Character information extraction based on CRFsuite

Jingzhong Wang[1], Zhongren Li[2,a], Wei Huang[2] and Ke Xiao[1]

[1]North China University of Technology, School of Computer Science, 100144, China
[2]North China University of Technology, School of Electronic Information Engineering, 100144, China

Abstract. By applying the Conditional Random Fields based on discriminant undirected graph to character information extraction, this paper proposes an automation character information extraction method based on CRFsuite. Through learning the known domain, this method extracts the feature leading words, position and means from the character information in the Internet to build up a character parameter. By using CRFsuite as a model, the method adopts it to data from the Internet, matches character information and builds up the structured character information database. The method proposed by this paper demonstrates the feasibility of the implement of automation extraction of character information in the mass Internet data, and provides an effective way to facilitate character information tracking and looking-up.
Keywords: CRFsuite; information extraction; machine learning.

1 Introduction

Information extraction is to extract information that is relevant or belongs to some specific types automatically and turn it into structured data, it is considered as one of the most important methods when dealing with mass data. Information extraction is an automation technique which helps people find the key part of the information in mass data. Dealing with the large scale of Internet information has eventually become an important part of natural language processing under big data era.

It classifies the objects of Information extraction to the following 3 types [1-4]: 1. Structured text, which is constantly generated by a fixed pattern, always has strong structural property. Such as the data in database. 2. Free Text, which means the text is in accord with natural language. It contains news, literatures, and government files, which are structural in some ways, but not strictly limited in structure. 3. Semi-structured Text, these type of content partially conforms to natural language rules. such as telegram messages or the images. Implementation of information extraction to this type of text needs the NLP technique.

In recent studies, J. Wang [5] proposed a method based on Maximum Entropy Hidden Markov Model. It is enhanced in accuracy and recall rate compared with the hidden Markov model, however it`s performance would be significantly affected when too many feature conditions taken into account, which makes the parameter identification complicated. W. Wei [6] proposed a text information extraction method based on second order Hidden Markov model. It is more accurate than the first order Hidden Markov Model, but when facing complicated texts, a more complex extraction model will be got which directly leads to the drop of the model parameters accuracy estimation and the decline of the information extraction accuracy.

[a] Corresponding author : lzrcoming@163.com

This paper brings up a character information extraction method using CRFsuite. And by means of learning the high frequency leading words and the position of hot spot of character information in the Internet, this method will manage a model with the extracted character features. By using the character feature model, this method matches and extracts character information in the mass data while checking, analysing and extracting those information of the large scale network data communication, and finally it is proposed to build a structured character database. It is an effective way to identify someone, and catch the information more comprehensively on a larger scale.

2 Extracting character information model from the mass data

2.1 Structure of information extraction

The structure of information extraction is shown in Figure1:

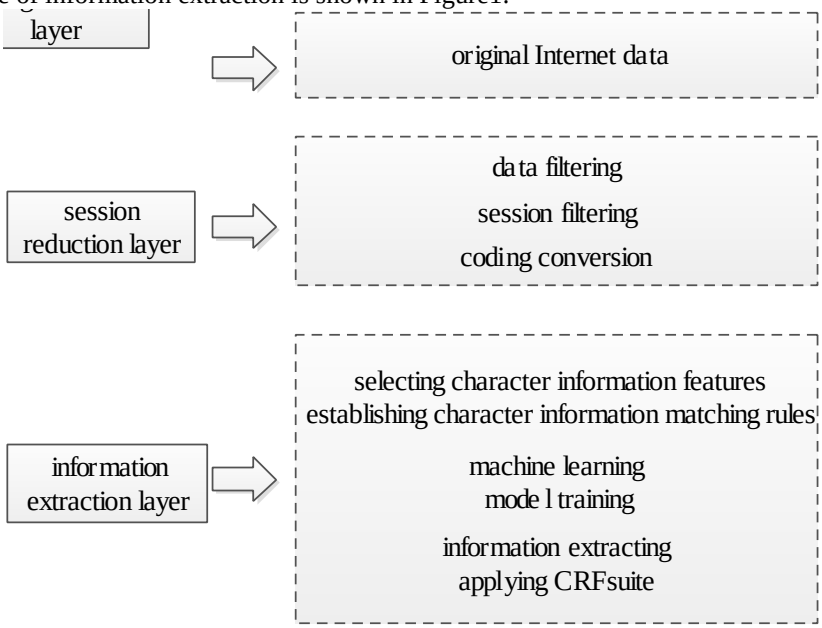

Figure 1. Structure of information extraction.

