An Improved Feature Weighting Strategy in Chinese Text Categorization

Jia Song[1,a], Sijun Qin[2,b *] and Pengzhou Zhang[3,c]

[1] The Faculty of Science and Technology, Communication University of China, Beijing, China

[2] New Media Institute, Communication University of China, Beijing, China

[3] The Faculty of Science and Technology, Communication University of China, Beijing, China

[a]songjia@cuc.edu.cn, [b]wingol88888@163.com, [c]zhangpengzhou@cuc.edu.cn

Keywords: Feature Weighting, Inverse Category Frequency, TFIDF-ICF, Text Categorization

Abstract. In the process of document formalization, feature weight algorithm plays an essential role. It greatly interferes with the performance of the classifier. To improve the classic TF-IDF algorithm on its shortcomings that ignore feature distribution among the classes, we develop a new strategy to weight feature based on ICF and traditional TF-IDF. We have conducted a series of experiments on two text corpuses, namely the TanCorpV1.0 and Sogou, to analyze the performance of our strategy, which are described in the paper. Experimental results demonstrate the proposed strategy can to some extent improve the performance of text categorization.

Introduction

With the advent of computers, storing large amounts of text in electronic form became an easy task. But it is difficult and significant to search, organize and manage these huge information resources. Therefore automatic text categorization (TC) otherwise identified as Text Classification is widely used to accomplish this task [1,2]. Document formalization founds the bases of language models, such as Vector Space Model (VSM) [3], Latent Semantic Analysis model, etc. It also consequently impacts on the accuracy of application technologies of Natural Language Processing, such as Information Retrieval, Text Categorization and so on.

In the process of document formalization, documents are represented by document vectors which are expected to indicate as much information of the documents as possible. To make the representation accurate, term weight strategy plays an essential role in the process. A large number of automatic term weighting strategies are available [4,5,6]. TF-IDF is the most widely used term weight algorithm nowadays in today's information retrieval systems. However, it has drawbacks as well. Therefore, an improved Feature Weight Strategy based on TF-IDF and ICF is put forward to address TF-IDF drawbacks in this paper.

Related Work

The Traditional TF-IDF Algorithm. The classic TF-IDF term weighting strategy is put forward by Salton [7,8,9]. The guideline is: term with higher term frequency and lower document frequency is assigned a higher term weight. It considers two aspects, one is TF(Term Frequency) which is used to represent term ability to describe the document content and second is IDF (Inverse Document Frequency) which is used to represent term ability to differentiate the document content. That is, term with higher term frequency has higher capability to describe the document content; term with lower document frequency has higher capability to differentiate the document content. TF-IDF is widely used in the field of IR (Information Retrieval) after many research and scientific proof. The classic TF-IDF term weight formula is as follows:

$$W(t_k,d)=\frac{tf(t_k,d)\times log_2\left(\frac{N}{n_k}+\alpha\right)}{\sqrt{\sum_{i=1}^{P}\left[tf(t_i,d)\times log_2\left(\frac{N}{n_i}+\alpha\right)\right]^2}} \quad (1)$$

Where, $tf(t_k,d)$ is the term t_k frequency in the document d, N is the text number of corpus, n_k is the text number given that the term t_k occurs, α is a constant (the value is 0.01 in general), $\log2(N/n_k+\alpha)$ is a function involved with inverse document frequency, P is the number of features and the denominator is the normalized factor.

However, the traditional TF-IDF method has its inevitable shortcomings. The simple structure of IDF is insufficient to reflect the distribution of features, especially the category distribution of features. Therefore, it is unable to perform well for feature weight adjustment. And, we are motivated to investigate other statistical characteristics of terms and found ICF (Inverse Category Frequency) discriminator after analyzing the category distribution of term statistically.

