


## Research Article

# Practical Skills of Business English Correspondence Writing Based on Data Mining Algorithm

Danqing Liu <sup>1,2</sup> and Hadina Habil<sup>1</sup>

<sup>1</sup>Language Academy, Universiti Teknologi Malaysia, Johor Bahru 81310, Johor, Malaysia

<sup>2</sup>School of English Language and Literature, Xi'an Fanyi University, Xi'an 710105, Shaanxi, China

Correspondence should be addressed to Danqing Liu; liudanqing@xafy.edu.cn

Received 14 February 2022; Revised 21 March 2022; Accepted 1 June 2022; Published 4 July 2022

Academic Editor: Ahmed Farouk

Copyright © 2022 Danqing Liu and Hadina Habil. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

English correspondence writing has become a necessary skill for every scientific researcher and high-tech talents. An English correspondence writing auxiliary writing system can help nonnative English speakers make up for the lack of professional expression. The key factor of business English correspondence writing system is the construction of knowledge base. To improve the business English correspondence writing knowledge base, we need to mine frequent patterns of sentences in each category. The purpose of this topic is to improve and supplement the knowledge base for the business English correspondence writing system and propose frequent pattern mining for sentences in each category, so as to improve the writing knowledge base for the business English correspondence writing system. Firstly, we crawl a large number of business English letters and telegrams from the Internet, extract the relevant summary information, then store it, and preliminarily construct a corpus based on sentences. Then, we do some research on the structure of business English correspondence abstracts, mark the sentences in the corpus and count the relevant information, and have a certain understanding of their writing methods. Finally, we mine frequent patterns for sentences in each category, so as to improve the knowledge base of summary writing for the business English correspondence writing system. In the experiment, we use the classical FP growth algorithm as the mining method. The experiment shows that the frequent patterns between 3 and 6 words have been mined to a certain extent. By gradually improving the mining strategy, the quality of mining results has been improved and the writing effect of business English correspondence of scientific researchers has been improved.

## 1. Introduction

With the growth of productivity and economy, more and more people begin to devote themselves to various scientific research in modern society, followed by the increasing number of papers, journal meetings, and papers published [1]. According to the research, the process of writing scientific research results into papers is more painful than theoretical derivation and experimental analysis. Similarly, students studying for a degree in China also have the need to publish papers. For nonnative English speakers, it is more difficult to write English papers. They are not only faced with the problem of how to express the experimental results clearly but also faced with problems such as English expression. To solve the above problems, this paper attempts to

write with the latest frequent mining pattern [2]. The goal of natural language generation task is to enable the machine to express and create on the basis of understanding text, numbers, structured data, charts, and other data. Therefore, we should focus on how to effectively serve users with automatic text generation technology in a systematic form. Under the above background, it is a pioneering attempt to apply the relevant technologies of English correspondence writing auxiliary writing system to the applied thesis writing system, which can effectively improve the quality and efficiency of thesis writing. This paper studies the automatic text generation system based on the nature-oriented English correspondence writing auxiliary writing system [3]. Based on a large amount of paper data in the field of scientific and technological papers, with the help of deep learning model,

the system uses document summary algorithm, graph structure based text generation algorithm, word vector algorithm, and text editing algorithm, and the user's writing experience can be improved from two aspects: automatically generating paper content and providing word suggestions. At the same time, in view of the continuous generation of new papers and the update of natural language generation model, the system provides a functional module for sustainable data capture and supports the iterative update of generation model. The system also provides a background management module for managing paper data and generating model, which provides strong support for the maintenance of the system. Based on the basic data of scientific and technological papers, the system designs a series of text generation models for paper writing, which can help users quickly and conveniently generate paper titles, abstracts, and touch-up sentences and reduce users' writing burden [4, 5]. The scientific contribution is the construction of the corpus. Since there is no ready-made corpus for writing abstracts of scientific papers, we need to build a corpus by ourselves. The main work includes web crawling, abstract information extraction of papers, corpus preprocessing, corpus research, and corpus labeling and storage. This work not only lays the groundwork for the follow-up work but also lays the foundation for other related works in the future. The system provides a user-friendly interaction mode, so that users can concentrate more on writing and provide work efficiency, as shown in Figure 1.

## 2. Literature Review

In the process of human-computer interaction, auxiliary writing system is a very broad concept. Kirubha and others found that many famous software can be classified into the category of auxiliary writing system, such as "MS Word" of Microsoft Corporation of the United States, "WPS" of China Kingsoft Corporation, and writing assistance software [6]. The original research orientation of natural language education is based on rule and conditional probabilities. Markov models are more representative. The Markov model regards the language generation process as different sequence states, calculates the transition direction of the next generation state based on the current word, and predicts the next word in the text in turn. However, with the in-depth research and theoretical development, especially in the 21st century, with the significant improvement of computer ability, learning natural language based on deep learning technology has gradually matured. In 2013, Wang and others proposed the word2vec algorithm and a bag-of-words model was constructed using neural network, and the word vector representation of the target language was calculated according to the context word distribution of the target language in a large-scale corpus [7]. Word2vec algorithm completes the transformation from text representation to static number vector. The algorithm maps words with similar meanings to the same region of the vector space and helps the computer understand the meaning of the text itself. When neural network is used to solve natural language problems, cyclic neural network transmits the information of one of the

