An Algorithm for Title Classification on Scientific News

Wujie Shao[1, a], Hongjian Zhu[2, b], Yunyang Yan[1, c], Quanyin Zhu[1, d*]

[1]Faculty of Computer and Software Engineering Huaiyin Institute of Technology, Huaian, 223005, China

[2]School of Communication Engineering, Chongqing University, Chongqing, 400044, China

[a]email:15950376029@163.com, [b]email:overless@foxmall.com, [c]email:yunyang@hyit.edu.cn, [b]email:hyitzqy@126.com

Keywords: Text Categorization; Scientific News; Term Frequency; Machine Learning; Weighted Eigenvalue

Abstract. Nowadays, with the rapid development of Internet, news is growing explosively. In order to improve the practical value of scientific news, an algorithm of text categorization which is used to classify scientific news is designed. The number of scientific news is 16034, which is gotten from web pages. During the experiment, scientific news of life, medical scientific news, scientific news of earth, mathematical and physical scientific news, chemical scientific news and informational scientific news respectively achieved 72.04%, 64.05%, 71.59%, 67.88%, 66.84% and 61.35% accuracy rate. This algorithm gets good effect and improves the value of the scientific news collected from Web and the accuracy of scientific news detailed classification.

Introduction

In this paper a new optimal design of soccer robot control system which is based on mechanical analyses and calculations on the pressure and transmutation states of chip kick mechanics, this new control system with high precision for speed control and high dynamic quality.

In the context of the information age, the internet is generating information all the time. Automatic text [1] classification for processing large amounts of data technology is becoming increasingly important. For survival and development of an enterprise, how to filter out valuable information from the mass of scientific information, produces an important influence. This method of detailed classification of scientific news [2] is an efficient approach to obtaining the classification of scientific news.

Text Categorization [5], whose core is to build a function from a single text to category, is an important technology of data processing [6], divided into supervised learning unsupervised, semi-supervised learning, enhance learning and learning [7]. This algorithm for title classification of scientific news is an algorithm of Chinese text categorization [8] and bases on title of scientific news.

The titles of scientific news almost contains all the information to be expressed [3], so a method of detailed classification of scientific news is used to handle headlines [4], not only does this save processing time, improve efficiency, but also get a better classification results.

Text Preparation

The scientific news is used in the experiment, all from these websites (http://www.most.gov.cn/, http://www.nsfc.gov.cn/, http://www.kejixun.com/, http://www.sciencenet.cn/), selected the time from 2012 to 2015.

Model Training

Frist of all, the science of news is divided into six categories news, as follow, scientific news of life, medical scientific news, scientific news of earth, mathematical and physical scientific news, chemical scientific news and informational scientific news. Secondly, establishing science

vocabulary corpus. Finally, the experimental scientific news are classified into a known six scientific news category.

Extracting scientific news titles. Setting full titles of scientific news sets TITLE, which contain any title of scientific news T_n, so that

$$TITLE = \{T_1, T_2,\ldots, T_n\}, n = 1,2,3,\ldots \quad (1)$$

EXPT is that each of the training scientific news has 1000 scientific news selected randomly one of scientific news from TITLE, obtains the titles set of scientific news of life, the titles set of medical scientific news, the titles set of scientific news of earth, the titles set of mathematical and physical scientific news, the titles set of chemical scientific news and the titles set of informational scientific news, that is,

$$EXPT = \{\{ET_{1,1}, ET_{1,2},\ldots, ET_{1,1000}\}, \{ET_{2,1}, ET_{2,2},\ldots, ET_{2,1000}\}, \{ET_{3,1}, ET_{3,2},\ldots, ET_{3,1000}\}, \{ET_{4,1}, ET_{4,2},\ldots, ET_{4,1000}\}, \{ET_{5,1}, ET_{5,2},\ldots, ET_{5,1000}\}, \{ET_{6,1}, ET_{6,2},\ldots, ET_{6,1000}\}\} \quad (2)$$

