# Research on Personal Education Background Extraction Using Rules

http://dx.doi.org/10.3991/ijet.v8i5.3029

Z.M. Zhong and C.H. Li
Huaihai Institute of Technology, Lianyungang, China

***Abstract*—With the explosive growth of Internet information, how to obtain the required information from the vast amounts of text information is becoming an important issue today. Extraction of personal attribute has made considerable progress, including name, sex, place of birth, date of birth, related events, etc. But the extraction of personal education background information does not arouse researcher's enough attention. The attribute of personal education background is very complex in text and involves all kinds of education structures such as primary school, middle school and university. We put forward a new method of extracting personal attributes from texts based on rules. Firstly, rules of personal education background are formulated after analyzing a lot of texts about people. Secondly, the related algorithm is designed to extract personal education background from unstructured texts based on rules. Finally, the experiment is implemented for 100 documents about people. The results show that the average precision of extracting personal education background is 0.898, the average recall is 0.8959, and the average F-measure is 0.8968.**

***Index Terms*—personal attribute extraction, personal education background extraction, personal education background rules**

## I. INTRODUCTION

In a deluge of Internet information, the personal information also shows explosive growth, but data rich and poor information. Nowadays, the main source of getting information is still text type data, and how to extract useful information from mass of texts is gradually becoming the hot issue of concern. There are some researches on personal attribute extraction, such as literature [1] and [2]. But for these literatures, education background is just a part of personal attributes, and there is no specialized research on education background. Personal education background attribute is very complex, and it is not easy to extract accurate education background from texts.

We formulate personal education background rules after collecting and analyzing a lot of character texts. The method proposed in this article can accurately realize personal education background attribute extraction.

Personal attribute extraction is the base for biographies and people search engines, and it directly affects the biographies and people search engines to "specialized, refined, and deep" the direction of development.

The research in this paper is the sub-direction of personal attribute extraction, which will provide paradigms and ideas for other attribute extraction of people.

## II. RELATED WORKS

Automatic summary has a history of over fifty years since H.P. Luhn [3] initiated automatic summary in 1958. In 2001, Schiffman et al. [4] used corpus statistics along with linguistic knowledge to select and merge descriptions of people from a document collection such as sex, date of birth and education background. Because the model of personal information is relatively fixed, some researchers constructed people ontology to guide personal information extraction. In 2007, Han et al. [5] applied OWL ontology description language to construct event ontology including fixed information and variable information, and an event is composed of people, time, location and content. Zhou et al. [6] designed a biographical summarization system using sentence classification and ideas from information retrieval. First, the sentences are classified into biographical and non-biographical. Using the methods of machine learning to classify the sentences, the sentences describing people are reserved. The importance of sentences is measured by words' ITF (inverse-term-frequency). Filatova et al. [7] applied the notion of atomic event to extract people's information, and generated biographical summarization according to people's occupations. They believe that the biography should contain important events, and events posses close relations with occupations. The important events relating to people are measured using statistic method.

The requirement of getting personal information is becoming increasingly important. Some search engines have been developed to provide people-oriented information service [8]. For example, Ask jeeves (http://www.ask.com) has expanded its Smart Search feature, adding "direct answers" with biographical information about famous people. ZoomInfo (http://www.zoominfo.com/people) is the most comprehensive source of business information on people. And Sogou (http://people.sogou.com) distinguishes the information of famous and non-famous people, for non-famous people, just returning the information registered by users.

Zhong et al. [9] present a method of identifying key people from a single document. After analyzing and quantifying three kinds of associative strengths between people and events, between people and between events, a people event map is constructed to represent a document, and the people's importance can be computed using the PageRank algorithm based on people event map.

## III. PERSONAL EDUCATION BACKGROUND ATTRIBUTE EXTRACTION USING RULES

### A. System Architecture

The System architecture of personal education background attribute extraction using rules is shown in Fig. 1.

