Proceedings of the 36th Chinese Control Conference
July 26-28, 2017, Dalian, China

Chinese resume information extraction based on semi-structured text

YAN Wentan, QIAO Yupeng

Key Laboratory of Autonomous System and Network Control, College of Automation Science and Engineering,

South China University of Technology, Guangzhou 510640, China
E-mail: ypqiao@scut.edu.cn

Abstract: A Chinese resume information extraction system (CRIES) based on semi-structured text is designed and implemented to obtain formatted information by extracting text content of every field from resumes in different formats and update information automatically based on the web. Firstly, ideas to classify resumes, some constraints obtained by analyzing resume features and overall extraction strategy is introduced. Then two extraction algorithms for parsing resumes in different text formats are given. Consequently, the system was implemented by java programming. Finally, use the system to resolve the resume samples, and the statistical analysis and system optimization analysis are carried out according to the accuracy rate and recall rate of the extracted results.

Key Words: semi-structured, Chinese extraction, regular expression matching

1 Introduction

With the wide application of Internet technology，social information increases exponentially. According to statistics, in these huge amounts of information, there are 60% ~ 70% in the form of electronic documents. According to the data structure feature, the documents can be classified into structured documents, unstructured documents, and semi-structured documents.

Resumes, one kind of common semi-structured document, usually contain valuable structured data hiding in personalized expression. The data may have predefined specifications, but information each file contains might vary, such as the different numbers of fields, containing different field names, field types or different nested format. This makes parsing the document a little more complex than it might be. Classification for them or handling them requires the user to manually open the files, read its whole structure, select the interested information and close them. This manual labor scales linearly with the number of target fields. But this data can be exploited when available as a relation table that we could use for answering precise queries or for running data mining tasks.

This paper introduces CRIES, a Chinese resume information extraction paradigm where the system makes a single data-driven pass over a handful of rule expression. It extracts resume fields just require putting the electronic document in the specified folder. There are two usage scenarios for this system. First, local Chinese resumes in interested fields can be automatically extracted. Second, the data can be updated dynamically based on the data from the internet.

At present, the number of Chinese Information extraction (IE) systems is smaller than English IE systems. Due to the complexity of Chinese semantics, there are still many difficulties in the Chinese IE research. We have analyzed some of the existing Chinese Resume IE system, one of these systems is [7], its IE accuracy reached 87%, another system [11], it achieved 89.3% accuracy and 75.5.0% recalls rate.

*This work is supported by National Natural Science Foundation (NNSF) of China under Grant 00000000.

The reminder of the paper is organized as follows. Section 2 presents the preliminaries of the problems we study. Section 3 provides the class function and process of information extraction. Section 4 gives the resume update demo. Section 5 gives the main test results. Section 6 makes a summary and introduces some optimization methods.

2 Preliminaries

This section lists some preliminary knowledge.

2.1 MVC framework

MVC (Model - View - Controller) architectural pattern is a kind of object-oriented design pattern introduced by Smalltalk-80. It separates the whole manipulate process mandatorily into three parts: input, operate and output, and is commonly used to create a reusable interface program. The architectural patterns have three sections, Model, View, and Controller, as shown in Figure1. The part of Model encapsulates the application problem of core data, logic relationship and business rules, provides the process of business logic. View just describes how to present the part inside Model in a form of visible interface. The Controller processes the request, operates the Model and selects a page to return to the user.

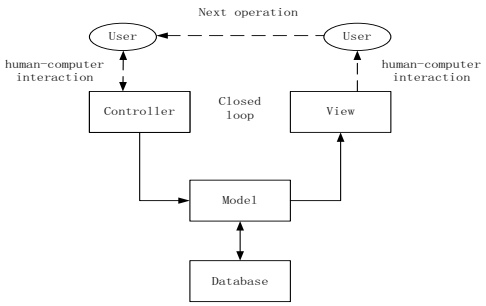

Fig. 1: MVC framework

2.2 Extraction rule expression

According to the expected information, extraction rule expression can be established to make a matcher based on regular expression matching to get their location information

and content information. The following table shows a part of these rule expression in detail.