earliest neurons in the network to the next neural cell. Because the text information in natural language is usually orderly, recurrent neural network is naturally suitable for the task of text sequence. Using recurrent neural network to create image subtitles and machine translation has produced satisfactory results [8]. In Rajesh's work, the simulation of text production style is realized by using long-term and short-term storage networks, and a dialogue system is constructed. Based on recurrent neural network [9], in the work of Duan and Gao, the encoder-decoder framework was first proposed when solving machine translation problems. At the encoder end, recursive neural network is used to encode the input text sequence into fixed-length meaning vector. At the decoding end, other networks are used to generate text sequence according to the output of meaning vector [10]. The powerful performance of attention mechanism has been paid attention to. In 2014, Nwet and Darren used the attention mechanism to solve the image classification problem in order to reduce the complexity of the task, allowing the model to process the pixels of the attention part [11]. In addition, the Google team proposed an automatic attention mechanism to replace the traditional neural network layout for the end of the model and a unit converter composed of multiple multihead and multifunctional mechanisms. Although the traditional recurrent neural network is suitable for the development of language sequence, it cannot calculate in parallel and capture the global structure information. Through the self-attention mechanism, the attention distribution of each word in the sentence and all words in the sentence is calculated, which can solve the problem of dependence at a distance. As it can calculate in parallel, the efficiency of training has also improved considerably. In addition, on the basis of the structure of the transformation module above, in the follow-up, you only need to use a small amount of downstream work data to get rewarding results on different tasks. Huang and others proposed paper model for biomedical and computer documents. Firstly, the model comprehends a large number of documents on the ground and builds a specific and complete graph of basic knowledge. Through the connection prediction algorithm, other nodes that can be associated with the graphics nodes are found [12]. In terms of the current popular pretraining language model, Yu used hundreds of millions of words of biomedical text mixed general language text to train the biomedical corpus [13]. Molchanova and others used 1.14 million biomedical and computer science and technology documents for the training of sciences, which is best suited to the management of scientific and technological documents in China; some researchers have formed the Chinese pretraining model of SciBERT CN using the corpus of Chinese scientific and technological documents [14]. Hasheminejad and Khorrami found that, at present, the researchers tried to produce documents using the pretraining model. The University of California, Berkeley, conducted an experiment to generate documents using the gpt-3i5j pretraining model. In the experiment, Researchers input the title and introduction of the experimental paper and generate the model and other paper contents [15]. The experimental results of Xin show that

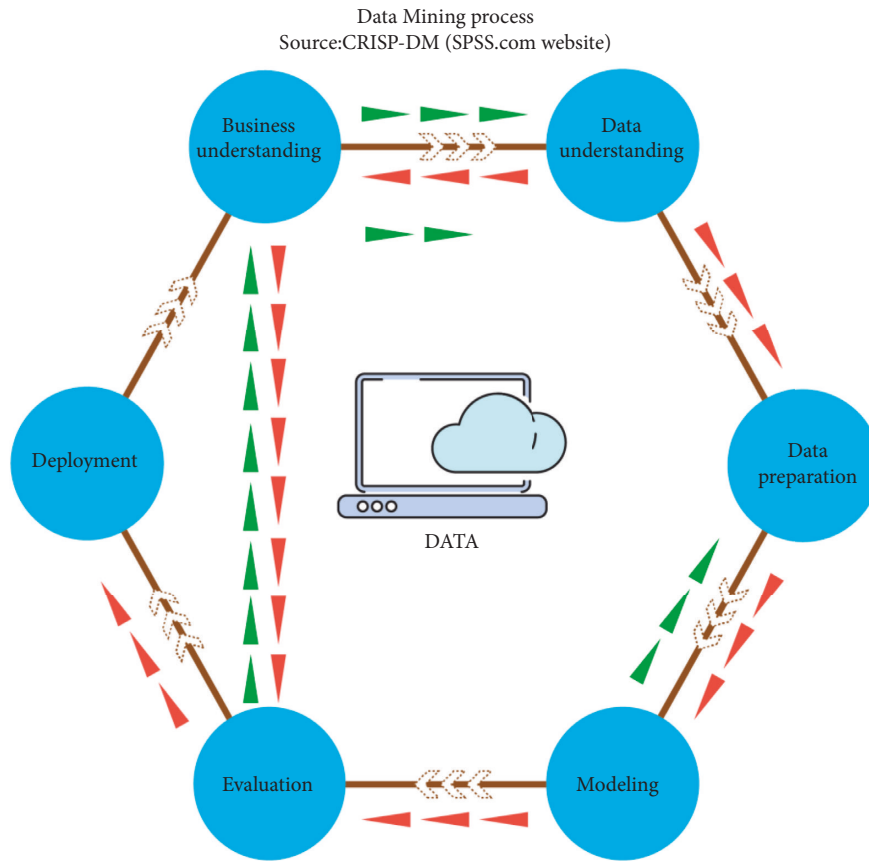


FIGURE 1: System flow.

even if the sentences generated by the model lack logic, they can use a large number of professional words to create sentences that confuse true and false [16]. Similarly, educational institutions allow graduates to write the same paper content with gpt-3. Finally, gpt-3 shows a level of writing similar to humans and reflects incredible creative ability in writing technical documents. To sum up, the use of natural language generation technology to produce documents is highly feasible and worthy of thorough research, but research in this direction, both at home and abroad, is still somewhat inadequate. Compared with other translation software, an English business-assisted writing system can help make up for the deficiencies in these aspects, and the construction of a knowledge base has become a key factor supporting an English-assisted writing system [17].