Inverse Category Frequency. In this paper the ICF (Inverse Category Frequency) concept, which can be used to measure the statistical correlation between features and categories, is presented. ICF higher values, greater independence between the words in the category. That is, the higher ICF value a feature has, the greater independence it has among classes. It is based on the following assumption: the lower ICF value a feature has, the more help for classifying a document that contains the feature. The ICF formula is as follows:

$$ICF(t_k)=\log_2\left(\frac{N_C}{n_C(t_k)}+\beta\right) \quad (2)$$

Where, N_C is the category number of documents in corpus, $n_C(t_k)$ is the category number given that the term t_k occurs, and β is a constant (the value is 0.1 in general). ICF considers various possibilities of the category number of characteristics and is able to perform well for feature weight adjustment. The validity of the feature weight and classification ability of feature is very good and stable.

TFIDF-ICF Algorithm. Above all, we proposed a new feature weighting strategy based on ICF and traditional TF-IDF to improve the classic TF-IDF algorithm on its shortcomings. The TFIDF-ICF feature weight formula is as follows:

$$W(t_k,d)=\frac{tf(t_k,d)\times log_2\left(\frac{N}{n_k}+\alpha\right)\times \log_2\left(\frac{N_C}{n_C(t_k)}+\beta\right)}{\sqrt{\sum_{i=1}^{P}\left[tf(t_i,d)\times log_2\left(\frac{N}{n_i}+\alpha\right)\times \log_2\left(\frac{N_C}{n_C(t_i)}+\beta\right)\right]^2}} \quad (3)$$

Where, $tf(t_k,d)$ is the feature tk frequency in the document d, N is the text number of corpus, n_k is the text number given that the term t_k occurs, N_C is the category number of documents in corpus, $n_C(t_k)$ is the category number given that the term t_k occurs, α and β are both constant (the value of α and β are 0.01 and 0.1 in general respectively), $\log2(N/n_k+\alpha)$ is a function involved with inverse document frequency, and P is the number of features and the denominator is the normalized factor.

Data set and Experimental Setting

Data Set. Data set of text classification is the premise and basis in this experiment. There is not a standard corpus for Chinese text categorization. In this paper we adopted TanCorpV1.0[10] and Sogou[11] corpus after analysis and comparison. To verify the feasibility of feature weighting strategy, we performed it on two distinct corpuses comparing to TF-IDF.

The TanCorpV1.0 corpus consists of 14120 documents, the number of training documents is 9814, and the number of test documents is 4306. The document number of 10 categories is different, and the ratio of training set to test set of each category is about seven for three. Detailed distribution of data set

is shown in Table 1. The Sogou corpus contains eight classes and 15920 documents. The document number of 8 categories is same, and the ratio of training set to test set of each category is also about seven for three. Detailed distribution of data set is shown in Table 2.

Five-fold cross validation is employed to generalize the experimental results. Every corpus is selected and divided into five parts evenly and randomly. Each time, one part of them is selected as test data and the results four are used as training data. It means the ratio of training set to test set of each category is also about four for one. The experimental results are averaged based on the five experiments.

Tab. 1 distribution of Tancorpv1.0 data set

Category	The number of training set	The number of test set
Talents	426	182
Sports	1964	841
Health	984	422
Entertainment	1050	450
House	654	281
Education	566	242
Cars	413	177
Computer	2060	883
Technology	728	312
Finance	573	246

Tab. 2 distribution of Sogou data set

Category	The No. of training set	The No. of test set
Sports	1393	597
Health	1393	597
IT	1393	597
Culture	1393	597
Job	1393	597
Travel	1393	597
Military	1393	597
Finance	1393	597

Experimental Setting.

Feature Selection. In this section, we present five commonly known feature selection (see Table 3) methods, Mutual Information (MI), $\chi 2$ statistics (CHI), Information Gain (IG), Odds ratio (OR) and GSS coefficient(GSS)[12]. In the interest of brevity, we have omitted their mathematical justification. Subsequently, we performed the term weighting scheme on these feature selection methods to verify our study. Using the five feature selection methods discussed in Table 3, 3000, 6000, 10000, 20000, and 35000 features are selected after stemming and stop words removing.

