Document Segmentation Method Based on Style Feature Fusion

Gang Liu\textsuperscript{1,2,a}, Kai Wang\textsuperscript{1,b}, Wangyang Liu\textsuperscript{2,c}, Xu Cheng\textsuperscript{2,d} and Tao Li\textsuperscript{1,e}

\textsuperscript{1}Harbin Engineering University, Harbin, Heilongjiang, China
\textsuperscript{2}CETC big data research institute company limited, Guiyang, Guizhou, China
\textsuperscript{a}liugang@hrbeu.edu.cn; \textsuperscript{b}2503613988@qq.com; \textsuperscript{c}liuwangyang@cetcbigdata.com;
\textsuperscript{d}chengxu@cetcbigdata.com; \textsuperscript{e}389083838@qq.com;

Abstract. Style crack refers to the position where the author's identity changes in the article completed by multiple authors. This paper summarizes the current situation and theory of related fields at home and abroad, and proposes a multi-feature based document segmentation method for plagiarism detection. Seven text style features are used for style crack recognition. Through the result of feature extraction, the combination of multi-feature fusion and unsupervised machine learning algorithm is used to classify the features based on extraction, and the clustering algorithm is used to cluster the style features so as to find the location of style cracks. Experiments show that the method is effective and scientific, and achieves good results.

1. Background and significance of the project

With the advent of the data age, people's research on essays and plagiarism detection are not only based on simple string judgments, but also on the author's writing style and writing habits. The writing style can not only see an author's writing habits, but also can be used in the plagiarism detection system and user portrait technology, and also helps the author's identification, and better plagiarism detection system. Providing a new perspective, new directions, and strong support for the authors of online anonymous articles[1]. Different people will travel their own unique styles when writing, mainly in terms of words, sentences, paragraphs, rhetoric, emotions, etc. These are the writing habits that the author has inadvertently developed, so through the writing style of the article. Feature extraction, inferring the attribution of an article through style features is effective[2].

2. Related theory and work basis.

The early style research mainly used statistical methods, mainly to count the laws of vocabulary, sentences and paragraphs, and use the laws of statistics to agree on a person's style. The extraction of style features was the first to study single features. As single features could not satisfy the experimental results, multi-feature fusion also came into being. In recent years, the development of machine learning and neural networks has introduced machine learning and neural network algorithms into style extraction and author recognition, and has achieved good results.[3]

Because of the variety and difficulty of Chinese, it is obviously more difficult to extract Chinese style than that of foreign language. Chinese needs to take into account the accuracy of word segmentation system and the complex sentence structure. Although the extraction of Chinese style is more difficult than that of foreign countries, the study of style has also received widespread attention.
3. Multi-feature extraction and fusion

3.1. One-dimensional style feature

3.1.1. Word length. In terms of Chinese style, it is possible to observe the author's usage habits of two-word words, three-word words, four-word words and four-word words in terms of word use according to the length of the vocabulary after the word segmentation. Some studies have found that the frequency of use of double-word words, Zhang Ailing's frequency of use is 0.17, and Lu Xun's frequency of use is as high as 0.43. From this we can see that the length of the word can be seen in the style of the article.

In English, the number of letters is statistically counted. The range of average word lengths in Chinese is relatively small, and the average word length is between 2-4. However, the experimental results can be seen by narrowing the comparison range. Finally, the word length is taken as a parameter of the final classification.

3.1.2. Average sentence length. The average sentence length is the usage habit of the statistical author for the length of the text sentence, and the frequency of use of the long and short sentences. The length of each sentence is counted, and the average sum is averaged. The average sentence length is marked with ",", "!", and "?", and the average value of the length words in the statistical sentence is taken as the parameter of the last dimension.

3.1.3. Emotional bias. Emotional analysis has always been the analysis of emotional words to analyze emotions. Fan Xia uses neural networks for emotional analysis and verifies their effectiveness and compares their results [4]. Sailunaz uses machine learning to calculate user impact scores based on various user-based and Twitter-based parameters, and to perform sentiment analysis on Twitter [5]. This paper uses the sentiment dictionary trained in the network to analyze the article emotionally.

3.2. Multidimensional style features

3.2.1. Lexical features. The use of words can completely extract a person's literary skills, the same can be based on the richness of the use of words to judge an author's writing style. The lexical features [6] can define the length of the word, the word frequency, the proportion and the density, and define the vocabulary characteristics as shown in Table 3.1:

| Numbering | Lexical features                        | Numbering | Lexical features                        |
|-----------|----------------------------------------|-----------|----------------------------------------|
| 1         | Total number of words                  | 5         | Self-created words/total words         |
| 2         | Number of words in two words           | 6         | Interjection / total number of words    |
| 3         | Number of words in three words         | 7         | Number of different words / total number of words |
| 4         | Number of words in four words          | 8         | Lexical density                        |

3.2.2. Special punctuation. Special punctuation features statistical colon, semicolon, thousand percent, unit symbol, left and right quotation marks, left and right brackets, exclamation mark, ellipsis, dash, question mark and comma.

