many problems such as the page load is too large, information display complex and saving information extraction is no good local preservation program. This paper took Baidu news as an example: 1) If the user can not be timely to search , extract, save or personalized process for effective information, it may lead to the loss of a certain amount of time-sensitive information resources of news ; 2) There are lot of drawbacks in manually searching, classifying, and collecting large number of Baidu news information, such as heavy workload, high repetition rate, low efficiency, inefficient, high error rate and some other drawbacks cannot be controlled.

The method of expert system of Web information search and extract is mainly based on the model themes [2] and ontology [3], the application and research of different professional field has developed rapidly, like web news events [4], commercial information extraction, document retrieval, scientific research personnel [5], health care [6] and so on. With the application fields of regular expression technology extends continuously, Web information extract based on it have been deeply studied by many scholars [7-8], meanwhile, it makes search positioning the target Web information more accurate more rapid by using keyword search optimization[9-10] and data noise cleaning, and improve Web information extraction on the space complexity, time complexity and response accuracy, what's more, Wei-Qing Cheng did a page classification of Web information extraction[11]. The extraction of web page information has been a qualitative leap in its quantity and integrity With the development of web crawler technology, and based on Web Crawler [12], Wen-Bing Wan made a research on Topic-Oriented Information Search. moreover, Qing-Qing Xiang put forward the news Extractor algorithm[13] which could extract the key information from the news.

The paper constructed the model of Web information extract based on search engine, and utilized, Baidu news as an example, we realized the algorithm of extract of key Baidu news information based on regular expression, on this basis, and we got more accurate and more standard news information for users by the data noise cleaning. According to the further individual needs of users, by setting the secondary information filter conditions, it made the operating of extracted information more convenient and efficient. The paper provided a structured, standardized, personalized and automatic processing decision for the extraction and preservation of Web information as the ultimate. Above all, it made the extraction of Web information and structured local preservation more convenient, fast, simple and economical.

2. Web Information Extraction Model Building

Search engine technology bases on keywords and semantics has been greatly improved with the deep research on the application of search engine technology [14], [15], and the search results with a high accuracy and high speed is relatively complete and informative. Based on the general design principles for the URL design(the general rule of site visit routing protocol)and HTML design standardization, different search engine news search results can be counted, summarized and analyzed to design a retrieval matching and structured extraction model of web information of search engine. The model is designed as follow:



Search Engine1(news)
Search Engine2(news)
Search Engine3(news)
Determine the search engine type
Analysis URL,URL feature and build mathematical expression
Analysis the HTML’s layout of News Search results and design regular expression
Read news information results
Display
News Match Results
Preserve relevant news information
Local database
Other search engines

Fig. 1. Search matching and structured extraction model of web information based on search engine.

At the same time, we analyzed the search results URLs under the environment of different browser, and the results showed that, access routes had been unified design standards strictly. So which could be cut into seven segments, and the mathematical expression is described as follow:

$$URL(\mathrm{i})_n = Appendstr(1)_n + Keyword(1)_{mn} + Appendstr(2)_n + Keyword(2)_{mn} + Appendstr(3)_n + pn(i)_n + Appendstr(4)_n \quad (1)$$

In the formula: Appendstr(1)$_n$,Appendstr(2)$_n$,Appendstr(3)$_n$,Appendstr(4)$_n$ are the search engine n’s searching results for public truncated strings of the URL; n is the search engine; i is the search result page number; pn(i)$_n$ is the paging function of the search engine n ;Keyword (i)$_{mn}$ is the search keys.

Generally speaking, the need to extract news information are News headlines, news links and news published time etc, by analyzing the different browser search layout features of news information results in the DOM tree, its label order by news links, news headlines and news release time, in this way, we can descript the regular expression match of news information as follow:

$$\begin{aligned}&\mathrm{Re}\, gExpressionStr(n) = \mathrm{Im}\, plodestr(1)_n + NEWSurl + \mathrm{Im}\, plodestr(2)_n + NEWStitle + \mathrm{Im}\, plodestr(3)_n \\ &+ NEWSpublis\, hTime + \mathrm{Im}\, plodestr(4)_n \end{aligned} \tag{2}$$

In the formula: n is the search engine; Implodestr(i)$_n$ is the function of the i-th truncated character corresponding to the search engine n; NEWSurl, NEWStitle, NEWSpublishTime represent news links, news headlines and news published time three extraction keywords.

