Improved term selection algorithm based on variance in text categorization

Ran Li[1, a] , Xianjiu Guo[2,b]

[1,2]Information engineering collegeDalian Ocean Univercity Dalian,China

[a]liran@dlou.edu.cn, [b]gxj@dlfu.edu.cn

Key words：variance; text classification; term selection

Abstract. This article improves the algorithm of term weighting in automated text classification. The traditional TFIDF algorithm is a common method that is used to measure term weighting in text classification． However, the algorithm does not take the distribution of terms in inter-class. In order to solve the problem, variance which describes the distribution of terms in inter-class and intra-class is used to revise TFIDF algorithm. This article mainly researched about the construction of LFHW term sets and new approaches to term weighting, These new approaches are also applied to the hierarchical classification system． Compared with traditional TFIDF algorithm ， the results of simulation experiment have demonstrated that the improved TFIDF algorithm can get better classification results．

Introduction

Text classification refers to determine a category for the text of each set of documents according to the predefined topic categories, Human knowledge has great growth, artificial knowledge classification has been difficult to adapt, so the technology of text classification is paid attention by the academic and industrial circles. With a simple and accurate method for the document representation into a form that the computer can process is based for the text classification, the classic text representation method is the vector space model.feature weighting is the core algorithm for Vector space model, feature weighting calculation plays a crucial role on text categorization effect. Based on the analysis of traditional feature extraction algorithm TFIDF, variance which describes the distribution of feature terms in inter-class and intra-class is used to describe the distributions of the feature terms inter-class and intra-class, finally the application of VSM algorithm is used for automatic text classification, the experimental results show that the improved TFIDF algorithm has better performance in text classification.

The traditional TFIDF

The traditional TFIDF algorithm method is proposed by Gerald Salton and McGill for document feature representation in the light of vector space information retrieval paradigm. In this method, the words appear in a document is called the feature words (Term), each term has a corresponding weighting, the weighting of words manifest important degree in document recognition, term weighting and term appear frequency in the document is proportional, with the frequency of term in all the documents is inversely proportional.

TFIDF is: TF ×IDF. The TF (Term Frequency) is called the word frequency, refers to the frequency of terms in the given document, if exclusion of forbidden and individual high-frequency words, the more frequency of terms in the given document , the stronger ability of terms for the document characterization; IDF (Inverse Document Frequency) is called the inverse document frequency, reflects the appear frequently of a term in a document set by document statistics, The smaller the number that document contain terms in document set, the more representative terms [1].

$$IDF（t_i）=log(N（C）／N(t_i，C))　　（1）$$

Where N (C) for all the documents, N (ti, C) for the documents that contain the terms.TFIDF algorithm mainly has 2 features as the following:

Published by Atlantis Press, Paris, France.

Inverse document frequency is purpose that the smaller the number that document contain terms is, the more representative terms is. N(C) for the total number of documents in the training set, without taking increase the number of documents in the same class into consideration. If the term ti appears frequently in the document of the same class, the term can be a good representative of this class, this term should be given a higher weighting. In formula (1), when the number N(ti, C) of document that contains term ti in class Ci increases, but IDF decreases. Therefore, such a kind of representative terms that appeared in the same class in while less appearance in other, the weighting in TFIDF algorithm not only has been strengthened, but weaken.

For the two different terms t1 ,t2 in the same class .the frequency of t1 in each document is more average than t2, so t1 is more representative than t2 in the ability of characterization of classes. If t2 appears great quantities in one or two document of the class, whereas appears rarely in the other document of the class, do not eliminate that the one or two document is the special case in the class, this term is not representative, the weights to be lower, but TFIDF algorithm can not distinguish for which.

Improvement feature weighting algorithm

Based on the analysis of the section second, in addition to frequency, weighting is closely related to the following two kinds of terms:

(1) The less the classes of terms distribution is and the more number the texts of the same class has, the stronger ability of terms distinguishing is, and also the bigger weighting.

(2) The more average the terms in each document of the class is, the more representative the terms has, the bigger weights would be given.According to this idea proposes, put forward improvement of feature weighting algorithm TFIDF-De-Di:

$$Weight_{TFIDF}(t_{ij})=TF(t_{ij})\times IDF(t_j)\times D_e\times(1-D_i) \quad (2)$$

Where, TF(tij) for word frequency, IDF(tj) for inverse document frequency, De for dispersion coefficient of inter-class, Di for dispersion coefficient of intra-class [2].

This article will look term ti as a random variable, the value of ti in inter-class will be representation by the frequency of ti in inter-class. By the definition of the variance, the distribution variance D(ti) of ti in inter-class reflects the degree of dispersion of ti in inter-class, the less D(ti) is, the more uniform the distribution of ti in inter-class is. That is if the term ti more evenly distributed in each class, D(ti) is smaller, the contribution to the classification of term ti is smaller, if, the term is uniformly distributed in inter-class, then D(ti) is 0, no contribution to classification. Using this feature of variance, this paper uses D(ti) to modified TFIDF formula, just can make up for the TFIDF that did not consider the distribution of term in inter-class.

From the above analysis, the value of ti in inter-class will be representation by the frequency of ti in inter-class, but the distribution probability of term ti in inter-class is very difficult to calculate, it is more difficult to calculate the variance, so this use the average variance of term ti approximation instead of D(ti).

