A Feature Weight Algorithm for Text Classification Based on Class Information

LI Yong-fei
Department of Computer
North China Institute of Science and Technology
Beijing, China
lyf518@ncist.edu.cn

***Abstract*—TFIDF algorithm was used for feature weighting in text classification. But the result of classification was not very well because of lack of class information in feature weighting. The known class information in the training set was used to improve the traditional TFIDF feature weight algorithm. Class distinction ability and class description ability were introduced, respectively expressed by inverse class frequency and term frequency in class, document frequency in class. A new feature weight algorithm based on class information, TF_IDT, was proposed. Naïve Bayes classifier was used to test the algorithm. The precision, recall and F1 measure were significantly increased. Macro F1 measure raise by 6.46%. It was proved to be useful for improving text classification to use class information in feature weighting. In addition, the computational complexity of the proposed algorithm was lower and more suitable for use in fields of limited computing capability.**

Keywords-text classification; feature weight; inverse class frequency; term frequency in class; document frequency in class

I. INTRODUCTION

Text classification referred to categorize the text content under the given class, which mainly included two phrases: text representation and classification algorithms [1]. Vector Space Model was a common way for text representation in practice. Text was represented as vector based on features and its weights, and then came classification learning and automatic classification. Feature selection and feature weight calculations were two major problems in text representation. Feature selection referred to select a part of feature items which could best reflect the statistical characteristics of the text in the original feature space to reduce the number of feature dimensions. Feature weight calculation referred to quantify the text content contribution and discrimination of feature items. A reasonable feature weight algorithm was able to highlight the class attribute of the text, and affected classification result heavily.

TFIDF algorithm was the most commonly used in text categorization feature weighting. The concept of TFIDF was proposed initially by Salton[2] in the field of information retrieval, which was used to calculate weight for describing the contribution to document content of index entry. When used in the field of text classification, TFIDF algorithm was not good enough because that only text representation and text discriminative [3] of feature items were involved, but not class information when calculating feature weights. There had been many improvements for TFIDF algorithm. For example, a new factor was added to reflect the class distinction of feature item in [4,5]. In [6], the concepts of class skewness and class dispersion were introduced to reflect inter-class and intra-class distribution of feature items, and the classification result was improved. In [7], the concepts of class distinction ability and documentation distinction ability were introduced, and feature weight was calculated based on the information gain of feature items, trying to more fully reflect the distribution of feature items, and a certain effect was also achieved. However, the computational complexity of these improvements was too high to be used in the fields of limited computing capacity.

A new feature weight algorithm was proposed in this paper. The original concept of TFIDF was analyzed, including document distinction ability and document description ability. The concepts of class distinction ability and class description ability were introduced, and expressed by inverse class frequency and term frequency in class, document frequency in class. The feature weight was calculated based on term frequency (TF), inverse class frequency (ICF), term frequency in class (TFc), and document frequency in class (DFc). Thus the class information was represented in weight calculation, and computational complexity was low to implement in the fields of limited computing capacity. The experimental result showed that significant improvement was achieved in text classification.

II. TFIDF ALGORITHM AND INADEQUACY

A. TFIDF Algorithm

TFIDF algorithm was the most important invention in the field of information retrieval, and was widely used in the fields related to search, document classification. It was measure of information retrieval correlation. The basic idea of TFIDF algorithm contained two aspects [8]: First, the higher the frequency of a term in a document was, the document description ability of the term was more powerful (TF, Term Frequency); second, in a given document set, the more the number of documents that contain a term was, the document distinction ability of the term was weaker (IDF, inverse document frequency). The classic TDIDF formula [2] was

$$W_{ij} = TF_{ij} \times IDF_j = TF_{ij} \times \log\left(\frac{N}{n_j}\right) \quad (1)$$

Published by Atlantis Press, Paris, France.

TF_{ij} was the number of occurrences of feature item (term) T_j in the document D_i; IDF_j was the reciprocal of the proportion of the number of documents containing feature item (term) T_j. N was the total number of documents, and n_j was the number of documents that contains feature item (term) T_j.

Based on the analysis about the basic idea of TFIDF, it was clear that TF expressed document description ability of the feature item, and IDF expressed document distinction ability of feature item. The feature weight calculated with TF and IDF was used to express document into vector in information retrieval. And then automation judgment of correlation between the search term and the subject document was achieved.