The special punctuation uses the simple statistical method. By using the statistics of the special punctuation above as the punctuation parameter of the final classification algorithm, it is of course necessary to delete the punctuation with the punctuation frequency of 0, so the text dimension is up to 11 dimensions.

3.2.3. Synonym. Synonym is a unique branch of Chinese style, which can extract the strength of Chinese literature. It can also extract the author's language skills and ability to control words in the use of synonyms. The usage habits of synonyms can also be seen in an author's habit of using words. This
feature is based on the use of synonyms, based on the use of synonyms, to summarize the usage habits of synonyms, and then summarize the author's writing habits.

3.2.4. Function word. This paper increases the number of virtual words, and uses the custom vocabulary table as a benchmark to calculate the use of the virtual words in the virtual vocabulary. First of all, the virtual vocabulary table is produced. The source of the vocabulary table is "Now the Chinese Dictionary of Function Words". There are 840 virtual words in the vocabulary list, which are the same as the synonym table. The virtual vocabulary table is too large. The vocabulary table contains some uncommon and uncommon function words, which will affect Calculation of results. Based on the news set, the TF-IDF statistics are performed on the virtual words of the virtual vocabulary, and the TF-IDF is deleted too low. Through multiple cleanings, it was finally streamlined to 230 virtual words. The choice of 230 virtual words can be controlled first in a reasonable dimension.

4. Style crack recognition

4.1. Style crack
Style cracking refers to the position where a text style changes. In other words, an article may be completed by different authors. Therefore, it is especially important to perform author-based segmentation techniques before author identification, that is, to find out each article in this article. The corresponding part of the author's text. By writing style segmentation, the goal is to find the style crack points, where the style and style change. Style crack recognition is mainly through the technique of feature extraction and classification algorithm, and the use of sliding window, dimension reduction and other techniques. Finally find out the style crack points.

4.2. Sliding window
The sliding window is composed of five sentences as a whole for style feature recognition. Each time you slide down a sentence, the sliding window contains five sentences at a time, and each window is styled. When the style changes, each style and the last result gradually change until the style is similar. The degree is close to the same, and there is a style crack in this position.

Because the length of the paper is small, the five sentences contain less information, and there is a large possibility that there will be an accidental phenomenon. It is assumed that there is a great possibility that the position of each style crack occurs at the end of each segment. Therefore, in order to improve the accuracy rate, only the recall rate can be sacrificed. It is assumed that each style crack will necessarily occur at the end of the paragraph, that is, each paragraph in the article is assumed to have one and only one author.

4.3. Parameter weighting
Aiming at the feature redundancy problem caused by the third chapter's style feature analysis, each parameter takes different weights in the process of style crack identification, so when looking for style cracks, we need to find the weights of each parameter, and then identify style cracks by adjusting the weights of parameters.

Algorithm description: Firstly, the data set is preprocessed, the data set is extracted, and the data set is stored in the file in disorder. In the case that the weights of other feature parameters are the same, the idea of controlling variable method is used first. The control word length parameter (WLP) is calculated from 0.01 to 0.99, and the other parameters are 0.5. The highest quality of the parameter WLP is obtained under the condition that other parameters are unchanged, and the highest quality is the lowest degree of similarity between the two texts. When the other parameters are unchanged, the average sentence length parameter (ASLC) is calculated from 0.01 to 0.99, and the optimal value of the ASLC is obtained, thereby traversing all the parameters, and then the optimal result of the previous parameter is used as a reference. Continue the above method to loop until the parameter optimal value is unchanged, and obtain the parameter weight recombination. The purpose of this step is to find the
validity of each parameter by the parameter weight method and delete the influence of the invalid parameter. It is found that some of the parameters weights are too small, and these characteristics have little positive impact on the results, but will affect the efficiency of the experiment, so delete this parameter.

5. Style clustering based on style features

5.1. Experimental idea
Text feature extraction is the main method of style recognition. This method classifies the style features hierarchically and extracts the features from the perspective of hierarchy. A mapping relationship between features and articles is added here. Feature extraction includes two kinds of features: one-dimensional feature and multi-dimensional feature. Parameter weight method is used to optimize feature. The results of each feature extraction are used as the input of the final k-means++ classifier. Style cracks are found through sliding windows. Style cracks are segmented by identifying style cracks.

