Hindawi
Mobile Information Systems
Volume 2022, Article ID 8343909, 12 pages
https://doi.org/10.1155/2022/8343909

Research Article

English Online Teaching Resource Processing Based on Intelligent Cloud Computing Technology

Wenjuan Zhang

Xinyang Vocational and Technical College, Xinyang 464000, China

Correspondence should be addressed to Wenjuan Zhang; zhangwenjuan@xyvtc.edu.cn

Received 29 April 2022; Revised 13 June 2022; Accepted 24 June 2022; Published 8 July 2022

Academic Editor: Wen Zhou

In order to improve the processing effect of English online teaching resources, this paper improves the efficiency of resource processing by building an intelligent cloud system. Aiming at the problem of insufficient acquisition of traditional teaching resources, this paper adopts the intelligent data acquisition method and proposes an English resource mining method based on crawler technology. Moreover, in view of the difficulty of English semantic recognition, this paper proposes a method for calculating weights of feature items weighted by mixed factors on the basis of traditional algorithms and constructs a system model based on the characteristics of online English teaching. In addition, in order to explore the system effect, the model proposed in this paper is verified by clustering experiments. Through the experimental research, it can be seen that the English online teaching resource processing system has a good effect in the processing of English teaching resources.

1. Introduction

At present, the pace of informatization construction in colleges and universities is also accelerating. Moreover, major colleges and universities have established their own campus networks to realize the connection with the Internet, which has laid a good hardware foundation for the development of computer-assisted classroom English teaching, distance English teaching, network English teaching, and building a lifelong learning system.

According to the scientific citation analysis theory, the interconnected data between documents contains rich and useful information, and web structure mining mainly derives information and knowledge from the web organizational structure and link relationships. Due to the complexity of the structure, the common search engine only regards the web as a collection of flat documents and ignores its structure information. Mining the structure of pages and web structure can be used to guide the classification and clustering of pages. Besides, hyperlinks, as an important feature of hypertext documents, provide valuable information for web information acquisition. Recently, web retrieval algorithms based on hyperlink analysis, such as PageRank, have greatly improved the retrieval accuracy compared with the word-based methods used by traditional search engines. Generally speaking, hyperlinks in web documents contain two kinds of information. First of all, they provide users with navigation information for browsing the web, just like the commonly used navigation bar is used to guide visitors to jump between pages. Secondly, the hyperlinks in the page are often the document author's recommendation for a certain document, and the recommended target document often has similar content and is recognized by the author. The latter forms the basis of link analysis. That is, the importance of a certain document is determined not by the content of the document but by the number of times it is linked (or referenced) by other documents. In web retrieval, in addition to the number of links by other documents, the quality of the linked source document is also a reference factor for evaluating the quality of the linked document. Documents linked or recommended by high-quality documents tend to be more authoritative. This method of link analysis of web pages is called web structure mining. Compared with the query result ranking algorithm based on word frequency statistics used by traditional search engines, the advantage of the algorithm based

on hyperlink analysis is that it provides an objective and less cheating web resource evaluation method (some web documents are used to trick traditional search engines by adding invisible strings). Therefore, link analysis algorithms are currently used in the ranking of documents in search engines.

Web log mining can automatically and quickly discover the browsing patterns of network users, such as frequently accessed paths, frequently accessed page groups, and user clustering [1]. Based on the recognition of user browsing patterns, one method is to manually improve the site structure according to these patterns, so as to facilitate the browsing of users. Another more effective and more automated method is to allow the site to dynamically adjust and customize the site structure and page content automatically according to the current user's browsing mode and provide personalized services according to the user's behavioral characteristics. Data preprocessing and log mining algorithms are the key technologies in web log mining [2]. The result of data preprocessing as the input of mining algorithm directly affects the quality of mining, and the selection and improvement of mining algorithm are an important factor to ensure the success of mining. Therefore, the research on web log mining technology focuses on these two aspects [3]. Web usage mining techniques can generally be applied to two fields: when used to analyze access logs of web servers, adaptive websites can be designed using the service model obtained from mining, and [4] when applied to a single user, by analyzing user access history to discover useful user access patterns. Because the data objects processed in web usage mining are usually the user's access history or the server's access log, the content represented by the data objects cannot be known, so the results obtained are generally rough. However, because this method is relatively mature and simpler to implement than content mining, it has also been widely used in personalized systems [5].

In today's era, a large amount of information is pouring in, which has also caused a lot of inconvenience to people's lives. The most troublesome thing for users is that they need to slowly select the information that is useful to them from the massive information; at the same time, for the disseminators of information, they hope that the information they publish or generate will be discovered by the vast number of information acquirers, so that being able to stand out has also become a difficult thing to do. acquirers [6]. Since then, it has become an important tool to solve these problems brought by the Internet. The recommendation system needs to understand the user's habits, and then the channel to understand the user's habits is to analyze the user's purchase, browsing, and other historical records to obtain the user's preferences, thereby proactively recommending information to users that they are interested in or need. There are two popular recommendation systems now: one is to send web links to users by analyzing their preferences [7]; the other is to directly recommend products to users on an e-commerce platform. This kind of recommendation can also be called an e-commerce personalized recommendation system. People are more inclined to personalized recommendation, because the personalized recommendation system can greatly improve the service efficiency and service quality of the website, and users will be very efficient when browsing [8]. Therefore, personalized customization can attract more users and enhance the stickiness of existing users. But any kind of recommendation system is composed of 3 parts, including the front page displayed in front of the user, the background log, and the recommendation algorithm. The background logs and recommendation algorithm users are invisible. The front page is directly facing the user and interacts with the user [9]. Background logs are used to record and store various historical behaviors of users. The recommendation algorithm generates recommendation results by analyzing the user's historical behavior logs and then performing a series of calculations. Finally, the recommendation result is returned to the front page to display to the user [10].

Literature [11] proposed the learning theory in the era of digital network learning. Relevance learning theory mainly focuses on four questions and specifically answers the questions: What is learning in the era of digital network learning? Why learn? How to learn? Where do you learn from? First of all, learning in the era of digital network learning is a dynamic, open, and continuous process. Learning is the process of connecting information sources or knowledge nodes, that is, the connection of learning networks [12]. Secondly, the purpose of learning is to obtain knowledge absorption and innovation by establishing organic connections between knowledge nodes; the core skill of learning is to be able to discover or connect different fields, viewpoints, and concepts. Again, both formal and nonformal learning are ways of learning, but future learning is closely linked to work activities, mainly achieved in practical activities, personal networks, or task completion. Finally, learning is based on technology-supported learning, and knowledge is developed dynamically and diversely, so it is more important to know where there are massive information resources than to have them [13].

