Journal of Information and Communication Engineering (JICE), 6 (1): 335-339, June 1, 2020
e-ISSN: 2432-5465, p-ISSN: 2186-9162
Applied Science and Computer Science Publications

Research on Text Classification Method based on PTF-IDF and Cosine Similarity

Yunxiang Liu
School of Computer Science & Information Engineering
Shanghai Institute of Technology
Shanghai, China
vxliu@sit.edu.cn

Qi Xu
School of Computer Science & Information Engineering
Shanghai Institute of Technology
Shanghai, China
xuqi.cn@outlook.com

Zhang Tang
School of Computer Science & Information Engineering
Shanghai Institute of Technology
Shanghai, China
tangzs808@foxmail.com

Abstract: Text classification is a foundational task in many NLP applications. The text classification task in the era of big data faces new challenges. We propose a Promoted TF-IDF (Promoted-TF-IDF) and cosine similarity method for text classification. In our model, with the pre-trained word segmentation tool, we apply PTF-IDF method to judge which words play key roles in text classification to capture the key components in category. We also apply Cosine Similarity algorithm to judge similarity between text and category. We conduct experiments on commonly used datasets. The experimental results show that the proposed method outperforms the state-of-the-art methods on several datasets.

Keywords: Text classification, TF-IDF, Computer Application, Natural Language Processing

I. INTRODUCTION

Text classification is an essential component in many applications, such as web searching, information filtering, and sentiment analysis [1]. Therefore, it has attracted considerable attention from many researchers.

Currently, the most commonly used algorithms for text classification are KNN, SVM, and Naive Bayes. The KNN algorithm was proposed by T. M. Cover in 1967[2]. Gongde Guo et al. applied KNN to text classification tasks [3], Researchers such as Jiang S have conducted in-depth research on KNN algorithm [4-5], KNN is not sensitive to outliers and has high algorithm complexity, so it is not suitable for large-scale text classification; Bayesian algorithm was proposed in the 1950s, and Kim et al. applied the Naïve Bayesian method to the field of text classification [6].Frank applied Naïve Bayes to unbalanced text classification[7], and Researchers had proposed many improvements based on Bayesian methods[8-9]. Naïve Bayes cannot handle group classification problems. Therefore, the naive Bayesian method is not suitable for large-scale text classification. Joachims made the proof of the SVM classifier in 1990[10], and researchers have conducted other research based on SVM methods [11-12]. Experimental results show that the combination of GHI feature selection algorithm, TF-IDF weight value calculation method and the SVM classifier method perform better [13], but the support vector machine performs well in the classification problem of small sample space. The TF-IDF algorithm is a commonly used algorithm in the field of information retrieval and data mining to evaluate the importance of a word in a document. Wu Yongliang used the traditional TF-IDF algorithm for text classification and the experimental results performs better than the above commonly used algorithms [14]. Ye Min et al. introduced the word position and word length information in the improved TF-IDF algorithm without considering the distribution of words in the document category [15]. Based on the above research progress, We optimize the TF-IDF algorithm, and propose a text classification method based on PTF-IDF and cosine similarity.

II. TF-IDF ALGORITHM

A. Traditional TF-IDF Algorithm

The idea of the TF-IDF algorithm is that if a word or

Corresponding Author: Qi Xu. School of Computer Science and Information Engineering, Shanghai Institute of Technology, China. E-mail: xuqi.cn@oulook.com

phrase appears frequently in one article and rarely in other articles, it is considered that the word or phrase has a good ability to distinguish categories and is suitable for classification. The traditional TF-IDF formula is shown in (1).

エラー! 参照元が見つかりません。 (1)

Where is the number of occurrences of the word in the document; represents the total number of feature items of the document; N represents the total number of documents; and represents the number of documents containing the term.

