International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 37953-37959

A feature weighting method based on category-distribution divergence (CDD)

Lu Yonghe[1, a], Ye Zeyuan[1,b], He Xinyu[1,c]

[1]School of Information Management, Sun Yat-sen University, Guangzhou, China
[a]luyonghe@mail.sysu.edu.cn, [b]yezeyuan@mail2.sysu.edu.cn, [c]sysuimpaper@163.com

Abstract

In vector space model (VSM), text representation is the task of transforming the content of a textual document into a vector in the feature space so that the document could be recognized and classified by a classifier. The feature weighting methods assign appropriate weights to the features to improve the performance of text categorization.TF-IDF method is by far the most versatile and widely used but problems also exist in it. Especially the distribution of features in inter-class and intra-class is not taken into full account when using classical TF-IDF method, which causes a negative effect on precision of categorization. Based on category-distribution divergence (CDD), this paper proposes a new feature weighting method which introduces features' degree of membership and degree of non-membership into TF-IDF. In experiment, this paper uses K nearest neighbor algorithm(KNN), Rocchio algorithm and support vector machines (SVM) to test the validity of the CDD algorithm. The results show that the CDD algorithm gets a better performance than classical TF-IDF measure especially when the number of features is high enough.

Keyword: Text categorization; VSM; feature weighting;TF-IDF; category-distribution divergence

1. Introduction

In the field of text categorization, because computers do not directly identify the unstructured text, therefore, before the text is classified into a pre-defined text category, it needs to be transformed into a structured text by using text representation methods.Currently, Vector space model (VSM), one of text representation methods, is widely used in text categorization [1].

In the vector space model (VSM), the content of a document is represented as a vector in the feature space, i.e. Equation(1).

$$d = \{(t_1, w_1), (t_2, w_2), (t_3, w_3), \ldots, (t_n, w_n)\} \quad (1)$$

Where d is the document, t_i is the feature of document, w_i is the weight of t_i, $1 \le i \le n$, n is the dimension of the feature space. When the data set of the feature space is determined, each feature corresponds to a dimension of feature space, the content of a document is expressed as Equation(2).

$$d = \{w_1, w_2, w_3, \ldots, w_n\} \quad (2)$$

Where w_i(usually between 0 and 1) is determined by feature weighting methods, it represents how much the feature t_i contributes to the semantics of document d,its value has a significant impact on the categorization results. TF-IDF method is one of the feature weighting methods widely used, but TF-IDF ignores the distribution of features in inter-class and intra-class, which causes a negative effect on precision of categorization,therefore, we presented a feature weightingmethodbased on category-distribution divergence (CDD).

2. Related Work

2.1. Classical TF-IDF weights algorithm

Traditional feature weighting methods include Boolean weighting, term frequency weighting (TF), the inverse document frequency weighting (IDF), TF-IDF weights

algorithm, and so on. TF-IDF method is by far the most versatile and widely used, and it is the base of many relevant researches [2].

Classical TF-IDF formula is represented as Equation(3).

$$w_{TF-IDF} = tf_{id} * \log\left(\frac{D}{df_i}\right) \quad (3)$$

Here, tf_{id} representsthe times that feature t_i occurs in document d, D is the total number of training documents, df_i is the number of documents where feature t_i occurs at least once.

Related research[3] analyzed the two basic assumptions of TF-IDF as follow:

- For a particular feature, as it is in different documents of the same category, its *TFs* are almost the same; as it is in different documents of the different categories, its *TFs* are quite different. So that *TF* can be used to tell whether documents are in the same category or not.
- The lower are the *DFs* of a particular feature, the better is it's distinguish ability for different categories. So that *IDF* is introduced into TF-IDF.

When these two assumptions are valid simultaneously, we can use TF*IDF as the weight of some dimension. In practice, these two assumptions are not valid simultaneously very often, which causes TF-IDF emphasizes that rare features are more important than frequent features [4]. Besides, because document set is dealt with as a whole and the distribution of features in inter-class and intra-class is not taken into full account when using TF-IDF method, it cannot get a high precision of categorization.

