Application Research of English Scoring Based on TF-IDF Clustering Algorithm

Saihanqiqige*
Chifeng University, Chifeng 024000, China
*Corresponding author e-mail:1461942076@qq.com

Abstract. In recent years, with the promotion and application of artificial intelligence technology in all fields, English has also had a lot of attention and automatic grading areas, but not much in the text of the characterization of breakthrough, the traditional technology based on latent semantic analysis of the text said, more latent semantic analysis technology can extract thematic information, and information is ignored. The purpose of this paper is to better reflect the content of the text, evaluate the English level more accurately, and improve students' English level better. This paper proposes a text representation method based on word vector clustering and a text representation method based on vector space model. A TF-IDF algorithm based on word vector is proposed. It is verified that the quadratic weighted Kappa value of the prediction results of the multi-model fusion algorithm based on word vector is better than the first place in Kaggle international English composition scoring competition, which verifies the effectiveness of the algorithm.

Keywords: TF-IDF, English Grading, Term Vectors, Clustering Algorithm, Vector Space

1. Introduction
With the development of computer technology, especially the development of artificial intelligence technology in recent years, computer technology is increasingly applied to people's study, work and life. Automatic composition scoring (AES) is one of them. In China, there are college entrance exams in English, and college students have cet-4 and cet-6[1-2]. In these tests, there is an English essay writing assessment, which is used to judge students' logical thinking ability and language ability. In addition to simple single choice scoring, English writing is a subjective expression, and artificial English essay scoring is also subjective, which may be beneficial to one reviewer but unsatisfactory to another. Due to the lack of clear criteria and the fact that each reader may have a subjective preference, the scores of different readers with the same level of English will vary [3-4]. In addition, due to the large number of students in the class, even ordinary writing tasks, a teacher needs to grade dozens of papers at a time, which requires a lot of effort from the teacher. If we can introduce automatic scoring system for English compositions, we can effectively solve the above problems [5-6].

To achieve this, we need to consider factors that reflect students' writing skills, such as the number
of common spelling mistakes, the proficiency with which clauses are used, and the relevance of writing to the subject. Secondly, after setting the scoring criteria, how to use the computer to extract relevant information from students' compositions accurately and automatically depends not only on the relevant research of English ontology, but also on the development level of natural language processing technology [7-8]. The essence of automatic composition scoring is that the computer matches the characteristic information of the training composition with the training composition scoring, then learns relevant model functions, and then predicts the process of composition scoring [9-10].

Although English composition has received the widespread attention and automatic grading areas, but not much in the text of the characterization of breakthrough, the traditional n - gramf or Latent Semantic Analysis based on the text said (Latent Semantic Analysis, abbreviated as LSA), and n - gramf grammar algorithm can only save the information, the last words of the adjacent word information will be ignored, it is easier to cause data sparse, and Latent Semantic Analysis technology can extract thematic information, and ignore information [11-12]. Therefore, this paper tries to study the representation of text content, and on this basis studies and designs a TF-IDF based on word vector, which is used in English composition automatic scoring clustering algorithm.

2. Method

2.1 Methods for Obtaining Original Data
First, the author, time and text of the comments related to Think PadE570c on the English composition are captured by the Python code on the web crawler and recorded in the person database. After comment fetching, the text content is screened and filtered. Eliminated all repetitive, copy-and-paste comments, changed typos, and broke some unpunctuated sentences. Complete the research data by using the jieba participle in Python to complete the comment text participle. In terms of comment attributes, combined with the collected data, the daily commodity price and the number of words of each comment were counted and recorded into the database to calculate the number of new comments, the average length of comments, the average score, the average number of replies and the average number of pictures.

Based on the above data, is proposed based on TF-IDF improved clustering algorithm sensitive information of text mining technology, through the TF - sensitive information of text, IDF method for network to obtain valuable sensitive information in the text of the sensitive information characteristics, using clustering algorithm, clustering analysis was carried out on the all sensitive information characteristics, complete network of sensitive information mining. The proposed method is efficient and accurate in mining network sensitive information. In terms of methods, TF-IDF algorithm, clustering and SPSS statistics have become popular research methods. The TF-IDF algorithm is used to extract the information, which improves the efficiency of classification and simplifies the follow-up query. Jieba segmentation in TF-IDF algorithm was used to extract the text content of WeChat public account of 985 university to understand the subject content that students need to push in different periods. Considering the defects in word meaning statistics when using TF-IDF algorithm to mine text content, some studies believe that when extracting keywords or keywords, the extraction precision of keywords can be improved by co-occurrence words and k-means clustering. Similarly, another study used Web text mining to extract the feature words of Chinese entries, and then used hierarchical clustering to get the conclusion. In addition, word frequency was counted by SPSS to extract hot text topics.

