An Integrated and Improved Approach to Terms Weighting in Text Classification

Jyoti Gautam[1] and Ela Kumar[2]

[1] School of Information and Communications Technology
Gautam Buddha University
Greater Noida, Uttar Pradesh(India)-201308

[2] School of Information and Communications Technology
Gautam Buddha University
Greater Noida, Uttar Pradesh(India)-201308

Abstract

Traditional text classification methods utilize term frequency (tf) and inverse document frequency (idf) as the main method for information retrieval. Term weighting has been applied to achieve high performance in text classification. Although TFIDF is a popular method, it is not using class information. This paper provides an improved approach for supervised weighting in the TFIDF model. The tfidf-weighting model uses class information to compute weighting of the terms. The model also assumes that low frequency terms are important, high frequency terms are unimportant, so it designs higher weights to the rare terms frequently. So, it uses rare term information along with class information for weighting. So, the paper proposes an improved approach which combines the benefits of the traditional kNN classifiers and Naïve Bayes supervised learning method.

Keywords: *Text classification, tf-idf, term weighting, kNN, feature selection*

1. Introduction

Text classification is the key technique in the data mining (DM) and information retrieval (IR) field and it has got a lot of interest in the recent decades. A lot of research has been done to improve the quality of text representation and develop high quality classifiers. Text classification (TC) is a task to categorize automatically text documents into categories from a predefined set. Most of the machines learning methods treat text documents as bag of words [1].

Vector space model [9, 10] is a classic method in which each document is represented as a vector of its words. Words are regarded as feature vectors. When applied to text categorization, the basic idea is to construct a prototype vector per category using a training set of documents. Given a category, the vectors of documents belonging to this category are given a positive weight, and the vectors of remaining documents are given a negative weight. By summation of the different weighted vectors, the prototype vector of this category is obtained. The method is easy to implement and efficient in computation.

There are two term weighting approaches, i.e. unsupervised and supervised term weighting methods [6] depending on the use of the known information on the membership of training documents. Unsupervised term weighting approaches, such as binary, tf, tfidf, in which binary tells whether a particular term appears in a document, tf indicates how frequently a term appears in a document, and tfidf calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TFIDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user[5]. The tf part can be regarded as a weight from intra-documents, and idf part is a weight from inter-documents. Unsupervised-based term weighting approach does not use the category information in the training set. Researchers have tried to replace the idf part with feature selection metrics, i.e. information gain, gain ratio, odds ratio and so on. The basis for these is information theory and supervised term weighting approaches. Unlike the case of unsupervised-based term weighting approach, supervised term weighting uses category information. It uses the inter and intra class information. In literature[12], interclass standard deviation(icsd), class standard deviation(csd) and standard deviation, were introduced to tf-idf model, the performance of classification is enhanced.

In yet another method of supervised term weighting,[13] the approach simply thinks low frequency terms are important, high frequency terms are unimportant , so it designs higher weights to the rare terms frequently. The approach presents an effective term weighting to avoid the deficiency of the traditional approach, and make use of KNN classifiers to classify over widely-used benchmark data set Reuters-21578.

In this paper, we propose yet another novel supervised term weighting approach. The approach combines a

weighting factor based on the frequency of the number of documents belonging to a category c_i where the term t_k occurs at least once. It also joins the inner and intra class information on a trained document set.

The rest of the paper is organized as follows: Section 2 discusses the traditional term weighting methods .Section 3 presents the shortage of the traditional term weighting scheme. Section 4 presents our improved approach. And finally Section 5 concludes this paper.

2. Term Weighting Approaches: Review and Analysis

The literature [7] provides a text classification method which is an improvement of the previous ones. Most of the traditional text classification methods utilize term frequency (tf) and inverse document frequency (idf) for representing importance of terms and computing weighting of ones in classifying a text document. Term weighting has a significant role to achieve high performance in text classification. The old tf-idf is a popular method, but it does not involve class information of the terms. The paper has provided with an improved tf-idf-ci model to compute weighting of the terms. The intra and inner class information are joined. The role of important and representative terms is raised and the effect of the unimportant feature term to classification is decreased. The F1 based on tf-idf-ci algorithm is higher than based on tf-idf model.

Literature [8] describes a new method which uses term co-occurrence as a measure of dependency between word features. A random-walk model is applied on a graph encoding words and co-occurrence dependencies, resulting in scores that represent a quantification of how a particular word feature contributes to a given text. The new scheme can be used as a text classification method.

The literature [14] provides a method for text categorization when one or more predefined categories are given. The paper reports a study conducted on 20 newsgroup dataset, using TFIDF in the context of document categorization. Feature selection is added to this result to improvise the categorization. The results achieved by this approach are very promising when compared to conventional methods with features chosen on the basis of bag-of-words text.