2.1. Improved TF-IDF weights algorithm

To solve the problem, many researchers have proposed different improved methods on feature weighting.

Huang X et al. [5] considered features' distribution in a certain category and presented an improved TF-IDF algorithm, which is effective and feasible in feature extraction. Bong Chih How and Narayanan K proposed a feature selection measure, namely, Categorical Descriptor Term (CTD) for text categorization, and they have found out thatCTD works well on collections with highly overlapped topics [6]. Zhi-Hong Deng et al. substituted Category Relevance Factors(CRF) for IDF in order to reflect the features' distinctiveness for categories [7]. Zhang Baofuet al. proposed an improved TF-IDF method which is combined with information entropies of features in inter-class and intra-class [8]. Li Yuan also used information entropy to calculate uncertainty measure of the features in the corpus [9]. Zhang Yu et al. introduced skew information among classes (SI), distribution information in classes (DI) and weight adjustment factor (WA) into the feature weight algorithm to show the distribution of features in inter-class and intra-class [10].According to the analysis of amount information of words which have a low frequency, Luo Xin et al. proposed a feature selection method based on word frequency differentia and an improved TF-IDF method [11]. Lu Jiausedvariance in inter-class and intra-class which describes the distribution of features to revise TF-IDF weight in order to make the algorithm effectively weigh the distribution proportion of features [12].Su Lihua et al. also believed that if a feature has a high frequency in a certain category while it has low frequency in other categories, then this feature can easily distinguish the category from the others. So Inter-class Standard Deviation was introduced into TF-IDF method [13].

Ko Y [14] proposed an effective feature weighting method using the category distributions of features: the log-odds ratio of positive and negative category distributions. And his method worked well in speech-act classification. Peng T et al. [15] presented a novel TF–IDF-improved feature weighting approach, which reflects the importance of features in the positive category and the negative category, respectively. GONG Jinget al. [16] proposed an improved TF-IDF algorithm which considered not only the distribution condition of feature in class, but also the semantic factors such as the position of the feature, length of the feature. LI Feng-gang et al. [17] also proposed a new feature weighting method which considered the features and categories of correlation based on TF-IDF algorithm.

Researches above are all taking the distribution of features in inter-class and intra-class into account, but they only considered the relationship between features and the category which these features belonged in(positive category), without thinking over the relationship between features and the categories which these features didn't belonged in(negative category). In order to solve this problem, based on category-distribution divergence (CDD), we proposed a new

feature weighting method which introduces features' degree of membership and degree of non-membership into TF-IDF.

3. A feature weighting method based on category - distribution divergence

3.1 The main idea of CDD

After referring to VSM model and its improved representation method, Zhang Aihua et al. [18]proposed that feature weight factor should have these basic features: 1) The feature can represent the documents which it belonged in. 2) The feature can distinguishes the documents which it belonged in from others.3) The feature can represent the category which it belonged in. Based on these assumptions, we classify these basic features into two types: the representativeness for category and the distinctiveness for category. The former is used to describe how strong the feature can represent its category and the latteris used to describe how strong the feature can distinguishes its category from others.

For a feature$t$, if it has a higher frequency in category c, which compares to other categories , then we consider feature $t$has a strong distinctiveness for category c,because t can easily distinguishes the category $c$from others. And for a category c, $t_i$and $t_k$are two features of category c, and they all have a strong distinctiveness for category c. In category c, if t_i has a higher frequency than t_k, then we consider $t_i$has a stronger representativeness than t_k . For examples, feature words *currency* and *stock exchange* have very high frequency in category *Finance* while they have very low frequency in other categories, then we can believe that *currency* and *stock exchange* all have a strong distinctiveness for category *Finance*. But *currency* has a higher frequency than *stock exchange* in category *Finance*. So currency is considered to have a stronger representativeness than *stock exchange* for category *Finance*.