2.2 Text Acquisition of Sensitive Information
TF-IDF method is usually used to extract network sensitive information text. By comparing the word frequency of network sensitive information, the words with high word frequency in network information content are collected to obtain the text of network sensitive information. The main idea of the TF-IDF method is that if the frequency TF of a word or phrase is high in one article but rarely
appears in other articles, it indicates that the word or phrase has good classification ability and can be used for classification. TF-IDF is the term frequency of TF and the reverse document frequency of IDF. TF represents the frequency of sensitive words generated in document d. The main idea of IDF is: if there are fewer documents containing the sensitive word t, that is, the smaller n is, the larger the IDF is, then the sensitive word has a good classification ability. The process of obtaining network sensitive information text by TF-IDF method is as follows:

\[ W_{ij} = T F_{ij} \times I D F_i \]  \hspace{1cm} (1)

\[ T F_{ij} = \frac{F_{ij}}{\max \{ F_{kj} \mid k = (1,2,\ldots,T) \} } \]  \hspace{1cm} (2)

\[ I D F_i = \log \left( \frac{N}{n_j} \right) \]  \hspace{1cm} (3)

Where \( W_{ij} \) represents the proportion of sensitive word \( k_i \) in the document \( d_j \), that is, the text of sensitive information in the English composition obtained; \( F_{ij} \) represents the frequency of the sensitive word \( k_i \) in the document \( d_j \), and the document \( d_j \) contains \( T \) keywords. \( N \) represents the total number of documents; \( n_j \) represents the total number of documents containing the sensitive word \( k_i \).

3. Experiment

This article selects data sets published by the "component scoring contest" on Kaggle, an international data mining platform. The constituent dataset consists of eight data subsets, each of which has a corresponding constituent topic. Students write according to the description requirements of the writing topic. For each subset of the training data, there were more than a thousand compositions learned by students and corresponding artificial scores, and the number of words in each composition ranged from 150 to 600. The score range of each data subset is also different. For example, the score range of data set 3 is 0-3, while the score range of data set 7 is 0-30, and the score range of data set 8 is 0-60.

The criteria chosen were Kaggle's "essay scoring competition" criteria, whose model and manual scores were measured by a quadratic weighted Kappa value. The value range for Kappa values is \([0,1]\). The higher the value, the higher the consistency between the two evaluators. When Kappa value is 0, this means that there is no correlation between the two scores and it is random. When Kappa is 1, the two scores are exactly the same. First, the quadratic weighted Kappa values of component fractions and artificial fractions of the model were calculated for each collection of papers, and then the Kappa values were transformed by fischer transform. Finally, the average Kappa value of all papers was taken as the Kappa value of the final English paper scoring model. Let the composition score have \( N \) grades : \( 1,2,\ldots,n \). There are two raters A and B, and the score of each essay e can be represented by an array of \((e_n, e_b)\), where \( e_n \) stands for the score of essay e by rater A(manual score), and \( e_b \) stands for the score of essay e by rater B(model score).

4. Discuss

4.1 Effect Verification of Text Features

Using the non-textual features, the author trained the model on 8 composition subsets and predicted the test set score, and calculated the corresponding quadratic weighted Kappa value. Among all composition data sets (1-8), the quadratic weighted Kappa value of random forest was the largest at
0.7734, followed by XGBoost at 0.7625, while the result of gradient ascending tree was only 0.7383. The author analysis, since each subset is only more than one thousand composition composition, all added up to more than ten thousand cases of sample composition subset, participate in the training model of the data is still not full, like the gradient promotion tree and XGBoost this two models are based on Boosting method, so in the case of a small amount of data, prone to excessive model fitting data, the model is relatively complicated, too much to think about the sample's personality, will overwrite the generality of the samples, leading to poor prediction effect. However, the random forest is realized based on Bagging method, and the results generated by multiple decision trees are used to generate the prediction results in the form of voting or taking the mean. In this way, the over-fitting phenomenon can be effectively avoided and the variance can be reduced even when the training sample data volume is small. Therefore, random forest can perform better in the case of small data volume.

On the basis of the non-text features and the deduction degree features, the models were separately practiced in all the test sets and the corresponding scores were predicted in the test sets, and the corresponding quadratic weighted Kappa values were calculated. The experimental results are shown in Table 1.

| Characteristics of the model | Random forests | XGBoost | Gradient lifting tree |
|------------------------------|----------------|---------|----------------------|
| Nontextual feature           | 0.7734         | 0.7625  | 0.7383               |
| Non-textual features and deduction features | 0.7806 | 0.7684 | 0.7412 |

4.2 Experimental Results of Content Text Features of Word Vector Clustering

It can be seen from table 1 that, on the basis of non-text features, the predictive indexes of each model were improved after the deduction degree feature was added. Among them, the random forest model showed the greatest improvement. After adding the deduction feature, the second-weighted Kappa value increased from 0.7734 to 0.7806, which verified the effectiveness of the deduction feature for the English composition scoring model.