### 3. Method

At present, the Internet technology is relatively developed and the network resources are very rich. The network has become an important source for obtaining corpus [18, 19]. Moreover, with the increasing standardization of HTML files, it is more convenient, accurate, and fast to obtain information from HTML files than from PDF, CAJ, DOC, and other formats. Therefore, obtaining corpus from Internet web pages is the most appropriate choice. The work of this paper to obtain the corpus is to use the special crawler tool “NetCrawler” to obtain the corpus web page from the famous foreign academic

retrieval website “<http://www.sciencedirect.com/>.” The advantage of crawling papers from the same academic website is that most URIs (uniform resource identifiers) have the same or similar field format, which is convenient for batch crawling. In addition, most HTML files have the same or similar organizational structure format, which is convenient for unified information extraction [20]. The corpus of this subject includes 15 university subjects such as computer science, life science, medicine, and chemistry. Each university subject also includes several specific research fields. Generally speaking, the number is relatively sufficient. The crawled corpus is a file in HTML format and needs further processing. Through the investigation of a large number of corpus files, it is found that the HTML tag organization of this website is relatively standardized. The information related to paper abstracts (including) is usually between the tags “<HR id =” a (a) bstract> “and” <HR>,” while the specific content of paper abstracts is usually between the tags “<p id = “\*\* ”>” and “</P>,” which brings great convenience to the batch extraction of summary information. The extraction method based on regular expression can be used for batch extraction easily, and the extraction accuracy is very high. Only a few HTML files cannot extract summary information because of the “uniqueness” of tag organization [21].

The manually labeled corpus of this subject comes from the abstract part of 400 papers, a total of 4555 sentences. After English word segmentation, there were a total of 127710 words (including punctuation, person name, special

symbols, misclassification, etc.), with an average of 28 words per sentence unit (including punctuation, person name, special symbols, misclassification, etc.), and a vocabulary size of 10345 (including punctuation, person name, special symbols, misclassification, etc.). The in-depth statistics are shown in Table 1.

In the words and thesaurus counted above, there are some redundant information and nonstandard information, such as word inconsistency caused by case of letters, word inconsistency caused by changes in singular and plural numbers, word inconsistency caused by changes in word form, and widely used punctuation marks. Therefore, further text standardization is needed [22]. The commonly needed text normalization processing steps include lowercase, stem extraction, and word form merging [23]. The related technologies listed above are introduced. After several steps of text normalization, such as lowercase, stem extraction, word form merging, and elimination of useless punctuation, the size information of the word list is shown in Table 2, and the word frequency distribution is shown in Figure 2.

Figure 2 lists the word frequency distribution of the words with the word frequency in the top 50. It can be seen that the overall word frequency distribution on the marked dataset generally obeys Zipf's law. Zipf's law states that, in a given corpus expressed in natural language, the frequency of a word is inversely proportional to its ranking from high to low in the corpus. We have the following formula:

$$R \times F \approx C. \quad (1)$$

Here,  $R$  is ranking of the frequency of a word in the corpus (from high to low),  $F$  is frequency of this word, and  $C$  is a constant.

Currently, research into the level of phrases is of great importance in the fields of natural language processing, information retrieval, automatic translation, and so on. The purpose is to divide the abstract sentences of English scientific papers in the corpus constructed by ourselves in the previous work into four categories: "research background," "subject content," "experimental method," and "results and conclusions," so that these classified abstract sentences of English scientific papers can be used for the next research work, that is, the frequent pattern mining of various categories of abstract sentences and the construction of knowledge base of English scientific paper auxiliary writing system. The sentence categories in the abstracts of 400 English scientific papers are manually labeled, with a total of 4555 abstract sentences, which are divided into two parts. 2400 labeled corpora in the first part are used as the training set of the classifier and 2155 labeled corpora in the second part are used as the test set. The specific statistical information of each category is shown in Table 3.

Selection and extraction of characteristics is one of the necessary steps in the classification. In the task of classifying the text, "word bag" and vector space model (VSM) are the most commonly used methods of representation of text [24]. The document frequency of a word is the number of documents in which the word appears in the corpus. This feature selection method assumes that rare words have little effect

TABLE 1: Information statistics of each category.

	Category 1	Category 2	Category 3	Category 4
Number of sentences	1055	780	1715	996
Proportion of sentences	0.24	0.17	0.38	0.2
Total words	29036	23126	47163	28395
Thesaurus size	4545	3965	6415	4675

and impact on classification. Therefore, a threshold is set in advance to filter out the words whose document frequency is lower than this threshold, so as to reduce the feature dimension. In addition, if a filtered word Yiqiao is a noisy word, this method will improve the classification effect. In general, document frequency is the simplest and least computational feature selection method. The information gain of a word is a measure of the information increment brought by the word to the classification, which reflects the ability of the word to distinguish this category from other categories [25]. It performs category prediction by calculating the number of instances in which a word is included and the number of instances in which it is not included. The calculation method is as follows:

$$IG(t) = - \sum_{i=1}^m p(c_i) + p(t) \sum_{i=1}^m p(c_i|t) \log p(c_i|t) + p(t) \sum_{i=1}^m p(c_i|t) \log p(c_i|t), \quad (2)$$

where  $M$  is number of categories,  $p(c_i)$  is probability of occurrence of class instances in the corpus,  $p(t)$  is the probability that instances containing word  $t$  appear in the corpus,  $p(c_i|t)$  is conditional probability that the instance containing word  $t$  belongs to category  $I$ ,  $p(t)$  is the probability that instances without word  $t$  appear in the corpus, and  $p(c_i|t)$  is conditional probability that an instance without word  $t$  belongs to category  $I$ .