Reference [14] defines the characteristics of learning with chaos, complexity, continuity, cocreation, specialization of connection, and stability of continuous expectation. Literature [15] mainly studies E-learning. In the process of designing learning resources, the authors began to pay attention to the development of relationalism theory and discussed the connotation of relationalism, the connotation of connected knowledge, and the relationship between learning network and connected knowledge. Relationship is discussed. Literature [16], on the basis of summarizing the characteristics and development dilemmas of the business English subject, proposes the principles of building a business English subject system based on the perspective of relational learning theory (i.e., the principles of overall development, inclusiveness, integration and innovation, and cultivating characteristics). The conception of constructing a business English subject system (i.e., the construction of a multitheoretical system and the construction of an online learning environment) is proposed. Literature [17], based on the analysis of the problems existing in the current college English autonomous learning of college students, discusses the new characteristics of college English autonomous

learning under the guidance of the relevant learning theory and summarizes college students' English autonomy in the era of digital network learning. Based on the guidance of constructivist learning theory and connectivity learning theory, [18] constructed a ubiquitous learning resource sharing platform for "three-duo" college English quality courses of "multimodal resources, multiterminal access, and multichannel interaction." And in practice, it breaks through the dilemma of lack of situational context in college English language schools and effectively alleviates the contradiction of insufficient college English learning resources.

The main contribution of this paper is as follows: in the process of English resource processing, the traditional algorithm ignores semantic mining. This paper considers the appearance semantics and frequency of words from two perspectives and proposes a mixed factor weighted feature item weight calculation method to improve text similarity. It can improve the accuracy of degree calculation and promote the processing efficiency of English online teaching resources.

From the above analysis, it can be seen that the current processing of English online teaching resources is mostly carried out through data mining at that time. By analyzing the literature research on the simple structure of the network, it can be found that the topology map has obvious convergence, that is, web pages with similar themes converge. Page theme groups, and this kind of grouping between pages, is called a network community. This makes the web pages on the Internet naturally form various types of link structures, each link structure is called a network community, and its member pages are basically related to a certain topic. This feature of the online community makes it very relevant to its internal topics and is easy to crawl. However, there may be relatively few links due to thematic differences between different online communities. This will reduce the crawling rate of the theme crawler, and the strategy of the theme crawler to traverse the tunnel and crawl the web page information efficiently has become one of the focuses of crawler research in recent years.

The main innovation of this paper is to process English online teaching resources by improving the crawler technology. By analyzing the VSM web page classification algorithm, the VSM web page classification algorithm is improved from three aspects: feature extraction, eigenvalue calculation, and class core vocabulary generation. In this paper, considering the semantics of the appearance of words, a method for calculating the weights of feature items weighted by mixed factors is proposed, which improves the accuracy of text similarity calculation.

This paper combines intelligent cloud computing technology to construct an English online resource processing system to improve the efficiency of English online resource processing and combines crawler technology to conduct data mining to create resources suitable for English teaching and improve the effect of English online teaching.

2. Cloud Computing-Based Crawler Technology

2.1. Key Technologies of Theme Crawler. The description of the crawling target by the topic crawler can be divided into three types: based on the characteristics of the target web page, based on the target data pattern, and based on the domain concept.

The objects crawled, stored, and indexed by crawlers based on the characteristics of target web pages are generally websites or web pages. According to the method of obtaining seed samples, it can be divided into the following:

(1) The first is to predetermine the initial crawling seed sample

(2) The second is to predetermine web page categories and seed samples corresponding to the categories

(3) The third is to determine the crawling target samples through user behavior, which is divided into crawling samples that display annotations during user browsing, and access patterns and related samples obtained through user log mining.

The crawler based on the target data pattern is aimed at the data on the web page, and the crawled data generally conforms to a certain pattern or can be transformed or mapped into the target data pattern. Another way of description is to establish target domain ontology, which is used to analyze the importance of different features in a topic from a semantic perspective.

When calculating the PageRank value of a certain web page, all backlinks should be considered. The calculation formula of the PageRank value of a certain page p is shown in the following equation:

$$PR(p) = (1 - d) + d * \sum_{v \in B(p)} \frac{PR(v)}{N_v}. \tag{1}$$

v represents the web page, P represents the page of the website, and the crawler data mining can be carried out in combination with the corresponding parameters. In the above formula, N_v represents the number of forward links of web page v; $B(u)$ represents the set of web pages with links directly pointing to page p; PR (p) represents the PageRank value of page p; $PR(v)$ means that web page v evenly distributes its PageRank value to its forward links; d is a damping constant factor, usually 0.85. In reality, it is impossible for Internet users to randomly jump to completely irrelevant pages when browsing Internet pages, and it is impossible to completely follow the links in the current page. Therefore, d actually represents the probability that the user follows the web page link to browse without generating random jumps. In order to ensure that the calculation results are always converged, a damping coefficient d is added. Although the PageRank algorithm considers both the existence of Sink web pages and the randomness of user access behaviors, it still ignores the relevance of most user access links and query topics and the purposeful factors in query.

The Hits algorithm calculates its Authority value and Hub value for each page that has been visited and then uses this to determine the order of link visits. The Authority and Hub values of page p are $A[p]$ and $H[P]$, respectively, and the Authority and Hub values are calculated according to the following formula:

$$\begin{cases} A[p] = \sum_{q:(q,p)\in E} H[q], \\ H[p] = \sum_{q:(q,p)\in F} A[q]. \end{cases} \tag{2}$$

In the formula, E is the set of all pages pointing to page p, and F is the set of pages pointed to by links in page p.

We utilize web page and topic relevance for fuzzy scoring. The relevance scores of extracted links are mainly affected by three factors, namely, link text information, link context information, and inheritance to parent nodes. The calculation of the correlation in the Shark-Search algorithm uses VSM (vector space model) and takes a real number between 0 and 1. The score of a single link in the linked list is calculated by the following formula:

$$\begin{aligned} \text{Potentialscordur}\gamma = {} & \gamma * \text{inheritedur} + (1-\gamma) \\ & * \text{neighborhod}(\text{url})\,(\gamma < 1). \end{aligned} \tag{3}$$

When the parent node topic is related, the relevance score inherited from the parent node inherited (child (url)) is calculated by the similarity between the topic q and the parent node web page according to the following formula:

$$\text{inherited}\left(\text{child}(\text{ur}\delta) = \begin{cases} \delta^* \text{sim}(q, \text{current}(\text{url})), \text{sim}(q, \text{current}(\text{url})) < 0, & (\delta < 1), \\ \delta^* \text{inherited}(\text{current}(\text{url})), & \text{otherwise.} \end{cases}\right. \tag{4}$$

Among them, δ is the attenuation factor.