B. *Inadequacy of TFIDF in Classification*

Formal representation of document was implemented by feature weight which was calculated with document description ability and document distinction ability. It was used widely to determine the relevance of the query term and the document in information retrieval. But when TFIDF was used for text representation in text classification, the effect was not so well because of the lack of class information. There were lots of researches on TFIDF algorithm improvement [4-7], where TFIDF formula was transformed by introducing some new statistical parameters to add class information and improve categorizing effect. However, the computational complexity of these improvements was too high to be used in the fields of limited computing capacity.

III. TF_ITD ALGORITHM BASED ON CLASS INFORMATION

Referring to the concept of document description ability and document distinction ability in traditional TFIDF algorithm, class description ability and class distinction ability was defied in the field of text classification, and was expressed by the similar statistics as traditional TFIDF. And the weight calculation formula was combined with such concept and term frequency.

A. *Class Distinction Ability*

Referring to IDF, which indicated the document distinction ability, class distinction ability was expressed as ICF (Inverse Class Frequency), defined as

$$ICF_j = \log\left(\frac{|C|}{Nc_j}\right) \tag{2}$$

ICF_j was reciprocal of the number of classes which contained feature item T_j. |C| was the total number of classes in training set. Nc_j was the number of classes which contained feature item T_j. The meaning of this factor was that a feature item was more useful in distinguish different classes if it only appeared in fewer classes. It meant that the class distinction ability of such feature item was stronger when feature item was in uneven distribution between different classes.

B. *Class Description Ability*

Class description ability meant the description ability of feature term for the class which the document where the feature term was in belonged to when feature item only concentrated in a few or even one class. There were two kinds of uneven distribution of feature item. The first was that feature item concentrated in the class where the document belonged to; and the second in other classes. Obviously class description ability of the previous was stronger. Referring to TF, which indicated document description ability, class description ability was expressed as TFc_{ij} (Term Frequency in class), defined as the number of occurrences of feature item T_j in class C_i. It meant that the more the feature item occurred in a class, the class description ability for the class was stronger.

However, further analysis showed that uniform distribution of feature item in this class should be considered when the item is concentrated in this class. The more evenly the feature item distributed in this class, the stronger class description ability was. If the feature item distributed unevenly and concentrated in a few documents in this class, its class description ability was weak. DFc (Document Frequency in class) was used to indicate it, defined as

$$DFc_{ij} = \log\left(\frac{n_{ij}}{N_i}\right) \tag{3}$$

DFc_{ij} was the proportion of the number of documents containing feature item T_j in class C_i. N_i was the total number of documents in class C_i, n_{ij} was the number of documents containing feature items T_j in class C_i.

C. *Improved Feature Weight Algorithm*

As previously mentioned, class distinction ability was expressed as ICF, and class description ability as TFc and DFc jointly. A new feature weight formula called TF_ITD was constructed with class distinction ability and class description ability, defined as

$$W_{kj} = TF_{kj} \times \left(ICF_j + TFc_{ij} + DFc_{ij}\right) \tag{4}$$

TF_{kj} indicated the number of occurrences of feature item T_j in the document D_k; and ICF_j indicated the reciprocal of the number of classes containing feature items T_j , defined in formula (2); TFc_{ij} indicated the number of occurrences of the feature items T_j in the class C_i which was the document D_k belonged to; DFc_{ij} indicated the proportion of the number of documents containing feature items T_j in class C_i and the number of all documents in C_i, defined in formula (3).

Class information was introduced into the new feature weight algorithm, which made the vector expression of text was more suitable for classification. And the calculation of the factors was relatively easy to implement, and had less computational complexity.

IV. EXPERIMENT AND ANALYSIS

A. *Experimental Conditions and Evaluation*

TanCorpV1.0 [9], a Chinese classification corpus from the Institute of Computing Technology of the Chinese Academy of Sciences, was used for experiment. There were 12 classes in the corpus. Experimental corpus was composed of 1200 documents including 100 documents for each class. Stop words were removed from the corpus. The experiment was conducted based on weka3.6, an open source data

mining platform. TF_ITD algorithm proposed in this paper was used for feature weight calculation. Naive Bayes algorithm was used as classification algorithm. The experimental corpus was equally divided into two parts, one as the training set, and the other as test data. Precision (P), Recall (R) and F1 measure were used to evaluate feature weight algorithm on classification effect in each class, while macro average F1 measure to evaluate the classification effect on the entire data set.