B. Analysis and Optimization of TF-IDF Algorithm

The traditional TF-IDF algorithm only considers the Frequency of feature items in the document and the distribution of feature items throughout the training set. However, the concentration of feature items throughout the categories and the dispersion within the class are not considered. Therefore, the optimization of the traditional IF-IDF algorithm starts from the following two aspects.

The fewer the number of text categories in which the feature item appears, the more representative it is, and the stronger the ability of the feature item to distinguish other different categories of text. We define the relationship between this feature item and the category as the degree of concentration between classes, denoted by. If a feature item is only in one class, then the feature item appears in a certain document, and the document can be exactly attributed to the class. The inter-class concentration of the above analysis feature items can be expressed as equation (2).

$$\mathrm{P}(t_i) = \frac{|C|}{|C_i|} \qquad (2)$$

Where is the total number of categories in the class training set and is the number of categories containing the feature item .

If a feature item appears in a small amount of text in a certain type of document, the feature item for the category is more representative. We define this feature term and the distribution relationship within the class as the intra-class dispersion, expressed as equation (3).

エラー! 参照元が見つかりません。 (3)

Where represents the total number of documents in a category in the training set, and represents the number of documents containing feature items in a category in the training set.

C. PTF-IDF Algorithm

According to the above analysis, the traditional TF-IDF algorithm is improved to obtain a PTF-IDF algorithm as shown in the formula (4).

$$\mathrm{PTF-IDF} = \mathrm{P}(t_i) \times \mathrm{Q}(t_i) \times TF\text{-}IDF$$

エラー! 参照元が見つかりません。 (4)

III. Cosine Similarity Algorithm

The idea of the cosine similarity algorithm is the similarity of vector is measured by the angle cosine of the two vectors. The closer the cosine value is to 1, the smaller the angle between the two vectors in space [9]. The calculation formula of cosine similarity is as shown in (5).

エラー! 参照元が見つかりません。 (5)

Where represents the dot product of vector **a** and vector **b**, and represents the product of the modulus of vector **a** and vector **b**.

IV. Text Classification Method Based on PTF-IDF and Cosine Similarity

A. Method Architecture

The architecture of the text classification algorithm proposed in this paper is shown in Figure 1.In the first stage, the data of the training set is pre-processed,and then the symbols (such as ",", ".") and stop words (such as "and", "or", "but") are removed. Finally, the category keywords are extracted using the PTF-IDF algorithm.

In the second stage, the PTF-IDF value of the keyword of the text is calculated after the test text is preprocessed.

In the third stage, calculating the similarity between text and each category, then classify the text with the category with the highest similarity.

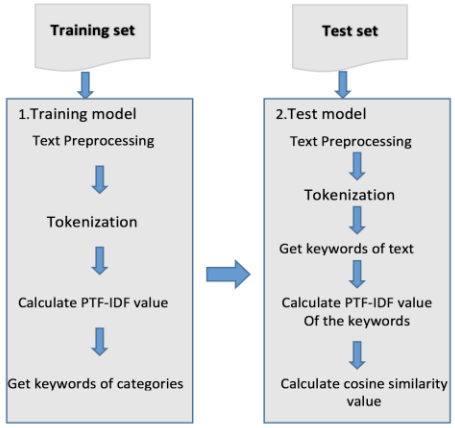

Fig. 1. Text classification algorithm architecture.

B. Optimal Category Keyword

The excessive dimension of the vector causes an increase in the amount of calculation, and too many keywords may result in a decrease in classification efficiency. When the category keywords reach a certain proportion, the classification accuracy remains stable. Taking the text corpus of Fudan University as an example, the test results are shown in Figure 2. Therefore, this article will select the top 75% of the keywords as analog keywords.

