Research on web character information extraction based on semantic similarity

Bao-Cheng Wang[1], Wei Huang[2], Zhong-Ren Li[2] and Ke Xiao[1]
1School of Computer Science, North China University of Technology
2 School of Electronic Information Engineering, North China University of Technology, Beijing, 100144, China
E-mail: endlesshw@sina.com

As for the loss of the comprehensiveness from the large amount of data when extracting information, this paper proposes a method of character information extraction based on semantic similarity algorithm to improve the comprehensiveness of the character information extraction of massive data in the network. The algorithm is put into the semantic tree to choose the synonyms of the word, and the character feature set which is extended by semantic similarity is applied to character information extraction. The results show that the recall reaches to 81.87% in the case of the accuracy rate being basically unchanged. Therefore, this method of character information extraction is obviously improving in comprehensiveness, and it can be used in network data.

Keywords: Semantic Similarity; Character Information Extraction; Machine Learning.

1. Introduction

Due to the rapid development and widely spread of the network, the network data generated by Internet has a substantial increase, including the user's information generated in the network registration. So how to use these resources more efficient and convenient? It is an urgent problem that we need to solve.

Information extraction is to extract information that is relevant or belongs to some specific types automatically and turn it into structured data[1]. There are many ways to extract information, such as Maximum Entropy Hidden Markov Model[2], Second Order Hidden Markov model[3] and Conditional Random Field[4-5]. However, the training template obtained by the model is limited and cannot contain all the information content of the unknown data. Therefore, the semantic similarity algorithm was applied to the character information extraction in this article to improve the comprehensiveness of the extraction on unknown domain based on the learning and labeling of the Conditional Random Field.

Research on semantic similarity is divided into 2 categories, one is based on the semantic distance, the most typical is the Wu&Palmer[6] algorithm and the

Leacock&Chodorow[7] algorithm. And the other one is based on information content, Resnik[8] algorithm and Pirro[9] algorithm is a representative algorithm based on information content. Now, the semantic similarity of the concept is mostly based on the combination of two above. This paper proposed a method of semantic similarity, which combined of information content and semantic tree.

2. Web Character Information Extraction

2.1. *Information Pretreatment*

The main purpose of information pretreatment was to get the rule description of each information element. The output of the module was used as the input of the following information extraction and the mark of algorithm.

We needed to register the corresponding information in different fields, and recorded the contents of registration information to text file. Then obtained the information from the text document which is converted from the data package.

Then the matching rules should be built, which included the prefix feature words, the value located in the packet and the method in the session, as shown in Tab. 1. For example, the phone number: 13269367286 in Fig. 1, the three prefix feature words included W[0]: mobile, W[1]: sendregmobile, W[3]: GET and the position: URL.

Tab. 1. The rules of character information extraction

Value	Feature1	Feature2	Feature3	Position
1	W[0]	W[1]	W[2]	URL
2	W[0]	W[1]	W[2]	Cookie
3	W[0]	W[1]	W[2]	Body

```
**********
GET /sendregmobilecode?mobile=13269367286&timesign=1417595622636&validatecode=k6c24
x-requested-with: XMLHttpRequest
Accept-Language: zh-cn
Referer: http://passport.58.com/reg
Accept: text/html, */*
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; .NET4.0
Host: passport.58.com
Connection: Keep-Alive
Cookie: 58home=bj; id58=05dzW1R+yXgay0vODHhiAg==; new_session=0; init_refer=; city=
CNZZDATA30017898=cnzz_eid%3D1690321483-1417594784-http%253A%252F%252Fbj.58.com%252F
PassportVerifyCode=PassPort58-4fe9ae04-7749-467e-a3c0-b89851d9cf67
```