Using IK Analyzer to handle six different titles sets and removing stop words, obtains six participle sets of scientific news of life, medical scientific news, scientific news of earth, mathematical and physical scientific news, chemical scientific news and informational scientific news, that is,

$$CORWORD = \{\{W_{1,1}, W_{1,2},\ldots, W_{1,a}\}, \{W_{2,1}, W_{2,2},\ldots, W_{2,b}\}, \{W_{3,1}, W_{3,2},\ldots, W_{3,c}\}, \{W_{4,1}, W_{4,2},\ldots, W_{4,d}\}, \{W_{5,1}, W_{5,2},\ldots, W_{5,e}\}, \{W_{6,1}, W_{6,2},\ldots, W_{6,f}\}\}, a=1,2,3,\ldots; b=1,2,3,\ldots; c=1,2,3,\ldots; d=1,2,3,\ldots; e=1,2,3,\ldots; f=1,2,3,\ldots \quad (3)$$

Word frequency is obtained by computing a word in the whole word set proportion, statistics of word frequency for Equation (3). Coming all the same words in the six words set CORWORD, make frequency of these words narrow 0.001 times, and the result which is six different scientific vocabulary corpus is given by Equation (5), so that,

$$WORDF = count(W_{I,J})/count(W_{P,Q}), I \in [1,P], J \in [1,Q], W_{P,Q} \in CORWORD \quad (4)$$

and

$$COR = \{\{(W_{1,1}, TF_{1,1}), (W_{1,2}, TF_{1,2}),\ldots, (W_{1,a}, TF_{1,a})\}, \{(W_{2,1}, TF_{2,1}), (W_{2,2}, TF_{2,2}),\ldots, (W_{2,b}, TF_{2,b})\}, \{(W_{3,1}, TF_{3,1}), (W_{3,2}, TF_{3,2}),\ldots, (W_{3,c}, TF_{3,c})\}, \{(W_{4,1}, TF_{4,1}), (W_{4,2}, TF_{4,2}),\ldots, (W_{4,d}, TF_{4,d})\}, \{(W_{5,1}, TF_{5,1}), (W_{5,2}, TF_{5,2}),\ldots, (W_{5,e}, TF_{5,e})\}, \{(W_{6,1}, TF_{6,1}), (W_{6,2}, TF_{6,2}),\ldots, (W_{6,f}, TF_{6,f})\}\} \quad (5)$$

Experimental Text

Experimental titles of scientific news are 50% of titles of scientific news, which is all the titles of scientific news wiped off the rest of the training titles of scientific news, so that

$$TET = \{\{AT_{1,1}, AT_{1,2},\ldots, AT_{1,u}\}, \{AT_{2,1}, AT_{2,2},\ldots, AT_{2,v}\}, \{AT_{3,1}, AT_{3,2},\ldots, AT_{3,w}\}, \{AT_{4,1}, AT_{4,2},\ldots, AT_{4,x}\}, \{AT_{5,1}, AT_{5,2},\ldots, AT_{5,y}\}, \{AT_{6,1}, AT_{6,2},\ldots, AT_{6,z}\}\}, u=1,2,3,\ldots; v=1,2,3,\ldots; w=1,2,3,\ldots; x=1,2,3,\ldots; y=1,2,3,\ldots; z=1,2,3,\ldots \quad (6)$$

Performing word processing for each title which comes from TET, show that,

$$TETW = \{\{TW_{1,1}, TW_{1,2},\ldots, TW_{1,r}\}, \{TW_{2,1}, TW_{2,2},\ldots, TW_{2,s}\},\ldots, \{TW_{m,1}, TW_{m,2},\ldots, TW_{m,t}\}\}, Z = u + v + w + x + y + z, m \in [1, Z], r=1,2,3,\ldots; s=1,2,3,\ldots; t=1,2,3,\ldots \quad (7)$$

Word frequency is obtained by computing a word in the whole word set proportion, statistics of word frequency for Equation (7), and the result is given by Equation (8), so that

$$WF = \{\{(TW_{1,1}, WTF_{1,1}), (TW_{1,2}, WTF_{1,2}),\ldots, (TW_{1,r}, WTF_{1,r})\}, \{(TW_{2,1}, WTF_{2,1}), (TW_{2,2}, WTF_{2,2}),\ldots, (TW_{2,s}, WTF_{1,s})\},\ldots, \{(TW_{m,1}, WTF_{m,1})TW_{m,1},\ldots, TW_{m,t}\}\} \quad (8)$$