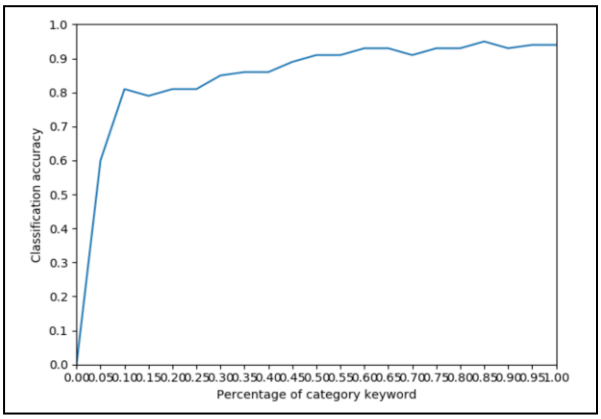

Fig. 2. Relationship between category keyword ratio and classification accuracy.

This paper will sort the PIF-IDF values of the feature items and select the best category keywords. Category keywords indicate category characteristics, and PTF-IDF values measure the importance of keywords. We display the top ten keywords' PTF-IDF value of the four categories in the BBC Corpus in Table 1.

TOP 10 KEYWORDS IN FOUR CATEGORIES IN THE BBC DATASET

Business/TF-IDF	Politics/TF-IDF	Tech/TF-IDF	Entertainment/TF-IDF
Growth/0.18100	Film/0.40002	Labour/0.321940	People/0.29197
Market/0.18000	Music/0.32023	Election/0.320042	Users/0.23078
Economy/0.17001	Show/0.27000	Government/0.266390	Software/0.19054
Company/0.16800	Actress/0.24003	Party/0.213720	Mobile/0.18912
Bank/0.16401	Festival/0.19003	People/0.196238	Technology/0.187
Business/TF-IDF	Politics/TF-IDF	Tech/TF-IDF	Entertainment/TF-IDF
			34
Economic/0.16400	Awards/0.19001	Tories/0.175112	Digital/0.16378
Firm/0.16202	Album/0.15340	Brown/0.160021	Music/0.14890
Yukos/0.15900	Band/0.13600	Minister/0.147715	Game/0.14007
Sales/0.15404	Award/0.12045	Chancellor/0.110239	Computer/0.13045
Government/0.14605	Chart/0.10034	Lib/0.107234	Microsoft/0.12076

V. EXPERIMENT

A. Dataset

BBC (BBC News Data) This data set is often used for text categorization tasks. The corpus contains 2,225 news texts. The five categories are business, entertainment, sports, technology, and politics. (https://www.kaggle.com/shineucc/bbc-newsdataset)

Reuters-21578 dataset, which contains 21,078 1987 Reuters news documents this data set is often used for text classification experiments. (http://archive.ics.uci.edu) Fudan University text classification corpus, which contains 20 categories of texts such as art、 education、 philosophy、 history、 communication、computer, etc. This dataset is often used in experiments about Chinese information processing.
(https://pan.baidu.com/s/1pxEN2OLAYOousdbhIQAAw)

B. Experimental Evaluation Indicators

In this paper, the accuracy, recall and F1 values are used for classification evaluation, and the classification indicators TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative) are defined.

The accuracy P is used as an indicator for the classifier to correctly recognize the text. The calculation formula is as shown in the formula (6).

エラー! 参照元が見つかりません。 (6)

The recall rate R refers to the ratio of the retrieved related documents to all relevant documents in the document library, and the calculation formula is as shown in equation (7).

エラー! 参照元が見つかりません。
(7)

The F1 value is a standard used to measure the accuracy of the classifier, and the calculation formula is as shown in equation (8).

エラー! 参照元が見つかりません。
(8)

Where P represents the above accuracy rate and R represents the above recall rate. Its maximum value is 1 and the minimum value is 0.

VI. Experiment Resalts

In this paper, the cross-validation method is adopted, and the F1 value is selected as the comprehensive performance index. 80% of the data is randomly selected as the training set, and 20% is used as the test set for the experiment. The experimental results of the traditional TF-IDF and cosine similarity text classification algorithms and the proposed PTF-IDF and cosine similarity text classification algorithms are shown in Figure3 on the Reuters-21578 dataset. The experimental results show that the PTF-IDF algorithm is better than the traditional TF-IDF algorithm on the Reuters-21578 dataset.