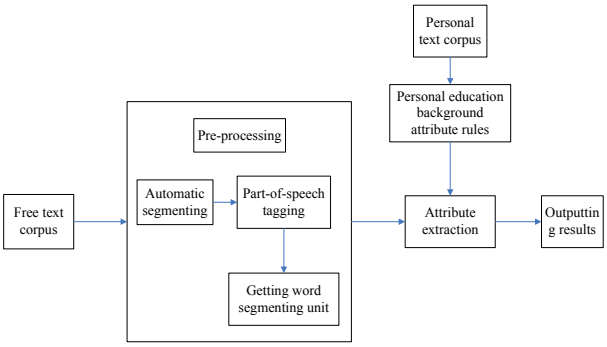


Figure 1. System architecture

The system architecture in Fig.1 mainly includes three parts: (1) collecting specialized personal corpus from Web such as chemist, mathematician and physicist, (2) constructing personal education background attribute rules, and (3) extracting personal education background attributes from texts.

### B. Formulating Personal Education Background Attributes Rules

ICTCLAS, developed by the institute of computing technique of Chinese academy of sciences, is a Chinese lexical analysis system containing the functions of words segmenting, part-of-speech tagging, and unlisted words identifying. The technique of Chinese personal name recognition is based on role tagging. The accuracy of word segmenting reaches 97.58%, and F_value of person recognition achieves 95.4% [10].

Definition 1. Word segmenting unit means a word and its part-of-speech. For example, 'Hu Jintao' and 'nr' (name) is a word segment unit.

There are two matching types for word segmenting unit. One is underlying word matching, and the other is part-of-speech matching. For example, a word '生于(birth)' can not only match with '生于(birth)' of segmenting word unit '生于 v', but match with 'v'.

Word segmenting unit supports 'OR' relation. For example, '.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校' can match with '..中学 n', '..附中 j' and '..学校 n'.

The rules of personal education background attributes are classified into two types.

(1) 'Search' with terminator and trigger

Search 'segmenting word unit 1' forward from 'segmenting word unit 2' ends with terminator;

Search 'segmenting word unit 1' backward from 'segmenting word unit 2' ends with terminator.

(2) 'Extract'

Extract 'segmenting word unit 1' between 'segmenting word unit 2' and 'segmenting word unit 3'.

After analyzing a lot of person corpus, the formulating personal education background rules are shown in Table Ⅰ, Table Ⅱ and TableⅢ.

TABLE I.
PRIMARY SCHOOL' RULES OF PERSONAL EDUCATION ATTRIBUTES

| No. | Rules | No. | Rules |
|---|---|---|---|
| 1 | Extract<br>.{0,}&.{0,}&.{0,}&.{0,}&.{0,}小学 Between<br>在\|进入 And 读书\|上学\|学习\|读 | 10 | Search<br>@ns..\|@nz..&.{0,}&.{0,}&.{0,}& .{0,}小学<br>Forward from 毕业&于 Ends with ，,。,！,. |
| 2 | Extract<br>.{0,}&.{0,}&.{0,}&.{0,}&.{0,}小学 Between 毕业&于 And.{0,} | 11 | Search<br>@ns..\|@nz..&.{0,} &.{0,}&.{0,}小学 Backward from 读书\|上学\|学习\|毕业\|读 Ends with ，,。,！,. |
| 3 | Extract<br>.{0,}&.{0,}&.{0,}&.{0,}小学 Between<br>在\|进入 And 读书\|上学\|学习\|读 | 12 | Search<br>@ns..\|@nz..&.{0,}&.{0,}&.{0,}小学<br>Forward from 毕业&于 Ends with，,。,！,. |
| 4 | Extract<br>.{0,}&.{0,}&.{0,}&.{0,}小学 Between 毕业&于 And .{0,} | 13 | Search<br>@ns..\|@nz..&.{0,}&.{0,}小学 Backward from 读书\|上学\|学习\|毕业\|读 Ends with ，,。,！,. |
| 5 | Extract<br>.{0,}&.{0,}&.{0,}小学 Between 在\|进入 And 读书\|上学\|学习\|读 | 14 | Search<br>@ns..\|@nz..&.{0,}&.{0,}小学 Forward from 毕业&于 Ends with ，,。,！,. |
| 6 | Extract<br>.{0,}&.{0,}&.{0,}小学 Between 毕业&于 And.{0,} | 15 | Search<br>@ns..\|@nz..&.{0,}小学 Backward from 读书\|上学\|学习\|毕业\|读 Ends with ，,。,！,. |
| 7 | Extract<br>.{0,}&.{0,}小学 Between 在\|进入 And 读书\|上学\|学习\|读 | 16 | Search<br>@ns..\|@nz..&.{0,}小学 Forward from 毕业&于 Ends with，,。,！,. |
| 8 | Extract<br>.{0,}&.{0,}小学 Between 毕业&于 And .{0,} | 17 | Search<br>.{0,}小学 Backward from 读书\|上学\|学习\|毕业\|读 Ends with，,。,！,. |
| 9 | Search<br>@ns..\|@nz..&.{0,}&.{0,} &.{0,}&.{0,}小学 Backward from 读书\|上学\|学习\|毕业\|读 Ends with ，,。,！,. | 18 | Search<br>.{0,}小学 Forward from 毕业&于 Ends with ，,。,！,. |