The structure shown in Figure1 is composed of 3 layers. The first layer is original data layer, which contains original Internet data waited to be extracted. The second layer is session reduction layer which contains preconditioning processes, including data filtering, session filtering and coding conversion. Data filtering is a process which identifies and filters the original data and identifies the segment of the data packet as well. And the filtering is abandoning the illegal data while returning the legal data as data stream. Session Filtering is mainly transmitted by TCP and UDP transmission protocol which is composed of five-tuple, and its target is internet data, it is a process which filters the HTTP session data and obtains the complete data after the legal data stream is acquired, or takes the single Request data in while the Response part is out. Coding conversion is to decompress and decode the data. The third layer is the information extraction layer which is composed of three parts: selecting the character information features and establishing the character information matching rules, machine learning, and model training with information extracting when applying CRFsuite. The first part is to select the frequently selected and high accuracy character features and extract them from the known domain. The second part is to train the model with the character information features extracted from the known domain by CRFsuite. The third one is to apply CRFsuite and the model to the character information, and extracted from the unknown domain which has been coded before.

2.2 Establishing character information matching rules

Restored and complete HTTP session data has been acquired after the preconditioning, including encoding conversion, data partitioning and website identification. Each session contains a request and a response of HTTP, or a single response. The character information matching rules is established based on them.

After the key information is extracted from the sample field using value matching method, it's found to be divided into 3 types, the position that character information appears in the Internet data, the means applied, and the feature leading words. The position contains the 3 following types: body, cookies, url. The means is the type of current session, Get or Post. It is to recognize the upstream GET and POST request data and the downstream Get and Post response data served as the boundary identifier of the information extraction. Feature leading words are the first three key words of the related character information sample values (most character information in data exists as the form "Segment name: Segment value" or "Segment name: values: Segment value"; so the first leading words describe the character information most accurately), which can be extracted using word segmentation filter method. For example, when a user registers his information on a website, we can acquire the feature leading words of the information, the position of it in the Internet data and the means of current session which uses the value matching method with the character information matching rules. Since the structure of the Internet data information is determined on the WEB server, the format of these data is similar to each other in a certain period. So if we establish a relatively accurate rule, it will be of high accuracy when analysing other information.

This paper develops the character featuring parameter system from the established character information matching method. The matching method forms a valid description of the character information in the mass Internet data through multiple locating of the key information, and is helpful to our analysis of character information and character model. Based on the character featuring parameter, this paper extracts the character information from the mass data, by means of modelling the trained features using CRFsuite. The established character featuring parameter is shown in Table1:

Table 1. Character featuring parameter.

Lead Word	Position	Method
1	Body	Post
2	Coolies	Get
3	Url	

2.3 Modelling of character information

Based on the system established before, this paper proposes the character information model by choosing some widely used character information features which describe someone most. The model is shown in Table2:

Table 2. Modelling of character information.

Number	Feature	Number	Feature
1	Name	7	Email
2	Phone	8	QQ
3	Address	9	MSN
4	Id_number	10	Username
5	Company	11	Password
6	Zip_code		

Since the difference of the format and the content of information transmitted in different network condition or different domain, when training character information model through CRFsuite, a synonym expansion of the former character information model will be needed (For example the title of someone`s phone number in a domain is PHONE while it is MOBILE in the other), to chase a higher accuracy and coverage. The WordNet [7-8] synonym expansion algorithm is adopted in this research. WordNet is an English dictionary based on a kind of cognitive linguistics, developed by psychologists, linguists and Computer engineers from Princeton University. It forms a 'words network' by the means of words rather than range them by alphabet. WordNet is a widely covered English words meaning network, and the nouns, verbs, adjectives and adverbs are organized as a network according to their means. This paper expands the existing model feature description by synonym utilizing WordNet, and helps the character information model to describe the character information more comprehensive, thus increasing the recall rate of the extraction.

3 Character information extraction based on CRFsuite

3.1 The theory of CRFsuite

CRFs was proposed by Lafferty et al in 2001. This theory, combined the features of Hidden Markov Model(HMM) and Maximum Entropy Model(MEM), not only overcome the output independence assumption problem of HMM, but also solve the label bias problem of MEM. This model is a discriminant undirected graph model [9]. The nodes in the graph indicate the random variables, while the branches show the relationship between these random variables. On the premise of the given observing sequences, this paper calculates the joint probability distribution of the marked sequence, and acquires the optimal sequence.