The information gain determines the category discrimination ability of the word to a certain extent. Using this feature selection method to select the word whose information gain is higher than a preset threshold as the feature can effectively reduce the feature dimension and improve the quality of the feature set. In a broad sense, mutual information refers to the correlation between two event sets. In the feature selection of text classification, it reflects the correlation between a word and a category. The calculation method is shown in the following formula:

$$I(t, c) \approx \log \frac{A \times N}{(A + C) \times (A + B)}, \quad (3)$$

where  $A$  is number of cases in the corpus containing word  $t$  and belonging to category  $C$ ;  $N$  is the total number of instances in the corpus;  $B$  is the number of cases in the corpus containing word  $t$  but not belonging to category  $C$ ;  $C$  is the number of instances in the corpus which do not contain word  $t$  but belong to category  $C$ .

TABLE 2: Thesaurus statistics after text normalization.

Statistical information	Population	Category I	Category II	Category III	Category IV
Number of words	114430	26016	20701	41955	25758
Thesaurus size	6310	2902	2695	4128	3128

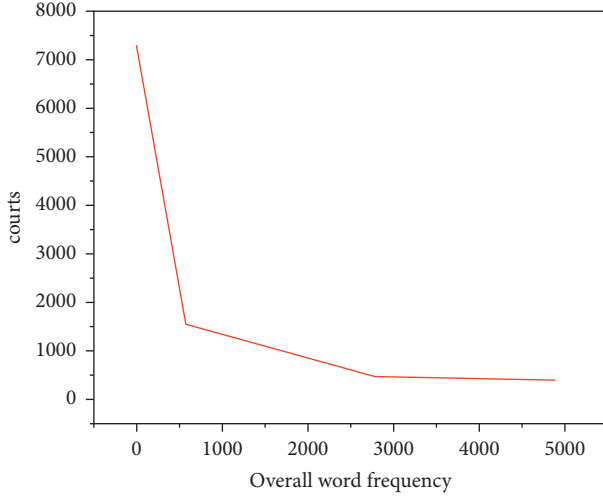


FIGURE 2: Overall word frequency distribution.

TABLE 3: Sample statistics of training set and test set.

	Training set	Test set
Total number of labeled samples	2405	2145
Number of samples in category I	535	520
Category I sample proportion	0.22	0.24
Number of samples in category II	415	365
Category II sample proportion	0.16	0.16
Number of samples in category III	920	795
Category III sample proportion	0.32	0.36
Number of samples in category IV	530	466
Category IV sample proportion	0.22	0.21

In the global feature selection of multiclass problems, the mutual information of a word is usually calculated by selecting the weighted average of the mutual information of the word in each class (formula (4)) or the maximum mutual information (formula (5)).

$$I_{\text{avg}}(t) = m \sum_{i=1}^m I(t, c_i), \quad (4)$$

$$I_{\text{max}}(t) = \max_{i=1}^m \{I(t, c_i)\}, \quad (5)$$

where  $M$  is number of categories,  $p(c_i)$  is the proportion of instances belonging to category  $I$  in the corpus, and  $I(t, c_i)$  is mutual information of word  $t$  in corpus for category  $I$ .

The size of mutual information reflects the relevance of a word and a category to a certain extent. However, its defect is that its value is greatly affected by the edge probability, which makes the mutual information value of a rare word greater than that of an ordinary word when the conditional probability of word  $t$  belonging to category  $C$  is the same. It is also an effective feature selection method to select words

whose mutual information value is higher than a preset threshold as features [26].

$X^2$  statistic is used to measure the lack of independence between a word and a category in text classification. The greater the value, the smaller the independence between the word and the category, that is, the greater the correlation. The calculation method of  $X^2$  statistics of a word for a category is shown in the following formula:

$$x^2(t, c) = \frac{N \times (AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)}, \quad (6)$$

where  $N$  is the total number of instances in the corpus;  $A$  is the number of instances in the corpus which contain word  $t$  and belong to category  $C$ ;  $B$  is the number of instances in the corpus which contain word  $t$  but do not belong to category  $C$ ;  $C$  is the number of instances in the corpus which do not contain word  $t$  but belong to category  $C$ ; and  $D$  is the number of instances in the corpus which do not contain word  $t$  and do not belong to category  $C$ .

In the global feature selection of multiclass problems, the  $X^2$  statistic of a word is usually calculated by selecting the weighted average (formula (7)) or maximum (formula (8)) of the mutual information of the word in each class.

$$x^2 \text{ avg}(t) = \sum_{i=1}^m P(c_i) x^2(t, c_i), \quad (7)$$

$$x^2 \text{ max}(t) = \max_{i=1}^m \{x^2(t, c_i)\}. \quad (8)$$

Like mutual information,  $X^2$  statistic reflects the correlation between words and categories to a certain extent. The difference is that  $X^2$  statistic is a normalized value. Therefore, for words in the same category, the correlation between words and categories can be compared according to the value, but, for low-frequency words, the confidence of this correlation is not strong enough. It is also an effective feature selection method to select words whose  $X^2$  statistics are higher than a preset threshold as features [27].