Table 1: Rule expressions

rule expressions	category	Is a block trigger word?
姓名 (Name)	姓 名 (Name)	N
性别 (Gender)	性别 (Gender)	N
户口所在地 (Registered residence)	户口 (Registered residence)	N
出生日期 (Birthday)	出生日期 (Birthday)	N
出生年月日 (Date of Birth)	出生日期 (Birthday)	N
教育背景 (Education background)	教育背景 (Education background)	Y
实习经历 (internship)	实习经历 (internship)	Y
工作经验 (Work experience)	工作经验 (Work experience)	Y
工作经历 (Work history)	工作经验 (Work experience)	Y

As shown in Table 1, a block trigger word indicates that the rule expression is the piecemeal basis of the resume content. And, apparently, after the block trigger word, it will be a piece of content containing the same type of information.

2.3 Extraction strategy

- Depth-first traversal
  By depth-first traversal, we start from the initial access node, which may have multiple adjacencies, visit the first adjacent node, which serves as another initial accessed adjacency, then access its first adjacent node. To sum up, every time after visiting the current node access to the current node's first adjacent node first. For example, the depth-first traversal order in Figure 2 is 1-2-4-8-5-3-6-7.
- Breadth-first traversal
  The breadth-first traversal is namely according to the level. It traverses the root node, and then the left child node and the right child node until traversing through all the children nodes of the current node. As shown in the Figure 2, the breadth-first traversal order is 1-2-3-4-5-6 -7-8.

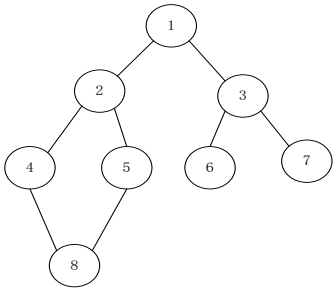

Fig. 2: Search path

2.4 HTML structure level

This article use java programming to achieve the system. For the document in HTML format, it is convenient to use jsoup (Java HTML parsing tools) to extract and operate data. In the HTML DOM, all things are regarded as nodes and jsoup defines what is node, element Etc. Above all, it makes HTML documents as native tree structure named as a node tree, and provides a set of available application programming interfaces (APIs) for developers.

- Node
  The Node is a basic and abstract model of the nodal points. Elements, the Documents, Comments, etc are the instances of node. According to the DOM, each element in the HTML document is a Node.
- Node level
  Nodes have a hierarchical relationship to each other. All nodes in the HTML document form a node tree. Each element, attribute, and text in the HTML document represents a node in the tree. The tree starts from the document node, and extends to all the text in the lowest level of the node tree.

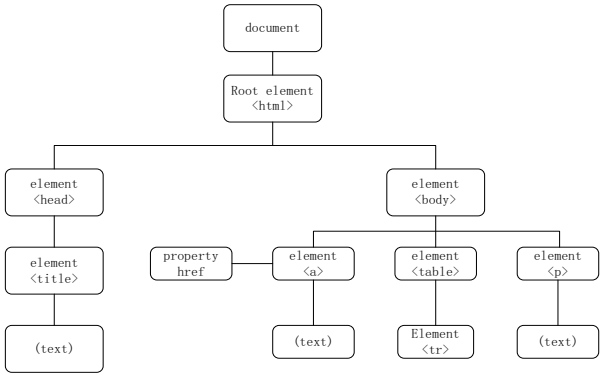

Fig. 3 HTML structure level

- Element
  An HTML element contains a tag name, properties and child nodes and other elements. From an element, we can extract data, traverse the nodes and manipulate the HTML document.

2.5 Some constraints

There are some constraints in the abstraction process.