Furthermore, if a feature$t$ has a low frequency in category c but a high frequency in other categories (non-category c), it means that t has a strong distinctiveness for non-category c and t can easily tell the documents which does not belong in category c. Besides, in non-category, the higher frequency of feature t is, the stronger representativeness of feature $t$is. For instance, feature words *currency* and *stock exchange* have a very low frequency in category *Sport* while they have a high frequency in non-category *Sport*, it means that these two feature words have a strong distinctiveness for non-category *Sport*. Moreover, if *currency* has a higher frequency than *stock exchange* in non-category *Sport*, then *currency* is considered to have a stronger representativeness than *stock exchange* for non-category *Sport*.

To sum up, for any categories, the representativeness for category and the distinctiveness for category are called degree of membership, meanwhile, the representativeness for non-category and the distinctiveness for non-category are called degree of non-membership. For any feature in certain category, if the feature has a high degree of membership and a low degree of non-membership, which is meant that the distribution of this feature in inter-class is quite different, then this feature should be paid more attention.

3.2 The CDD Algorithm

We assume $t_i$is the *ith* feature, $c_j$is the *jth* category, and $c_j$is called as positive category while$\overline{c_j}$ is called as the negative category of c_j. Then we assume that N_{11} refers to the number of documents which include feature t_i, and these documents belong to category c_j; N_{10} refers to the number of documents which include feature t_i, and these documents don't belong to category c_j; N_{01} refers to the number of documents which don't include feature t_i, and these documents belong to category c_j; N_{00} refers to the number of documents which neither include feature $t_i$nor belong to category c_j(see in table 1).

Table 1 the relationship among parameters

	c_j	$\overline{c_j}$
t_i	N_{11}	N_{10}
$\overline{t_i}$	N_{01}	N_{00}

Based on the above reasons, we assume M as the total number of categories and N as the total number of documents, then the distinctiveness of t_i for category c_j is computed as Equation(4).

$$\mathrm{diff}(t_i, c_j) = \log_2\left(\frac{N_{11}}{N_{10}+1} + 1\right) \quad (4)$$

The representativeness of t_i for category c_j is computed as Equation (5).

$$\mathrm{repr}(t_j, c_j) = N_{11} \quad (5)$$

Then the degree of membership of t_i for category c_j is given by Equation (6).

$$\mathrm{belong}_{positive}(t_j, c_j) = \mathrm{diff}(t_j, c_j) * \mathrm{repr}(t_j, c_j) = N_{11} \log_2\left(\frac{N_{11}}{N_{10}+1} + 1\right) \quad (6)$$

Similarly, the degree of non-membership of t_i for category c_j (the degree of membership of t_i for category $\overline{c_j}$) is defined by Equation (7).

$$\mathrm{belong}_{negative}(t_j, c_j) = N_{10} \log_2\left(\frac{N_{10}}{N_{11}+1} + 1\right) \quad (7)$$

Hence, by combining Equation (6) and (7), we get the Equation (8).

$$\mathrm{belong}(t_j, c_j) = N_{11} \log_2\left(\frac{N_{11}}{N_{10}+1} + 1\right) - N_{10} \log_2\left(\frac{N_{10}}{N_{11}+1} + 1\right) \quad (8)$$

Finally, the category-distribution divergence function of t_i for category c_j is defined by Equation (9).

$$\mathrm{diver}(t_j, c_j) = \begin{cases} \log_2(\mathrm{belong}(t_j, c_j) + 1), & \mathrm{belong}(t_j, c_j) > 0 \\ 0, & \mathrm{belong}(t_j, c_j) \le 0 \end{cases} \quad (9)$$

Furthermore, we use function Log on the parameter tf in TF-IDF, then we get the Equation (10).