Literature [13] provides an improved method of term weighting for text classification. Traditional algorithm of term weighting only considers about tf, idf and so on, and this approach simply thinks low frequency terms are important, high frequency terms are unimportant, so it designs higher weights to the rare terms frequently. In this paper, an effective term weighting approach is provided to avoid the deficiency of the traditional approach, and make use of KNN classifiers to classify over widely-used benchmark data set Reuters-21578. The experimental results have proved that the new approach can improve the accuracy of classification.

One of the difficulties in text classification is the high dimensionality of the feature space. How to reduce the dimensionality of the feature space and improve the effectiveness and accuracy of classification has become the main problem to be solved in automatic text classification. So feature selection is often considered a critical step in text classification. Feature selection methods keep a certain number of words with the highest score according to a measure of word relevance. The quality of features will influence accuracy of text classification. According to the literature [2], traditional scoring measures for feature selection have come from the domains of information theory and information retrieval.

This paper [4] by Thorsten Joachims mainly analyses the traditional Rocchio algorithm with TFIDF classifier. The Rocchio classifier, its probabilistic variant with TFIDF classifier and a standard Naïve Bayes classifier are compared. The results provided the information that the probabilistic algorithms are preferable to the heuristic Rocchio classifier. Bag-of –words data is used for analysis. The algorithm proposed in the paper uses feature selection for better processing. Paper proposes a probabilistic classifier based on TFIDF. It makes use of probabilistic indexing paradigm which offers an elegant way to distinguish between a document and the representation of a document. This proposed probabilistic TFIDF classifier offers a theoretical justification for the vector space model and the TFIDF word weighting heuristic for text categorization.

The paper [11] proposes an improved approach named tf.idf.IG to remedy this defect by Information Gain from Information Theory. The paper overcomes the limitations of old tf.idf. The idf can't well reflect discriminative and importance of feature, weight adjustment method is put forward in which the IDF function is replaced by evaluation function used in feature selection.

Berger and Lafferty in the year 1999[1] proposed a probabilistic framework that incorporates the user's mindset at the time the query was entered to enhance their approximations. They suggest that the user has a specific information need G, which is approximated as a sequence of words q in the actual query. By accounting for this noisy transformation of G into q and applying Bayes' Law to equation (1), they show good results on returning appropriate documents given q.

Let us assume that we have a set of documents D, with the user entering a query $q = w_1, w_2, \ldots, w_n$ for a sequence of words w_i .Then we wish to return a subset D* of D such

that for each d ϵ D*, we maximize the following probability:

P (d | q, D) (1)

3. The Shortage of Traditional Term Weighting Scheme

The focus of this paper [7] is term-weighting based text classification. So, how to assign appropriate weight to the given feature is of utmost importance. An effective feature can not only represent the content of category belonging to, but also provides discrimination with other categories. The tf-idf approach proposes three basic assumptions. They are:

- Rare terms are no less important than frequent terms- idf assumption.
- Multiple appearances of a term in a document are no less important than single appearance- tf assumption.
- For the same quantity of term matching, long documents are no less important than short documents- normalization assumption.

The Table1 below shows the relation of term t_k and category C_i.

Table1: relation of term t_k and category C_i.

	C_i	$\overline{C_i}$
t_k	A	B
$\overline{t_k}$	C	D

A indicates the number of documents belonging to category C_i where the term t_k occurs at least once; B indicates the number of documents not belonging to category C_i where the term t_k occurs at least once; C denotes the number of documents belonging to category C_i where the term t_k does not occur at least once; D denotes the number of documents not belonging to category C_i where the term t_k does not occur at least once.

The tf-idf term weighting scheme assigns higher weights to the rare terms frequently because of idf assumption. Thus, it will influence classification performance. For some a category c_i, the terms which are distributed uniformly in the intra-category should be assigned higher weights, but not rare terms. This weighting is not included in the tf-idf approach. For some a category, the higher value of the proportion of A and C is a good feature. So, accordingly, we have to assign appropriate weighting.

In addition to term frequency and inverse document frequency, [8] the weighting is affected by other factors also. The term weighting joins class information also.

- The weighting should be large when the numbers of the classes distributed is more, but the document numbers in one class is far greater than the sum in others.
- The weighting should be small when the numbers of the classes distributed is more, and the document numbers in all classes is large.
- The weighting should be large when the numbers of the classes distributed of the term is very small, even the numbers is one, and the term distributed in the documents of the class is average.
- The weighting should be small when the numbers of the classes distributed of the term is very small, even the numbers is one, but the term distributed only little documents of the class, even one or two documents.

The objective of this paper is to provide an improved weighting algorithm which joins class information along with the weighting assigned to the rare terms.