Tab. 3 feature selection methods used in this paper

Algorithm	Denoted by	Formula
Information Gain	$IG(t_k, c_i)$	$\sum_{c\in\{c_i,\overline{c_i}\}}\sum_{t\in\{t_k,\overline{t_k}\}} P(t,c)\log\frac{P(t,c)}{P(t)P(c)}$
Mutual information	$MI(t_k, c_i)$	$\log\frac{P(t_k,c_i)}{P(t_k)P(c_i)}$
Chi-square	$\chi\ (t_k, c_i)$	$\frac{N[P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(t_k,\overline{c_i})P(\overline{t_k},c_i)]}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$
Odds ratio	$OR(t_k, c_i)$	$\log\frac{P((t_k\|c_i))(1-P((t_k\|\overline{c_i})))}{P((t_k\|\overline{c_i}))(1-P((t_k\|c_i)))}$
GSS coefficient	$GSS(t_k, c_i)$	$P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(t_k,\overline{c_i})P(\overline{t_k},c_i)$

Feature Selection. In this section, we present five commonly known feature selection (see Table 3) methods, Mutual Information (MI), $\chi 2$ statistics (CHI), Information Gain (IG), Odds ratio (OR) and GSS coefficient(GSS). In the interest of brevity, we have omitted their mathematical justification. Subsequently, we performed the term weighting scheme on these feature selection methods to verify our study. Using the five feature selection methods discussed in Table 3, 3000, 6000, 10000, 20000, and 35000 features are selected after stemming and stop words removing.

Classifier. To evaluate the performance of the proposed TFIDF-ICF term weighting strategy in the field of text categorization, K-Nearest Neighbour (KNN) [13] is utilized as the text classifier for the reason that it is one of the most widely used yet a very simple algorithm. The K value is set to 20 in the experiments.

Performance Evaluation. For evaluating the performance of a text classifier, the standard measures – precision, recall and F1, as well as those used in conventional information retrieval, is used. From the perspective of probability, precision is defined as the conditional probability that given a category c, the probability that assign the category to a test document d is correct. The recall is also defined as a conditional probability that if d ought to be assigned c, this decision is taken [14]. Given the contingency table of category C_i as shown in Table 4. In this table, a is the number of documents correctly assigned to C_i, b is the number of documents incorrectly assigned to C_i, c is the number of documents incorrectly rejected by C_i, d is the number of documents corrected rejected by C_i. The precision (P_i), recall (R_i), and F1 measure ($F1_i$) of category C_i are calculated as follows:

$$P_i = \frac{a}{a+b},\quad R_i = \frac{a}{a+c},\quad F1_i = \frac{2\times P_i \times R_i}{(P_i + R_i)} \tag{4}$$

Tab. 4 the contingency table for category ci

Category c_i		**Expert Judgement**	
		Yes	**No**
Classifier Judgement	**Yes**	a	b
	No	c	d

In this paper, we aggregated these measures over all categories by this method called Macro-averaged. Macro-averaged F1 measures as formula (5) below, where C expresses the number of categories.

$$Macro_{F1} = \frac{\sum_{i=1}^{C} F1_i}{C} \tag{5}$$

Experimental Results and Analysis

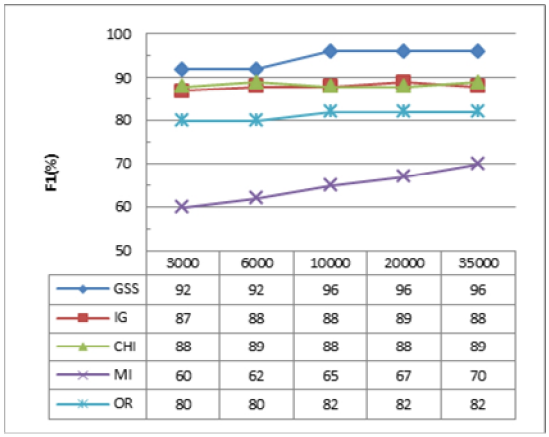


Fig. 1 Micro-averaged F1 values of KNN (TFIDF) for five feature selection methods on Sogou