- English characters are case insensitive.
- Use the greed mode of regular expression match.
- In the rules expressions, the long string should prevail. For example, if there is "professional skills" in the original text, program will find out two trigger words, one is "professional skills", the other is "skills", process will be subject to the "professional skills".

3 Class functions and extraction algorithm

3.1 Software environment

The method described in the article is realized by java programming, it makes use of HTML DOM APIs, and the hardware platform is the PC where the operating system is Windows.

Development environment: JDK: SUN JDK1.7

Application Server: Tomcat7
Software development environment: MyEclipse10
Database platform: MySql5.5
Network requirements: could access to external networks

3.2 Class functions

This part mainly introduces the class used to handle resume, lists the class name and their primary function. Finally, the direction of data flow will be shown in the figure 4.

As shown in the figure 4, the local resume process thread read resume from the folder declared by the class named constant, then class "HandResumeThread" start to deal with the resume content by invoking class "ContentHandler" and class "ContentFetcher", after grabbing the resume content according to file suffix, class "ContentValidator" is called to distribution these different texts content to different content parser. Finally, the program will put the data into the database.

Table 2: class function

Class name	Class function
QueryResumeThread	The entrance of parsing the local resume
Constants	Read configuration information
ResumeThreadManager	Thread management
HandResumeThread	Resume processing thread
ContentHander	resume processing class
ContentFetcher	Abstract the content of the document
ContentValidator	classify the resume
RuleManager	Text document parser
SpecialContentParse	Process special case
ZhiContentParse	HTML document parser
IfChenContentParse	PDF document Parse
51ContentParse	HTML document parser
Utils	String handling

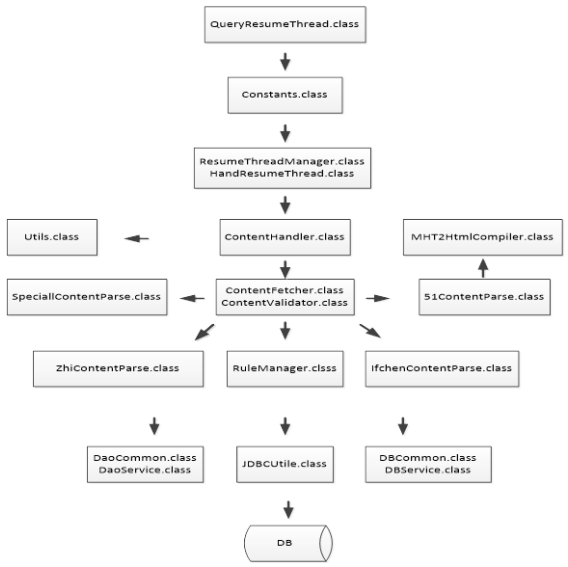

Fig. 4 class relationship

3.3 Extraction algorithm

As shown in the figure4, for different types of documents we use different parsing methods. These documents can be divided into two categories, documents in PDF format or Word format, according to their suffix of filename. We can grab the text of the documents by programming according to the characteristics of their, text content can be divided into simple and plain text and text in HTML format. Consider that the document data vary widely, high degree of its freedom, and high complexity of the test, we use the extraction strategy mentioned in the part 2 to mainly analysis the primary structure of the resume. As for deeper level parsing, such as personal information, we have done a more detailed parsing.

3.3.1 HTML document abstraction algorithm

As shown in the figure3, the HTML document have a tree-like hierarchical, and java jsoup provides DOM element analysis tools. It mapping the document elements, element content, element attributes and processing instructions to different nodes. Therefore, the method of extracting this type of document is different from the simple text flow, and the mainly task is analyzing the tree structure level, and then combine the matching based on regular expression in part 2 we can extract the target information.