TABLE II.
'MIDDLE SCHOOL' RULES OF PERSONAL EDUCATION ATTRIBUTES

| No. | Rules | No. | Rules |
|---|---|---|---|
| 1 | Extract .{0,}&.{0,}&.{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 在|进入 And 读书|上学|学习|求学|读 | 10 | Search @ns..|@nz..&.{0,} &.{0,}& .{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Forward from 毕业&于 Ends with ，,。,！,. |
| 2 | Extract .{0,}&.{0,}&.{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 毕业&于 And .{0,} | 11 | Search @ns..|@nz..&.{0,} &.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Backward from 读书|上学|学习|毕业|求学|读 ends with ，,。,！,. |
| 3 | Extract .{0,}&.{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 在|进入 And 读书|上学|学习|求学|读 | 12 | Search @ns..|@nz..&.{0,} &.{0,}&.{0,}大学|.{0,}学院 Forward from 毕业&于 Ends with，,。,！,. |
| 4 | Extract .{0,}&.{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 毕业&于 And .{0,} | 13 | Search @ns..|@nz..&. {0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Backward from 读书|上学|学习|毕业|求学|读 Ends with ，,。,！,. |
| 5 | Extract .{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 在|进入 And 读书|上学|学习|求学|读 | 14 | Search @ns..|@nz..&.{0,}& .{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Forward from 毕业&于 Ends with ，,。,！,. |
| 6 | Extract .{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 毕业&于 And .{0,} | 15 | Search @ns..|@nz..&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Backward from 读书|上学|学习|毕业|求学|读 ends with ，,。,！,. |
| 7 | Extract .{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 在|进入 And 读书|上学|学习|求学|读 | 16 | Search @ns..|@nz..&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Forward from 毕业&于 Ends with ，,。,！,. |
| 8 | Extract .{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Between 毕业&于 And .{0,} | 17 | Search .{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Backward from 读书|上学|学习|毕业|求学|读 Ends with ，,。,！,. |
| 9 | Search @ns..|@nz..&.{0,}&.{0,}&.{0,}&.{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Backward from 读书|上学|学习|毕业|求学|读 Ends with ，,。,！,. | 18 | Search .{0,}中学|.{0,}高中|.{0,}附中|.{0,}学校 Forward from 毕业&于 Ends with ，,。,！,. |

