A Method of Text Dimension Reduction Based on CHI and TF-IDF

HaiBo Tang[1, a], Lei Zhou[1, b], Xu Chengjie[1,d], Quanyin Zhu*[1, d]

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

[a]1508thb@sina.com, [b]zl_gxy@163.com, [c]1589206823@qq.com, [d]hyitzqy@126.com

Keywords: Dimension extraction; CHI; TF-IDF; Text classification

Abstract. In order to improve the result of text dimension extraction, a text dimension reduction method which is based on CHI and TF-IDF is designed and realized. News websites on the Internet provide news articles, through the page analysis based on SVM and application method designed based on CHI and TF-IDF, good results achieved by extracting text dimensions from news website. According to the news articles from NetEase and ChinaNews, the proposed algorithm model of the text dimension extraction is designed and achieved 81.2% accuracy of the text classification, which provided the data support for the method designed. This method can make up a low frequency word defect of CHI and get efficient result on text classification.

Introduction

With the development of the Internet, a large number of document data emerged on the Internet. The technology of automatic text classification [1] used to process massive data becomes more and more important; it has gradually become key technologies to organizing and processing large amount of document data. As regard to text categorization, text preprocessing has been a bottleneck of the classification, result of the text's preprocessing affects the performance of the classifier [2] directly.

Prescribing test and TF-IDF are the common means applied in text classification. Prescribing test is a commonly used text categorization feature selection [3] methods. The method is only concerned about the relationship between words and categories [4], without considering the association between words; so that the selected feature set have greater redundancy. TF-IDF text feature word extraction method is based on statistics [5], it treats words as independence unit, to determining a characteristic word of the text, it count the frequency of the words appears in texts and the number of text which contains the same words in the collection of the text. Although this method gets achievements in reducing the time of calculation and simplifying the extraction [6] steps of the feature words. However, this method didn't take the relationship between words into consideration and ignore the higher discrimination of low-frequency words, which limits the accuracy of text feature extraction. This paper proposes a method which takes advantage of the TF-IDF and CHI and improves the accuracy of the text categorization. By pretreatment of the CHI [7], experiment managed to make use of the TF-IDF to make up the deficiencies of CHI.

Training Process

Definition. The news collected from websites is $News = \{D_1,D_2,\ldots,D_i\}$. D_i has one part, include content. The *News* is divided into two classes that will be trained by SVM, which is $SCN = \{\{ D_1,D_2,\ldots,D_w \},\{ D_1,D_2,\ldots,D_k \}\}$ $(w+k=n)$, D_i can be described in $D_i=\{word_1, word_2, \ldots word_n\}$ $(i \in (j+z))$; Get all the different words from *SCN* of one category is $AW_m = \{word_1,word_2,word_3,\ldots word_m\}$ and the frequency of each word in each article is $WF_n = \{f_{word_1}, f_{word_2} \ldots f_{word_n}\}$. Furthermore, the frequency of each word in AW_e that appears in different article category is : $WT_{AW_m} = \{ST_{AW_m}, NST_{AW_m}\}$, Set the CHI value of each word that belongs to one area is given in Eq 1:

$$CT_{word_{pn}} \frac{(ST_{AW_m}(n - NST_{AW_m}) - NST_{AW_m}(n - ST_{AW_m}))^2}{(ST_{AW_m} + NST_{AW_m})(2n - NST_{AW_m} - ST_{AW_m})} \quad (1)$$

The TF-IDF value of each word in one article is given in Eq 2:

$$WT_{word_n} = \frac{f_{word_n}}{\sum_n n_{n,n}} * \log \frac{|D|}{1+|j : word_n \in D_i|} \quad (2)$$

$\sum_n n_{n,n}$ Represents the appearance times of word which is most frequent, |D| represents the number of articles that in corpus and *j* represents the number of articles that contain $word_n$ in the corpus. The sum of each word's TF-IDF is $SWT_{word_m} = WT_{word_1} + WT_{word_2} + \ldots + WT_{word_r}$. Set the new weight of each word for one area is given in Eq 3, p = 1,2,…m:

$$WW_m \frac{WT_{word_m} - \min(WT_{word_m})}{\max(WT_{word_p}) - \min(WT_{word_p})} * CT_{word_m} \quad (3)$$

