Finding of People in Economic Field based on Personal Profile

Lei Chun[1], Yang Fengfan[2] and Lou Liang[2]
[1]*School of Economics and Management, Southwest Jiaotong University, Chengdu, China*
[2]*College of Electronics and Information Engineering, Sichuan University, Chengdu, China*
{leichunsc, yangfengfansc, louliangsc}@163.com

Keywords: People in Economic Field, Field Trie Tree, Chinese Text Classification, Bayesian Classifier.

Abstract: There are a lot of personal profile information on the Internet. If we can automatically collect these information on the Internet, classify them and extracting personnel information from them, it can provide some help for the finding of people in specific field. In the research, we focus on study of Finding of People in Economic Field based on Personal Profile from the Internet. We proposed a field dictionary building method base on the paper database, and make an improvement to the text segmentation technology used in people classification. We designed a hybrid segmentation method based on Field Trie Tree and HMM-Viterbi model. On the basis of Bayesian classifier, we build a model to find people in Economic field based on Personal Profile. Experiments shows that the method has a high rate of the recognition of people in economic field.

1 INTRODUCTION

Field people recognition means judging the people's working field by the tabs of people or description text of the people and select these people working in specific areas. For the information of people on the Internet are various, it is not easy to find the people working in economic field. It not only require proper data source, but also need to make comprehensive analysis of these information and determine the likelihood of people belonging to the target fields. The current method used for field people recognition is mainly based on text classification. Text classification is the process to analyze the text and then put the text into the pre-defined categories according to the results of the analysis (K. Aas, L. Eikvial, 1990). Text categorization research originated in the 1950s, H. P. Luhn is the first researchers in the field, he realized the classification of the text by word frequency statistics method, followed by the emergence of automatic text categorization and classification method based on knowledge engineering (Fabrizio Sebastiani, 2002). After the 1990s, with the development of knowledge of statistics and machine learning, classification based on knowledge engineering had gradually been replaced by some new classification algorithm based on machine learning which mainly including Rocchio algorithm, Bayesian classifier algorithm, K-nearest neighbor algorithm, support vector machine algorithm and etc. (N. Fuhr, S. Hartmanna, G. Lustig, M. Schwantner, K. Tzeras, 1991). Chinese text classification started in the 1980s, mainly based on the study and improvement of English text classification, applying the English text classification algorithm to Chinese text and forming the Chinese text classification. The acquisition and pre-processing of personal profile information on the Internet can be completed by the crawlers. The research mainly focused on the method of field classification of collected personal profile information.

2 IMPROVEMENT OF CHINESE WORD SEGMENTATION TECHNOLOGY

Workflow of text classification based on machine learning has three main processes: text preprocessing, text categorization classifier training and text classification. Text preprocessing process is to make text segmentation, remove stop words, extract feature words and etc. (Feng Guo-he, 2007). The training process is to input a number of artificial accurate classified documents into the computer, allowing the computer to analyze each category

Fengfan Y., Chun L. and Liang L.
Finding of People in Economic Field based on Personal Profile.
DOI: 10.5220/0006028004360439
In *Proceedings of the Information Science and Management Engineering III (ISME 2015)*, pages 436-439
ISBN: 978-989-758-163-2

Finding of People in Economic Field based on Personal Profile

document and extract some characteristics and rules which can effectively distinguish among each category. The features and rules are used by the classifier. The classification process is to classify the input text, the classification is determined by the trained classifier (Yang Wen-chuan, 2013) Fig. 1 is a complete text classification process.

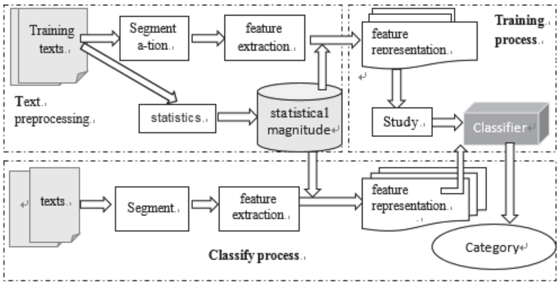

Figure 1: The basic process flow of text classification.

Text segmentation is the basis of Chinese text classification process, which directly impacts on the choosing of the characteristics, and thus affects the results of classification. There are some professional terms in the personnel profile, which are essential characteristics to field recognition. According to these characteristics, we propose a hybrid segmentation method based on field Trie tree dictionary and HMM models, which can improve the recognition rate of professional terms in the text.