According to formula (5), the similarity between the topic q and the link text information sim (q, anchor (context (url))) can be obtained by simply calculating the similarity score of the link text. The score of the neighbor link neighborhood (url) and the context and text content of the link text are calculated as shown in formula (6).

$$\begin{aligned} &\text{anchor}(\text{context}(\text{url})) \\ &= \begin{cases} 1, & \text{anchor}(\text{url}) > 0, \\ \text{sim}(q, \text{anchor}(\text{text})), & \text{otherwise}, \end{cases} \end{aligned} \tag{5}$$

$$\begin{aligned} &\text{neighborhood}(\text{url}) \\ &= \beta * \text{anchor}(\text{url}) + (1-\beta) \\ &\quad * \text{anchor}(\text{context}(\text{url}))\,(\beta < 1). \end{aligned} \tag{6}$$

The crawling depth d of the Shark-Search algorithm is prespecified by the user. In one algorithm, the user needs to preset 4 parameter values: d, γ, δ, and β.

The topic-related pages on the Internet have obvious convergence; that is, web pages of the same type converge into page groups. There is a structure called Web Community in the network. That is, web pages on the Internet naturally form various link structures, and each link structure is called a network community, and the member pages in it are roughly related to a certain topic. This feature of the online community makes it very relevant to its internal topics and is easy to crawl. However, there may be relatively few links between different online communities due to differences in topics. The network community and network tunnel existing on the network are shown in Figure 1.

In the link relationship of each web page in Figure 1, we can see that when the theme crawler completes crawling all theme-related pages of an online community A, it enters the theme-independent link path and then enters other online communities B, because the linked pages are multiple. The topic is not related to the page, so it may be pruned on a certain link path, so there is a possibility that the network community B cannot be reached. This reduces the crawling rate of the theme crawler. A lot of topic-related resources are left out. Another situation is that the link relationship between two online communities is one-way; that is, an online community A contains a page that links to online community B, and online community B does not have a link to online community A. Particularly, web pages of affiliation in the network are a typical example of this situation. For example, the lower-level department webpages of some administrative departments have links to the higher-level departments, and most of the upper-level departments will not link to the huge lower-level department websites. In the case of one-way links, according to the principle of branch reduction of the crawling trajectory of the theme crawler, the possibility of loss of web page information of lower-level departments will increase. The unrelated links between related pages are called "network tunnels."

The space vector model is an algebraic model of correlation, applied to information filtering, information extraction, indexing, and evaluation. In this model, the document is regarded as a multidimensional vector space formed by keywords, and the set of index words is usually the phrases that appear at least once in the document. When searching, the input search term is also converted into a vector similar to the document. This model assumes that the degree of correlation between the document and the search term can be obtained by comparing the cosine angle deviation between the document and the search term.

The central idea of this model is that text information basically contains some keywords to express or reveal the content-independent attributes of the text, each independent attribute can be regarded as a dimension of the concept space, and these independent attributes are called text feature items. Text can be represented as a collection of these feature items, namely, $D = \{t_1, w_1; t_2, w_2; \cdots; t_n, w_n\}$.

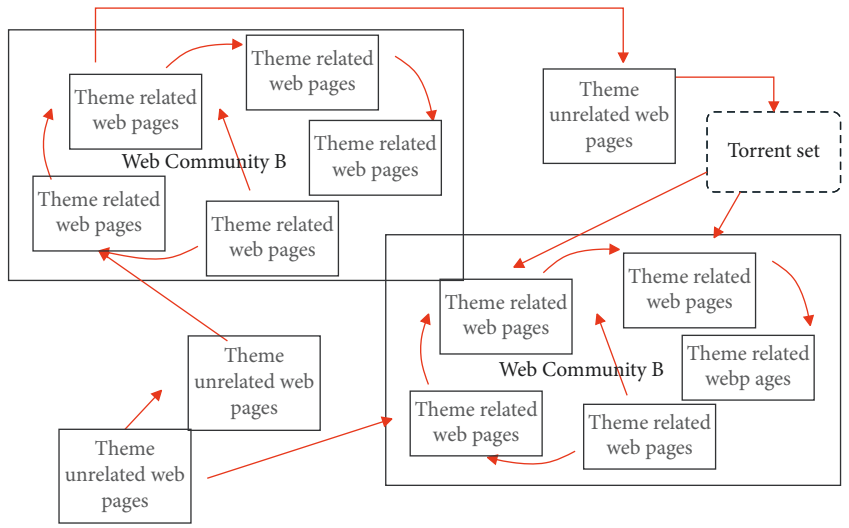

Figure 1: Web Community and network tunnel.

Among them, t_i is the feature word, and w_i is the weight of the feature word. In this way, the realization of the text information representation is transformed into the representation of the space vector. The cosine similarity can be used to measure the similarity between them. When calculating the similarity between two target vectors, the larger the cosine value, the higher the text similarity. The cosine similarity calculation formula is shown as follows:

$$\mathrm{sim}\left(D_i, D_j\right) = \frac{\sum_{k=1}^{n}\left(w_{k,i} \times w_{k,j}\right)}{\sqrt{\sum_{k=1}^{n} w_{k,i}^2} \times \sqrt{\sum_{k=1}^{n} w_{k,j}^2}}. \tag{7}$$

Through the abovementioned vector space model, text data is transformed into structured data that can be processed by computers, and the similarity between two documents is transformed into the similarity between two vectors. Usually, it is easier to calculate the cosine between the included angle vectors than to directly calculate the included angle. The cosine of zero means that the search term vector is perpendicular to the document vector; that is, there is no match; that is, the document does not contain this search term. However, VSM also has shortcomings. It regards text as a set of feature items for similarity calculation, which simplifies the calculation and ignores some important information.

Currently, the main weight calculation methods are as follows:

(1) Square root function: $w_i(d) = \sqrt{tf_i(d)}$

(2) Logarithmic function: $w_i(d) = \log(tf_i(d) + 1)$

(3) Boolean function: $w_i(d) = \begin{cases} 1, & tf_i(d) \geq 1, \\ 0, & tf_i(d) = 0, \end{cases}$

(4) TF-IDF function: $w_i(d) = t_i(d) * \log(N/n_i)$.