The corpus selects news corpus, and extracts 1300 news articles from 20 people and 40 people respectively. The subjects of the corpus include news about tourism collected from People's Daily and news about sports selected from Hupu. Because there are some impurities in news collection, such as time, picture, picture introduction and photographer's name. Firstly, impurities are dealt with and the text is selected. Store the news in a CSV file and label it with the author's name. The 1300 news articles were divided into 1150 training sets and 150 test sets. The ratio of training sets to test sets was about 9:1. In order to verify the accuracy of small pages, 150 test sets are divided into 100 pages stored by length and 50 pages stored by paragraph. 50 pages are about 215 news pages.

5.2. Weight optimization of parameters
In the corpus, 100 articles of 5 authors were randomly selected to form a small sample training set for training the parameter weight method. Firstly, the training set is preprocessed, and each author's document set is put down into a txt file, and the style feature extraction stroke style feature vector is performed on each author's document set. The first step is to extract the average sentence length parameter, and then perform the word segmentation process. After the word segmentation process, the average word length, lexical features, and special punctuation marks are extracted. Next, the virtual word is extracted to perform the function word tf-idf algorithm, and the synonym is filled with the synonym vector. Calculate the emotional bias of the training set. The experimental results are shown in Table 5.1:

| Author 1 and author 2 | Word length | Average sentence length | emotion analysis | Lexical features | Special symbol | Synonym | Function word similarity | Text similarity result |
|-----------------------|-------------|-------------------------|------------------|-----------------|----------------|---------|------------------------|----------------------|
| Author 1 and author 3 | 0.27        | 0.27                    | 0.12             | 0.52            | 0.42           | 0.89    | 0.65                   | 74.6586              |
| Author 1 and author 4 | 0.18        | 0.32                    | 0.13             | 0.62            | 0.21           | 0.82    | 0.88                   | 54.7158              |
| Author 1 and author 5 | 0.45        | 0.17                    | 0.07             | 0.87            | 0.55           | 0.92    | 0.88                   | 82.9728              |
| Author 2 and author 3 | 0.31        | 0.31                    | 0.21             | 0.62            | 0.52           | 0.82    | 0.94                   | 62.3541              |
| Author 2 and author 4 | 0.52        | 0.12                    | 0.12             | 0.43            | 0.31           | 0.85    | 0.84                   | 76.2359              |
| Author 2 and author 5 | 0.64        | 0.08                    | 0.14             | 0.76            | 0.66           | 0.76    | 0.76                   | 92.8413              |
| Author 3 and author 4 | 0.12        | 0.07                    | 0.08             | 0.62            | 0.97           | 0.65    | 0.65                   | 10.3281              |
| Author 3 and author 5 | 0.54        | 0.34                    | 0.12             | 0.45            | 0.29           | 0.67    | 0.78                   | 87.3741              |
| Author 4 and author 5 | 0.21        | 0.31                    | 0.11             | 0.77            | 0.19           | 0.98    | 0.90                   | 59.3951              |

Table 5.1 Parameter Weight Method Results
The results of the parameter weight method can be seen that the author's four words are obviously different from other authors in the length of the word. After the experiment, it is found that the author's average length of the word is 3.2132, the three characters and the four words occupy a large proportion, while the other authors are in the 2-3 between 3. The length of the sentence has little effect on the result. The lengths of the second, fourth and fifth sentences are similar, and the gaps between the other sentences are also small. The reason why sentiment analysis has the least impact on the article is that the five authors are all objective news, and the emotional shift to the supervisor is small. There are no obvious rules for the parameters of lexical features, indicating that the influence weights of multiple features are different, but it is definitely a positive influence on the recognition of style cracks. The use of special punctuation is more even, and only the special punctuation of author three is similar to everyone. The effects on synonyms and virtual words are obvious, the parameters are large, and the effect on the results is also large. From the experimental results, synonyms and function words have a greater impact on the results, but other characteristics can also reflect their writing style in special circumstances. The style similarity calculation was carried out by extracting the results from the writing style, and it was found that the authors had higher similarity of one, three and five.

5.3. Style crack recognition results and analysis

5.3.1. News set experiment. The style crack identification data set randomly extracts 20 news articles of 5 authors, splits according to paragraphs, and uses paragraphs as a part. At the beginning of the experiment, sliding window technology is used. Each time you slide down a sentence, the number of sentences per window For the five experiments, the clustering results are poor in the results of K-means++ clustering, because each change is a sentence, the magnitude of the change is small, and each change is not obvious, leading to clustering. The window results are less biased. In addition to the position where the pattern breaks at the end of the paragraph, the accuracy will pick up. In the clustering process, the result of k is uncertain, which is caused by the inaccuracy of the K-means++ algorithm. Therefore, many clustering errors occur. The test results are shown in Table 5.2. The visualization is shown in Figure 5.1.