3. The Method of Structured Web Information Extraction

Regular expression can be used to automatically match, replace, and extract valid key strings and tags in files such as HTML \ XML[16-17].The main application areas are as follows: test string pattern, replace text, extract required field information from strings, crawler, site validation, and log analysis.

The paper chose the news of Baidu search engine search sector as a study object, and typed in the search keywords(whitespace separated), according to the search results, we can determine the standard mathematical expression for the URLs and obtain the DOM tree structure of HTML forms, at the same time, we analyzed the news label layout in the DOM tree(news links, news headlines, news release time) for the target news information regular expression design basis, in addition, it provides feasibility analysis for searching, matching, extracting the target label attribute information.

3.1. Design of URL Mathematical Expression and Regular Expression

Development test environment: 1)browser: 360 browser version 4.0 or later; 2)search engine: Baidu search engine, news search sector; 3)network bandwidth: 2MB/s.

Through the Baidu news search sector typed in two keywords “浙江农林大学” as keywords1 and “徐爱俊” as keywords2(input format is keywords1 + space + keywords2), according to search result, we got search results of the DOM tree structure information as shown follow:

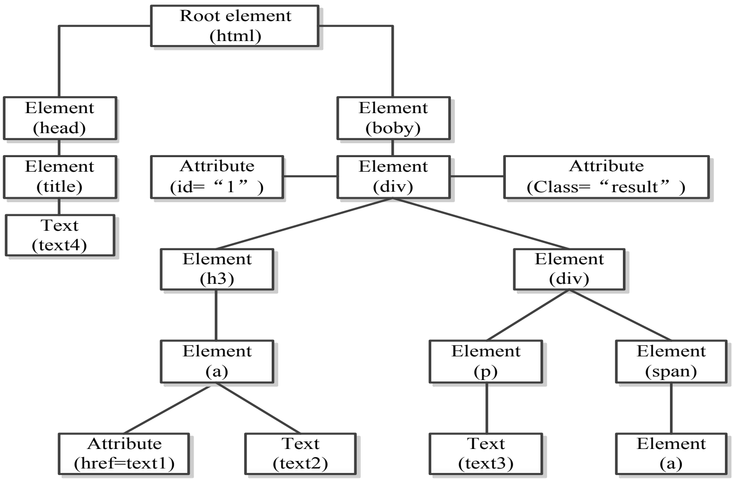


Fig. 2. The DOM tree structure of news search results.

text1:news link URLs;
text2:news headlines;
text3:news source and news release time of the synthesis of the string;
text4search engine_keywords1 keywords2;

We can come to the conclusion that the target information label layout of the HTML DOM tree follows the w3c standards strictly in the news search results[18-19], and the visit routing protocol standard is stable, so we can inject them into Search matching and structured extraction model of web information(Fig.1).

Appendstr(1)$_{\text{Baidu}}$= “http://news.baidu.com/ns?word=” ;
Appendstr(2)$_{\text{Baidu}}$=” %20” ;
Appendstr(3)$_{\text{Baidu}}$=” &pn=” ;
Appendstr(4)$_{\text{Baidu}}$=” &ct=1&tn=news&ie=utf-8&bt=0&et=0” ;
Keyword(1)$_{\text{Baidu}}$=keywords1;
Keyword(2)$_{\text{Baidu}}$=keywords2;

The value of pn is proportional to the page number,and its mathematical relation derives the formula as follows:

$$pn(i)_{\text{Baidu}} = (i-1)*20 \tag{3}$$

Further, we can obtain the Baidu search URL expression:

$$\begin{aligned} URL(i)_{Baidu} = {} & Appendstr(1)_{Baidu} + Keyword(1)_{Baidu} + Appendstr(2)_{Baidu} + Keyword(2)_{Baidu} \\ & + Appendstr(3)_{Baidu} + pn(i)_{Baidu} + Appendstr(4)_{Baidu} \end{aligned} \tag{4}$$