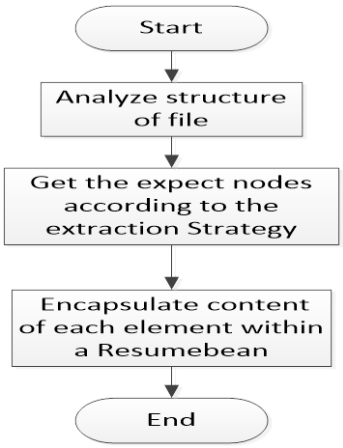

Fig. 5 HTML document parsing process

3.3.2 Plain text abstraction algorithm

The simple plain texts do not have a tree structure, it is impossible to use the above methods to parsing them, but we can establish regular expression that represents the key information of the resume. According to this expression, we can do the extraction steps as follows in figure 6:

1. According to the rule expression of the section 2 to select the trigger word, then get the trigger word position in the document, which is used to delimit the block. Specifically, use the greed mode of regular expression match to iterate over the rule expressions, so we can get the end position and start position of the line that the trigger laid, similarly, get the position information of the start point and end point of the trigger.
2. Use the location information in part 1 for sorting the trigger words to get a trigger words list.
3. There may be error trigger word in the list, we can remove it according to the rules below:

- Remove all other rule expression in the line.
- Remove the following characters:

“和”, ”及”, ”（空格）”, ”基本”, ”其他”, ”其它”, ”专业”, ”社会”, ”学校”, ”校园”, ”学生”, ”个人”, ”情况”,”基础”, ”简介”, ”信息”, ”在校”, ”校内”, ”校外”, ”经历”, ”期间”, ”外语”,”英语”,”计算机”, ”(”, ”)”.

(“and”, “with”,” (space)”,” basic”, “other”, “other”,” major”,” social”,” school”,” campus”,” student”,” personal”,” situation”,” basis”,” introduction”,” information”,” during school”,”in school”,” off campus”,” experience”,” period”,” foreign language”,” English”,” computer”,” (”,”)”).

4. According to the real trigger words obtained in the part3, we can capture the corresponding content from the text, parse rule is that the content between the two trigger word belongs to the former.

5. Parse the block line by line to get the expected field encapsulated them into resume Bean.

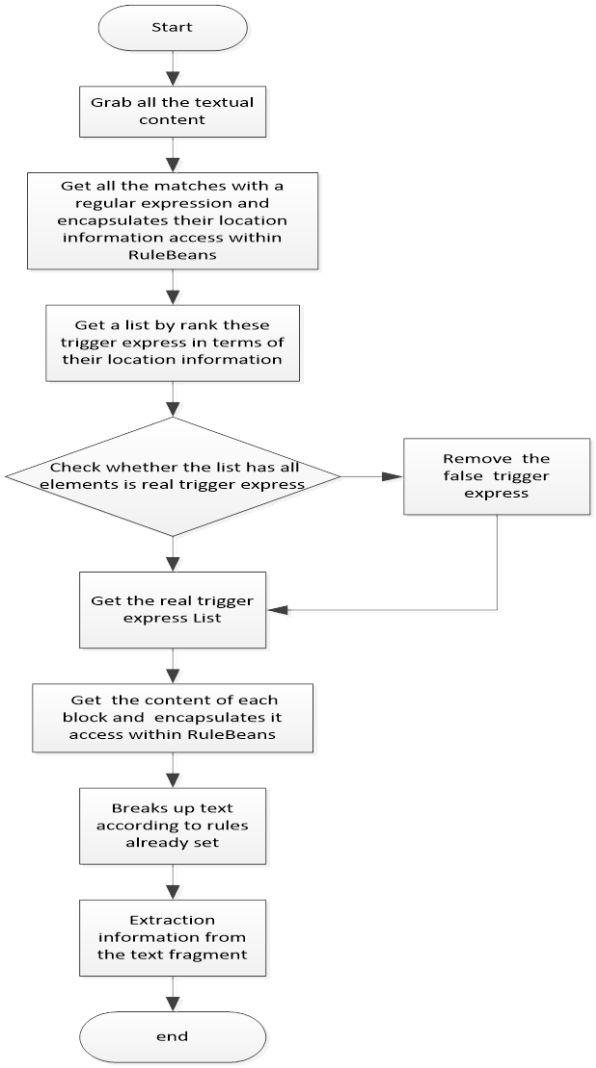

Fig. 6 text parsing process

4 Resume Update

When we extract resume information, once the network is changed, the resume data in the database will not be reliable, so it is necessary to update the data in a timely manner according to the latest record of network.