Fig. 1. The text from the data decompression

The prefix feature words of these part contain most of the character attributes and different describe of the attributes. However, not all of the prefix feature words were useful, it may produce some error characters, which would

reduce the extraction accuracy. Therefore, it is necessary to make further screening of these prefix feature words, such as the garbled, punctuation and the meaningless item which is only one letter or more than a certain threshold. The character information extraction rules of the known data were shown in Fig. 2, w means the prefix word, and pos is the abbreviation of position.

```
name    w[0]=value      w[1]=txtTrueName        w[2]=text       pos=EntityBody
phone   w[0]=value      w[1]=txtPhone   w[2]=text       pos=EntityBody
QQ      w[0]=value      w[1]=txtQQ      w[2]=maxlength  pos=EntityBody
phone   w[0]=mobile     w[1]=success    w[2]=type       pos=EntityBody
phone   w[0]=mobile     w[1]=confirmstatus      w[2]=step2      pos=EntityBody
phone   w[0]=mobile     w[1]=step3      w[2]=mobile     pos=EntityBody
name    w[0]=email      w[1]=authmailsent       w[2]=com        pos=EntityBody
e_mail  w[0]=email      w[1]=authmailsent       w[2]=com        pos=EntityBody
name    w[0]=email      w[1]=authmailsent       w[2]=location   pos=EntityBody
name    w[0]=txtEmail   w[1]=ff752a73a0c0       w[2]=9fcd       pos=EntityBody
e_mail  w[0]=email      w[1]=authmailsent       w[2]=location   pos=EntityBody
```

Fig. 2. The character information extraction rules from the known data

2.2. *Character Information Extraction Based on Semantic Similarity*

2.2.1. *Semantic similarity algorithm*

Semantic similarity has been widely used in many fields such as Natural Language Processing, semantic web, information retrieval and so on[10]. At present, most of the semantic similarity methods are based on the WordNet semantic dictionary database. WordNet is a semantic dictionary database (Miller, 1985), which is developed by lexicology professors and linguistics professors at Princeton university.

As the word usually has multiple semantic sense, so a necessary step before the expansion by semantic similarity algorithm was to semantic disambiguation, and obtain a group of exact semantic sense set.

In this paper, we used Semantic Interconnections Structural algorithm[11] to obtain the Sense set of the word set. The input was a word list $X = \{x_1, x_2, \ldots, x_n\}$ and a Context(Y), Y was corresponding to a disambiguation word set. While fetching a word from the list X, for each sense of the word, calculated its similarity with the Context(Y), and then chose the maximum similarity of the sense, add words and the sense to Context(Y). The iteration was ongoing until the word is not ambiguous. The semantic disambiguation formula is as follow:

$$S(w) = \arg \max_{S_i \in S(w)} \sum_{S_j \in Y} Sim(S_i, S_j). \quad (1)$$

In formula 1, Sim (S_i, S_j) means the similarity of two sense, used the Lin algorithm[12] to computed the similarity of any two sense vector. At the same time, used an open source Python packet WordNet: Similarity to completed the similarity algorithm. The calculation method, as shown in the formula 2:

$$Sim(S_i, S_j) = \frac{2I_{max}(S_\Delta)}{I(S_i) + I(S_j)}. \quad (2)$$

S_i and S_j is relevant concepts, S_Δ is the nearest common parent node from S_i and S_j. I (S) is the information entropy.

After having a word set with the exact semantic sense, a synonym set is constructed based on the semantic tree in WordNet. The more abstract of the nodes in the semantic tree, the higher the position is. So we calculated the similarity of the two sememes, which was to calculate the length and depth between the sememes. The semantic distance formula was shown in formula 3, and the semantic depth formula was shown in formula 4:

$$Sim(S_i, S_j) = \frac{\omega}{Dis(S_i, S_j) + \omega}. \quad (3)$$

Among the formula, ω is the relative average distance between two concept words in semantic tree and the set, it is an adjustable parameter, Dis (S_i,S_j) indicate the shortest distance between the S_i and S_j.

$$Dis_s(S_i, S_j) = \frac{K(S_i) + K(S_j)}{K_{max}(S_i) + K_{max}(S_j)}. \quad (4)$$

K (S_i) and K (S_j) mean the node's depth between S_i and S_j, K_{max} (S_i) and K_{max} (S_j) represent the maximum depth of the node in the semantic tree.