The corresponding truncated character is brought into equation (2):
Implodestr(1)$_{\text{Baidu}}$=<div[^>]*?class=""result""[^>]*?><h3 .*?><a[^>]*?href="";
Implodestr(2)$_{\text{Baidu}}$=""[\s\S]*?>;
Implodestr(3)$_{\text{Baidu}}$=</a></h3><div[^>]*?>[\s\S]*?<pclass=""c-author"">.*?&nbsp;&nbsp;
Implodestr(4)$_{\text{Baidu}}$=\d+:\d+</p>;
NEWSurl=(?<NEWSurl>.*?);
NEWStitle=(?<NEWStitle>.*?);
NEWSpublishTime=(?<NEWSpublishionTime>.*?);
The regular expression of Baidu news search is:

$$\begin{aligned} \mathrm{Re}\, gExpressionStr(\mathrm{Baidu}) = {} & \mathrm{Im}\, plodestr(1)_{Baidu} + NEWSurl + \mathrm{Im}\, plodestr(2)_{Baidu} \\ & + NEWS\mathrm{title} + \mathrm{Im}\, plodestr(3)_{Baidu} + NEWSpublis\, hTime + \mathrm{Im}\, plodestr(4)_{Baidu} \end{aligned} \tag{5}$$

3.2. The Design of Implementation Scheme

According to the search keywords, we can predefine the URL queue, after that, URLs response the Web server through HttpWebRequest to obtain search results form text, and we put HTML forms into a form buffer pool(AllStrHtml) for unified management, after the search, by the design of regular expression matching, we can retrieval and extract the target news information and process the results for more needs. The realization of the technology roadmap as shown follow:

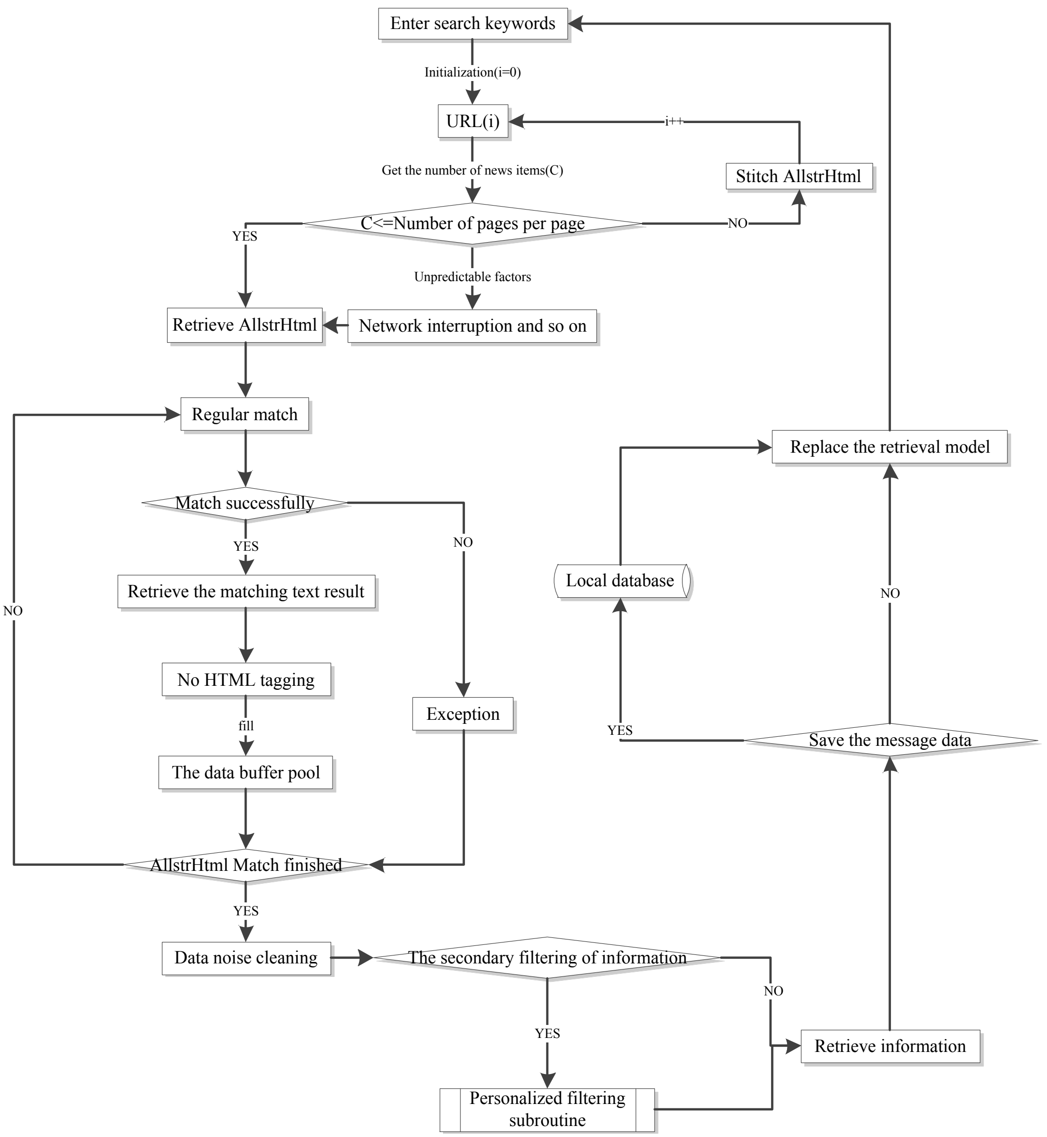


Fig. 3. The technology roadmap of structured extraction method of Baidu news information.

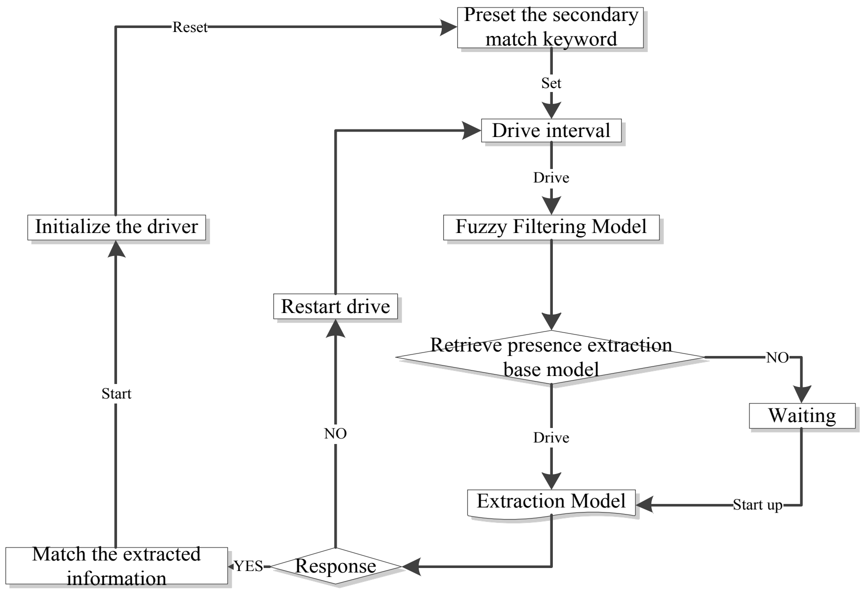


Fig. 4. The concrete implementation process.

We can search, match and extract the Baidu news by the basic model, based on it, after the optimization algorithm, we can locate the information what you need more quickly, the concrete implementation process as follow:

3.3. Algorithm

We type in keywords$(1)_{Baidu}$ and keywords$(2)_{Baidu}$, after standard the input values, according to the search results, we can get the formula (4) to predefine the Baidu news URL of queue, and then make the URL queue response (POST form) Web server and extract HTML form with HttpWebRequest, at the same time, put the form into the buffer pool AllStrHtml temporarily, if the judgment to extract the termination condition is true, the extraction is stopped, in the end, we can use the regular expression match to extract target news information. The implementation process is described as follow:

a)Type in the keywords: keywords$(1)_{Baidu}$ and keywords$(2)_{Baidu}$, and standardize them;

b)Keyword $(1)_{Baidu}$ and Keyword $(2)_{Baidu}$ are converted to the string data type;

c)Get the number of news items displayed per page from the search results(MN);

d)Bring the results of b) and c) into formula (1),initialize the URL $(i)_{Baidu}$, i = 0, and we can get the formula (3) and (4) for the next steps;

e)The URL$(0)_{Baidu}$ obtained from d) can be used to drive HttpWebRequest, and response the web sever by POST method and extract the URL $(0)_{Baidu}$'s HTML (0) form;

f)Analyze the DOM tree structure of HTML(0), according to the news information retrieval model and design regular match expression by equation (5);

g)Use the regular match expression from f) to retrieve the HTML (i) form, get the number of news match (C (i));

h)Store the HTML(i) in the AllstrHtml(AllstrHtml+=HTML(i));

i)C(i)<=MN, if false i++, return g);else terminate the POST command, get AllstrHtml’ s size n = i, go to j);

j)Extract HTML form Terminated, no HTML tagging for AllstrHtml;

k)Regular match expression (3) is embedded in the retrieval model of news information retrieval;

l)Match the results from i) and j),obtain the extracted news information results(NEWSurl,NEWStitle and NEWSpublishionTime);

m)Make perform data drain and data noise cleaning for l)’ s results;

n)Repeat l) and m) Until it exits the current news search model;

o)Drive optimization model;

p)Return a);

4. Implementation results and analysis

4.1. Experiments Settings

Browser: 360 browser version 4.0 or later;

Search engine: Baidu search engine news sector;

Network bandwidth: 4MB/s;

Testing time: Before Junde21,2016;

Search keywords: keywords$(1)_{Baidu}$ and keywords$(2)_{Baidu}$ are the university reach the top 450 of Chinese University Comprehensive Ranking in 2016 (Data were divided into three groups : 1-150 as the first group of data , 151-300 as the second group of data , 301-450 as the third group of data , and each group has at least one test data . 10 pieces of test data are selected).

4.2. Extraction Results and Analysis

The test results data as shown in Table 1.

Table 1. Baidu News Search and Extraction News Results Compare Table

University	President	Baidu(S)	Extraction(R)	Match(M)	Ratio(R\S*100%)	Rate(M/S*100%)
Zhejiang A & F University	ZHOU**	119	118	118	99.16	99.16
Minzu University of China	HUANG**	39	39	39	100.00	100.00
Fudan University	XU**	261	261	261	100.00	100.00
Jiangxi Science and Technology Normal University	ZUO**	27	27	27	100.00	100.00
Northeastern University	ZHAO**	130	129	129	99.23	99.23
Xi`an Jiaotong University	WANG**	397	397	397	100.00	100.00
Shandong University	ZHANG**	504	504	504	100.00	100.00
Dalian University of Technology	GUO**	99	98	98	98.99	98.99
Huazhong University of Science and Technology	DING**	126	126	126	100.00	100.00
Southwest Jiaotong University	XU**	224	221	220	98.66	98.21
Average					99.60	99.56

Notice: In order to respect the individual’ s privacy, this paper had processed the real name in the table.

Baidu(S): The total research results of Baidu News API.

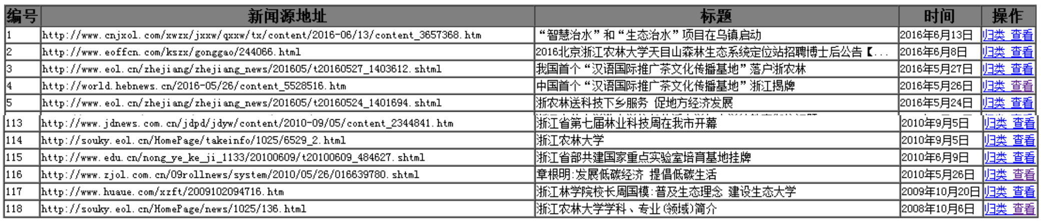
Extraction: The total research results by using this method.