TABLE III.
'UNIVERSITY' RULES OF PERSONAL EDUCATION ATTRIBUTES

| No. | Rules | No. | Rules |
|---|---|---|---|
| 1 | Extract .{0,}&.{0,}&.{0,}&.{0,}&.{0,}&.{0,} 大学|.{0,}学院 Between 在|进入 And 读书|上学|学习|读 | 10 | Search @ns..|@nz.. &.{0,} &.{0,} &.{0,}&.{0,} 大学|.{0,}学院 Forward from 毕业&于 Ends with ，,。,！,. |
| 2 | Extract .{0,}&.{0,}&.{0,}&.{0,}&.{0,}&.{0,} 大学|.{0,}学院 Between 毕业&于 And.{0,} | 11 | Search @ns..|@nz..&.{0,} &.{0,}&.{0,} 大学|.{0,}学院 Backward from 读书|上学|学习|毕业|读 Ends with，,。,！,. |
| 3 | Extract .{0,}&.{0,}&.{0,}&.{0,} 大学|.{0,}学院 Between 在|进入 And 读书|上学|学习|读 | 12 | Search @ns..|@nz..&.{0,} &.{0,}&.{0,} 大学|.{0,}学院 Forward from 毕业&于 Ends with，,。,！,. |
| 4 | Extract .{0,}&.{0,}&.{0,}&.{0,} 大学|.{0,}学院 Between 毕业&于 And.{0,} | 13 | Search @ns..|@nz..&.{0,} &.{0,} 大学|.{0,}学院 Backward from 读书|上学|学习|毕业|读 Ends with，,。,！,. |
| 5 | Extract .{0,}&.{0,}&.{0,} 大学|.{0,}学院 Between 在|进入 And 读书|上学|学习|读 | 14 | Search @ns..|@nz..&.{0,} &.{0,} 大学|.{0,}学院 Forward from 毕业&于 Ends with，,。,！,. |
| 6 | Extract .{0,}&.{0,}&.{0,} 大学|.{0,}学院 Between 毕业&于 And.{0,} | 15 | Search @ns..|@nz..&.{0,} 大学|.{0,}学院 Backward from 读书|上学|学习|毕业|读 Ends with，,。,！,. |
| 7 | Extract .{0,}&.{0,} 大学|.{0,}学院 Between 在|进入 And 读书|上学|学习|读 | 16 | Search @ns..|@nz..&.{0,} 大学|.{0,}学院 Forward from 毕业&于 Ends with，,。,！,. |
| 8 | Extract .{0,}&.{0,} 大学|.{0,}学院 Between 毕业&于 And.{0,} | 17 | Search .{0,} 大学|.{0,}学院 Backward from 读书|上学|学习|毕业|读 Ends with ，,。,！,. |
| 9 | Search @ns..|@nz..&.{0,}&.{0,}&.{0,}&.{0,} 大学|.{0,}学院 Backward from 读书|上学|学习|毕业|读 Ends with ，,。,！,. | 18 | Search .{0,} 大学|.{0,}学院 Forward from 毕业&于 Ends with ，,。,！,. |

## IV. EXPERIMENTS AND EVALUATION

### A. Experimental Corpus

The experimental data are three kinds of people: scientist, physical star and politician. We select famous people for each kind of people and collect 100 documents manually in total. Accordingly, these data and reference answers are creditable. The detailed corpus is shown in Table Ⅳ.

TABLE IV.
SELECTED CORPUS

| Personal corpus types | Number of corpus | Answer number of primary school | Answer number of middle school | Answer number of university |
|---|---|---|---|---|
| scientist | 40 | 0 | 9 | 47 |
| physical star | 30 | 2 | 15 | 21 |
| politician | 30 | 4 | 10 | 19 |
| Total | 100 | 6 | 34 | 87 |

### B. Evaluating Indexes

We use F_value to evaluate experimental results, and the method of computing F_value is shown in Equation (1):

$$F = \frac{P \times R \times 2}{P + R} \tag{1}$$

$$P = \frac{U}{V} \tag{2}$$

$$P = \frac{U}{V} \tag{3}$$

In Equation (2) and (3), U is the number of extracted correct education background attributes, V is extracted education background attributes, and W is the sum of personal education background attributes of standard answers.

### C. Experimental Results

Using the method proposed in this paper for 100 texts, the obtained results are shown in Table Ⅴ.