$$w_{TF-IDF} = \log(tf) * idf = \log(tf) * (N_{11} + N_{10}) \quad (10)$$

In summary, we defined our CDD method as Equation (11).

$$w_{CDD}(t_j, c_j) = w_{TF-IDF} * \mathrm{diver}(t_j, c_j) = \log(tf) * (N_{11} + N_{10}) * \mathrm{diver}(t_j, c_j) \quad (11)$$

4. Experimental Evaluation

4.1. Experimental setting

To verify our study, we use Sogou Lab Data[19] as experimental corpora. We select 18,000 news stories from 9 categories in Sogou Lab Data, including Car, Finance, IT, Health, Sports, Tourism, Education, Recruitment and Military. Each category consists of 200 documents. In this experiment, we randomly divided the data corpora into training and testing split as proportions of 1:1.In pre-processing stage, we use Lucene to process word segmentation and word frequency statistics, CHI_{max} model as our experiment's feature selection measure and VSM model as the text representation methods.

In order to eliminate the influence of documents length on the feature weight, we use function Cosine to normalize w_{id}, and it is shown as Equation(12) [20].

$$w_{id} = \frac{w_{id}}{\sqrt{\sum_{i=1}^{N}(w_{id})^2}} \quad (12)$$

Where w_{id} is the weight assigned to feature t_i in the document d, and N is the dimension of the feature space. In our experiment, we use K nearest neighbor algorithm (KNN), Rocchio algorithm and support vector machines(SVM) to test the validity of the CDD algorithm. The parameter k is equal to 7 in KNN and the parameter cost is equal to 8 and the parameter gamma is equal to 0.038125in SVM. Besides, cosine is used to compute the similarity between two documents by using Equation(13).

$$\mathrm{sim}(d_j, d_j) = \cos\alpha = \frac{\sum_{k=1}^{n}(w_{ik} w_{jk})}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{jk}^2}} \quad (13)$$

Where $\mathrm{sim}(d_i, d_j)$ is the similarity between document d_i and document d_j, w_{ik} is the kth feature in document d_i, n is the dimension of the feature space. Finally categorization effectiveness will be evaluated by Macro-averaged F-measure(MF).

4.1. Experimental results

We do comparison experiments using classical TF-IDF and our CDD at different dimensions. The experimental results are shown in table 2, table 3and table 4, the corresponding visual results are shown in figure 1, figure 2 and figure 3.

Table 2 The changes of value MF using KNN algorithm

Dimensions	TF-IDF	CDD	Change
360	0.861088	0.869729	1.00%
720	0.867383	0.873028	0.65%
1080	0.863268	0.876681	1.55%
1440	0.874079	0.875667	0.18%
1800	0.869721	0.874497	0.55%
2160	0.865671	0.881639	1.84%
2520	0.862303	0.88528	2.66%
2880	0.871957	0.891824	2.28%
3240	0.861402	0.889273	3.24%
3600	0.867098	0.889094	2.54%

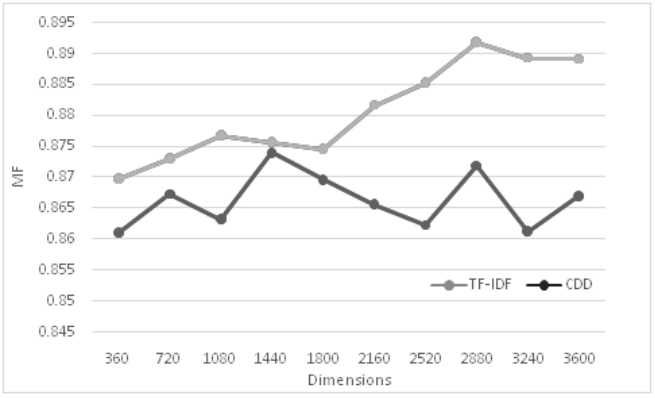

Figure 1 The changing curve of value MF using KNN algorithm

Table 2 and Figure 1 show the experimental results of TF-IDF and CDD using KNN algorithm at different dimensions: CDD gets a better performance than TF-IDF.