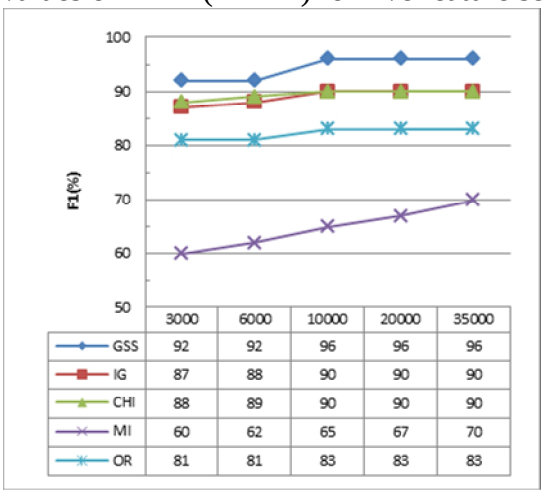


Fig. 2 Micro-averaged F1 values of KNN (TFIDF-ICF) for five FS methods on Sogou

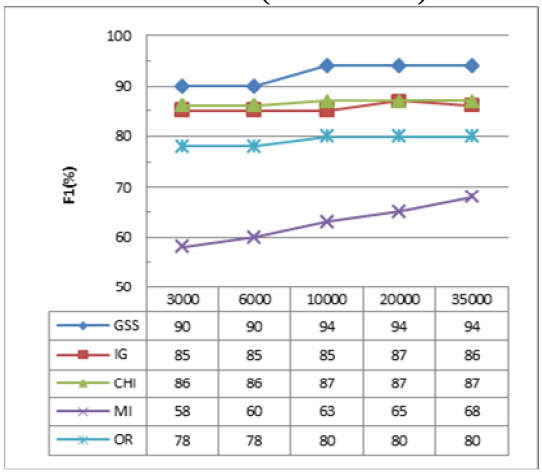
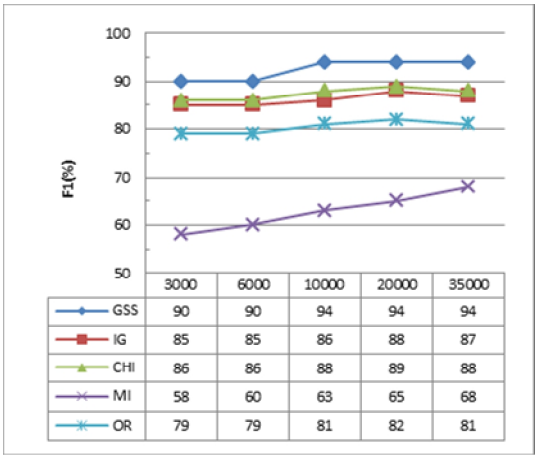


Fig. 3 Micro-averaged F1 values of KNN (TF-IDF) for five FS methods on TanCorpV1.0



	3000	6000	10000	20000	35000
GSS	90	90	94	94	94
IG	85	85	86	88	87
CHI	86	86	88	89	88
MI	58	60	63	65	68
OR	79	79	81	82	81

Fig. 4 Micro-averaged F1 values of KNN (TFIDF-ICF) for five FS methods on TanCorpV1.0

Fig. 1 and Fig. 3 present the micro-averaged F1 for the five feature selection algorithms on Sogou and TanCorpV1.0 respectively. In Fig.1, the micro-averaged F1 over the five different numbers of features in descending order are 94.4% for GSS coefficient; 88.4% for chi-square; 88.0% for information gain; 81.2% for Odds Ratio; and it is worth to note that mutual information is ranked at last with a low averaged F1 value, 64.8%.