This paper not only introduces the extraction of deductive features but also introduces the method and steps of text feature generation based on word vector clustering. This paper choose the wikipedia English corpus as the training corpus, and use word2vec training language model, the window size is set to 5, word vector dimension is set to 400, truncation word number is set to 5, there are two kinds of models, respectively is CBOW model and Skip - "gramm model, when the parameter sg is set to 0 (the default is 0) represents the selection CBOW model, sg is set to 1 when the selection is the Skip -" gramm model. Set the number of clustering centers of word vector clustering to 20. Based on non-text features and deductive features, content text features extracted based on word vector clustering were added to the model for training. The results are shown in Figure 1.
As can be seen from Figure 1, after adding content text features based on word vector clustering, the effect of each model was greatly improved, especially the gradient lifting tree model, which was improved by about 0.04. The XGBoost model also has a big boost, from 0.7684 to 0.7983, and its prediction performance has surpassed that of the random forest, performing best of the three models. The XGBoost model does not predict as well as a random forest when only non-text features are present, but it does better when more features are added. In addition, it can also be seen from the experimental results that the prediction effect of skip - Gram is better than that of CBOW on the whole. The author believes that this is because in the calculation, CBOW will add up the text content, and the prediction effect will be greatly reduced when rare words are encountered, while skip -- Gram can predict the usage environment of rare words. In addition, in general, CBOW is applicable to small training corpus, while skip.gram is more applicable to large corpus. The wikipedia English corpus selected in this paper is more than 14G, so it is still a relatively large training corpus, so Skip -- Gram also performs better.

5.Conclusion
In view of the low efficiency of text representation in the current English paper scoring system, this paper proposes the text representation clustering method based on the word vector TF-IDF and the text representation method based on the vector space model, which can not only fully represent the word information, but also give consideration to the degree of representation. This paper extracts a large number of non-text features that can reflect the quality of students’ compositions from the side. At the same time, it summarizes a set of machine learning model parameter adjustment methods based on its own experience in adjusting parameters, so that the model parameters can be adjusted to the optimal in a directional and efficient way, and the results of model prediction can be further improved. On this basis, a TF - based IDF algorithm for word vector English composition is designed.

References
[1] Hong Kang, Zhiguo Yu, Yang Gong. Initializing and Growing a Database of Health Information Technology (HIT) Events by Using TF-IDF and Biterm Topic Modeling[J]. AMIA. Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 2018, 2017:1024-1033.
[2] Inbal Yahav, Onn Shehory, David Schwartz. Comments Mining With TF-IDF: The Inherent Bias and Its Removal[J]. IEEE Transactions on Knowledge & Data Engineering, 2018, PP(99):1-1.
[3] Zigui Jiang, Rongheng Lin, Fangchun Yang. A Fused Load Curve Clustering Algorithm
Based on Wavelet Transform[J]. IEEE Transactions on Industrial Informatics, 2018, 14(5):1856-1865.

[4] Yinhao Lu, Buyang Cao, Cesar Rego. A Tabu Search based clustering algorithm and its parallel implementation on Spark[J]. Applied Soft Computing, 2017, 63:97-109.

[5] Tengfei Zhang, Fumin Ma. Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function[J]. International Journal of Computer Mathematics, 2017, 94(4):663-675.

[6] W. Li, J.-Y. Zhao, T.-S. Yan. Improved K-means clustering algorithm optimizing initial clustering centers based on average difference degree[J]. Control & Decision, 2017, 32(4):759-762.

[7] Tao Lei, Xiaohong Jia, Yanning Zhang. Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering[J]. IEEE Transactions on Fuzzy Systems, 2018, PP(99):1-1.

[8] Xiaofeng Mai, Jie Liu, Xiong Wu. Stokes space modulation format classification based on non-iterative clustering algorithm for coherent optical receivers[J]. Optics Express, 2017, 25(3):2038.

[9] Ali Feizollah, Nor Badrul Anuar, Rosli Salleh. Evaluation of Network Traffic Analysis Using Fuzzy C-Means Clustering Algorithm in Mobile Malware Detection[J]. Advanced Science Letters, 2018, 24(2):929-932.

[10] Orkid Coskuner, Vladimir N. Uversky. Tyrosine Regulates β-Sheet Structure Formation in Amyloid-β42: A New Clustering Algorithm for Disordered Proteins[J]. Journal of Chemical Information & Modeling, 2017, 57(6):1342-1358.

[11] Chunyong Yin, Lian Xia, Sun Zhang. Improved clustering algorithm based on high-speed network data stream[J]. Soft Computing, 2017, 22(4):1-11.

[12] Hu Zhang, Yongke Wu. Optimization and Application of Clustering Algorithm in Community Discovery[J]. Wireless Personal Communications, 2018, 102(2):1-12.