3 FIELD DICTIONARY BUILDING METHOD BASE ON THE PAPER LIBRARY WANFANG

To construct specialized field dictionary, we must first get the specific field professional terms. This paper presents a growing method based on word seeds. The method is mainly depend on the keywords searching ability provided by the paper database WanFang (http://www.wanfangdata.com.cn). First, build some professional term lists as seed lists manually. It is much better to choose these professional terms which can distinguish themselves from other field professional terms most. In the research, we focus on the economic field, so we choose "Macroeconomic science", "audit", "finance", "insurance" and other words can be identified in the field. Then, we use the crawler to crawl the papers with the keywords in seed list and extract the keyword in the paper which has same keywords in the seed list. The extracted keywords can be used as new keywords in the targeted field. With the process goes over and over, we can get more and more specific field professional terms dictionary. The whole acquisition process shown in Fig. 2.

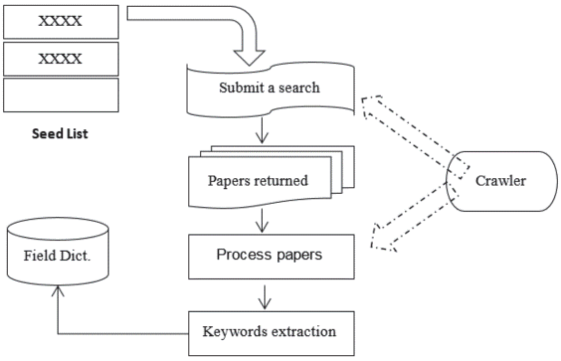

Figure 2: Automatic field dictionary building method.

4 A HYBRID SEGMENTATION METHOD BASED ON FIELD TRIE TREE AND HMM-VITERBI MODEL

Trie tree, also known as dictionary tree, prefix tree, is a tree structure. The statistics, sorting and searching operation of strings stored with the Tire structure is very fast (Shang Wen-qian, 2007). Its advantages are: strings with the same prefix share the same storage for prefix, so not only can reduce storage costs with the same prefix of the string, but can also locate the matching string of characters directly from the individual, thus effectively reducing the number of matches, shorten find time. In this paper, we the find the professional terms in the text with Trie tree.

HMM-Viterbi (Hidden Markov Model) segmentation can improve the recognition rate of professional terms, place names, organization names and etc. But the figures introductory text is relatively short, with small vocabulary, recognition rates vocabulary, which can directly affects the classified by field. The study designed a hybrid segmentation method based on word dictionary and probability. First, obtain the dictionary of the target areas and build the Trie tree with the dictionary. Second, using a forward maximum matching method to extract the terms of specific field with Trie tree. Then, use HMM-Veterbi model to handle the remained text for the terms of specific field. The entire segmentation process consists of pretreatment and segmentation

437

process. The process of hybrid segmentation method is shown in Fig. 3.

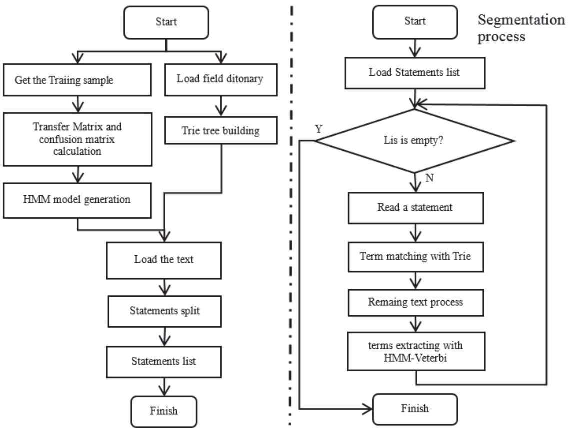

Figure 3: The process of hybrid segmentation method.

5 ECONOMIC FIELD PEOPLE FINDING BASED ON BAYESIAN CLASSIFIER

After the actual test and compare, chi-square test was selected as a text feature selection algorithm, and the number of features was set to 2000. The Bayesian classification has some advantages that the necessary parameters is calculated with a small amount of training data, it runs fast and it is easy to implement. So, the Bayesian classification algorithm is introduced.

5.1 Field Classification Standard

According to the regular research or working field classification, treat the similar fields as a big category. First, people’s professional field was classified into "style art", "computer communication", "natural science", "economics area" and other big categories, and then classified into sub-category like "political economy" "management economics", "money and Banking", "finance" and others, as shown in Fig.4. The advantage of such a classification is that classification with big category can put all the text into proper category and have a better correct rate, and with a sub-category it can make the classification more accurate.