Table 3 The changes of value MF using Rocchio algorithm

Dimensions	TF-IDF	CDD	Change
360	0.863428	0.853386	-1.16%
720	0.859932	0.868646	1.01%
1080	0.867288	0.877372	1.16%
1440	0.863999	0.879292	1.77%
1800	0.866259	0.882649	1.89%
2160	0.867854	0.880741	1.48%
2520	0.864687	0.882988	2.12%
2880	0.864996	0.882384	2.01%
3240	0.865763	0.883349	2.03%
3600	0.866921	0.881302	1.66%

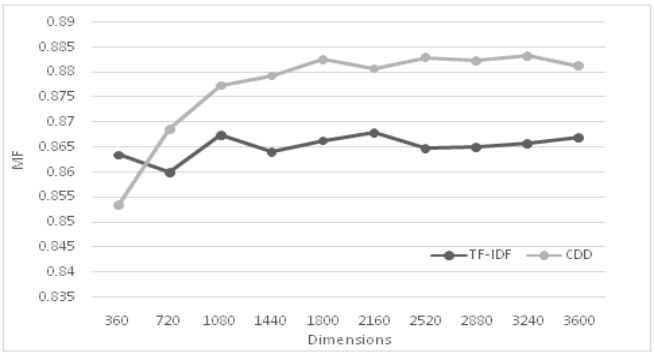

Figure 2 The changing curve of value MF using Rocchio algorithm

Table 3 and Figure 2 show the experimental results of TF-IDF and CDD using Rocchio algorithm at different dimensions: when the number of features is high enough, CDD gets a better performance than TF-IDF in text categorization.

Table 4 The changes of value MF using SVM algorithm

Dimensions	TF-IDF	CDD	Change
360	0.856182	0.852498	-0.43%
720	0.86292	0.86235	-0.07%
1080	0.85942	0.864653	0.61%
1440	0.868896	0.877307	0.97%
1800	0.867904	0.877799	1.14%
2160	0.866629	0.878054	1.32%
2520	0.867347	0.88205	1.70%
2880	0.865885	0.885968	2.32%
3240	0.864527	0.88424	2.28%
3600	0.86454	0.882057	2.03%

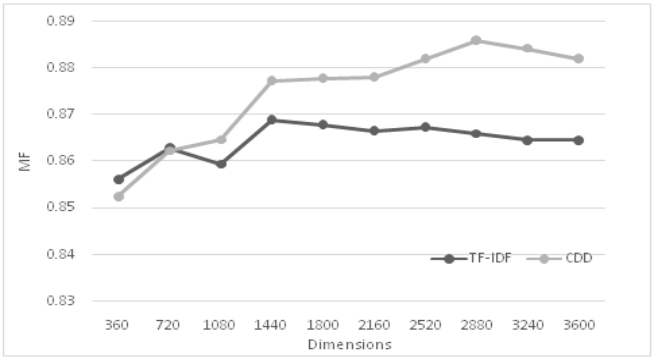

Figure 3 The changing curve of value MF using SVM algorithm

Table 4 and Figure 3 show the experimental results of TF-IDF and CDD using SVM algorithm at different dimensions: when the dimensions is high enough, CDD has a higher categorization precision than TF-IDF.

5. Conclusion

The paper describes a new feature weighting method based on category-distribution divergence(CDD),which use degree of membership and degree of non-membership in feature weighting calculation. Our research shows that CDD all works well in KNN, Rocchio and SVM algorithm and can get a higher categorization precision than classical TF-IDF. Since there are

many similarity algorithms between two documents, in this paper we only use function Cosine, so we cannot sure using other similarity algorithms can get the similar results. Therefore, the further studies will focus on using other similarity algorithms to test the validity of CDD.

6. Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 71373291).

