Chinese Researcher Profile Annotation Based on Conditional Random Fields with Semantic Rules

Sun Jian, Xu Jungang, and Cen Zhiwang

***Abstract*—There are many researcher homepages on Web, if one wants to process researcher information for search engine, building a semantic profile for the academic researcher to identify and annotate information is an effective method. In this paper, we label Chinese researcher information with Conditional Random Fields (CRF) model, which has achieved good performance on Named Entity Identification. We proposed a hybrid annotation method which combines Conditional Random Fields and semantic rules, considering some features such as suffix, prefix, and semantic features of named entity at the same time. The comparison experiments show that this method can correctly extract the real content of the Chinese researcher homepages and assign a suitable category label to each part of the contents simultaneously.**

***Index Terms*—Semantic web, ontology, conditional random fields, feature selection, annotation**

I. Introduction

IF we plan to design a researcher search engine, we must parse and annotate the researcher profile from researcher homepages firstly. Traditionally, personal profile annotation is viewed as an engineering issue and is conducted manually. Some annotation tools provide a criterion environment for users, and they can create label content according to his/her profile. Other methods are used semantic annotation technology label different types of information in a separated fashion with the given redefined rules or special machine learning models [1].

Normally, researcher information is described as a hierarchical structure with two layers. The first layer is composed of general information blocks such as personal information, education background, work experience, academic achievements, award, project and etc. The second layer is detailed pieces in each general information block, such as name, address, research area, paper and etc. , see table 1.

Table 1. Two layers of researcher information.

The First Layer	The Second Layer
Personal information	Name, gender, birthday, address, zip code, Phone, fax, mobile, email, affiliation, degree, position, research area
Education background	Date, institution, department, major, degree
Work experience	Date, institution, department, position
Academic achievement	Authors, title, journal, volume, pages, conference, location, date, editors, book, publisher, technology, patent
Award	Date, title, institution, prize, rank
Project	Date, title, institution, category, position

Extracting information from researcher homepage is not an easy task. In spite of constituting a restricted domain, profile can be written in multitude of formats (e.g. structured tables, list or plain texts) and in different languages (e.g. Chinese and English). Moreover, written styles could be much diversified. A hybrid label method based on Conditional Random Fields is proposed in this paper. We collect the education background and academic achievement block information as the experiment data. We consider the Chinese words and semantic features. And the label predicting result is higher than the baseline method.

The remainder of this paper is organized as follows. Section II describes the related work. Section III introduces the principles of Conditional Random Fields. Section IV explains the process of experiments and the result of comparison. Finally, Section V summarizes and proposes the future work.

II. Related Work

It's a beneficial for many web applications to extract researcher profile information in search engine. Several researcher efforts have been made. Y. Kun, G. Gang, and Z. Ming have proposed resume information extraction with cascaded hybrid model [2]. Y. Limin, T. Jie, and L.Juanzi

Manuscript received March 9, 2011; revised April 4, 2011. This work was supported in part by the National Hi-Tech Research and Development Program (863 Program) of China (No. 2009AA01Z128) and the President Fund of Graduate University of Chinese Academy of Sciences (GUCAS) (No. Y05101DY00)

Sun Jian is with the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences, Beijing, 100190 China (e-mail:jiansun6000@gmail.com).

Xu Jungang was with Tsinghua University, Beijing, 100084 China. He is now with the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences, Beijing, 100190 China (phone: 0086-10-86717689;fax:0086-10-82649882;e-mail:xujungang@gmail.com) .

Cen Zhiwang is with the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences, Beijing, 100190 China (e-mail:cenzhiwang@gmail.com).

have proposed a unified approach to researcher profiling [3]. The previous extracting information method can mainly be divided into automatic and statistical methods.

The automatic method is suitable for the structured data and rules. This annotation method is based on predefined rules, such as KIM (Knowledge and Information Management) [4], which have a good efficiency for annotating document, but just work when processing single type of document in a single field. Balog and Rijke employed heuristic rules to extract contact information from emails [5]. F. Ciravegna proposed an adaptive algorithm for information extraction from web-related texts [6], and developed an automated semantic annotation module named Amilcare. However, this method has some defects, such as low efficiency, poor quality in rule learning.

The statistics method can obtain good result in named entity identification. Machine learning methods such as hidden Markov model [7], the maximum entropy model [8], and support vector machine [9] have already been applied in named entity identification. And CRF method [10]–[13] has achieved good results in solving some problems of named entity extraction. Zhang Suxiang used Conditional Random Fields to recognize person with multi-features [14].