2.2. Improvement of VSM Web Page Classification Algorithm. Generally, general texts contain a large number of words, and these words have different effects on text classification. If all vocabulary is used for calculation, it will cause a large amount of calculation. The so-called feature extraction is to select those items with a large degree of discrimination of the text as the features of the text for classification. This can not only reduce the amount of computation but also improve the effect of classification. There are many methods for text feature extraction, and the commonly used are document frequency, information gain, mutual information, CHI, expected cross entropy, text evidence weight, odds ratio, feature selection method based on word coverage, etc. In the literature, these methods are compared. In this paper, the improved mutual information feature extraction method proposed in the literature is selected. The mutual information definitions of terms and categories are shown in the following formula:

$$\mathrm{RMI}\left(T, C_i\right) = \log\left[\frac{P(T|C_i)}{P(T)}\right]\frac{1}{R(t)}. \tag{8}$$

The value of the correction factor is shown in the following formula:

$$R(i) = \frac{N(i)}{\sum N(j)}. \tag{9}$$

The first 200 words with large mutual information between terms and categories are selected as text features.

The position weight is set to σ_t and its value is shown in the following formula:

$$\sigma_t = \begin{cases} 0.8, & \text{if } t \text{ in title}, \\ 0.6, & \text{if } t \text{ in head of paragraph}, \\ 0.4, & \text{if } t \text{ in end of paragraph}. \end{cases} \tag{10}$$

S_t is the number of times the word appears in the corresponding position, and the word weight calculation formula with the position weight added is shown as follows:

$$\mathrm{weight}_i(d) = \mathrm{weight}_i(d) * \frac{\sum s_{t_i} * \sigma_{t_i}}{\sum s_{t_i}}. \tag{11}$$

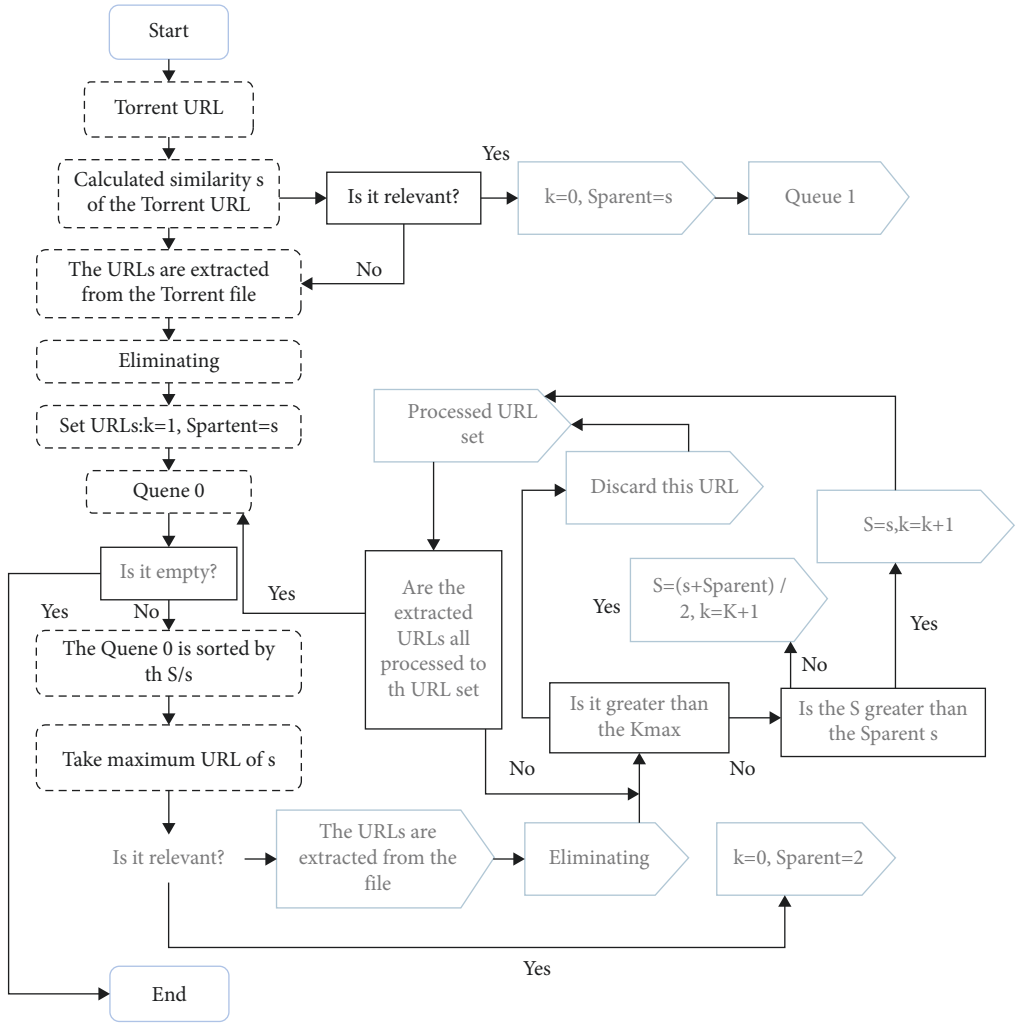

Figure 2: Crawler crossing tunnel algorithm based on dynamic adjustment theme.

Long words should have higher weight. In this way, after taking the Chinese idea into account, we calculate the weight of the vocabulary according to the following formula:

$$\text{weight}_i(d) = \frac{a}{a+1}\text{weight}_i(d). \tag{12}$$

In the formula, a represents the length of the vocabulary t_i.

The total number of times the word t_i appears in the text d is S_i, that is, the word frequency $t_f(t_i, d)$, and the total number of times the word t_j appears in the text d is S_j, that is, the word frequency $t_f(t_j, d)$. The cooccurrence frequency of the word t_i and the word t_j is recorded as S_{ij} (the count of nonrepetition in the sentence), and it can be known that $S_{ij} = S_{ji}$.

$$\begin{aligned} P_{ij} &= \frac{S_{ij}}{S_{ii} + S_{jj} - S_{ij}} \\ &= \frac{S_{ii}}{S_i + S_j - S_{ij}}. \end{aligned} \tag{13}$$

Among them, P_{ij} is the cooccurrence frequency of words t_i and t_j. Furthermore, it can be seen that $P_{ij} = p_{jj}$, $p_{ii} = 1$.