7. References

[1] Tasi C S, Huang Y M, Liu C H, et al. Applying VSM and LCS to develop an integrated text retrieval mechanism[J]. Expert Systems with Applications, 2012, 39(4): 3974-3982.

[2] Gautam J, Kumar E, Khatoon M. Semantic Web Improved with IDF Feature of the TFIDF Algorithm[C]//Proceedings of the International Multi Conference of Engineers and Computer Scientists. 2014, 1.

[3] J Thorsten. A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for Text Categorization[C] // Proc of 14th Int. l Conf on achine Learing (ICML. 97) , 1997:143-151.

[4] Lin Yongmin, Lu Zhenyu, Zhao Shuang, Zhu Weidong. Research on Feature Weighting in VSM[J]. Journal of Information.2008.3:5-10

[5] Huang X, Wu Q. Micro-blog commercial word extraction based on improved tf-idf algorithm[C]//TENCON 2013-2013 IEEE Region 10 Conference (31194). IEEE, 2013: 1-5.

[6] HOW B C,ARAYANAN K.An empirical study of feature selection for text categorization based on term weightage[C] // Proceedings of the 2004 IEEE /W IC /ACM International Conference on Web Intelligence Washington. DC: IEEE Computer Society.2004:599-602

[7] Deng ZH, Tang SW, Yang DQ, etc.A Comparative Study on Feature Weight in Text Categorization[J]. ADVANCED WEB TECHNOLOGIES AND APPLICATIONS.2004.3007:588-597

[8] Zhang Baofu, Shi Huaji, Ma Suqin. an Improved Text Feature Weighting Algorithm Based on TF-IDF[J].Computer Applications and Software.2011.28(2):17-20

[9] Li yuan. Research on Word Segmentation and Feature Selection of Chinese Text Chinese Text Categorization [D].College of Computer Science and Technology of Jilin University.2011

[10] Zhang Yu, Zhang De-xian. Improved Feature Weight Algorithm[J]. Computer Engineering.2011. 37(5): 210-212

[11] Luo Xin,Xia De-lin,Yan Pu-liu. Improved feature selection method and TF-IDF formula based on word frequency differentia[J].Computer Applications. 2005.25(9):2031-2033

[12] Lu Jia. Improved Feature Selection Algorithm Based on Variance in Text Categorization[J]. Computer Engineering and Design. 2007. 28(24):6039-6041

[13] Su Li-hua, Zhu Zhang-hua, Bai Wen-hua. Term Weighting Algorithm in Text Categorization Based on VSM[J]. Computer Knowledge and Technology. 2010.6(33):9327-9329

[14] Ko Y. New feature weighting approaches for speech-act classification[J]. Pattern Recognition Letters, 2015, 51: 107-111.

[15] Peng T, Liu L, Zuo W. PU text classification enhanced by term frequency–inverse document frequency- improved weighting[J]. Concurrency and Computation: Practice and Experience, 2014, 26(3): 728-741.

[16] GONG Jing, HU Ping－xia, HU Can. Improvement of Algorithm for Weight of Characteristic Item in Text Classification[J]. Computer Technology and Development, 2014, 24(9): 128-132.

[17] LI Feng- gang, LIANG Yu, GAO Xiao- zhi, ZENGER Kai. Research on text categorization based on LDA-wSVM model[J]. Application Research of Computers. 2015, 32(001): 21-25.

[18] Zhang Aihua,J ing Hongfang, Wang Bin, Xu Yan. Research on Effects of Term Weighting Factors for Text Categorization[J]. Journal of Chinese Information Processing. 2010.24(3):97-104

[19] Sogou Lab. Sogou Lab Data [R/OL]. [2015-1-20]. http://www.sogou.com/labs/dl/c.html.

[20] Howard R. Turtle, W. Bruce Croft. A Comparison of Text Retrieval Models[J]. The Computer Journal, 1992, 35(3):279-290.