Table 5.2 Style crack recognition results using sliding window

| Author   | Accuracy/% | Recall/% | F-Measure/% |
|----------|------------|----------|-------------|
| Author one | 66.1       | 87.5     | 75.3        |
| Author two | 61.6       | 81.4    | 70.1        |
| Author three | 58.4       | 75.3     | 65.8        |
| Author four | 68.8       | 71.7    | 70.2        |
| Author five | 56.9       | 80.7     | 66.7        |

Figure 5.1 Visualization based on sliding window K-means

Finally, the slide window is discarded. The paragraph conversion symbol is used to treat each paragraph as an author. The style feature extraction is performed in units of each paragraph. The k-means clustering algorithm is performed according to the extracted style features. See Table 5.3, the visualization is shown in Figure 5.2.
Table 5.3 Style crack identification based on paragraphs

| Author    | Accuracy/% | Recall/% | F-Measure/% |
|-----------|------------|----------|-------------|
| Author 1  | 74.3       | 90.5     | 81.6        |
| Author 2  | 75.6       | 84.7     | 79.9        |
| Author 3  | 66.8       | 85.0     | 74.8        |
| Author 4  | 71.5       | 81.7     | 76.3        |
| Author 5  | 67.9       | 92.7     | 78.4        |

Figure 5.2 Paragraph-based K-means visualization

Although the sliding window is proposed to find all the style crack points as much as possible, since the change of the sentence is not obvious each time a sentence is changed, the style clustering effect is general. The effect of pattern crack recognition based on paragraphs is better than the use of sliding window, which has improved in accuracy and recall rate, and can also increase the evaluation value by 10 percentage points.

5.3.2. Fiction experiment. The original detection of 40 times after the Dream of Red Mansions has always been the main object of the discussion of the writers. The final experiment of segmentation crack recognition is to use the Dream of Red Mansion as the background, and each time as a whole, the style feature extraction, in which the virtual words are no longer used in this paper. The vocabulary table uses 22 vocabulary words. The style clustering based on style feature extraction is performed on 120 times. The result statistics are divided into the first 40 rounds, the middle 40 rounds, and the last 40 rounds. The k-value of the k-means algorithm is 2, the results are shown in Table 5.4, and the visualization is shown in Figure 5.3.

Figure 5.3 Visualization of Dream of Red Mansions Result Analysis

Table 5.4 Analysis of the results of the Dream of Red Mansions

|                      | Xueqin Cao | Er Gao | F-Measure/% |
|----------------------|------------|--------|-------------|
| First 40 chapters    | 36         | 4      | 90.0        |
| 40 chapters in the middle | 33     | 6      | 82.5        |
| latter 40 chapters   | 8          | 32     | 82.5        |
It can be seen from the results that the accuracy of the first 80 times is relatively high, and the last 40 times is relatively low. Through the experimental results of individual features, it is found that the average sentence length, sentiment analysis and virtual words have a great influence on the results. There are obvious differences in the length of the sentence after the first 80 rounds. The reason for the emotional analysis is that the first 80 times Positive, the latter 40 negative, the virtual word has the greatest impact on the experimental results, 22 virtual vocabulary has a positive impact on the experimental results.

6. Conclusion
In this paper, style crack recognition is the result of the extraction of style features, which is a combination of multi-feature fusion and unsupervised machine learning algorithms. Among them, multi-feature fusion is to extract the author's style features better, and machine learning is based on sliding windows or paragraphs. Based on the extracted features, the clustering algorithm is used to cluster the style features to find the style cracks. position. Experiments on the news corpus and the novel corpus are carried out respectively. It is concluded that the segment-based crack recognition is 10% higher than the experimental results based on the sliding window. Therefore, the experimental idea based on the sliding window still needs further improvement.

Reference
[1] Franco-Salvador M, Rosso P, Montes-y-Gómez M. A systematic study of knowledge graph analysis for cross-language plagiarism detection[J]. Information Processing & Management, 2016, 52(4): 550-570.
[2] Juola P, Mikros G K, Vinsick S. Correlations and Potential Cross-Linguistic Indicators of Writing Style[J]. Journal of Quantitative Linguistics, 2019, 26(2): 146-171.
[3] Tschuggnall M, Gerrier T, Specht G. StyleExplorer: A Toolkit for Textual Writing Style Visualization[C]//European Conference on Information Retrieval. Springer, Cham, 2019: 220-224.
[4] Fan Xia, Zhi Zhang, Study of text emotion analysis based on deep learning[J]. IEEE. 2018 13th: 2158-2297p.
[5] Sailunaz, Kashfia, Emotion and Sentiment Analysis from Twitter Text. PRISM, 2018: 107533.
[6] Jin Q, Li C, Chen S, et al. Speech emotion recognition with acoustic and lexical features[C]//2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015: 4749-4753.