Experimental News Text Classification

Considering two sets OS_i and SE_j, where OS_i is the element $TF_{K,L}$ of the set COR, $K \in [1, 6]$, if K=1, $L \in [1, a]$, else if K=2, $L \in [1, b]$, else if K=3, $L \in [1, c]$, else if K=4, $L \in [1, d]$, else if K=5, $L \in [1, e]$, else K=6, $L \in [1, f]$. SE_j is the element $WTF_{C,D}$ of the set WF, $C \in [1, m]$. When two

words which one comes from COR and the other one comes from WF are the same, OS_q and SE_q are their $TF_{K,L}$ and $TW_{C,D}$, so that,

$$SIM = \frac{\sum_{q=1}^{Q}(OS_q * SE_q)}{\sqrt{\sum_{i=1}^{M}(OS_i)^2} * \sqrt{\sum_{j=1}^{N}(SE_j)^2}} \quad (9)$$

Specifically, for example, experimental title of scientific news $AT_{1,1}$, its participle is A = {$TW_{1,1}$, $TW_{1,2}$,…, $TW_{1,r}$}, and its word frequency set is B = {($TW_{1,1}$, $WTF_{1,1}$), ($TW_{1,2}$, $WTF_{1,2}$),…, ($TW_{1,r}$, $WTF_{1,r}$)}, which is compared with the six different scientific vocabulary corpus, the molecular of formula for the product of word frequency, when the six corpus words and word focused words are the same, so that,

$$Sim_1 = \frac{\sum_{q=1}^{Q}(TF_{1,q} * TW_{1,q})}{\sum_{i=1}^{a}(TF_{1,i})^2 * \sum_{j=1}^{r}(TW_{1,j})^2} \quad (10)$$

$$Sim_2 = \frac{\sum_{q=1}^{Q}(TF_{2,q} * TW_{1,q})}{\sum_{i=1}^{b}(TF_{2,i})^2 * \sum_{j=1}^{r}(TW_{1,j})^2} \quad (11)$$

$$Sim_3 = \frac{\sum_{q=1}^{Q}(TF_{3,q} * TW_{1,q})}{\sum_{i=1}^{c}(TF_{3,i})^2 * \sum_{j=1}^{r}(TW_{1,j})^2} \quad (12)$$

$$Sim_4 = \frac{\sum_{q=1}^{Q}(TF_{4,q} * TW_{1,q})}{\sum_{i=1}^{d}(TF_{4,i})^2 * \sum_{j=1}^{r}(TW_{1,j})^2} \quad (13)$$

$$Sim_5 = \frac{\sum_{q=1}^{Q}(TF_{5,q} * TW_{1,q})}{\sum_{i=1}^{e}(TF_{5,i})^2 * \sum_{j=1}^{r}(TW_{1,j})^2} \quad (14)$$

$$Sim_6 = \frac{\sum_{q=1}^{Q}(TF_{6,q} * TW_{1,q})}{\sum_{i=1}^{f}(TF_{6,i})^2 * \sum_{j=1}^{r}(TW_{1,j})^2} \quad (15)$$

and

$$S = \{Sim_1, Sim_2, Sim_3, Sim_4, Sim_5, Sim_6\} \quad (16)$$

Getting the max value of elements in set S (16), that is,

$$MAX = \max(Sim_1, Sim_2, Sim_3, Sim_4, Sim_5, Sim_6) \quad (17)$$

If MAX = Sim_1, experimental scientific news to be classified belongs to scientific news of life, else if MAX = Sim_2, experimental scientific news to be classified belongs to medical scientific news, else if MAX = Sim_3, experimental scientific news to be classified belongs to scientific news of earth, else if MAX = Sim_4, experimental scientific news to be classified belongs to mathematical and physical scientific news, else if MAX = Sim_5, experimental scientific news to be classified belongs to chemical scientific news, else if MAX = Sim_6, experimental scientific news to be classified belongs to informational scientific news. And the other experimental scientific news to be classified are classified into six different scientific news by this algorithm.

Test Results

Table 1 is description of the total number of scientific news, which is obtained from the Web pages. The total number of scientific news is 16034. Specifically, the number of scientific news of life, the number of medical scientific news, the number of scientific news of earth, the number of mathematical and physical scientific news, the number of chemical scientific news and the number of informational scientific news are 3332, 2451, 2421, 2643, 1765 and 3422.