CRFs model is trained by the schema of sequence label, which can't understand the content, but semantic technology may solve this problem. Ding Shengchun and Jiang Ting have proposed the comment target extraction with Conditional Random Field and domain ontology [15].

III. The Theory of Conditional Random Fields

A. Conditional Random Fields

The model of Conditional Random Fields is an undirected graph. The primary advantage of CRFs over Hidden Markov Model (HMM) is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models.

For our information annotating, we define an CRF model, implementing a kind of mapping from word, phrase or sentence sequence $X=\{x_1,x_2,\cdots,x_n\}$ (generally used as sentence) to label sequence $Y=\{y_1,y_2,\cdots,y_n\}$, y_i represents the key value of hidden state variable.

The probability of the label sequence Y with the observation sequence X is defined as

$$p(y\mid x)=\frac{1}{z(x)}\exp(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(y_{i-1},y_i,x,i)) \quad (1)$$

Where $z(x)$ is a normalization factor, which make the sum of all possible probability of labeling sequence is 1.

The formula of $z(x)$ is defined as

$$z(x)=\sum_{y\in Y}\exp(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(y_{i-1},y_i,x,i)) \quad (2)$$

Y is the set of all possible probability of labeling sequence, n indicts the length of given sequence, $f_k(y_{i-1},y_i,x,i)$ is the feature function, which describes any dependent characteristics, both edge and vertex feature of undirected graph, λ_k is the weight factor of the k^{th} feature function.

We separate the observation features from making the definition of feature selection for convenience. Taking the location i as an example, we define the observation feature as follows.

$$o(x,i)=\begin{cases}1 & if\ x_i="获得(achieve)"\ and\ x_i="博士学位(Ph.D\ degree)"\\ 0 & otherwise\end{cases} \quad (3)$$

Where x_i represents the word of location i. So we can define the feature of edge e and vertex v as follows.

$$e(y_{i-1},y_i,x,i)=\begin{cases}o(x,i) & if\ y_{i-1}="O"\ and\ y_i="B-degree"\\ 0 & otherwise\end{cases} \quad (4)$$

$$v(y_i,x,i)=\begin{cases}o(x,i) & if\ y_i="B-degree"\\ 0 & otherwise\end{cases} \quad (5)$$

Different features with different weights can be obtained from training corpus. Assumed that one CRF model is defined according to formula (1), the most possible mark sequence can be defined according to formula (6) as follows.

$$y^*=\arg\max_{y\in Y}p_\lambda(y\mid x) \quad (6)$$

B. The construction of researcher profile

The data resource of research work is researcher profile, including concept and its properties. Our statistical study on one thousand researchers shows that 82.4% of the profiles are from the institutions and others are from universities.

According to the characteristics of researcher profile, the researcher ontology will be constructed, including concepts, properties and relations, see Fig. 1.

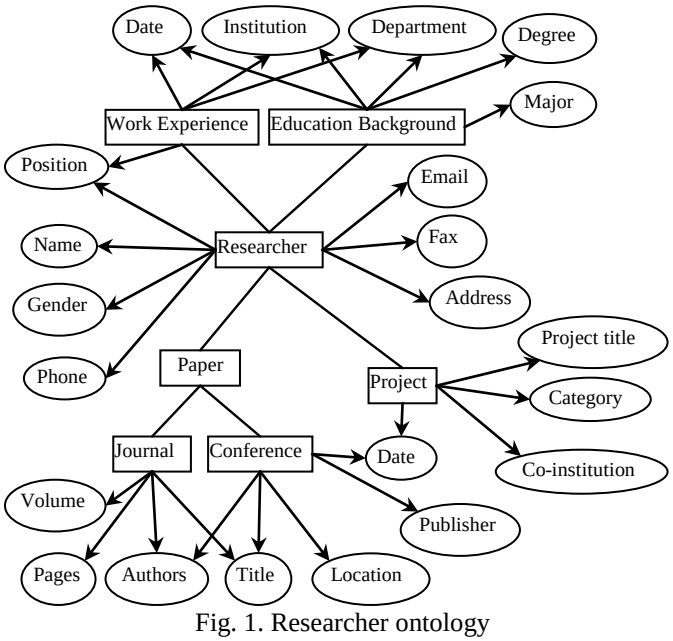

Fig. 1. Researcher ontology

C. Feature Selection

1) Base feature

a) BIO-label

We used the BIO chunk method to identify named entity. Every word (character) will be classified three categories.