Supervised statistical learning method is widely used in classification tasks. Its basic idea is to use labeled data to train the classifier and then use the classifier to predict the target samples of unknown categories. There are many models and methods that can be used for text classification, such as naive Bayesian model, k-nearest neighbor model, maximum entropy model, artificial neural network model, and support vector machine model. Among them, support vector machine model is recognized as the most suitable model for text classification task. At present, nonlinear support vector machines have been widely used in various classification tasks, especially in text classification tasks [28]. The nonlinear support vector machine takes the labeled samples as the input and solves the class II classification

problem by constructing the target classification decision function, as shown in

$$\min_{\partial} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \partial_i \partial_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N \partial_i, \quad (9)$$

$$s.t. \sum_{i=1}^N \partial_i y_i = 0, \quad (10)$$

$$0 \leq \partial_i \leq C, \quad i = 1, 2, \dots, N, \quad (11)$$

$$K(x, \Xi) = \exp(-\sigma \cdot \|x - \Xi\|^2). \quad (12)$$

SVM uses kernel function to map the feature space to a higher dimensional space. Therefore, when using SVM model, it is generally not necessary to reduce the dimension to control the complexity of the model. But this does not mean that thesaurus can be used as the full feature in the classification of abstract sentences of scientific papers in this paper, because, as described in the feature selection section above, if some noise words are not filtered, the classification effect will be affected. The data used in the classification experiment of abstract sentences here are from the corpus of abstract sentences of English scientific papers constructed by ourselves in the previous work, including 4555 annotated abstract sentences and a large number of unmarked abstract sentences. The 4555 labeled data samples are divided into two parts, of which 2400 data samples are used as the training set and 2155 are used as the test set. Firstly, we use the supervised learning model support vector machine to do a preliminary control experiment in order to have a preliminary understanding of the problems and challenges. The experimental method adopts the word bag method used in the general text classification experiment, takes the words after lowercase, word stemming, and word form merging as the feature set, selects the features with the word frequency threshold and  $\chi^2$  statistic threshold, respectively, takes the value of 0-1 as the feature value, classifies with the SVM model, trains with all 4550 labeled data samples, and makes a 50-fold cross validation. The highest accuracy results obtained are shown in Table 4.

The frequent pattern mining in this paper is different from the general frequent pattern mining or frequent itemset mining. Common frequent pattern mining tasks, such as the classic frequent pattern mining and association rule mining in supermarket shopping, generally only need to investigate the cooccurrence relationship of different items and the association support. In the frequent pattern mining of various abstract sentences of scientific papers in this chapter, in addition to mining the cooccurrence relationship between words in various sentences, we should also consider the collocation order of cooccurrence words and whether they are representative of categories. The so-called collocation sequence refers to, for example, the fact that, through frequent pattern mining, it is found that words A, B, and C appear at the same time with high support, but the total arrangement order of these three words is the six cases "ABC," "ACB," "BAC," "BCA," "CAB," "CBA." What kind

of situation is more in line with the actual expression is a problem that needs to be investigated. The so-called category representativeness means that the mined frequent patterns are easier to distinguish from other categories. The smaller the intersection between the frequent pattern sets of various categories, the better. For example, in the "background meaning" sentence of English correspondence writing, there are more symbolic expressions such as "the background o...," "recent years...," and "have been proposed"; in the "subject content" sentences of English correspondence writing, there are more symbolic expressions such as "in this paper/article," "this paper introduction/descriptions," and "here we present/propose." In the "experimental method" sentence of English correspondence writing, there are more symbolic expressions, such as "we use...," "the data/datasets," and "step one/two/three." In the sentence of "results and conclusions" in English correspondence writing, there are more symbolic expressions, such as "the results of for," "the result(s) indicate that...," and "we conclude/demonstrate that..." All of the previously mentioned statements have great category representativeness. On the other hand, according to Zipf's law and the previous statistical information on the distribution of words in the corpus, the frequency of function words (including stop words, some prepositions, articles, etc.) is high both in the types of abstract sentences of English scientific papers and in the overall expression of English language. The frequency of cooccurrence between them is much higher than that between other meaningful words, and the regularity of cooccurrence is poor. Therefore, the frequent pattern mining of these function words has little significance. Since the number of various sentences involved is not very large, the establishment of inverted index can be completed in memory. The memory based inverted index establishment algorithm is shown in Table 5 [29]. For the full arrangement of words in a frequent pattern, you can count the number of times of each full arrangement by returning to the sentence containing the frequent pattern and finally decide the arrangement with the largest number of votes by voting, which is the prototype of the frequent pattern. In addition, by calculating the expectation and variance of the relative distance between words in the sample, it can also be used to determine the prototype of frequent patterns.

At present, there is no unified evaluation index for the result quality of frequent pattern mining. According to different specific problems, evaluation indicators that can be roughly applicable to specific problems can be formulated. For the frequent pattern mining of words in the sentence set, on the one hand, the support of the discovered frequent pattern needs to be investigated, because the support is about equal to the number of occurrences of the frequent pattern in the sentence set. The greater the support, the greater the probability of cooccurrence of each word item in the frequent pattern. On the other hand, we should also examine the stability of its structure. The so-called stability refers to whether the frequency distribution of the full arrangement of words in a frequent pattern is stable according to the order of occurrence; that is, if each full arrangement appears more evenly in the sentence set, it becomes more unstable. On the contrary, if it appears in

TABLE 4: Comparison of accuracy under different dimensions and different feature selection methods.