Table 1. The total number of six dig scientific news

Classification of scientific news	Number of scientific news
scientific news of life	3332
medical scientific news	2451
scientific news of earth	2421
mathematical and physical scientific news	2643
chemical scientific news	1765
informational scientific news	3422

Table 2 is that the number of training scientific news is 6000, in other words, each of the training scientific news has 1000 scientific news selected randomly one of scientific news. Experimental scientific news are randomly selected 50% of scientific news that is the remainder of the training titles of scientific news. The number of correct classification is the scientific news classification correct number.

Table 2. Experimental result description

Classification of scientific news	Number of training scientific news	Number of experimental scientific news	Number of correct classification	Accuracy rate
scientific news of life	1000	1166	840	72.04%
medical scientific news	1000	726	465	64.05%
scientific news of earth	1000	711	509	71.59%
mathematical and physical scientific news	1000	822	558	67.88%
chemical scientific news	1000	383	256	66.84%
informational scientific news	1000	1211	743	61.35%

Table 3 is another set of experimental result. The experimental scientific news is are randomly selected 50% of scientific news that is the remainder of the training titles of scientific news. The number of correct classification is the scientific news classification correct number.

Table 3. Experimental result description

Classification of scientific news	Number of experimental scientific news	Number of correct classification	Accuracy rate
scientific news of life	1166	835	71.61%
medical scientific news	726	457	62.95%
scientific news of earth	711	511	71.87%
mathematical and physical scientific news	822	546	66.42%
chemical scientific news	383	259	67.62%
informational scientific news	1211	732	60.45%

Conclusion

This algorithm for title classification of scientific news is used to classify scientific news into six categories. First of all, selected randomly six different scientific news, any category has 1000 titles, and processing them, forming the six different corpus. Second, processing scientific news to be classified, and using Equation (9) to calculate between scientific news to be classified and any

corpus. Finally, Comparing results, and scientific news are classified into one of classification of scientific news. The experiment achieved good effect and improves the value of the scientific news collected from Web and the accuracy of scientific news detailed classification.

Acknowledgement

This work is supported by the Key Research Plan of Jiangsu Province. China (BE2015127), the University Science Research Project of Jiangsu Province (15KJB520004), the Science and Technology Projects of Huaian (HAG2015060, HAG2014028), the Project of National Undergraduates Innovation under the Grant No.201511049013Z and Scientific Foundation Project of Huaiyin Institute of Technology (HGC1412).

Corresponding Author

Quanyin Zhu, Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, 223005, China.

References

[1] Xiao-Jun Tong, Ming-Gen Cui, Research on Chinese Text Automatic Categorization Based on VSM, International Conference on Wireless Communications, Networking and Mobile Computing [C], 2007, 3863 – 3866.

[2] N. Ishii, T. Murai, Text Classification by Combining Grouping, LSA and kNN, 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-B [C], 2006, 148 – 154.

[3] Ching-hao Mao, Semantic Similarity Measurement of Chinese Financial News Titles Based on Event Frame Extracting, IEEE International Conference on e-Business Engineering [C], 2006, 229 – 236.

[4] Xin Liu, Gao Rujia, Song Liufu, Internet news headlines classification method based on the N-Gram language model, International Conference on Computer Science and Information Processing [C], 2012, 826 – 828.

[5] Wang Fei, Li Cai-Hong, A Two-Stage Feature Selection Method for Text Categorization by Using Category Correlation Degree and Latent Semantic Indexing [J], Journal of Shanghai Jiaotong University, 2015, 44 – 50.

[6] Lu Pan, Quanyin Zhu, An Identification Method of News Scientific Intelligence Based on TF-IDF, International Symposium on Distributed Computing and Applications for Business Engineering and Science [C], 2015, 501 – 504.

[7] Gao Huang, Shiji Song, Jatinder N. D. Gupta, Semi-Supervised and Unsupervised Extreme Learning Machines [J], IEEE Transactions on Cybernetics, 2014:44 (12): 2405 – 2417.

[8] Yaohong Jin, Wen Xiong, Cong Wang, Feature selection for Chinese Text Categorization based on improved particle swarm optimization, International Conference on Natural Language Processing and Knowledge Engineering [C], 2010, 1 – 6.