B-entity refers to the first character or word of one entity, I-entity refers to the other word or character of the entity, and O refers to the irrespective word. Here, entity recognition can be converted into classifying B-entity, I-entity and O task. The entities we will recognize include institution, department, major, paper, title, journal, conference and etc.

For example, there is one sentence like "张海在中国科学院自动化所获得博士学位", so segmentation words include '张', '海', '在', '中国科学院', '自动化所', '获得', '博士', '学位'. So, the BIO-labels include '张/B-person 海/I-person', '在/O', '中国科学院/B-institution', '自动化所/I-institution', '获得/O', '博士'/B-degree, '学位/I-degree'.

b) Context feature

The context information can help us identify named entity. For example, "李浩/博士/从事/人工/ 智能/研究 (Ph.D. Li Hao is engaged in the artificial intelligence research)", the relation between the context information and implied word (or character) used for named entity is particularly important. We have to specify the feature templates in advance. This file describes which features are used in training and testing. Each line in the template file denotes one template. In each template, special macro *%x [row, col]* will be used to specify a token in the input data, where *row* specifies the relative position from the current focusing token and *col* specifies the absolute position of the column. In this paper, we set the value of observation window as 3: -1, 0 and 1. Feature template is listed as follows, see Table. 2.

Table 2. Context feature template

%x[-1,0]
%x[0,0]
%x[1,0]
%x[-1,0]/%x[0,0]
%x[0,0]/%x[1,0]
%x[-1,0]/%x[1,0]

2) Semantic feature

a) Prefix and suffix features

The feature of prefix and suffix can help identify person name, institution, department, major, degree, supervisor or group leader. "赵, 钱, 孙, 李" as surname of Chinese person name will be used to recognize the prefix of one person's name. Some Chinese suffix feature phrase, such as "大学, 研究所, 系, 专业, 学位, 导师, 组长" can identify the named entity defined in researcher information.

b) Context semantic feature

Chinese named entity usually exists in the context at the same time, such as "研究员(research fellow)", "导师(supervisor)" and so on. This will be considered as the same semantic type. So this context information can be used better to improve the recall rate of named entity.

c) Relation feature extraction

There are three kinds of relation features in template.

(i) When the semantic feature and the Chinese surname appear at the same time, such as "博士张帆 (Ph.D. Zhang Fan)", the value of this kind of feature is set 1.

(ii) When different type (except name) of entities appear at the same time, such as "在中科院获得硕士学位 (… achieved the master degree in Chinese Academy of Sciences)", their values of this kind of feature are set 1.

(iii) When the sentence hasn't the relation between different type (except name), such as "北京大学 教授 (Peking University professor)", the value of this kind of feature will be set 1.

IV. Experiments

The preliminary task of extracting researcher information is to annotate the researcher profile as experiment data, which is labeled and checked manually by the employed domain experts.

We consider one sentence as one chunk unit, including date, several named entities and some other words. Each sentence must be a whole unit with corresponding syntax, neither simple phrase nor tedious paragraph. The date is saved as a text file in XML format by HTMLParser and some html tags.

One example is list as follows.

<resume><ri><date>1990 年 9 月 -1994 年 7 月 </date><institution>山东大学</institution><department>计算机科学系</department><major>软件专业</major>,获<degree>工学学士学位</degree></ri>

<ri><date>2001年4月至今</date> 于<institution>中科院计算所</institution><department>生物信息学研究组</department></ri>

<ri><date>2006年4月-2008年7月， </date>于加拿大<institution>滑铁卢大学</institution><department>李明教授实验室 </department><position> 访问学者 、</position><position>博士后</position></ri></resume>

Personal homepages crawled down from more than 10 institutions of Chinese Academy of Sciences is processed as data set, which is parsed as plain text format. We use 5000 sentences existed in 400 documents as our experiment data, 90% of which is training set, 10% of which is testing set. We choose the "Education background", "Work experience" and "Academic achievement" general information tag as the experiment in this paper.

The number of type in the training and testing set is listed in Table 3.