Match: The count of matching between Baidu News research results and extraction results by using this method.

From table 1, we can learn that, by the extraction method, has illustrated that, the minimum extraction ratio is 98.66%, while its maximum reaches 100.00%, and average results are 99.60%. Hence, we can conclude that this method has performed relative well in its practical applicability.

To further illustrate the convenience of the extraction results , we choose first test data as an example, and the Baidu API search results are that each page shows 20 items of news information , 119 items in total .

The extraction results of Baidu news by using this method is shown as the Fig. 5 :



编号	新闻源地址	标题	时间	操作
1	http://www.cnjxol.com/xwzx/jxxw/qxxw/tx/content/2016-06/13/content_3657368.htm	“智慧治水”和“生态治水”项目在乌镇启动	2016年6月13日	归类 查看
2	http://www.eoffcn.com/kszx/gonggao/244066.html	2016北京浙江农林大学天目山森林生态系统定位站招聘博士后公告【...	2016年6月8日	归类 查看
3	http://www.eol.cn/zhejiang/zhejiang_news/201605/t20160527_1403612.shtml	我国首个“汉语国际推广茶文化传播基地”落户浙农林	2016年5月27日	归类 查看
4	http://world.hebnews.cn/2016-05/26/content_5528516.htm	中国首个“汉语国际推广茶文化传播基地”浙江揭牌	2016年5月26日	归类 查看
5	http://www.eol.cn/zhejiang/zhejiang_news/201605/t20160524_1401694.shtml	浙农林送科技下乡服务 促地方经济发展	2016年5月24日	归类 查看
113	http://www.jdnews.com.cn/jdpd/jdyw/content/2010-09/05/content_2344841.htm	浙江省第七届林业科技周在我市开幕	2010年9月5日	归类 查看
114	http://souky.eol.cn/HomePage/takeinfo/1025/6529_2.html	浙江农林大学	2010年9月5日	归类 查看
115	http://www.edu.cn/nong_ye_ke_ji_1133/20100609/t20100609_484627.shtml	浙江省部共建国家重点实验室培育基地挂牌	2010年6月9日	归类 查看
116	http://www.zjol.com.cn/09rollnews/system/2010/05/26/016639780.shtml	章根明:发展低碳经济 提倡低碳生活	2010年5月26日	归类 查看
117	http://www.huaue.com/xzft/2009102094716.htm	浙江林学院校长周国模:普及生态理念 建设生态大学	2009年10月20日	归类 查看
118	http://souky.eol.cn/HomePage/news/1025/136.html	浙江农林大学学科、专业(领域)简介	2008年10月6日	归类 查看

Fig. 5. The result of structured extraction method.

There are 118 items of extracted Baidu news. To a certain extent, the method has optimized in extracting Baidu News , not only in the realization of displaying multiple page results in signal page, but also in the locally saved proposal for users and original web query function .

4.3. Optimized Effect of Structural Extraction Method

The paper has realized the target extraction algorithm, and by the drive optimization model, it can be to secondary fuzzy query, query time orientation and keyword positioning query, which based on the basic algorithm model. It can more quickly and more effectively meet the results information extraction of the personality operation requirements, and realizes the goal to position the news information of search results fast, based on the basic extraction mode of optimization design. Therefore, the retrieval matching and extraction algorithm achieves its expected results.

4.4. Summary

This method is the first time to put forward the search matching and structured extraction model of web information based on search engine to achieve the goal to search and extract Web multi-page and get the structured news information, and which has advantages in the aspects such as amount, integrity etc, than previous research. The method doesn’t use the mechanical application or third-party API calls, but predefines the sites URL queue, which is more economical and getting more target information than Mechanize+Beautiful Soup method, and has the characteristics of universal usability. In addition, this method is different from Web crawler algorithm strictly, based on the predefined URL queue, which can make full use of the third party server capabilities to reduce the pressure of the server parsed and time efficiency improved. It can keep the target information searched quickly and high integrity at the same time.