4. Our Improved Approach

The tf-idf term weighting scheme has been used extensively and has become the default choice in text classification. Our improved approach joins class information combined with the weighting assigned to the rare terms.

$$W(t_k,d_j,c_i) = (1-\alpha).tfidf_{k,j} + \alpha.\ tfidf_{k,j} \times \text{weighting} \quad (2)$$

α is called a balance factor, which lies between , $0 \le \alpha \le 1$.

When α = 0, equation (2) becomes classic TFIDF approach, and when α = 1, equation (2) becomes our newly improved approach. Using balance factor, we can get better classification results.

Where, the weighting is the class information along with the weighting assigned to the rare terms. The weighting is:

$$\text{Weighting} = CI \times \frac{Ai}{Ci} \quad (3)$$

Where, Ai indicates the number of documents belonging to category c_i where the term t_k occurs at least once and Ci denotes the number of documents belonging to category c_i where the term t_k does not occur at least once.

Where, the term $CI = Cit \times Cii$ (4)

The class information consists of two parts. One is intra class information, and the other is inner class information. That is:

Where, cit is intra class information, cii is inner class information.

Intra Class Information

$$Cit = \frac{P(ti|Cj)}{\sum_{k=1,\ k\neq j}^{m} P(ti|Ck)} \quad \text{if} \quad \sum_{k=1,\ k\neq j}^{m} P(ti|Ck) \neq 0 \quad (5)$$

$$Cit = \frac{P(ti|Cj)}{\beta} \quad \text{if} \sum_{k=1,\ k\neq j}^{m} P(ti|Ck) = 0 \quad (6)$$

Where, $P(ti|Cj)$ is the probability of documents containing term t_i in the class Cj of the training set. The value of the parameter β is determined through actual situation. Generally, β = 0.001. The number of classes taken is m.

It is quite evident that the Cit is a monotone increasing with the number of documents in the class Cj containing the term t_i increasing. The Cit is increasing with the sum of documents beyond the class Cj containing the term t_i decreasing. Only when the documents of two classes containing the term t_i , and the document numbers of containing the term t_i are almost same the value of Cit approximate to 1. The largest value is achieved when the sum of documents beyond the class Cj containing the term t_i is zero.

Inner Class Information

Inner class divergence can be represented by the term Cii. It is very important to classify when the term t_i appears evenly in the documents of one class.

$$Cii = \frac{\mathbf{tfavg\ (ti\ ,Cj)}}{\sum_{k=1}^{N(Cj)}[\,|\,\mathbf{tfik - tfavg\ (ti\ ,Cj)}\,|\,]}$$

$$\text{If} \quad \sum_{k=1}^{N(Cj)}[\,|\,\mathbf{tfik - tfavg\ (ti\ ,Cj)}\,|\,] \neq 0 \quad (7)$$

$$Cii = \frac{\mathbf{tfavg\ (ti\ ,Cj)}}{\gamma}$$

$$\text{If} \sum_{k=1}^{N(Cj)}[\,|\,\mathbf{tfik - tfavg\ (ti\ ,Cj)}\,|\,] = 0 \quad (8)$$

Where, γ is a parameter, the value is determined which is based on the actual situation. Generally, γ = 1. The tf_{ik} is the frequency of the term t_i in document k. The $tf_{avg}(t_i, Cj)$ is the average term frequency of the term t_i in the documents of the class C_j :

$$tf_{avg}(t_i,Cj) = \frac{\sum_{k=1}^{N(Cj)} \mathbf{tfik}}{\mathbf{N(Cj)}} \quad (9)$$

The largest value of Cii is achieved when the term t_i appears evenly in the documents of the class C_j. If the difference of the term t_i appeared in the documents of the class C_j is larger, then the denominator of the function (6) is larger, then the value obtained of Cii is less. So, it is representative and important for classification purposes when the term t_i appears evenly in the documents of one class.

Normalization

It is normalized for decreasing the high frequency term inhibited to low frequency term. The tf-idf-weighting function is normalized as follows:

$$W_{\text{tf-idf-weighting}}(tij) = \frac{tf \times idf \times weighting}{\sqrt{\sum_{i=1}^{n}(tf \times idf \times weighting)^2}} \quad (10)$$

There are different statistics classification and machine learning technologies which are used in text classification. This paper combines the benefits of kNN algorithm and Naiive Bayes algorithm. The kNN algorithm is a method for documents classification based on closet training examples in the feature space. It is amongst the simplest of the machine learning algorithms. And it is an effective method. The core of kNN is, first of all, giving a trained document set, then, for a new preclassified document(that is, a test document), finding most relevant articles of the k documents from the training documents, finally, in accordance with k documents' category information, classifying the test document. Whereas Naiive Bayes is the most effective heuristic learning method.