In Fig. 3, the micro-averaged F1 over the five different numbers of features in descending order are similar with it in Fig. 1.

Fig. 2 and Fig. 4 present the micro-averaged F1 when using TFIDF-ICF term-weighting strategy to classify documents respectively. Compared with the results presented in Fig. 1 and Fig.3, the data in the figure reveal that TFIDF-ICF term-weighting method can effectively improve the performance of KNN when features are selected by chi-square, information gain, and Odds Ratio, whereas produce the equivalent results when mutual information and GSS coefficient feature selection methods are employed. The improvements based on the three feature selection algorithms are about 2.07% and 2.0% on Sogou and TanCorpV1.0 respectively. The overall averaged F1 over all different numbers of features are 80.48% and 77.2% if TF-IDF is used on two corpuses, and 81.72% and 78.40% when TFIDF-ICF is employed. On average, the improvements are 1.24% and 1.20% respectively.

Conclusion

The paper describes a new feature weighting method based on TF-IDF term weighting strategy. The possibility that it can be used to modify the feature weight has been proven in this study. It has suggested that the use of positive feature weight strategy can be comparatively effective. We have found out that TFIDF-ICF works well in the field of text categorization. The experimental result has shown that the weight of important and representative feature is raised and the effect of the unimportant feature to classification is decreased .And the new strategy improves the performance of the text classifiers obviously. Next, we will adopt other classification algorithms such as support vector machine (SVM) to verify the feasibility of feature weighting strategy.

Acknowledgment

The work was supported by the project of the Xinhua News Agency scientific research in 2014 and the project of National Key Technology R&D Program (2012BAH15B01-1, 2014BAK10B01-01, and 2014BAK10B01-02).

References

[1] D. Cai and X. He, Manifold adaptive experimental design for text categorization, IEEE Trans. Knowl. DataEng., vol. 24, no. 4, pp.707 -719 2012.

[2] S. Qin, J. Song, P. Zhang, and Y. Tan, Feature Selection for Text Classification Based on Part of Speech Filter and Synonym Merge, Proceeding of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'15),2015,pp.717-721.

[3] D. L. Lee, H. Chuang, and K. Seamons, Document ranking and the vector-space model, IEEE Software, 14(2):67-75, March 1997.

[4] Xuxiang Tang, KETW Key Terms Extraction and Term Weighting for Newsgroup Message Classification, Software Engineering, (2009)360-364.

[5] Erenel, Z., Altincay, H., Varoglu, E., A symmetric term weighting scheme for text categorization based on term occurrence probabilities, In Proceedings of 15th European Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, (2009)1-4.

[6] Xia Tian, Wang Tong, An improvement to TF: Term Distribution based Term Weight Algorithm, Second International Conference on Networks Security, Wireless Communications and Trusted Computing, (2010)252-255.

[7] Salton , G., Fox , E., Wu , H., "Extended Boolean information retrieval", Communications of the ACM, vol. 26, no.12, 1983, pp.1022-1036.

[8] Salton G, Buckley C. Improving retrieval performance using relevance feedback. Computer Science Technical Report TR88-898, Department of Computer Science, Cornell University, Ithaca, N.Y., 1988:513 -523.

[9] SALTONG, CLEMENTTY. On the construction of effective vocabularies for information retrieval[C].Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval NewYork:ACM,1973:11.

[10] Tan Song-bo, Wang Yue-fen, Chinese text classification corpus-TanCorp V1.0, http://www.searchforum.org.cn/tansongbo/corpus.htm.

[11] Information on http://www.sogou.com/labs/dl/c.html.

[12] YAO Xu, WANG Xiao-dan, ZHANG Yu-xi, QUAN Wen, Summary of feature selection algorithms, Control and Decision, 2012(02).

[13] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Automatic modulation classification using combination of genetic programming and knn," Wireless Communications, IEEE Transactions on, vol. 11, no. 8, pp. 2742-2750, august 2012.

[14] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys, vol. 34, no.1, pp.1-47 2002.