2.2.2. *Character information extraction*

Extracted the character information by the dictionary matching which was based on semantic similarity, the words in the dictionary was also called leading words or trigger words, provided by the first prefix feature words in data mining, used the new extraction rules which was formed after semantic extension to trigger the information extraction.

Due to the network packet content was complicated, not only the information content from network transmission, but also the tags on the Web page. So after the new rules, the redundant information would lead to the less accurate data and increased the useless information. Accordingly, we needed to filter the information after extraction. This part we considered the regular expression to filter out some interference.

Firstly, we extended the trigger words(W[0]) in dictionary to get a more comprehensive dictionary information. Then, we matched the trigger words in the dictionary. Extracting two words before the trigger word and the contents in a certain range behind the matching words. Finally, preliminary screening the extraction information by regular expressions.

2.3. *Algorithm Marking*

CRFs algorithm was used in the part of the algorithm predictive marker. the main idea of the algorithm model was based on Maximum Entropy Model, combined with the Hidden Markov Model[13]. It not only solves the label bias of the Maximum Entropy Model, but also overcome the output independence assumption of strict requirements by HMM. Assuming the observed sequence is $x = \{x_1, x_2, \ldots, x_i\}$, the state sequence is $y=\{y_1, y_2, \ldots, Y_i\}$, the conditional probability of state alignment in CRF model is shown as formula 5:

$$P\langle y \mid x\rangle = \frac{1}{Z(x)} \exp\{\sum_{i=1} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)\} \quad (5)$$

Among the formula: $f_j(y_{i-1}, y_i, x, i)$ as the feature function, it is an arbitrary value, λ_j is the weight of $f_j(y_{i-1}, y_i, x, i)$, it can be obtained by training. Z (x) is the probability normalization factor, which takes the sequence as the condition.

$$Z(x) = \sum_{y} \exp\{\sum_{i=1}^{m} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)\} \quad (6)$$

In this paper, the marked content belongs to the maximum conditional probability of the sequence. According to the content, position and the leading words in the text, it is a sequence labeling problem, which can be used in CRFs model. In the model, the selected of the feature is very important. So after the systematic research on the structure characteristics and basic features of the character information in massive network data, the three prefix words and the position from extraction rules in section 3 were chosen to be the feature.

3. Experimental Results and the Analysis

The experimental results were verified by accuracy rate and recall rate. By browsing different main-stream webpages manually, this paper collected 200 Internet data packet from different domains, and was separated into 2 parts, 150 packet for one and 50 for the other. The 150 one was then used for training of the model CRF_Model while the 50 one was used for verification of the accuracy rate, recall rate and F1. Accuracy rate is the ratio of related data in retrieval and total number of retrieved, it measures the precision of a retrieval system. Recall rate is the ratio of related data in retrieval and total related number of the packet, it measures the recall of a retrieval system. F1 is the weighted harmonic mean of Accuracy and Recall.

Experiments were divided into two groups. Firstly, we experimented the character information extraction method without the semantic similarity and

recorded the results, the same experiment was done with the method in semantic similarity. To keep off the accidental event, three experiments were carried out in each group, 150 domains were randomly selected among the 200 domains as the training set, and the other 50 domains were the test set. Finally compared the accuracy and the recall of two groups.

Tab. 2 showed the comparison of the average of three experiments about the 10-character information elements.

Tab. 2. Extraction number of the character information feature marking

Num.	Feature	Character information extraction	Character information extraction based on semantic similarity algorithm
1	Name	5437	6581
2	Sex	4253	5003
3	E_mail	7462	8143
4	Phone	6784	7832
5	Address	5613	6418
6	Id_num	5120	5872
7	QQ	6345	7046
8	Passwd	4210	5036
9	Account	8461	9867
10	Company	3594	4513
11	Sums	57279	66311

According to the experimental result, we can know that the extracted numbers of semantic similarity algorithm is 9032 more than without the semantic similarity. For example, before the extended, it may only have ‘phone’ and ‘number’ to matching the information in data files, but after the extended, it added the synonyms such as ‘mobile’,’telephone’,’tel.’, thus to match more comprehensive information.