Set up a total of N class, TF(ti) representing the occurrence frequency of the term in class Ci, $\overline{TF(t_i)}$ represents the average frequency of term in each class, the calculation formula is: $\overline{TF(t_i)}=\frac{1}{n}\sum_{i=1}^{n}TF(t_i)$. If square of average variance of t1 in inter-class for the De, The formula of square of average variance of t1 is

$$D_e=\frac{1}{n}\sum_{i=1}^{n}(TF(t_i)-\overline{TF(t_i)})^2 \quad (3)$$

Using De modified TFIDF, the formula is as follows

$$Weight_{TFIDF}(t_i)=TF(t_i)\times IDF(t_i)\times D_e \quad (4)$$

Apparently, when ti uniform distribution in inter-class, since De=0, so WeightTFIDF(ti)=0, the term ti has not contribute to classification.

The following analysis is the distribution situation of term ti in intra-class , if total document number of class Ci is m, The frequency of ti in each document will be looked as the value of ti in each document, $\overline{TF(t_i)}$ for the average frequency of ti in class Ci, The calculation formula is as follows

$$\overline{TF(t_i)} = \frac{1}{m}\sum_{j=1}^{m} TF_{ij}(t_i) \quad (5)$$

Di for the square of average variance of ti in class Ci, then

$$D_i = \frac{1}{m}\sum_{j=1}^{m}(TF(t_{ij}) - \overline{TF(t_i)})^2 \quad (6)$$

In order to facilitate for representation, adding a denominator to Di, so that its value would be less than l

$$D_i = \frac{\frac{1}{m}\sum_{j=1}^{m}(TF(t_{ij}) - \overline{TF(t_i)})^2}{\frac{1}{m}\sum_{j=1}^{n}(TF(t_{ij}))^2} \quad (7)$$

With the above analysis, the more distribution uniform the term ti in documentation of the class Ci is, the less Di is, while ti can more represent the Ci class, the corresponding 1-Di is greater, So 1-Di can be used to modified TFIDF formula [3,4].

$$Weight_{TFIDF}(t_{ij}) = TF(t_{ij}) \times IDF(t_j) \times D_e \times (1-D_i) \quad (8)$$

The results and analysis of experiment

The corpus from the Fudan University corpus, The training samples and test samples respectively have ten categories, a total of 4000 documents. The training samples are a total of 2600 documents, test sample are a total of 1400 documents.

Experiment use angle cosine formula of vector space model VSM (Vector Space Model) to calculate the similarity between the vectors. Assumption the standard vector of class is C, unclassified document vector is d, the similarity between the two can be measured using angle cosine value of the two vector, the calculation formula is as follows:

$$sim(d_i, c_j) = \frac{\sum_{k=1}^{n} W_{ik} \times W_{jk}}{\sqrt{\left(\sum_{k=1}^{n} W_{ik}^2\right)\left(\sum_{k=1}^{n} W_{jk}^2\right)}} \quad (9)$$

Where, W_{ik} , W_{jk} respectively description weights of the term k in text di and class cj.

Three experiments tested the algorithm, The traditional TFIDF algorithm, the new algorithm combined with the variance and not only combined with variance but also LFHW algorithm. First calculated wij of each term, then calculated the vector of various class and vector of unclassified document in test set. Single classification system requires the calculation vector of ten kinds of class, through the similarity calculation to obtain the class of unclassified document. Hierarchical classification system requires the calculation vector of ten large class and three small class of finance and economics, and then, the first comparison with vector of large class to get large categories that unclassified document belongs to, if belong to the finance and economics class, comparison with vector of small class to get small categories that unclassified document belongs to,

so the hierarchical classification system required test a total of thirteen classes [5].

In experiments using R (Recall) and P (Precision) of two commonly used text classification assessment test value, recall is the ratio between the number of retrieved relevant documents and the number of relevant documents in the document set, Precision is the ratio between the number of retrieved relevant documents and the total number of documents retrieved. In all of the algorithm, the traditional TFIDF algorithm got the worst classification accuracy, TFIDF-De-Di algorithm was superior to the TFIDF algorithm, the algorithm that base on the algorithm of TFIDF and consider De, Di and LFHW factor algorithm got the highest classification accuracy. At the same time, the more test article, the higher classification accuracy. Improved weighting algorithm also applicable to hierarchical classification system[6].

Conclusions

Feature weighting algorithm has a great effect on the accuracy of text automatic classification system. In this study, some disadvantages of the traditional TF-IDF algorithm were analyzed, one kind improved weighting algorithm that based on the variance and consider distribution information in inter-class and intra-class. The comparison of experimental results proves that the improved weighting algorithm has a good performance in the precision of classification.

Acknowledgment

This paper is supported by Nature Science Foundation of Liaoning Province (No. 201202021) and Science and Technology Department of Liaoning Province (No.2012216012) and Key Laboratory of Marine Information Technology of Liaoning Province.

References

[1] J.Lu, “Improved feature selection algorithm based on variance in text categorization” J. Computer Engineering and Design, Vo1.28 No.24,pp.6039-6041,December 2007

[2] J. Bai, J. Nie, G. Cao, “Integrating compound terms in Bayesian text classification,” 2005 IEEE/WIC/ACM international Conf. France,p.598,2005.

[3] J. Yan,N. Liu,B. Zhang,OCFS: Optimal orthogonal Centroid Feature Selection for test catagorization[M]. Brazil：SIGIR,2005

[4] F. Xu, Z. Luo, “An Improved Approach to Term Weighting in Automated Text Classification,” J. Computer Engineering and Applications, vol. 41, pp.181-184,January 2005

[5] G. Chen, D. Huang, “Feature Selection Model of TFIDF Text Categorization Based on Information Entropy,” J. Hubei Institute for Nationalities (Natural Sciences), vol. 26, pp.401-404, December 2008.

[6] F. Sebastiani, “Machine learning in automated text categorization,” J. ACM Computing Survey，vol. 34,PP.41-47, January 2002.