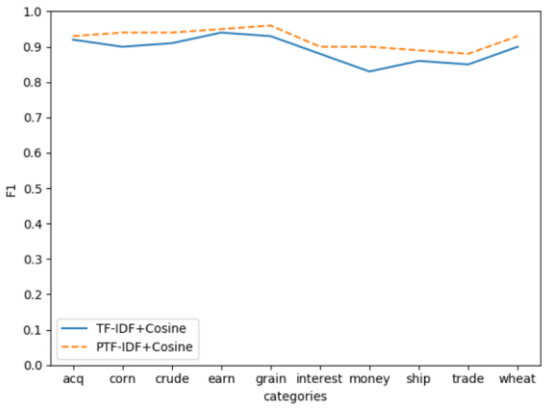

Fig.3 Comparison of experimental results on the Reuters-21578 dataset

VII. Conclution and Future Work

This paper analyzes the shortcomings of the traditional TF-IDF algorithm and proposes an improved PTF-IDF algorithm. Furthermore, a text classification algorithm based on PTF-IDF and cosine similarity is proposed. And compared with the traditional TF-IDF algorithm; and based on the experiment of finding the optimal keyword, the paper finds that the accuracy of text classification reaches a stable value when the category keywords reach a certain proportion. Subsequent research includes: more in-depth study of text sequence information and adding text sequence information to the classification process will greatly help the accuracy of text classification.

References

[1] Aggarwal, C. C., and Zhai, C. 2012. A survey of text classification algorithms in mining text data. Springer. 163-222.

[2] Cover Y, Hart P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27

[3] Guo G, Wang H, Bell D, et al. Using kNN model for automatic text categorization[J]. Soft Computing, 2006, 10(5): 423-430.

[4] Jiang S, Peng G, Wu M, et al. An improved K-nearest-neighbor algorithm for text categorization[J]. Expert Systems with Applications, 2012, 39(1):1503-1509.

[5] Soucy P, Mineau G W. A simple KNN algorithm for text categorization[C]. Proceeding IEEE International Conference on. IEEE, 2001: 647-648

[6] Kim S B, Han K S, Rim H C, et al. Some effective techniques for naïve bayes text classification[J]. Knowledge and data Engineering, IEEE Transactions, 2006,18(11): 1457-1466.

[7] Frank E, Bouckaert R R. Naïve bayes for text classification with unbalanced classes[M]. Knowledge Discovery in Databases PKDD 2006. SpringBerlin Heidelberg, 2006: 503-510.

[8] Wang S, Jiang L, Li C. Adapting naïve bayes tree for text classification[J]. Knowledge and Information Systems, 2015, 44(1):77-89

[9] Rennie J D, Shih L, Teeven J, et al. Tackling the poor assumptions of naïve bayes text classifiers[C] Proceeding of the ICML, 2003, 3612-3623

[10] Joachims T.Transductive inference for text classification using support vector machines[C]. Proceedings of the International Conference on Machine Learning.1999(99): 200-209.

[11] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. The Journal of Machine Learning Research, 2002(2): 45-66

[12] Kim H, Howland P, Park H. Dimension reduction in text classification with support vector machines[J]. Journal of Machine Learning Research, 2005: 37-53

[13] LongMao Hu. Research on comparison of Chinese text classification techniques[J]. Journal of Anqing Teachers Colleges(Social Science Edition), 2015(2): 49-53

[14] Yongliang Wu, Shuling Zhao, et al. Text Classification Method Based on Tf-IDF and Cosine Similarity[A]. Journal of Chinese Information Processing. 2107.

[15] Ye Min, Shiping Tang, Zhendong Niu. An Improved Chinese text classification algorithm based on Multiple Feature Factors[A]. Journal of Chinese Information Processing. 2107.