5. Conclusion

This paper puts forward a search matching and structured extraction model of web information based on search engine, and takes Baidu news as an example, to provide a structured, automatic, standard, massive and personal solution to extract and classify news information, which makes the news information extraction and local preservation more accurate, economic, fast, efficient and practical. The mean extraction rate of Baidu news searching reached 99.60% around , while matching rate reached 99.56% . It has a better practical value and provide an effective way to let users quickly and effectively apply web information under large data condition . This method has performed well in its extensibility which may structured extract more search engine’s information resource . In addition , it make a further contribution to structured extraction and locally saving of web information , with the advantages of intelligence , diversity , and generality . Fitting and integrating more of the current search engine information resource , and realizing web information’s integrated structural extraction and locally saved scheme , it introduce a more comprehensive web information sharing platform .

Reference

[1] Wang, Y. Z., Jin, X. L., & Cheng, X. Q. (2013). Network big data: Present and future. *Chinese Journal of Computer*.

[2] Zhang, Y., Chen, M., & Liao, X. F. (2013). Big data applications : A survey. *Journal of Computer Research and Development*.

[3] Zhang, T. T., Liu, K., & Wang, W. J. (2015). The mode of automatically crawling web data and its open source solutions for researchers. *Journal of Information Resources Management*.

[4] Li, R. J., Zhang, J., Zhang, X. M., & Gui, X. Q. (2016). Web information extraction in health field. *Journal of Computer Applications*.

[5] Zhou, H. M., & Xi, J. Q. (2011). Design and realization of template—Based web crawler. *Computer Technology and Development*, *11*.

[6] Wu, H. L. (2010). Study on the web information extraction technology based on the ontolgy and DOM

tree. *Information Science*.

[7] Tang, H. l., & Zheng, X. M. (2013). Research of regular expressions and application in web. *Computer Technology and Developmen*.

[8] Hu, J. W., Qin, Y. Q., & Zhang, W. (2011). Regular expression and its applications to web information extraction. *Journal of Beijing Information Science and Technology University*.

[9] Zhou, J. F., & Meng, X. F. (2012). Keyword search on XML data : A survey. *Chinese Journal of Computers*.

[10] Ren, J. H., Zhou, J., Meng, X. F., & Wei, K. (2013). Results ranking approach of XML keyword search based on keyword's structural relationships. *Computer Scienc*.

[11] Cheng, W., Qing, Y. J., Yang, J., *et al*. (2013). Web information extraction research based on page classification. *Computer Technology and Development*.

[12] Wan, W. B. (2015). Research on web crawler information collection strategy based on topic search. *Software Guide*, *11*, 68-70.

[13] Xiang, J. J., Geng, G. G., & Li, X. D. (2016). Key information extraction algorithm of news web pages. *Journal of Computer Applications*.

[14] Ma, S., Zhao, W., Yuan, C. Y, *et al*. (2013). Research on critical technologies of semantic retrieval based on rule reasoning. *Acta Electronica Sinica*.

[15] Zhang, W. Z., Zhang, H. L., Xu, X., & He, H. (2012). Distributed search engine system productivity modeling and evaluation. *Journal of Software*.

[16] Introduction of regular expression. Retrieved from: https://msdn.microsoft.com/zh-cn/library/28hw3sce(v=vs.80)

[17] Muzammil, S., Phil, M., & Mark, S. (2015). Automatic generation of valid and invalid test data for string validation routines using web searches and regular expressions. *Science of Computer Programming*, 405-425.

[18] Zeng, J. (2013). Usability analysis of HTML web page development. *Art and Design*.

[19] Yang, C. L., Liu, N. T., Lin, Y., Shao, K. (2013). Domain-oriented structured analysis of web text. *Journal of Hefei University of Technology*.

Li Ji was born in 1991 in Taikang, Zhoukou, Henan Province, China, He received master's degree in engineering from Anyang Institute of Technology in 2014,Currently enrolled in master's Zhejiang Agriculture and Forestry University, the main research areas of Resources and Environment Information System, etc.

Xu Aijun was born in 1976 in Anqing, Anhui Province, China, Ph.D, He is engaged in resource and environment information systems and forest resources information management research, etc.