TABLE V.
EXPERIMENTAL RESULTS OF EXTRACTING PERSONAL EDUCATION BACKGROUND

| Degree Stages | Number of correct answers | Number of extracted answers | Number of extracted correct answers | P | R | F_value |
|---|---|---|---|---|---|---|
| Primary School | 6 | 6 | 5 | 0.8333 | 0.8333 | 0.8333 |
| Middle School | 34 | 33 | 31 | 0.9394 | 0.9118 | 0.9254 |
| University | 87 | 89 | 82 | 0.9213 | 0.9425 | 0.9318 |
| Average P, R and F_value | | | | 0.8980 | 0.8959 | 0.8968 |

From Table Ⅴ, the performance of extracted education background is ideal for three stages: ‘primary school’, ‘middle school’ and ‘university’, and the average P, R and F_value are 0.8980, 0.8959, and 0.8968 respectively. From the experiment, the ‘primary school’ attribute is obviously less than that of ‘middle school’ and ‘university’, and ‘university’ attribute is the majority.

## V. CONCLUSIONS

We propose a new method of extracting personal education background attribute using rules. The concept of word segmenting unit is defined in order to simplify the algorithm process. Word segmenting unit is a fine operation for each word. The research result in this paper is just a sub-area of personal attribute extraction. The future word includes: multi-personal attribute extraction from a document, integrating personal attributes from multi-documents, and other attributes such as marriage, hobby, contact method and related events.

## ACKNOWLEDGMENT

This research (paper) is supported by the National Natural Science Foundation of China (No.60975033), and the Science and Technology Foundation of Lianyungang (CG1121).

## REFERENCES

[1] Z. M. Zhong, C. H. Li, and L. Qiao, “A Method of Personal Attribute Extraction Combining Rules and Statistics”, *AISS*, Vol.4, pp.79-86, 2012. http://dx.doi.org/10.4156/aiss.vol4.issue18.10

[2] B. Wang and C. Gao, “Automatic Summarization Method for Chinese Document based on Comprehensive Background Concept Lattice”, *JCIT*, Vol.6, pp.390-399, 2011. http://dx.doi.org/10.4156/jcit.vol6.issue6.40

[3] H. P. Luhn, “The Automatic Creation of Literature Abstracts”, *IBM Journal of Research and Developments,* Vol.2, pp. 159-165, 1958. http://dx.doi.org/10.1147/rd.22.0159

[4] S. Barry, M. Inderjeet, and C. Kristian, “Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics”, *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL’2001)*, New Brunswick, New Jersey, pp. 450-457, 2001.

[5] Y. J. Han, “Reconstruction of People Information based on an Event Ontology”, *In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, NLP-KE’07, Beijing, China, pp. 446-451, 2007.

[6] L. Zhou, M. Ticrea, and E. Hovy, “Multi-document Biography Summarization”, *In Proceedings of EMNLP*, Barcelona, Spain, pp. 434-441, 2004.

[7] E. Filatova and J. Prager, “Tell Me What You Do and I’ll Tell You What You Are: Learning Occupation-Related Activities for Biographies”, *In Proceedings of HLT/EMNLP 2005,* Vancouver, Canada, pp. 49-56. 2005.

[8] N. Ren, "Personal Position and Title Information extraction in Large-Scale Real Texts", *Dissertation of Beijing Language and Culture University*, pp. 4-7, 2008.

[9] Z. M. Zhong, Z. T. Liu, C.H. Li, and W. Zhou, "Identifying Key People from A Single Document Using People Event Map", *Journal of Computational Information Systems*, Vol. 6, pp. 17-23, 2010.

[10] H. P. Zhang, Q. Liu, "Automatic Recognition of Chinese Personal Name based on Role Tagging", *Journal of Computers*, Vol. 27, pp. 85-91, 2004.

## AUTHORS

**Z. M. Zhong** is with School of Computer Engineering, Huaihai Institute of Technology Lianyungang, China (e-mail: zhongzhaoman@163.com).

**C. H. Li** is with School of Computer Engineering, Huaihai Institute of Technology Lianyungang, China (e-mail: cli@hhit.edu.cn).

Manuscript received 09 August 2013. Published as re-submitted by the authors 13 October 2013.