And then, the result of each word weight that belong to one area is : $SWW_m = \{WW_1, WW_2, \ldots WW_m\}$

Experimental Environment. In this paper, The data of articles that are used in the experiments, all from these websites (http://www.chinanews.com, http://news.163.com/,http://news.sina.com.cn/,http://news.qq.com/, http://www.most.gov.cn/), select the time from June 2015 to August 2015. Every article is made up of one part, which is its content. The new method of text dimension extraction is based on CHI and TF-IDF, The machine is Window 7, RAM is 6G. The tool of Jieba of participle is used to build the proposed model (http:// www.iteye.com/news/26184-jieba). The news is collected from website using PyQuery from Python 2.7.

The General Process of Experiment. Obtain 300 tech-news and 300 non-tech news articles from website is *News = {{D_{x1}, D_{x2},…,D_{xn}},{D_{x1},D_{x2},…,D_{xn}}}*, each article is made up of one part that is its content, and set *SCN = News.*

Text pretreatment for *SCN*, include text participle, remove stop-words and name or special word. Get *Wi = {W1, W2 ,…Wi}.*

Get different words from two classifications: W_p *= {{W_1, W_2, …W_j},{ W_1,W_2,…W_k }}.*

Get the word frequency of each article is : $WF_n = \{f_{word_1}, f_{word_2} \ldots f_{word_n}\}$

Get all the different word from *SCN*, the result is: AW_m *= {$word_1$, $word_2$, $word_3$, … $word_m$}.*

According to Equation (1), the CHI value of different words are gotten in :

$$CT_{word_{pn}} \frac{(ST_{AW_e}(n - NST_{AW_e}) - NST_{AW_e}(n - ST_{AW_e}))^2}{(ST_{AW_e} + NST_{AW_e})(2n - NST_{AW_e} - ST_{AW_e})}$$

According to WF_n, D_i , get the TF-IDF of the word in one article is:

$$WT_{word_{pn}} = \frac{f_{word_{pn}}}{\sum_n n_n, n} * \log \frac{|D|}{1+|j : word_{pn} \in D_i|}$$

According to the TF-IDF above and get the former 65 percent words of each article and collect the words from all the articles from *SCN* and need a process of deduplication and get the sum of the word's TF-IDF according to Equations(10):

$$WTIF_m = \{wtif_1, wtif_2, \ldots wtif_m\} \quad (13)$$

According to Eq 3, each value in the set needs to be normalized, and then get Equation (10) is as follow:

$$WE_m = \{we_1, we_2, \ldots we_m\} \quad (14)$$

Get the new weight of each word which was gotten in Step 6.

$$NW_m = \text{WCF}m * \text{WTIF}m \quad (15)$$

And Get the first 1000,800,750 dimensions from Equation (12) and get tree dimensions Set DM1, DM2, DM3.

According to Equation (8), Get first 1200,900,850,800,750 words of each classification are as follows:

CW = {{TEW_f}, {NTW_j}} f, j = 1200,900,850,800,750

Save model parameter, include *DM1*, *DM2*, *DM3*, *SCN*, *CW*.

Algorithmic introduction

Step 1: Pretreat for each article, include text participle, removing stop-words and name or special word. The words are all from one specific classes.
Step 2: Get all the TF-IDF value of each word that appears in each article and remove the word whose value is the latter one-third.
Step 3: Get the words gotten from **Step 2** deduplication which are all from the same class and calculate the sum of its TF-IDF.
Step 4: Normalize the sum of TF-IDF of different words gotten in **step 3.**
Step 5: Get all the CHI of each word that gotten in **step 2.**
Step 6: Get the new weight of each word that was gotten in **Step 3**, which is its normalized TF-IDF * CHI gotten in **step 5.**
Step 7: Extraction the dimension of news whose weight is in the first two-thirds that get in **Step 6**.

Experimental Results

This experiment use Sogou Chinese Corpus and text categorization algorithm is SVM. In this part, list 3 improved models' result.