Finally, in a text, we can get a cooccurrence probability matrix between words in a word space, which is a symmetric matrix with n rows and n columns, which represents the number of text feature items.

$$P_{n\times n} = \left[p_{ij}\right]. \tag{14}$$

When using this matrix to modify the weight (t_i, d), the weight of the characteristic item t_i is modified as

$$w_i(d) = \sum_{j=1}^{n} pi_j \times w_i(d). \tag{15}$$

After processing in this way, words that often modify other words or are modified by other words have a high probability of cooccurrence, and the weight of words with a high probability of cooccurrence is strengthened. Words with high cooccurrence probability are generally more important words, reflecting the theme of the text. Also, not only the weight of the word but also the weight of the word it

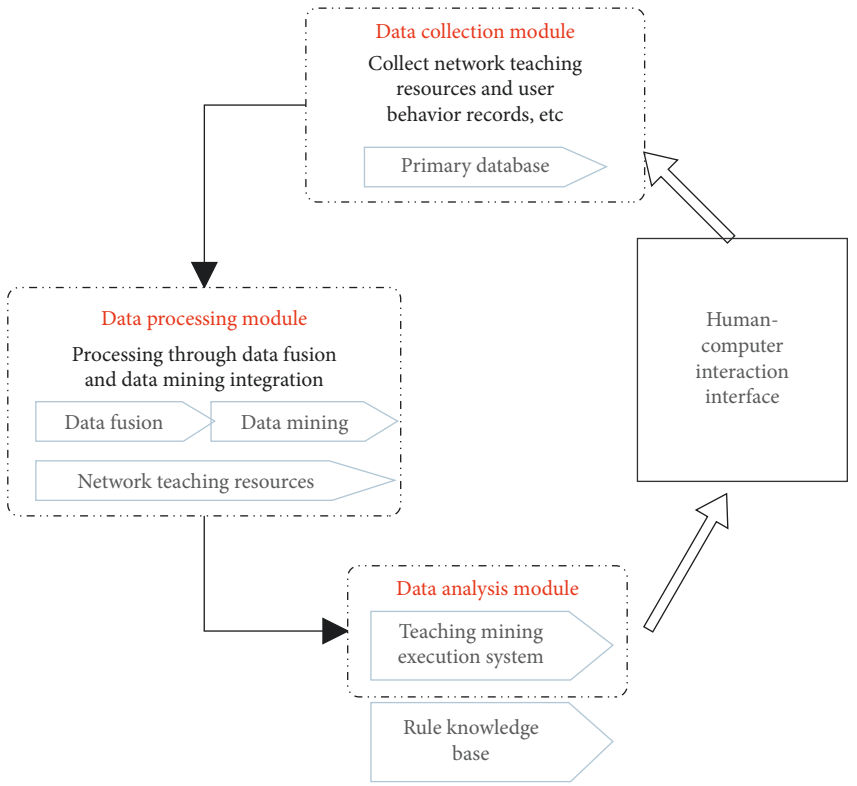

Figure 3: Model framework of network English teaching resource system based on data fusion and data mining.

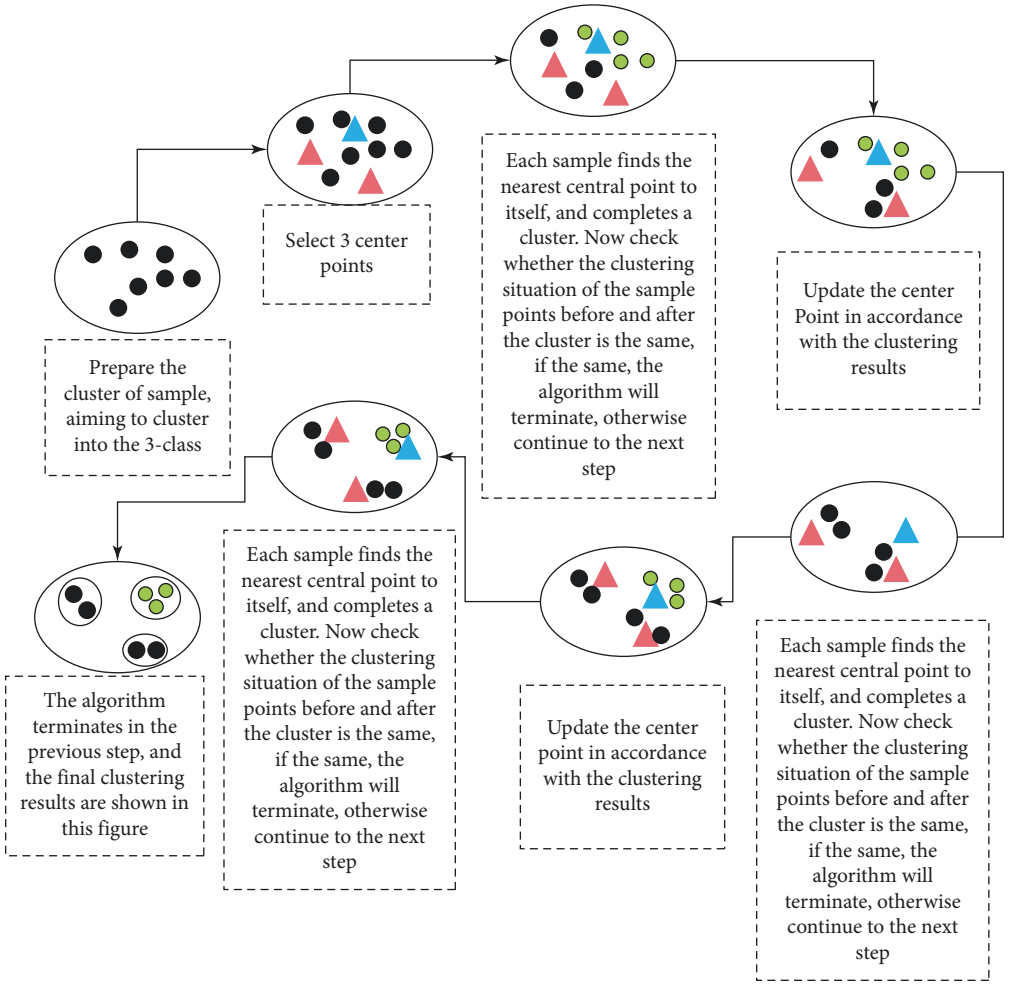

Figure 4: Clustering algorithm process.

wants to associate should be strengthened. The text features obtained in this way can summarize the text content, conform to people's thinking habits, and reflect the semantic information of the text.

The steps of the improved classification algorithm in this paper are as follows:

Step 1: the algorithm trains the core vocabulary group. Since each text category is a specific field set by people, the text of this category is a description of the category. Concept words of this class will appear repeatedly in the text of that class. For example, words such as programs, viruses, and drivers appear in computer texts higher than other categories. In the subject crawler research of this paper, the search information is mainly analyzed and extracted, and the class core vocabulary group Class Words $(D_j) = \{w_1, w_2, \ldots, w_m\}$ related to the search information is generated.