Definition of Conditional Random Fields [10]: orders G=(V,E) presents an undirected graph, and $Y=(Y_V)_{v\in V}$. The elements in Y are in one-to-one correspondence with the nodes in the undirected graph G. Under the condition X, the conditional probability distribution of Y_V obeys the Markov features of the graph G:

$$p(Y_v \mid X, Y_w, w\neq v) = p(Y_v \mid X, Y_w, w\sim v) \quad (1)$$

Where w, v indicates the nodes of the undirected graph, and w~v indicates the branches of the graph G. Now we call (X, Y) a Conditional Random Field.

The CRFs algorithm [11-13], structured as a first order chain, only takes the former one feature element into account. The model takes the n nodes as the global condition input, where $x=\{x_1, x_2, \ldots, x_n\}$. This paper builds the character featuring parameter based on the sequence x, which we call it the observation sequence. Meanwhile, there are n nodes taken as outputs of the model, where $y=\{y_1, y_2, \ldots, y_n\}$. The sequence y, which was called marker sequence in this paper, indicates the types of character information, such as names and phone numbers, as shown in Figure2.

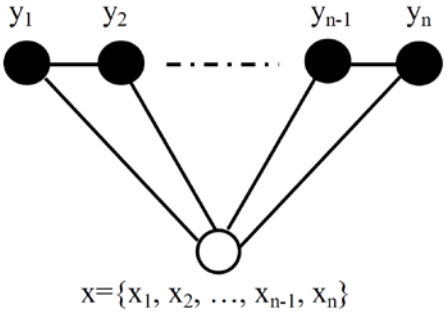

Figure 2. Structure of CRFs model

The algorithm calculates the most probable type of xi based on the given observation sequence, which provides the feature types, y= {y_1, y_2, …, y_n}, of the character featuring parameter space vector, x= {x_1, x_2, …, x_n}. Based on the CRFs, the condition probability of parameter collection Λ= {λ_1, λ_2, …, λ_n} correspondent to the observation sequence x can be calculated as Formula (2):

$$p_\Lambda (y/x, \lambda) = (1/Z_\Lambda) * [\Sigma_i \Sigma_m \lambda_m y_m (y_{i-1}, y_i, x, i)] \quad (2)$$

Among the Formula, $f_m(y_{i-1},y_i,x,i)$ is a Boolean characteristic function, in which transfer characteristic function t_m(yi-1,yi,x,i) and states characteristic function $s_m(y_{i-1},y_i,x,i)$ unified. λ_m is the weight of the characteristic function and can be trained and structured to be the parameters of first-order chain CRFs model. Z_Λ is the normalization factor, and can be calculated as Formula (3):

$$Z_\Lambda = \Sigma_y \exp[\Sigma_j \Sigma_m \lambda_m f_m (y_{i-1}, y_i, x, i)] \quad (3)$$

Therefore, the key of character information extraction based on CRFs is the level of representation of the trained model in the domain and the accuracy of character information description. So this paper takes some authority statistic domain as the training domain, while for the accuracy of the character information description, the paper expands the character information description with the synonym expansion algorithm, and unifies the expanded synonyms by means of description.

3.2 Character information extraction model based on CRFsuite

CRFsuite is an implementation of sequence data tagging based on CRFs. Compared with other implementations of CRFs algorithm, CRFsuite excels other implementations due to its speed [14]. The size of corpus and the number of corpus for training features affect the training speed most. The following figures list the comparison of training speed and accuracy rate of CRFsuite [15], Wapiti, CRF++ and MALLET when the number of features is in 450000 to 600000, as shown in Figure3, Figure4.