B. Experimental Analysis

Traditional TFIDF and TF_IDT algorithm based on class information were respectively used for calculate feature weight, and naive Bayes algorithm for classification experiment. The experimental results were shown in table 1 for comparison.

TABLE I. COMPARISON BETWEEN EXPERIMENTAL RESULTS OF 2 ALGORITHM(%)

Class	TFIDF			TF_IDT		
	P	***R***	***F1***	***P***	***R***	***F1***
Finance	0.787	0.86	0.822	1	1	1
Region	0.923	0.923	0.923	1	0.962	0.98
Computer	0.879	0.967	0.921	1	0.983	0.992
Estate	1	0.915	0.956	1	0.979	0.989
Education	0.559	0.786	0.653	0.667	1	0.8
Sci & Tech	1	0.958	0.979	1	1	1
Car	1	0.979	0.989	1	1	1
Talent	0.837	0.554	0.667	1	0.846	0.917
Sports	1	1	1	1	1	1
Health	0.98	0.98	0.98	1	0.98	0.99
Art	0.944	0.962	0.953	1	0.887	0.94
Entertainment	0.98	1	0.99	1	1	1

As what could be seen from the experimental results, the classification effect of TF_IDT about Precision, Recall and F1 measure was better than traditional TFIDF algorithm on almost all class. F1 value was improved on talent class by 25%, which was the best one. Only recall of art class decreased by 7.5%, resulting in its F1 value decreased. By analyzing the confusion matrix, it was found that the number of documents which belonged to art class but wrongly classed into education class increased from two to six. The reason was that there were considerable overlaps between the feature items of the two classes. However, misclassification of other classes was greatly improved. The macro average F1 measure of TF_IDT algorithm was 96.73%, whereas the TFIDF only 90.27%. The increase of macro average F1 measure, 6.46%, was achieved on TF_IDT algorithm. It can be concluded that the overall performance of TF_IDF was much better than TFIDF.

Experimental result showed that the improved weight algorithm made the classification result better due to application of class information in the training corpus when calculating the weight.

V. CONCLUSION

After analyzing the inadequacy of traditional TFIDF algorithm due to lack of class information, the concepts of class distinction ability and class description ability were introduced, and statistics was used to express them by the similar statistics as traditional TFIDF. On this basis, TF_ITD algorithm based on class information was proposed, and verified by Naive Bayes classification algorithm. The experiment results showed that the new algorithm significantly improved the classification effect. The next work was to study statistics which could more accurately represent the class information, and to construct a more reasonable weight formula.

REFERENCES

[1] Franca D and Fabrizio S. “Supervised Term Weighting for Automated Text Categorization”, Proceedings of the 18th ACM Symposium on Applied Computing. Melbourne: ACM Press, 2003, pp:784-788.

[2] Shi Cong-ying, Xu Zhao-jun and Yang Xiao-jiang. “Comprehensive Research on TFIDF Algorithm”, Computer Application, vol. 29, Jun. 2009, pp. 167-170(in Chinese).

[3] Zhang Ai-hua, Jing Hong-fang and Wang Bin etc. “Study on function of feature weighting factor in Text classification”, Chinese Information Processing, vol. 24, Mar. 2010, pp. 97-103(in Chinese).

[4] ZHANG Yu-fang, Peng Ming-shi and Lyu Jia. “Improvement and application of TFIDF method in text classification”, Computer Engineering, vol.32, Oct. 2006, pp. 76-78(in Chinese).

[5] Shen Zhi-bin and Bai Qing-yuan, “improvement of feature weight algorithm in Text classification”, Journal of Nanjing Normal University (Engineering and Technology), vol.8, Apr. 2008, pp.95-98(in Chinese).

[6] Zhang Yu and Zhang De-xian, “An improved feature weight algorithm”, Computer Engineering, vol.37, May. 2011, pp.210-212(in Chinese).

[7] Li Kai-qi, Diao Xing-chun and CAO Jian-jun, “improved text feature weight algorithm based on information gain”, Computer Engineering, vol.37, Jan. 2011, pp.16-18, 21(in Chinese).

[8] Liu Ting, Qin Bing and Zhang Yu etc., “Introduction to Information Retrieval System”, Beijing: Mechanical Industry Press, 2008(in Chinese).

[9] Tan Song-bo and Wang Yue-fen, Chinese text classification corpus-TanCorpV1.0, Http://www.searchforum.org.cn/tansongbo/corpus.htm. 2011-12-20(in Chinese).