We use the design pattern introduced in section 2, class “InitServlet” in Figure 7 is used to initialize configuration parameters, “QueryServlet” provide the operation method for the resume, which include query, view details, update the data, and put the data into the database. As shown in figure 7 the view expressed human-computer interaction interface.

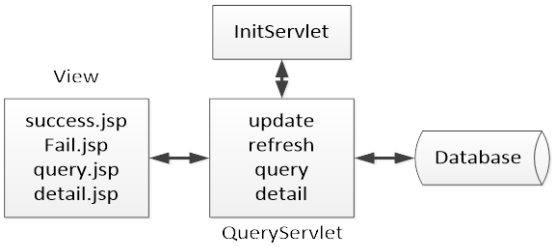

Fig. 7 resumes update structure

5 Main result

5.1 local resume abstraction test

We can see the part of rule expression in Figure 8, as shown in this paper, the 96 resumes of different sources and different format is parsed in a predetermined required format, then we validate the analytical ability and the quality of the results. The part of extraction results is shown in the figure 9 and figure 10.

id	type	resumecol	express	istrigger	createtime
20	姓名	name	姓名	0	2015-06-28 07:31:58
21	性别	sex	性别	0	2015-06-28 07:32:16
22	户口	registeredResidence	户口所在地	0	2015-06-28 11:36:10
23	出生日期	birthday	出生日期	0	2015-06-28 07:32:56
24	出生日期	birthday	出生年月日	0	2015-06-28 07:32:25
25	婚姻状况	isMarried	婚姻状况	0	2015-06-28 07:42:13
26	毕业时间	graduationTime	毕业时间	0	2015-06-29 09:47:19
27	毕业院校	university	毕业院校	0	2015-06-29 13:56:24
28	简介	(Null)	简介	0	2015-06-25 16:42:20
29	学历	degrees	学历	0	2015-06-29 10:01:06
30	专业	major	专业	0	2015-06-28 19:22:48
31	学位	academicDegrees	学位	0	2015-06-29 16:53:37
32	期望工作地点	expectedWorkingplace	期望工作地点	0	2015-06-28 19:20:10
33	工作年限	workingLife	工作年限	0	2015-06-29 10:10:34
34	联系电话	mobileNumber	联系方式	0	2015-06-28 07:34:11
35	联系电话	mobileNumber	tel	0	2015-06-28 07:35:33
36	户口	registeredResidence	生源地	0	2015-06-28 11:36:10
37	户口	registeredResidence	户口	0	2015-06-28 11:36:10
38	出生日期	birthday	生日	0	2015-06-28 07:40:14
39	是否应届	(Null)	是否应届	0	2015-06-25 16:44:10
40	联系地址	currentResidence	居住地	0	2015-06-28 07:33:12
41	联系地址	currentResidence	现住地	0	2015-06-28 07:33:13
42	联系地址	currentResidence	现住址	0	2015-06-28 07:33:14

Fig. 8 rule expressions

As shown in Figure 8, there is part of pre-established rule expression in the database.

workinglife	annualsalary	expectedsalary
9年工作经验	(Null)	25000元/月以上
10年以上工作经验	15-30万人民币	20000-29999/月
3-4年工作经验	(Null)	10000-14999/月
7年工作经验	(Null)	不显示职位月薪范围
1年工作经验	6-8万人民币	10000-14999/月
8年工作经验	(Null)	10001-15000元/月
1年工作经验	5-6万人民币	6000-7999/月