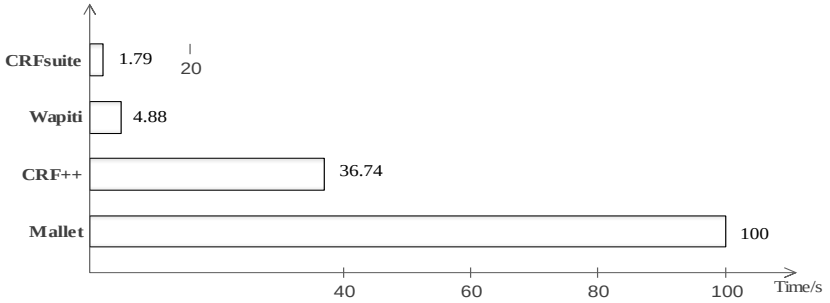

Figure 3. Comparison of Training Speed of CRF

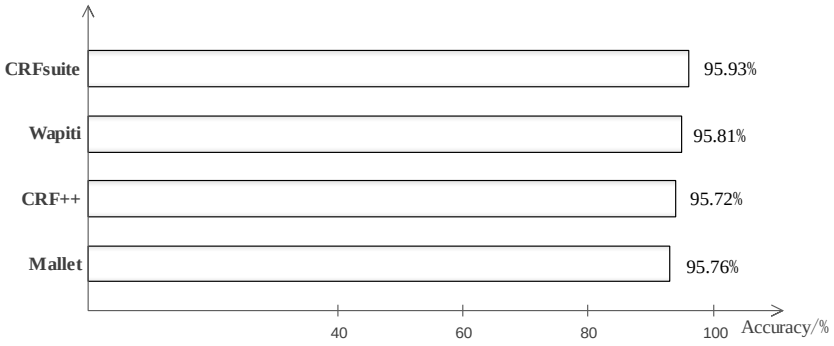

Figure 4. Comparison of Training Accuracy of CRF

When dealing with mass data, the speed and accuracy are both important indexes. Therefore, we eagerly need a fast and accurate method. CRFsuite is better than other algorithms in both two ways, so this paper trains the character model by CRFsuite in consideration of its excellent performance. Through the segmentation and encoding conversion of the Internet data packet, this paper builds the character featuring model according to character information matching rules using value matching method. Then export the model information in form of algorithm interface by a certain program, and train it to the character information model applying CRFsuite. Finally, the model is applied to the backbone network to analysis and extract the character information.

3.3 Extraction algorithm implementation based on CRFsuite.

The following steps are technological processes of the character information extraction from Internet data packet based on CRFsuite:

Step1: Pre-processing, which is segmentation and coding transmission of the massive network data, along with reduction of it to a single file.

Step2: Acquire the character feature information by the value matching method, and note it with featuring information in the character information model. Then decrypt it as the text file info_train.

Step3: Processing. Transform the character feature information into the form of interface for CRFsuite. Name the outcome with features train.

Step4: Machine learning.

Step5: Bring the features train CRFsuite in, and train the character information model with CRFsuite. Then the character information featuring model is acquired and named as CRF_MODEL.

Step6: Segmentation and encryption of the mass data from the Internet.

Step7: Acquiring the first leading word, the position of first showing up, etc. of character feature information by bringing the info_train in Step2 and utilizing the character matching rules (Character information exists in the form “Segment name: Segment value” in most domain. So the first leading word describes the character information most accurately). The outcome is named as dict.

Step8: Matching the dict acquired in Step7 and the network data analysed in Step6, to acquire the first, second and third leading words, their positions and the following character information value. And the outcome is named as info_test.

Step9: Take info_test to Step3 to process the information. The outcome in form of CRFsuite is then named as features_test.

Step10: Mark the features_test with the trained character information model CRF_MODEL, then output the character information in the form “feature: character information value”.

4 Verification and analysis of the experimental result

Firstly, the accuracy and the recall rate of the Internet character information extraction using CRFsuite was verified. By browsing different main-stream webpages manually, this paper collects 200 Internet data packet from different domains, and separates them into 2 parts, 150 packet for one and 50 for the other. The 150 one is then used for training of the model CRF_MODEL while the 50 one is used for verification of the accuracy and the recall rate of CRF_MODEL.

The CRF_MODEL is trained by using CRFsuite with the training set which has been segmented and transcoded before, and then applied to the verification set to tag it. The tagging is then compared with the known one (Y), and the accuracy and the recall rate is calculated.

To avoid the contingency of the experimental result, this verification adopts the random collocation of the sample field. Five times of experiment were taken in total, and a random collection of 150 samples was selected to be the training set while the rest 50 for verification for every experiment. The result is shown in Table3.

Table 3. Result of character information extraction in CRFsuite.