	Threshold = 0.05 (%)	Threshold = 0.03 (%)	Threshold = 0.02 (%)	Threshold = 0.01 (%)
Word frequency	48	53	57	59
$X^2$ statistic	53	58	59	62

TABLE 5: Memory based inverted index establishment algorithm.

Input	Sentence set
	Inverted index based on sentence set
	(1) Initially traverse the sentence set. For each word, count the number of sentences containing the word $f_w$ .
Output	(2) Create an array of length $\sum f_w$ , and for each word $W$ , generate a pointer $p_w$ to the beginning of its record table block.
	(3) Traverse the sentence set again, for each word $w$ in each sentence $d$ , add the sequence number of sentence $d$ to $p_w$ , and move $p_w$ backward.

one arrangement, then this arrangement can be considered a stable frequent pattern.

We directly use the *FP* growth method to mine frequent patterns of summary sentences of various classes, and the experimental results show that the effect is poor. Most of the frequent patterns with high support are composed of function words (including stop words, prepositions, numerals, etc.), as shown in Table 6. Even if some words with actual meaning appear in some frequent patterns with low confidence, the frequent patterns are usually composed of only a small number of notional words and a large number of functional words, as shown in Table 7.

Among the frequent patterns mined, the three frequent itemsets have increased significantly. Because the sentence representation is in the form of bigram, it can be said that the frequent patterns between 3 and 6 words have been mined to some extent, which is better than before. From the perspective of English assisted writing, the more words the frequent pattern contains, the more it can reflect the value of assisted writing. In general, using bigram sentence representation method for frequent pattern mining, the mining effect and results have been greatly improved, but there is still a certain gap with the expectation, which is related to the limited size of sentence set and the limitation of mining methods.

#### 4. Experiment and Discussion

As can be seen from the experimental results in Figure 3, with the increase of feature dimension, the classification accuracy increases, but the increase tends to be gentle. In addition, the effect of feature selection using  $X^2$  statistics is better than that using word frequency, which proves that words with strong classification ability play a better role in classification. However, from the perspective of accuracy value, even if cross validation is adopted on the training set itself, the results are still not ideal [30]. This is directly related to the feature sparsity of sentences. Then, the following groups of experiments are carried out using SVM model to verify the impact of feature selection on classification effect. The results are shown in Table 8.

In this experiment, the kernel function of support vector machine adopts linear kernel function, and the penalty term factor  $C$  is set to 1. The parameters  $P$  and  $f$  shown in Table 4

TABLE 6: Examples of frequent patterns with support greater than 200 in category 1.

Support	Frequent mode
530	The of
400	Of and
365	The and
305	Of the and
365	Of the
205	The of
355	Of to
275	Of the to
200	The and to

TABLE 7: Examples of frequent patterns containing the notional word "information" in category 1.

Support	Frequent mode
42	The is information
40	Of is information
30	To is information
35	The of is information
36	And in information
32	Of in information
25	To in information

represent different feature selection methods.  $P$  mainly selects different methods from the perspective of text features, and its significance is shown in Table 9.  $F$  selects different methods from the perspective of main feature representation methods, and its significance is shown in Table 10.

From the above results, it can be seen that the representation of text features has a great impact on the accuracy of classification, while different feature selection methods have little impact on the accuracy of classification, and there is almost no significant difference. Among them, the classification accuracy of retaining stop words ( $P=0, 1, 2, 3$ ) is significantly higher than that of removing stop words ( $P=4, 5, 6, 7$ ). On this basis, the classification accuracy of bigram text feature ( $P=1, 3$ ) is higher than that of unigram ( $P=0, 2$ ). Without word stemming and word form merging ( $P=0, 1, 4, 5$ ), the effect is not more obvious than that of word stemming and word form merging ( $P=2, 3, 6, 7$ ). To some extent, this is due to the common writing methods and expression characteristics of various parts of the abstract

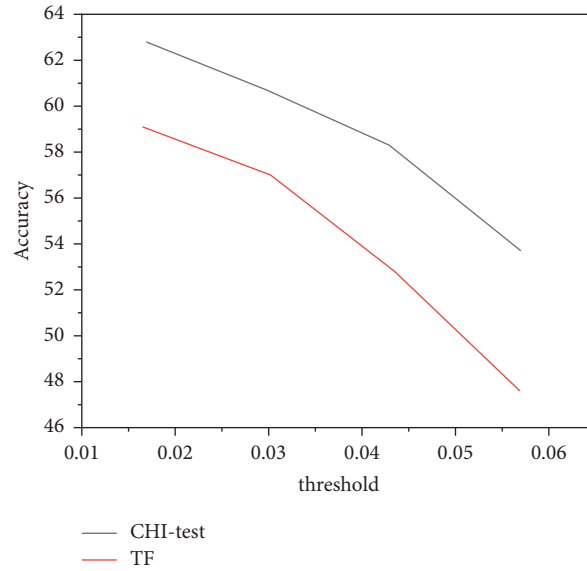


FIGURE 3: Comparison of preliminary experimental results.