5. Conclusion and Future Work

Term weighting plays an important role to get high performance in text classification. The traditional tf-idf algorithm is a popular method for document representation and feature selection. But, it is not joining class information. Here, we proposed a supervised term weighting scheme, which makes use of a kind of information ratio to judge a term's contribution for category along with class information. The improved approach of using information ratio has become a new way to compute term's weights to avoid assigning higher weights to rare terms. It has been proved through experiments that the improved approach of using information ratio is an effective solution to improve the performance of text classification. The approach using class information uses intra and inner class information. When using class information, the numbers of the feature term is decreased when the threshold is given. The experimental results showed that the performance is enhanced. The role of important and representative terms is raised and the effect of the unimportant feature term to classification is decreased. So, an improved approach for

term weighting which joins information ratio together with class information has been proposed. The approach can be implemented and performance analysis can be done.

References

[1] Berger, A & Lafferty, J., "Information Retrieval as Statistical Translation", Proc. of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIR), 1999, 222-229.

[2] Elena Montanes, Irene Diaz, Jose Ranilla, Elias F.Combarro, Javier Fernandez, "Scoring and Selecting Terms for Text Categorization" , Journal of IEEE Intelligent Systems, Vol. 20, Issue 3, 2005, pp. 40-47.

[3] G.Salton and M.J.McGill, "An Introduction to Modern Information Retrieval" , McGraw Hill, 1983.

[4] Joachims, Thorsten, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization" , Computer Science Technical Report CMU-CC-96-118, Carnegie Mellon University, 1996.

[5] Juan Ramos, "Using TF-IDF to determine word relevance in document queries" ,Department of Computer Science, Rutgers University.

[6] M.Lan, C.L.Tan, and H.B.Low, "Proposing a new term weighting scheme for text categorization" , Proc. of the Twenty-First National Conference on Artificial Intelligence, 2006, pp. 763-768.

[7] Ma Zhanguo, Feng Jing, Chen Liang, Hu Xiangyi, Shi Yanqin, "An improved approach to terms weighting in text classification" , in proc. of the International Conference on Computer and Management, 2011, pages1-4.

[8] S. Hassan, C. Banea, "Random Walk term weighting for improved text classification" , Proc. of Text Graphs: 2nd Workshop on Graph Based Methods for Natural Language Processing, ACL, 2006, pp. 53-60.

[9] Salton, G., & Buckley, C., "Term-weighting approaches in automatic text retrieval", Journal of Information processing and Management, Vol. 24, Issue No. 5, 1988, pages 513-523.

[10] Salton, G., Wong, A., & Yang, C.S., "A vector space model for automatic indexing" , Communications of the ACM, Vol.18, Issue No. 11, 1975, pages 613-620.

[11] S. Lu, X. Li, S. Bai, S. Wang, " An improved approach to weighting terms in text" , Journal of Chinese Information Processing, 2000, 14(6), pp. 8-13.

[12] V. Lertnattee, T. Theeramunkong, "Effect of term distributions on centroid-based text categorization" , International Journal on Information Sciences – Informatics and Computer science, vol. 158,Issue 1,2004, pp. 89-115.

[13] Xin Hu, Hua Jiang, Ping Li and Shuyan Wang proposed, "An improved method of term weighting for text classification", appears in *International Conference on Intelligent Computing and Intelligent Systems*,Vol. 1, IEEE, 2009, pages 294-298.

[14] Yi-hong Lu, Yan Huang, "Document Categorization with Entropy based TFIDF classifier", in proc. of the WRI Global Congress on Intelligent Systems, vol.4, 2009, pages 269-273.

Jyoti Gautam completed B. Tech. (Instrumentation and Control Engineering) in the year 1997 from Delhi University. Afterwards, she did her M.Tech. (Computer Technology and Applications) in the year 1999 from Delhi University. Currently, pursuing PhD in the area of Semantic Web from Gautam Buddha University, Greater NOIDA. Presently employed as Associate Professor in the Department of Computer Science and Engineering in JSS Academy of Technical Education, NOIDA. One of the published papers titled 'An Improved Framework for Tag-Based Academic Information Sharing and Recommendation System' occurs in the proceedings of the WCE 2012 VOL II.

Dr. Ela Kumar completed B.E.(Electronics and Communication) in the year 1988 from IIT, Roorkee. Afterwards, she did her M.Tech. (Computer Science and Technology) in the year 1990 from IIT, Roorkee. Later, she did Ph.D. (Computer Science and Technology) from Delhi University in the year 2003.Worked as Asstt. Professor at YMCAIE,Faridabad. Presently, working as Dean and Associate Professor in the school of ICT, Gautam Buddha University, Greater Noida. She won Rashtriya Gaurav Award by IIFS society in NOV 2010 for meritorious activities in the field of IT. She has participated as member Advisory Board in various conferences. She has participated in various seminars as Speakers. She has participated as expert for Course Designing. Her publications include 10 research papers in International Referred Journal, 6 in National Referred Journals, 10 in International Conferences and 15 in National Conferences. She has penned 4 books. Her expertise include many other activities.