Step 2: the algorithm classifies the text D_i and extracts the core vocabulary Core Words $D_i = \{w_1, w_2, \ldots, w_m\}$; then the influence of Core Words D_i on the text D_i belonging to the core vocabulary group of D_j is $V(D_i, D_j)$. The calculation formula is shown as follows:

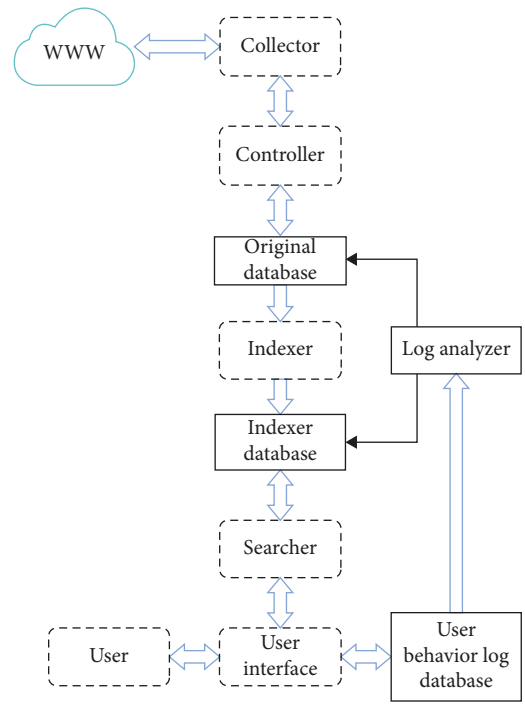

Figure 5: The architecture of the search engine for English teaching resources.

$$V(D_i, D_j) = \frac{\sum_{i=0, w_i \in D_i}^{n} \text{weigh}(w_i) + \sum_{i,j=,i \neq j, w_j, w_j \in D_i}^{n} \text{weigh}(w_i) \times \text{weigh}(w_j)}{\sum_{i=0}^{n} \text{weigh}(w_i) + \sum_{i,j=,i \neq j}^{n} \text{weigh}(w_i) \times w_{ij}(w_j)}. \quad (16)$$

Among them, weight (w) is the weight of w, and this paper uses the weight of the feature item weighted by the mixed factor for calculation.

Step 3: the influence value obtained above is weighted and combined with the cosine angle of the traditional VSM, and the final score is obtained as follows:

$$\text{SCORE}(D_i, D_j) = \alpha \times V(D_i, D_j) + \beta \times \text{sim}(D_i, D_j). \quad (17)$$

Among them, α, β is the weight, $\alpha + \beta = 1$.

Next, we will apply this algorithm to the subject crawl out of the tunnel algorithm.

2.3. The Design of the Crawler Crossing Tunnel Algorithm Based on Dynamically Adjusting Themes. The traditional idea of tunnel technology is a heuristic global optimal algorithm. When a crawler using tunnel technology encounters an irrelevant web page, it does not stop immediately but continues to explore K steps forward on this path, and the size of K is manually set. This allows spiders to jump from one Web Community to another, even though there is no link between the two Web Communities. If the distance between the two Web Communities is not large, it is possible to improve the crawling rate of the web page. However, this way of manually setting the K value is inflexible, and resources are wasted in vain when the distance between Web Communities is slightly larger or larger.

The idea of dynamically adjusting the crawler to pass through the tunnel is as follows: when the topic crawler enters the tunnel from a Web Community, the crawled pages that are not related to the topic use the above improved VSM text classification algorithm to calculate the topic similarity. After that, according to the similarity of the text, according to the idea of "Better Parents Have Better Children," the genetic factors of the parents are considered to judge the similarity of the theme. At the same time, the number of steps K to be advanced is dynamically judged according to the similarity prediction value, so that the K value is flexibly set to cross the tunnel. This eliminates the defect that the fixed K value is too large or too small. This paper proposes a crawler traversal tunnel algorithm based on dynamically adjusting themes.

The algorithm steps are as follows:

Step 1: according to the search topic, the algorithm trains two sample sets of topic-related and topic-independent offline training.

Step 2: the algorithm calculates the topic relevance s of the seed URL, extracts the URL links of the seed document D, and calculates the relevance s of the topic of the document D corresponding to these URLs and

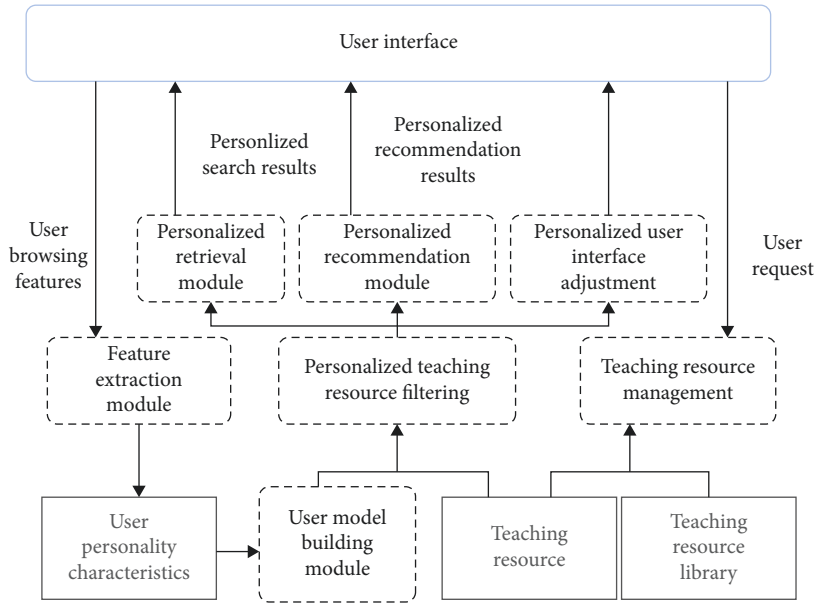

Figure 6: Retrieval model structure.

sets its parent $S_{\text{parent}} = s$, and the path $K = 0$ enters the crawling queue Queue0.

Step 3: the algorithm sorts the URLs in Queue0 according to the size of the relevance s. If Queue0 is not empty, the algorithm extracts the most relevant URL1 and judges its type according to the offline training sample set.

Step 4: if the document corresponding to the URL is a subject-related document,

(1) the URL enters the Queue1 queue
(2) the algorithm extracts the URLs of this article D and "eliminates duplicates"
(3) the algorithm calculates the topic relevance s of the extracted URLs one by one and sets its $K = 0$. If the document corresponding to the URL is a topic-irrelevant document, the algorithm determines whether the $K1$ of the URL is less than $K_{\max}$. If it is greater than Kmax and greater than the critical relevance value scritical, the K value is cleared to 0, and if it is less than the critical relevance value, the URL is discarded.