TABLE 8: Influence of feature selection on classification accuracy.

<i>P/F</i>	0	1	2	3
0	62.3	62.4	52.4	60
1	66	65.4	65.4	64.3
2	61.2	61.6	61.4	59.5
3	65	64.8	64.8	65
4	56.5	55	55	55.8
5	57.3	58	58	58.2
6	55.2	53.8	53.8	54.2
7	57.3	55.8	55.8	57.9

TABLE 9: Text feature pattern represented by parameter *P*.

<i>P</i>	Text feature mode
0	Retain stop words without word stemming and morphological merging, unigram
1	Retain stop words without word stemming and morphological merging, bigram
2	Retain stop words, carry out word stemming and word form merging, unigram
3	Retain stop words, carry out word stemming and word form merging, bigram
4	Remove stop words without word stemming and morphological merging, unigram
5	Remove stop words without word stemming and morphological merging, bigram
6	Remove stop words, carry out word stemming and word form merging, unigram
7	Remove stop words, perform word stemming and word form merging, bigram

TABLE 10: Feature representation represented by parameter *F*.

<i>F</i>	Feature representation method
0	Binary
1	Word count
2	TF
3	TF-IDF

sentences of English scientific papers, such as “in this paper.” The emergence of collocations such as “the aim of” indicates that the abstract sentence is likely to describe the subject content of the paper. If the abstract sentence is expressed in the form of words and some stop words are filtered out, the text features obtained may not be fully representative of the

category; in addition, the appearance of the symbol “%” is likely to indicate that this sentence is describing the result of English correspondence writing. In addition, using bigram as a feature doubles the amount of text information compared to using unigram as a feature, and retaining stop words also increases the amount of text information in a disguised manner, so the classification effect is improved [31].

## 5. Conclusion

The main work is to mine frequent patterns in all kinds of English correspondence writing abstracts. Firstly, this paper briefly introduces the frequent pattern mining task of



abstract sentences and expounds its purpose, significance, characteristics, and the difference from ordinary frequent pattern mining tasks. Then, the paper briefly introduces the relevant knowledge of frequent pattern mining. When training support vector machine, selecting a larger penalty factor, that is, when paying more attention to noise points or outliers, we can obtain higher accuracy in the test set, but the negative impact is lower recall rate and  $F$  value, and the generalization is affected to a certain extent. The final accuracy reached more than 70%. Finally,  $FP$  growth algorithm is used to mine frequent patterns of summary sentence sets of various categories, and the mining results are studied and analyzed. In view of the shortcomings, the mining strategy is gradually adjusted to improve the quality of mining results. However, the result of mining is limited by the size of sentence set and mining algorithm, and there is still a certain gap with the expected goal.  $FP$  growth algorithm is used to mine frequent patterns in the collection of summary sentences of various categories. By using stop word filtering, bigram representation of sentences, and quality evaluation method combining support and stability, the mining effect is gradually improved and gradually becomes close to the expected goal. A large number of English correspondence writing corpora are obtained, the abstract sentences are divided into four categories: “research background,” “subject content,” “experimental method,” and “results and conclusions,” and some frequent patterns in each category are excavated, which supplements and improves the knowledge base of English assisted writing system in abstract writing of scientific papers.

## Data Availability

No data were used to support this study.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

## Acknowledgments

This work was supported by first-class undergraduate major construction project in Shaanxi Province: Business English Major in Xi'an Fanyi University. It was also supported by scientific research project at Xi'an Fanyi University with Shaanxi Danjian Experimental Equipment Co., Ltd.: “A Display Tool for English Vocabulary Teaching” Utility Model Patent Implementation Permit (21XYH232).