Table 1. Experimental result description

Dataset name	Number of samples	Category name	Experiment group
Science-Data	**200**	***Tech-news***	**1**
Non-Science-Data	**200**	***Non-tech-news***	**1**
Non-Class	**200**	***Test data***	**1**
Science-Data	**250**	***Tech-news***	**2**
Non-Science-Data	**250**	***Non-tech-news***	**2**
Non-Class	**200**	***Test data***	**2**
Science-Data	**300**	***Tech-news***	**3**
Non-Science-Data	**300**	***Non-tech-news***	**3**
Non-Class	**200**	***Test data***	**3**

Table 1 is the description of experimental data set. Which is consist of tech news and non-tech news, each article has only one dimension that is content. The number of samples represents the number of texts that belongs to each category.

Table 2. Experimental result description

Experiment name	Dimension	Dimension extraction	Accuracy
Experiment 1	1200	CHI	79%
	900	CHI	80%
	850	New Method	80.20%
Experiment 2	900	CHI	80%
	750	CHI	80.60%
	750	New Method	81%
Experiment 3	1000	CHI	80%
	800	CHI	80.40%
	800	New Method	81.20%

Analysis

Sample Set is consist of 500 articles which is made up of 250 science- news and 250 non-science article, 400 articles which is made up of 200 science- news and 200 non-science article, 600 articles which is made up of 300 science- news and 300 non-science article. After manual testing,

the accuracy of new method is more than the result taken by CHI with fewer dimensions, which plays an important role in Text Dimension Reduction; meanwhile, this method can improve the efficiency of classifier, making it to be more sensitive to different text.

Conclusion

Our research work of text dimension reduction method of text extraction comes from our related work. The website has a large number of long articles about different area. The design has realized the dimension reduction method of long text. CHI is only concerned about the relationship between words and categories; the new method has improved the result of text Dimension extraction. On this basis, the accuracy of text classification by SVM has reached 81.2%, This experiment has achieved good effects and improve the result of text classification that are from website.

Acknowledgement

This work is supported by the National Sparking Plan Project of China (2011GA690190), the Key Research Plan of Jiangsu Province. China (BE2015127), the fund of Huaian Industry Science and Technology. China (HAG2014023, HAG2014028)

Corresponding Author

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

Reference

[1] Shuigeng Zhou, Tok Wang Ling, Jihong Guan, Jiangtao Hu, Aoying Zhou. Fast text classification: a training-corpus pruning based approach. Proceedings of IEEE International Conference on Database Systems for Advanced Applications. IEEE Computer Society, Kyoto, Japan, 2003. 127 – 136

[2] Bogdanov, A.V. Neuroinspired Architecture for Robust Classifier Fusion of Multisensor Imagery. Proceedings of IEEE Transactions on Geoscience and Remote Sensing. 2008. 1467 - 1487

[3] ErHeng Zhong, Sihong Xie, Wei Fan, Jiangtao Ren, Jing Peng, Kun Zhang. Graph-Based Iterative Hybrid Feature Selection. Proceedings of IEEE International Conference Data Mining. IEEE Computer Society, Pisa, 2008. 1133 – 1138

[4] Mingmin Xu, Liang He, Lin Xin. A Refined TF-IDF Algorithm Based on Channel Distribution Information for Web News Feature Extraction. Processing of 2010 Second International Workshop on Education Technology and Computer Science. 2010. 15 - 19

[5] Na Wang, Pengyuan Wang, Baowei Zhang. An improved TF-IDF weights function based on information theory. Proceedings of IEEE International Conference on Computer and Communication Technologies in Agriculture Engineering. IEEE Computer Society, Chengdu, 2010. 439 – 441

[6] Xiaodong Huang. A novel video text extraction approach based on Log-Gabor filters. Proceedings of IEEE International Conference on Image and Signal. IEEE Computer Society, Shanghai, 2011, 474 – 478

[7] Ming-zhen Liu; Ya-Feng Liu. Tai Chi Thought and research on software problems. Proceedings of International Joint Conference on Computer Science and Software Engineering. 2015. 126 – 131