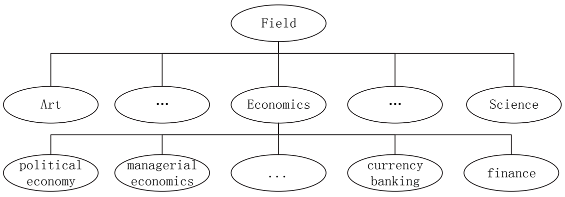

Figure 4: Field Classification.

5.2 Framework of Economic Field People Recognization

Recognition framework of the economic field people contains a total staff of four modules: data acquisition sub-module (DAC), pretreatment sub-module (PRE), field identification sub-module (REC) and database operations sub-module (DBT). The four modules collaborate with each other to finish the process of the collection of information and the field recognition. PRE and REC are called by DAC, and DBT is calls by REC. Framework of economic field people identification is shown in Fig. 5. DAC module mainly work as a data collector and preprocessor, and read the configuration file of the framework. PRE mainly reads the word and text classification required training sample, as well as dictionaries and other documents from the main module, and training. It finally builds the HMM & NBM model used for segmentation and classification. The HMM & NBM model is called by the REC. PRE module runs only one time at the start time of the system. After generating the model, PRE notice REC modules to take field reorganization. REC handles the clean personal profiles received from the master module and perform classification recognition with the HMM & NBM model.

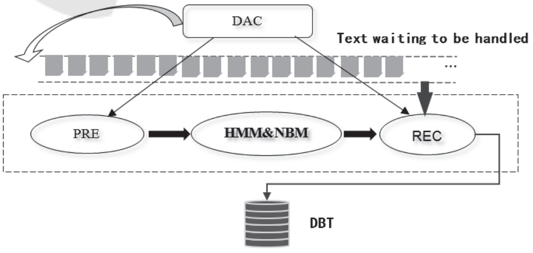

Figure 5: Framework of economic field people’s identification.

6 EXPERIMENT AND ANALYSIS

6.1 Data Source

The experiment of reorganization of people in

Finding of People in Economic Field based on Personal Profile

economic field uses the data from Baidu Encyclopedia and profiles of University Teachers' Introduction web page. The training samples and testing sample are those with good quality.

6.2 Training Nums

According to the classification criteria above, for sub-category of major category we collected 100 Biography text as the initial classification of training samples. Entire distribution shown in Fig 6.

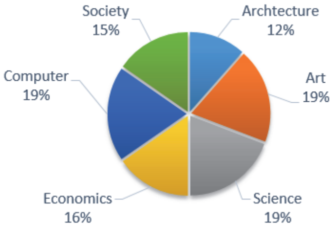

Figure 6: The classification.

6.3 Method

To test the accuracy of the method, we use confirmatory test method, which is to select a certain amount of the economic field profile texts as a basis for expansion, and randomly selected text not in economic field. The selected profile texts formed the experimental data.

6.4 Result and Analysis

The recognition of people in economic field conducted three experiments with a different number of economic field samples and a different total number of samples. The experiments results obtained are shown in Table 1.

The results showed that the reorganization method works well with the experiment samples and it is possible to recognize a large number of people in the economic field with the personnel profile. The recognition rate is about 80%.

Table 1: Result list.

	Economic samples	Total samples	Recognized	Correct	Accuracy rate	Recall rate	F1
Group 1	20	100	17	15	0.8823	0.7500	0.8107
Group 2	40	150	46	36	0.7826	0.9000	0.8372
Group 3	60	200	53	48	0.9056	0.8000	0.8495

7 CONCLUSIONS

The improvement of Chinese word segmentation technology proposed in this paper can improve the tech adapted in the field reorganization based on text classification. The framework proposed in this paper works well and has a good recognition rate. In the future study, and we will do further improvement in the recognition rate and the unstructured information extraction from the profile and form the structured information of the people.

REFERENCES

K. Aas, L. Eikvial. Text Categorization: a survey. Technical Report, Norwegian Computing Center, June 1999.

Fabrizio Sebastiani. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.

N. Fuhr, S. Hartmanna, G. Lustig, M. Schwantner, K. Tzeras. AIRX--a rule-based multistage indexing systems for large subject fields[C], Proceedings of RIAO’91 Conference, 1991:606-623.

Feng Guo-he. Automatic text classification technology research [J]. Journal of Information, 2007(12): 108-111.

Shang Wen-qian. Text classification and related technology research [D]. Beijing Jiaotong University, 2007

Yang Wen-chuan. Research on Chinese Text Segmentation based on double-arry Trie Tree [J]. Journal of Computer Engineering and Science. 2013, 35(9): 127-130.

439