## References

- [1] J. Zhang, S. O. Williams, and H. Wang, “Intelligent computing system based on pattern recognition and data mining algorithms,” *Sustainable Computing: Informatics and Systems*, vol. 20, no. DEC, pp. 192–202, 2018.
- [2] H. Esmaily, M. Tayefi, M. Ghayour-Mobarhan, and A. Amirabadizadeh, “Comparing three data mining algorithms for identifying the associated risk factors of type 2 diabetes,” *Iranian Biomedical Journal*, vol. 22, no. 5, pp. 303–311, 2018.
- [3] A. Goel and S. Srivastava, “Study of data mining algorithms in the context of performance enhancement of classification,” *International Journal of Computer Application*, vol. 134, no. 9, pp. 1–5, 2016.
- [4] T. Hong, W. Zhao, R. Liu, and M. Kadoch, “Space-air-ground IoT network and related key technologies,” *IEEE Wireless Communications*, vol. 27, no. 2, pp. 96–104, 2020.
- [5] A. Onan, “Sentiment Analysis on Product Reviews Based on Weighted Word Embeddings and Deep Neural Networks,” *Concurrency and Computation: Practice and Experience*, vol. 1, 2020.
- [6] V. Kirubha, S. M. Priya, and S. M. Priya, “Survey on data mining algorithms in disease prediction,” *International Journal of Computer Trends and Technology*, vol. 38, no. 3, pp. 124–128, 2016.
- [7] Q. Wang, J. Huang, Y. Feng, and J. Fei, “Efficient data mining algorithms for screening potential proteins of drug target,” *Mathematical Problems in Engineering*, vol. 2017, no. -3-2, pp. 1–10, Article ID 9852063, 2017.
- [8] S. Celik and O. Yilmaz, “Comparison of different data mining algorithms for prediction of body weight from several morphological measurements in dogs,” *Journal of Animal and Plant Sciences*, vol. 27, no. 1, pp. 57–64, 2017.
- [9] P. Rajesh, “A comparative study of data mining algorithms for decision tree approaches using weka tool,” *Advances in Natural and Applied Sciences*, vol. 11, no. 9, pp. 230–241, 2017.
- [10] J. Duan and R. Gao, “Research on English movie resource information mining based on dynamic data stream classification,” *Security and Communication Networks*, vol. 2021, no. 4, pp. 1–10, 2021.
- [11] K. T. Nwet and S. Darren, “Machine learning algorithms for Myanmar news classification,” *International Journal on Natural Language Computing*, vol. 8, no. 4, pp. 17–24, 2019.
- [12] R. Huang, A. Sato, T. Tamura, J. Ma, and N. Y. Yen, “Towards next-generation business intelligence: an integrated framework based on dme and kid fusion engine,” *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 11509–11530, 2017.
- [13] H. Yu, “Online teaching quality evaluation based on emotion recognition and improved aprioritid algorithm,” *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 7037–7047, 2021.
- [14] E. I. Molchanova, E. N. Korzhova, T. V. Stepanova, and V. V. Kuz'min, “Analysis of the samples with an unknown matrix using data mining algorithms,” *Inorganic Materials*, vol. 53, no. 14, pp. 1454–1457, 2017.
- [15] S. M. H. Hasheminejad and M. Khorrami, “Clustering of bank customers based on lifetime value using data mining methods,” *Intelligent Decision Technologies*, vol. 14, no. 4, pp. 507–515, 2021.
- [16] Y. Xin, “Analyzing the quality of business English teaching using multimedia data mining,” *Mobile Information Systems*, vol. 2021, no. 12, pp. 1–8, Article ID 9912460, 2021.
- [17] A. Onan, “Mining opinions from instructor evaluation reviews: a deep learning approach,” *Computer Applications in Engineering Education*, vol. 28, no. 1, pp. 117–138, Article ID 9912460, 2020.
- [18] L. Qiao, Y. Li, D. Chen, S. Serikawa, M. Guizani, and Z. Lv, “A survey on 5G/6G, AI, and Robotics,” *Computers & Electrical Engineering*, vol. 95, no. 2021, Article ID 107372, 2021.
- [19] Y. Li, Y. Zuo, H. Song, and Z. Lv, “Deep learning in security of internet of things,” *IEEE Internet of Things Journal*, vol. 99, p. 1, 2021.

- [20] M. M. Yatskou, V. V. Skakun, and V. V. Apanasovich, "Method for processing fluorescence decay kinetic curves using data mining algorithms," *Journal of Applied Spectroscopy*, vol. 87, no. 2, pp. 333–344, 2020.
- [21] P. Pandey and I. Singh, "Improving accuracy using different data mining algorithms," *International Journal of Computer Application*, vol. 150, no. 10, pp. 10–13, 2016.
- [22] Ö. Gülsüm Uzut and S. Buyrukoglu, "Veri Madenciliği Algoritmaları İle Gayrimenkul Fiyatlarının Tahmini," *Euroasia Journal of Mathematics, Engineering, Natural and Medical Sciences*, vol. 8, no. 9, pp. 77–84, 2020.
- [23] Y. Liu and G. Fu, "Emotion recognition by deeply learned multi-channel textual and EEG features," *Future Generation Computer Systems*, vol. 119, pp. 1–6, 2021.
- [24] N. Moqbel, "Analysis and implementation some of data mining algorithms by collecting algorithm based on simple association rules," *International Journal of Computer Application*, vol. 138, no. 4, pp. 20–26, 2016.
- [25] V. Subeesh, E. Maheswari, G. R. Saraswathy, A. M. Swaroop, and S. S. Minnikanti, "A comparative study of data mining algorithms used for signal detection in fda aers database," *Journal of Young Pharmacists*, vol. 10, no. 4, pp. 444–449, 2018.
- [26] R. Balasubramaniam, "A detailed analysis of different data mining algorithms with hypothyroid data set," *Innovative Food Science & Emerging Technologies*, vol. 5, no. 9, pp. 250–255, 2018.
- [27] E. Balraj and D. Maalini, "A survey on predicting student dropout analysis using data mining algorithms," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 8, pp. 621–626, 2018.
- [28] S. Rosenthal, M. Veloso, A. K. Dey et al., "Is someone in this office available to help me?" *Journal of Intelligent and Robotic Systems*, vol. 66, no. 1-2, pp. 205–221, 2012.
- [29] Z.-G. Liu, Y. Yang, and X.-H. Ji, "Flame detection algorithm based on a saliency detection technique and the uniform local binary pattern in the ycbcr color space," *Signal, Image and Video Processing*, vol. 10, no. 2, pp. 277–284, 2016.
- [30] G. E. Sakr and I. H. Elhajj, "Vc-based confidence and credibility for support vector machines," *Soft Computing*, vol. 20, no. 1, pp. 133–147, 2016.
- [31] M. Torres and G. Qiu, "Habitat image annotation with low-level features, medium-level knowledge and location information," *Multimedia Systems*, vol. 22, no. 6, pp. 767–782, 2016.