If it is less than $K_{\max}$, the algorithm judges whether it is greater than the minimum subject similarity $s_{\min}$. If it is less than $s_{\min}$, the URL will be discarded; otherwise, the size of its relevance s_1 and the parent's parent will be determined first.

If the correlation value s_1 is less than the parent's correlation value apparent, then according to "Better Parents Have Better Children," consider the genetic factors of its parents. If the correlation value s_1 is greater than or equal to the parent correlation value apparent, it is set to the correlation degree of s_1. The calculation formula is shown in formulas (18) and (19).

$$s = \begin{cases} \dfrac{s_1 + s_{\text{parent}}}{2}, & \left(K < K_{\max}, s_1 < s_{\text{parent}}\right), \\ s_1, & \left(K < K_{\max}, s_1 >= s_{\text{parent}}\right), \end{cases} \tag{18}$$

$$K = K_{\text{parent}} + 1\left(K_1 < K_{\max}\right), \tag{19}$$

Step 5: the algorithm determines whether the crawling queue Queue0 is empty. If it is not empty, the algorithm goes to Step 3, and if it is empty, the algorithm ends. The algorithm flowchart is shown in Figure 2.

3. English Online Teaching Resource Processing Model Based on Intelligent Cloud Computing Technology

This paper constructs a network English teaching resource system model based on data fusion and data mining. The model is mainly divided into four parts: human-computer interaction interface, data collection module, data processing module, and data analysis module. The model frame is shown in Figure 3.

We set a resource to be the sample points to be clustered, and the goal is to cluster the sample points into three categories. For each point, the algorithm calculates the center point that is closest to itself among all the center points, which can be defined as the same cluster. After one iteration, the center point of each cluster class is recalculated, and then the center point closest to itself is found again for each point. In this way, the loop continues until the cluster class of the two iterations before and after no longer changes. The algorithm process is shown in Figure 4.

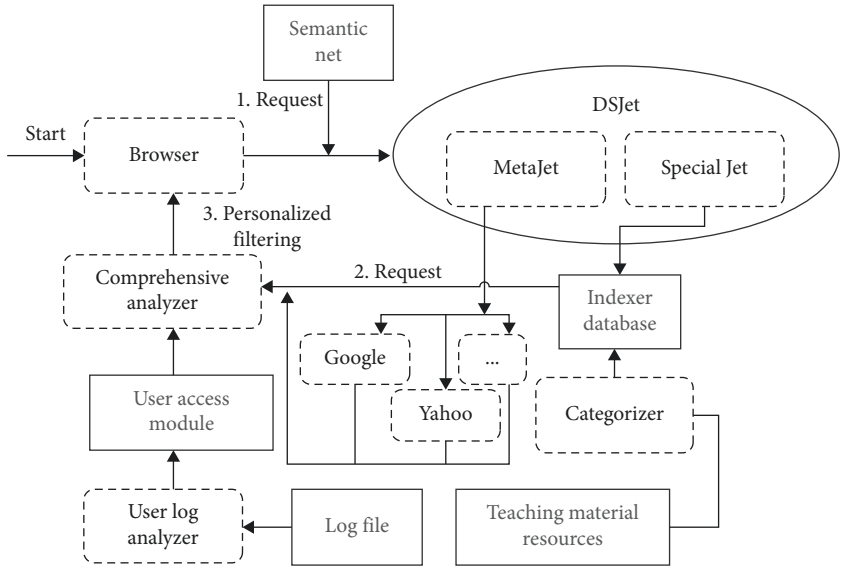

Figure 7: English teaching resource processing plan.

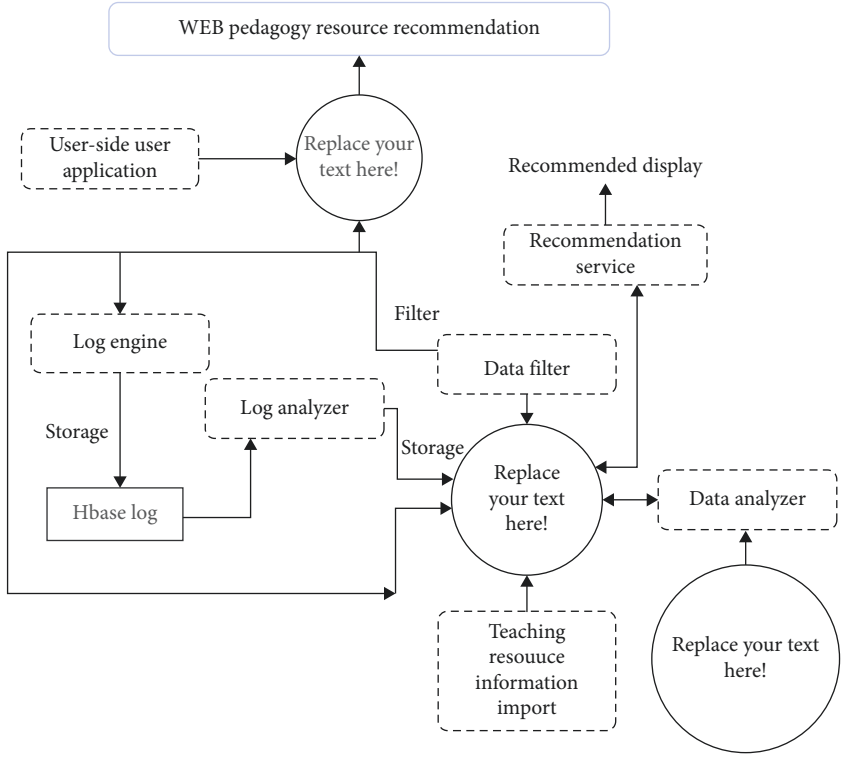

Figure 8: Schematic diagram of personalized English teaching resource recommendation system.

The block diagram of this system is shown in Figure 5, which is mainly composed of key parts such as collector, controller, original database of English teaching resources, indexer, retriever, and user interface.

According to the user's personalized interest characteristics, the English teaching resources are filtered to help the user quickly and accurately retrieve and recommend the English teaching resources he is interested in in the massive network English teaching resource database system. The structure is shown in Figure 6.

This system is mainly composed of four parts: user interface part, search engine part, database part, and user access mode personalized interface part (equivalent to four agent systems). The index user interface section incorporates the relevant feedback principle. The English teaching resource processing plan is shown in Figure 7.

The search engine part fully combines the advantages of professional search engines and metasearch engines and can better meet the needs of users in terms of precision and recall.