Table 3. The number of type in the training and testing set

Type	Sum	Training set	Testing set
Author	2789	2510	279
Area	119	107	12
Conference	1335	1201	134
Date	4585	4126	259
Degree	691	622	69
Department	702	632	70
Editor	59	56	3
Institution	2193	1974	219
Journal	1720	1548	172
Location	890	801	89
Major	235	211	24
Pages	1786	1607	179
Position	1545	1390	155
Publisher	261	235	16
Title	3218	2896	322
Volume	1419	1277	142

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The Features we choose are base feature and semantic feature separately. The semantic feature includes 329 concepts and 27 relations. CRF++ [16] is used as CRFs training and set software. The comparison results are shown in Table 4.

Table 4. The comparison results between base feature and semantic feature

Template	Base feature			Semantic feature		
Type	Precision	Recall	F-measure	Precision	Recall	F-measure
Author	95.32%	95.87%	94.33%	96.59%	96.59%	96.59%
Area	0.00%	0.00%	0.00%	62.50%	41.76%	50.01%
Conference	66.89%	61.64%	64.16	74.71%	73.06%	73.86%
Date	86.62%	75.77%	80.83%	92.43%	89.85%	91.12%
Degree	93.75%	78.95%	85.72%	100.00%	85.71%	92.31%
Department	85.00%	68.00%	75.56%	89.47%	70.83%	79.07%
Editor	79.00%	72.73%	75.73%	79.00%	72.73%	75.74%
Institution	59.68%	42.53%	49.67%	77.78%	71.01%	74.24%
Journal	74.21%	73.44%	73.82%	78.31%	75.00%	76.62%
Location	73.15%	64.50%	68.55%	81.67%	61.25%	70.00%
Major	70.00%	58.33%	63.63%	70.00%	58.33%	63.64%
Pages	84.43%	82.14%	83.30%	84.91%	86.40%	85.65%
Position	81.25%	57.78%	67.53%	83.76%	80.34%	82.01%
Publisher	89.29%	65.79%	75.76%	89.29%	83.33%	86.21%
Title	81.94%	82.66%	82.30%	83.61%	84.87%	84.24%
Volume	84.34%	84.11%	84.22%	86.86%	85.88%	86.36%

The precision, recall and F-measure between base feature and semantic feature are shown in Fig. 2, Fig. 3 and Fig. 4.

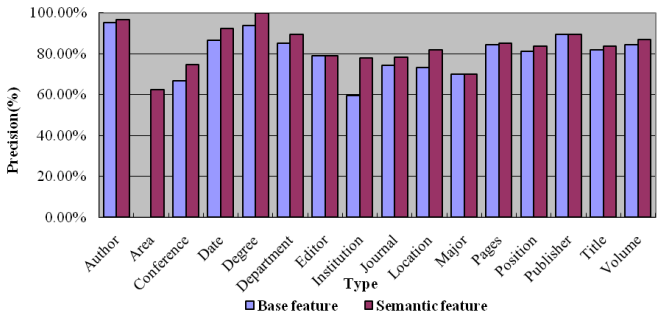

Fig. 2. Precision of base feature and semantic feature

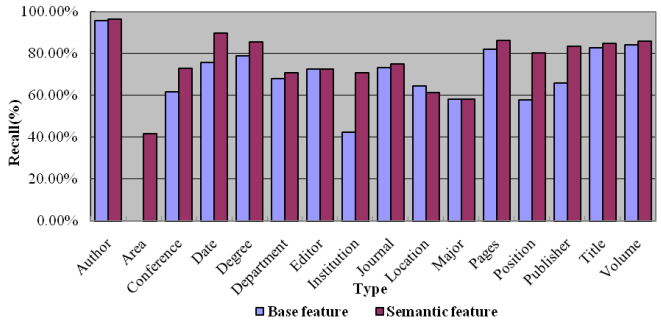

Fig. 3. Recall of base feature and semantic feature

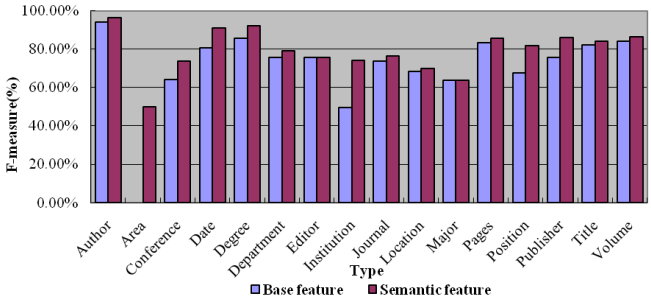

Fig. 4. F-measure of base feature and semantic feature

From Fig.2, Fig.3 and Fig.4, We can see that the precision, recall rate and F-measure value in CRF with semantic feature is higher than that in CRFs with base feature.