Fig. 9: fields in result

school	major
(Null)	(Null)
西北工业大学	计算机科学与技术
安徽科技学院	计算机科学与技术
(Null)	(Null)
西安邮电大学	信息与计算科学
(Null)	(Null)
东华理工大学	软件工程

Fig. 10: fields in result

Figure 9 and Figure 10 are part of extraction result set, and each column represents a field of resume. In order to facilitate extract quality measure, define the following variables and assessment indicators:

- Fields Number (FN): Fields number for all the items in the Resume.
- Effective Fields Number (EFN): FN subtracts the number of the fields that is not defined in the standard resume library field.
- Analytic fields number (AFN): The number of fields abstracted from the resume.
- Effective Analytic fields number (EAFN): The number of reasonable fields viewed from the semantics point.
- Recall rate：EAFN/FN *100%
- accuracy rate：AFN/FN*100%

So, we get the following results shown in Table 3 and 4.

Table 3: Variable Value

Resume id	Is it parsed?	FN	EFN	AFN	EAFN
1	YES	26	18	18	17
2	YES	30	20	19	18
3	YES	29	21	21	20
4	YES	28	21	19	18
5	YES	28	24	21	20
6	YES	26	21	20	19
7	YES	29	21	21	20
8	YES	27	19	19	18
…	…	…	…	…	…
20	YES	28	21	21	20
26	YES	30	21	21	21
27	YES	29	20	20	20
…	…	…	…	…	…

Table 4: indicators

Resume type	Number	Recall rate	accuracy rate
Zhilian	19	99.72%	99.72%
51job	10	96.63%	91.83%
IfChang	23	92.02%	90.43%
Other	44	88.19%	85.3%
Total row	96	93.75%	91.71%

On the whole, the resume parsing quality is good; especially the parsing of resume coming from big recruitment website is of high quality. What affect the overall quality is part of the resume in the "other" category, include layout and style, diversified and ambiguous item name and other causes.

5.2 Network update test

For this reason that recruitment website have security restrictions, the Resume update test work is restricted to login here temporarily by using semi-automatic pattern to update resume in the test environment when the recruitment website has any updated data.

As shown in the figure 11, the first URL is the entrance of query information.

```
1 http://localhost:8080/TencentResume/queryservlet?func=query
2 http://localhost:8080/TencentResume/refresh.jsp
```

Fig. 11 URLs of entrance

Then check the details of resume as shown in figure 12.

Resume field	Text content
到岗时间 (Report Time)：	我目前在职，正考虑换个新环境 (I am currently working and are considering changing jobs)
当前居住地 (Current Residence)：	北京 朝阳区 (Chaoyang District,Beijing)
期望工作地 (Expect Workplace)：	广州 (Guangzhou)
个人简介 (Personal Profile)：	test...
…	…
语言技能 (Language)：	英语：读写能力一般|听说能力一般 (English：Reading and writing ability in general | Listening and speaking ability in general)
…	…

Fig. 12 resumes details before update

Change the Residence and Language in the website as shown in figures 13 and 14.

Fig. 13 modify address

Fig. 14 modify language

Then, click the second URL in the figure 11 then we can see the interface in the figure 15.

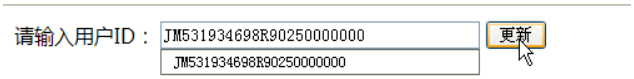

Fig. 15 update interface

Finally, we can obtain the refreshed resume information as shown in the figure 16.

Resume field	Text content
到岗时间 (Report Time)：	我目前在职，正考虑换个新环境 (I am currently working and are considering changing jobs)
当前居住地 (Current Residence)：	深圳 南山区 (Nanshan District.Shenzhen)
期望工作地 (Expect Workplace)：	广州 (Guangzhou)
个人简介 (Personal Profile)：	test...
...	...
语言技能 (Language)：	英语：读写能力一般|听说能力一般 (English:Reading and writing ability in general | Listening and speaking ability in general) 日语：读写能力熟练|听说能力熟练 (Japanese:Excellent reading and writing skills | Excellent listening and speaking skills) 韩语：读写能力熟练|听说能力熟练 (Korean:Excellent reading and writing skills | Excellent listening and speaking skills)
...	...