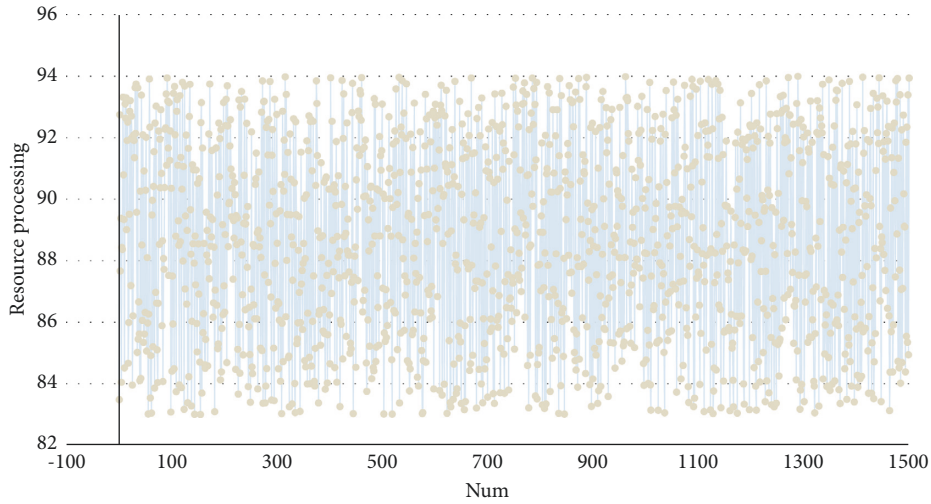

Figure 9: Clustering effect of English online teaching resource processing system based on intelligent cloud computing technology.

The schematic diagram of the complete personalized English teaching resource recommendation system is shown in Figure 8.

Through the clustering method, the effect of the English online teaching resource processing system based on intelligent cloud computing technology proposed in this paper is evaluated, and the obtained clustering results are shown in Figure 9.

From the above clustering results, the English online teaching resource processing system based on intelligent cloud computing technology proposed in this paper has a good effect in English teaching resource processing and can effectively improve the efficiency of English online teaching.

4. Conclusion

While doing a good job in the construction of hardware facilities, many colleges and universities have successively invested a lot of human, material, and financial resources in various ways to build a comprehensive English teaching resource library. These English teaching resources are currently mainly integrated and centrally stored and managed through the English teaching resource library and are provided to teachers and students in the form of portal websites. However, information from many aspects shows that the effect of these English teaching resources in English teaching practice is not ideal, and they have not fully played their due role in terms of the depth and breadth of application. This paper combines intelligent cloud computing technology to build an English online resource processing system to improve the efficiency of English online resource processing and combines crawler technology to perform data mining. From the clustering results, the English online teaching resource processing system based on intelligent cloud computing technology proposed in this paper has a good effect in the processing of English teaching resources and can effectively improve the efficiency of English online teaching.

Data Availability

The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is sponsored by Xinyang Vocational and Technical College.

References

[1] Z. Sun, M. Anbarasan, and D. J. C. I. PraveenKumar, "Design of online intelligent English teaching platform based on artificial intelligence techniques," *Computational Intelligence*, vol. 37, no. 3, pp. 1166–1180, 2021.

[2] H. Kim, "The efficacy of Zoom technology as an educational tool for English reading comprehension achievement in EFL classroom," *International Journal of Advanced Culture Technology*, vol. 8, no. 3, pp. 198–205, 2020.

[3] A. E. P. Atmojo and A. Nugroho, "EFL classes must go online! Teaching activities and challenges during COVID-19 pandemic in Indonesia," *Register Journal*, vol. 13, no. 1, pp. 49–76, 2020.

[4] A. Yuliansyah and M. Ayu, "The implementation of project-based assignment in online learning during covid-19," *Journal of English Language Teaching and Learning*, vol. 2, no. 1, pp. 32–38, 2021.

[5] S. Liu and J. Wang, "Ice and snow talent training based on construction and analysis of artificial intelligence education informatization teaching model," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 3421–3431, 2021.

[6] E. Apriani and J. Hidayah, "The ICT used by the English lecturers for non English study program students at STAIN curup," *Vision: Journal of Language and Foreign Language Learning*, vol. 8, no. 01, pp. 26–37, 2019.

[7] R. G. Jones, "Second language writing online: an update," *Language, Learning and Technology*, vol. 22, no. 1, pp. 1–15, 2018.
[8] P. P. Paredes, C. O. Guillamón, and P. A. Jiménez, "Language teachers' perceptions on the use of OER language processing technologies in MALL," *Computer Assisted Language Learning*, vol. 31, no. 5-6, pp. 522–545, 2018.
[9] Y. Bin and D. Mandal, "English teaching practice based on artificial intelligence technology," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 3, pp. 3381–3391, 2019.
[10] R. W. Todd, "Teachers' perceptions of the shift from the classroom to online teaching," *International Journal of TESOL Studies*, vol. 2, no. 2, pp. 4–16, 2020.
[11] M. A. AlGhamdi, "Arabic learners' preferences for instagram English lessons," *English Language Teaching*, vol. 11, no. 8, pp. 103–110, 2018.
[12] Z. Xu and Y. Shi, "Application of constructivist theory in flipped classroom - take college English teaching as a case study," *Theory and Practice in Language Studies*, vol. 8, no. 7, pp. 880–887, 2018.
[13] G. A. Toto and P. Limone, "Motivation, stress and impact of online teaching on Italian teachers during COVID-19," *Computers*, vol. 10, no. 6, pp. 75–94, 2021.
[14] C. M. yue, L. Dan, and W. Jun, "A study of college English culture intelligence-aided teaching system and teaching pattern," *English Language Teaching*, vol. 13, no. 3, pp. 77–83, 2020.
[15] Y. Rinantanti, S. Z. B. Tahir, and A. Suriaman, "The impact of EFL senior high school teachers' performance in papua, Indonesia toward the students' English learning achievement," *Asian EFL Journal*, vol. 23, no. 3.3, pp. 431–447, 2019.
[16] R. Shadiev and M. Yang, "Review of studies on technology-enhanced language learning and teaching," *Sustainability*, vol. 12, no. 2, pp. 524–532, 2020.
[17] M. A. S. Khasawneh, "An electronic Training Program on Developing the Written Expression Skills among a Sample of foreign language learners EFL who are at-risk for Learning disabilities during the emerging Covid-19," *Academy of Social Science Journal*, vol. 7, no. 10, pp. 1974–1982, 2021.
[18] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading," *Scientific Data*, vol. 5, no. 1, pp. 180291–180313, 2018.