Tab. 3. Results of the character information extraction

Number	Precision	Recall	F1	Time
1	78.48%	75.15%	76.78%	7.7s
2	83.23%	69.91%	75.99%	7.6s
3	80.66%	73.95%	77.16%	7.9s
AVG	80.79%	73.00%	76.64%	7.7s

Tab. 4. Results of the character information extraction based on semantic similarity

Number	Precision	Recall	F1	Time
1	75.31%	83.17%	79.05%	8.2s
2	80.65%	77.66%	79.13%	8.6s
3	78.12%	84.78%	81.31%	8.3s
AVG	78.03%	81.87%	79.83%	8.3s

The experimental results of the character information extraction methods without the semantic similarity and the experimental results based on semantic similarity was listed in Tab.3 and Tab. 4.

Compared the result of table 4 and table 5, the character information extraction based on semantic similarity algorithm improved 8.87% in the recall rate. The spend time on the method in this paper was 0.6s longer than the method without the similarity algorithm but within an acceptable range. As the extracted information increased, and not all of them was used, so it will produce some information redundancy, which lead to the accuracy of the data extraction decreased by 2.76%. Further study and improvement will be done for that. The result indicated that most of the increased information were useful and correct. It can be used in the character information extraction on the Web.

4. Conclusion

This paper proposed a method of character information extraction based on semantic similarity algorithm. 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 rules. Finally, we extracted the character information on the network data, and achieved a better result.

To keep off the accidental event, this paper carried on two groups of experiments, one was the extraction method without the semantic similarity, the other was the method with the semantic similarity. Each group has been carried on for three times, each time different data was applied to training and labelling. The results showed the method in this papers had increased 8.87% in recall and 3.19% in F1 compared with the method without semantic similarity algorithm. It was proved that the method proposed in this paper could be used in the project.

Acknowledgements

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

References

1. Q. Liu, H. Jiao and H. Jia, Research on approaches of information extraction System, Appl. Res. Com. **24**, 6 (2007).

2. Y. P. Lin, Y. Liu and S. Zhou, Using Hidden Markov Model for text information extraction based on Maximum Entropy, Act. Elec. Sini. **33**, 236(2005). (In Chinese).
3. S. X. Zhou, Y. Lin and Y. Wang, Text information extraction based on the Second-Order Hidden Markov Model, Act. Elec. Sini. **35**, 2227(2007).
4. Y. Y. Luo and D. Huang, Chinese word segmentation based on the marginal probabilities generated by CRFs, Jour. Chin. Inf. Pro. **23**, 3(2009).
5. X. D. Han and C. Zhou, Conditional random fields theory review, Chin. Sci. Tech., 10(2010). (In Chinese).
6. X. P. Meng and Y. Gao, *Power System Analysis*, (Higher Education Press, 2004). (In Chinese).
7. Z. Wu and M. Palmer, Verbs semantics and lexical selection, (*ACL'94 Pro. 32nd meeting on Association for Computational Linguistics*), (Stroudsburg, USA, 1994).
8. Z. X. Bian, Research on model of IC parameter for semantic similarity of concept in WordNet, Com. Eng. Appl. **47**, 128(2011). (In Chinese).
9. X. Mei, X. Meng and J. Chen, SSCM: a scheme for calculating semantic similarity, Chin. Hig. Tech. Let. **17**, 458(2007). (In Chinese).
10. E. Cambria, Y. Song, H. Wang and N. Howard, Semantic multidimensional scaling for open-domain sentiment analysis, IEEE Int. Sys. **29**, 44(2014).
11. Y. S. Wang, A similar words algorithm automatically classifying different senses, Com. Appl. Soft. **32**, 258(2015).(In Chinese).
12. D. Lin, An information-theoretic definition of similarity, *Fift. Int. Conf. On Machine Learing*, (San Francisco, USA, 1998).
13. K. Williams, L. Li and M. Khabsa, A web service for scholarly big data information extraction, *IEEE Int. Conf. on Web Services (ICWS)*, (Anchorage, AK, 2014).