Fig. 16 resumes details after update

By observing contents of these two fields Residence and Language, we can see that resume on the website is parsed successfully and it is viable to update the local resume.

6 Conclusion and optimization

We report on experiments over 96 Chinese resume in different forms that all files are resolvable, in other words, resume parsing ratio is 100%. We report statistics on 2784 available resume fields, and show that the overall recall rate is 93.75%, and the accuracy rate is 91.71%. According to these two indicators, we obtained better results when compared to the method employed in [7,11]. In addition, we have realized the goal of update local resume data. However, there are many optimization problems exist, such as extending the resume structure model, adjusting the rules expression and doing the further analysis.

References

[1] Amato F., Boselli R., Cesarini M., et al. Challenge: Processing web texts for classifying job offers[C]//Semantic Computing (ICSC), 2015 IEEE International Conference on. IEEE, 2015: 460-463.

[2] Cafarella Michael J., Banko Michele, Etzioni Oren. Open information extraction from the web [P].：US8938410, 2015-01-20.

[3] Gaikwad S V, Chaugule A, Patil P. Text mining methods and techniques[J]. International Journal of Computer Applications, 2014, 85(17).

[4] Gupta R. Journey from data mining to Web Mining to Big Data [J]. arXiv preprint arXiv:1404.4140, 2014.

[5] Gong Yiguang, Mei Ping.Research on a Combined Ontology-based Text Information Extraction Technology [A]. //Proceedings of 2010 International Conference on Broadcast Technology and Multimedia Communication (Volume 4) [C]. International Communication Sciences Association, Hong Kong: 2010:4: 129-132.

[6] Hobbs J R. Information extraction [M]//Handbook of Natural Language Processing, Second Edition. Chapman and Hall/CRC, 2010: 511-532.

[7] Jiang ZhiXiang, Identification of the Semi-structured text [D]. Beijing University of Posts and Telecommunications, 2009.

[8] Kumar L, Bhatia P K. Text Mining: Concepts, Process and Applications [J]. Journal of Global Research in Computer Science, 2013, 4(3): 36-39.

[9] Liu Qian，Jiao Hui， Jia Huibo，Research on Approaches of Information Extraction System [J]. Application Research of Computers，2007， 24(7):6-9.

[10] MacLennan C J. Text based schema discovery and information extraction: U.S. Patent 7,930,322[P]. 2011-4-19.

[11] Mu YunHe, Research on Resume and SRS Information Extraction Method [D], Shanghai Jiao Tong University,2010.

[12] Pajić V, Lažetić G P, Pajić M. Information extraction from semi-structured resources: a two-phase finite state transducers approach[C]//International Conference on Implementation and Application of Automata. Springer Berlin Heidelberg, 2011: 282-289.

[13] Sukanya M, Biruntha S. Techniques on text mining[C]//Advanced Communication Control and Computing Technologies (ICACCCT), 2012 IEEE International Conference on. IEEE, 2012: 269-271.

[14] Sleiman H A, Corchuelo R. Tex: An efficient and effective unsupervised web information extractor [J]. Knowledge-Based Systems, 2013, 39: 109-123.

[15] Sleiman H A, Corchuelo R. A class of neural-network-based transducers for web information extraction [J]. Neurocomputing, 2014, 135: 61-68.

[16] Strohmeier S, Piazza F. Artificial Intelligence Techniques in Human Resource Management—A Conceptual Exploration [M]//Intelligent Techniques in Engineering Management. Springer International Publishing, 2015: 149-172.

[17] Venkateswaran D C J, Murugan S, Radhakrishnan D N. An Useful Information Extraction using Image Mining Techniques from Remotely Sensed Image (RSI)[J]. International Journal on Computer Science and Engineering, 2010, 2(08): 2577-2580.