Number	Features	Precision	Recall	F1
1	11	85.22%	81.35%	83.24%
2	11	84.37%	83.15%	83.76%
3	11	85.20%	83.08%	84.13%
4	11	86.79%	80.14%	83.33%
5	11	84.17%	83.73%	83.95%
AVG	11	85.15%	82.29%	83.68%

Secondly, put the other three methods in the experience, the comparison is shown in Table4:

Table 4. Comparison of different methods**.**

Number	Method	Precision	Recall	F1
1	Wapiti	83.75%	81.72%	82.72%
2	CRF++	84.93%	80.25%	82.52%
3	Mallet	83.36%	81.89%	82.62%
4	CRFsuite	85.15%	82.29%	83.68%

From the experimental result, 85.15% of the Character information was extracted by the CRFsuite while maintaining a relatively high accuracy and recall rate. We did a several experiments, and took the average, greatly reduce the experiment results affected by chance.

The character information marking result is shown in Table5.

Table 5. Result of character information mark in CRFs.

Features	**Total**	**Correct**	**Error**	**Correct Rate**
Name	4860	4138	722	85.15%
Phone	6245	6114	131	97.90%
Address	4215	3059	1156	72.58%
Id_number	3015	2742	273	90.95%
Company	2648	2027	621	76.53%
Zip_code	5679	5261	418	92.64%
Email	6147	6064	83	98.65%
QQ	4589	3875	714	84.44%
MSN	1547	1286	261	83.13%
Username	5025	3711	1314	73.85%
Password	3067	2303	764	75.11%

From the result, we can find that CRFsuite has a higher identifying rate to some certain types of character information. For example, cell phone number, email phone number and zip code for the title of these information are simple. But to address, company and username, the identifying rate is low because of the various titles of these information. For example, we can use our email or our phone number as the username, and the form of the address is more diversified, it may also have missed some information while decoding the packet. Therefore, the extraction method to these kinds of information need to be enhanced.

5 Conclusion

This paper propose a method of character information extraction based on CRFsuite. Through the systematic research on the structure characteristics and basic features of the character information in massive network data, we concluded the character information features which could exactly located a person, and created the character information model. Finally, we extracted the character information by the CRFsuite on the network data, and achieved a better result.

To keep off the accidental event, this paper has carried on five experiments, each of them take different data to training and labelling, the average of the accuracy rate, recall rate and F1 is 85.15%, 82.29% and 83.68%, respectively. The same experiment was done with the Wapiti, CRF++ and Mallet, and it was demonstrated that the accuracy rate in CRFsuite was the highest while the recall rate was almost same. It was proved that the method proposed in this paper was better than others, and could be used in the project.

In the future study, we will keep on improving the accuracy rate of the extraction. On the one hand, optimization of the extract precision of address and company, on the other hand, through improving the annotation accuracy of low marginal probability to enhance the recognition ability of the features, so as to advance the learning ability of CRFsuite.

Acknowledgements

This work was supported by a special fund of School (No. XN060)

References

1. Q. Liu, H. Jiao, H.B. Jia, Appl. Res. Com, **24**, 6(2007)
2. Y. P. Lin, Y. Z. Liu, S. Zhou, Chin. J. Elec, **33**, 236(2005)
3. S. X. Zhou, Y. P. Lin, Y. Nan, Chin. J. Elec, **35**, 2227(2007)
4. Q. Chen, W. Zhu, C. Ju, *ICNC2014*, 780(2014)
5. J. Wang, J. Wu, Y. Zhang, G. He, *ISCID2014*, 501(2014)
6. W. Wei, S. Shi, Y. Liu, H. Wang, C. Yuan, *WISA2013*,65(2013)
7. Y. Qu, K. Wang, D. Zhang, Chin. J. Elec, **3**, 617(2012)
8. J. Wang, W. Zuo, Z. Yan, J. Com. Res. Dev., **7**,1499(2014)
9. D. Birkhed, B. Sundin, SI. Westin, *ICDE2015*, 13(2015)
10. L. Qiao, C. Li, Z. Zhong, J. Wang, D. Liu, JNNU(ETE), **35**,134(2012)
11. M. Li, S. Wang, H. Wang, Hig. Tech. Let., **12**, 1040(2015)
12. X. Sun, S. Lin, R. Liu, CA. Tra. Int. Sys., **23**, 80(2009)
13. S. Hori, M. Murata, M. Tokuhisa, Q. Ma, *SCIS2014*, 1102(2014)
14. K.R. Rahem, N. Omar, *ICIMU2014*, 250(2014)
15. D. Birkhed, B. Sundin, SI. Westin, *IEEE2015*, 1534(2015)