V. Conclusion

In this paper, we discussed researcher profile annotation with CRFs model. According to the semantic feature in researcher profile domain, we improved the CRF model, which can identify the concept, properties and relation named entities. The experiment results show that precision, recall and f-measures in our method is higher than baseline method. However, the semantic feature depends on domain ontology, which needs to appoint specific domain, such as academic domain. And our future work is to process the other kinds of data source, such as DBLP, project database and so on, to prove the effectiveness of our method.

References

[1] H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, and N. Shadbolt, "Automatic ontology-based knowledge extraction from web documents," *IEEE Intelligent Systems*, vol. 18, pp. 14-21, Jan-Feb. 2003.

[2] Y. Kun, G. Gang, and Z. Ming, "Resume information extraction with cascaded hybrid model," in *Proc. the 43rd Annual Meeting on Asoociation for Computational Lignuistics (ACL 2005)*, Stroudsburg, PA, USA, 2005, pp.499-506.

[3] Y. Limin, T. Jie, and L. Juanzi, "A unified approach to researcher profiling," in *Proc. the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2007)*, Fremont, CA, Nov. 2-5, 2007, pp. 359-366.

[4] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, "KIM–semantic annotation platform," in *Proc. the 2nd International Semantic Web Conference (ISWC 2003)*, Florida, USA, Oct. 20-23, 2003, pp.834-849.

[5] K. Balog and M. Rijke, "Finding experts and their details in E-mail corpora," in *Proc. the 15th International Conference on World Wide Web (WWW 2006)*, Edinburgh, Scotland, May. 23-26, 2006, pp. 1035-1036.

[6] Y. Jian-hua and S. Shun-hong, "Towards automatic concept hierarchy generation for specific knowledge network," in *Proc. the 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2006)*, Annecy, France, Jun. 27-30, 2006, pp. 982-989.

[7] Y. Hongkui, Z. Huaping, and L. Qun, "Chinese named entity indentification using cascaded markov model," *Journal on Communications*, vol. 27, pp. 229-235, Feb. 2006.

[8] O. Bender, F. Josef Och, and H. Ney, "Maximum entropy models for named entity recogination," in *Proc. the 7th Conference on Natural Language Learning (CoNLLl-2003)*, Edmonton, Canada, May. 31-Jun. 1, 2003, pp. 148-151.

[9] I. Hideki and K. Hideto, "Efficient support vector chassifiers for named entity recognition," in *Proc. the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, Aug. 24-Sep. 1, 2002, pp. 73-78.

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[10] S. Jingwei, Y. Jing, Z. Jianping, and Y. Yonghong, "Chinese prosody structure prediction based on conditional random fields," in *Proc. the 5th International Conference on Natural Computation (ICNC 2009)*, Tianjin, China, Aug. 14-16, 2009, pp. 602-606.

[11] F. Lei, X. Ying, M. Yao, and Y. Hao, "Conditonal random fields model for web content extraction," in *Proc. the 5th International Multi-conference on Computing in the Global Information Technology (ICCGI 2010)* , Valencia, Spain, Sep. 20-25, 2010, pp. 30-34.

[12] T. Yongmei, W. Xu, and C. Yong, "Chinese semantic role labeling using CRFs and SVMs", in *Proc. 2009 International Conference on Natural Laguage Processing and Knowledge Engineering (NLP–KE 2009)*, Dalian, China, Sep. 24-27, 2009, pp. 1-5.

[13] S. Xiao and N. Xiaoli, "Chinese base phrases chunking based on latent semi-CRF model", in *Proc. 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2010)*, Beijing, China, Aug. 21-23, 2010, pp. 1-7.

[14] Z. Suxiang, "Combining multi-features with conditonal random fields for person recognition," in *Proc. 2010 International Conference on Asian Language Processing (IALP 2010)*, Harbin, China, Dec. 28-30, 2010, pp. 178-181.

[15] D. Shengchun and J. Ting, "Comment target extraction based on conditional random field & domain ontology," in *Proc. 2010 International Conference on Asian Language Processing (IALP 2010)*, Harbin, China, Dec. 28-30, 2010, pp. 189-192.

[16] Taku Kudo (2005). CRF++: A CRF toolkit. Available: http://crfpp.sourceforge.net/.

ISBN: 978-988-19251-5-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)