Hindawi
Complexity
Volume 2021, Article ID 5575417, 10 pages
https://doi.org/10.1155/2021/5575417

Research Article

Design and Implementation of English Intelligent Communication Platform Based on Similarity Algorithm

Yujie Chai

School of Foreign Languages, Henan Institute of Technology, Xinxiang, Henan 453003, China

Correspondence should be addressed to Yujie Chai; venuschai@hait.edu.cn

Received 6 February 2021; Revised 15 March 2021; Accepted 19 March 2021; Published 30 March 2021

Academic Editor: Wei Wang

Intelligent communication processing in English aims to obtain effective information from unstructured text data using various text processing techniques. Text vector representation and text similarity calculation are important fundamental tasks in the whole field of natural language processing. In response to the shortcomings of existing sentence vector representation models and the singularity of text similarity algorithms, improved models and algorithms are proposed based on a thorough study of related domain technologies. This paper presents an in-depth and comprehensive study of text vectorization representation and text similarity calculation algorithms in the field of natural language processing. The existing text vectorized representation models and text similarity computation algorithms are described, and their shortcomings are summarized to provide a basis for the background and significance of this paper, as well as to provide ideas for improvement directions. It is experimentally verified that the sentence vector model proposed in this paper achieves higher accuracy than the SIF sentence vector model for text classification tasks. In the task of text similarity computation, it achieves better results in three evaluation metrics: accuracy, recall, and F1 value. The algorithm also improves the computational efficiency of the model to a certain extent by removing feature words with low feature contribution. The algorithm first improves the deficiencies of the traditional word-shift distance algorithm by defining multifeature fusion weights and realizes a text similarity calculation algorithm based on multifeature weighted fusion with better similarity calculation results. Then, a linear weighting model is constructed to further combine the similarity calculation results of the hierarchical pooled IIG-SIF sentence vectors to realize the multimodel fusion text similarity calculation algorithm.

1. Introduction

As the 21st century enters people's vision, network communication technology develops rapidly, the era of big data gradually enters people's vision, the complicated data information fills the Internet, and the amount of information carried by the Internet is growing. This huge information gradually becomes an important source to answer users' questions [1–5]. Most of the traditional search engine information retrieval methods still search by keywords. Although this retrieval method can help users search information, and it is, to a certain extent, feedback to the user in a large number of relevant and irrelevant search results, it is difficult for users to find their desired answers quickly [6]. How to accurately obtain the information users need most from the massive information sources is a common goal pursued by researchers in the information age. In recent years, the emergence of question-and-answer systems has gradually attracted the attention of scholars [7]. The design of the system makes a hierarchical description of the system's overall architecture, achievement management architecture, and orchestration system architecture and demonstrates database design and interface design of this system. Intelligent question-and-answer systems are a new research hotspot in the field of natural language processing and information retrieval, which allows users to ask questions in the form of sentences in natural language and automatically returns concise and accurate answers to users using natural language processing techniques, and its emergence reflects people's exploration and pursuit of fast and accurate information retrieval. Compared with traditional search engines, intelligent Q&A systems not only

conform to the expression form of users' questions, but also have obvious advantages over search engines in terms of questioning intent in terms of keyword matching and can directly refine the answers users want and present them to users in a structured way [8]. This intelligent question-and-answer approach plays an important role in answering questions and solving problems in distant education.

The rapid development of information technology has led to a geometric increase in the volume of data. Only about 20% of this massive data belongs to structured data, and the remaining 80% belongs to unstructured data. Text is the most common unstructured data, the main carrier of messages, and the source of important information and knowledge. How to process and understand text data and obtain useful knowledge from the huge amount of unstructured text data so that it can better serve human society is a question worthy of consideration and research [9]. In the external request processing and external services, WEB Handler is employed to define the properties and the compiler options of the HTTP processing program (.ashx) file. Client UI adopts WPF presentation. The complexity, ambiguity, and diversity of textual information itself lead to the fact that textual information is easy to be understood by humans but difficult to be processed by computers [10]. To solve this problem, natural language processing, as an important direction in the field of computer science, focuses on various theories and methods that can achieve effective communication and exchange between humans and computers in a natural language way, so that computers can better process text and provide effective solutions to various real-life scenario application problems. Text similarity computation is one of the fundamental research contents in the field of natural language processing, which is widely used in common research fields and scenarios such as search, recommendation, dialogue, data mining, and machine translation applications. In information retrieval systems, the core technique is to compute the similarity between the content of the text to be retrieved and the set of texts in the database [11]. In dialogue systems, the main purpose is to compute the questions in the question-and-answer database that are consistent with the semantic information of the user's question, then extract the answer to that question, and output it to the user to complete the dialogue process. In the recommendation system, the content data associated with the input text similarity is output, as shown in Figure 1.

Therefore, breakthroughs in text similarity computation can not only promote the flourishing development of data processing and natural language processing fields, but also advance the progress of general artificial intelligence to support the efficient work of human beings. The research of text similarity computation mainly contains text vectorization representation and text distance computation. Among them, text distance is usually measured by vectorizing the text representation and then using the distance between vectors [12]. Therefore, obtaining text vectors that can fully express the text information is also one of the important elements of the research. Text representation is the conversion of unstructured textual linguistic symbols into a computer-computable form, which is usually a vector, i.e., a vectorized representation of text. Text vectors can be applied not only for similarity computation but also as a basis for common text computing tasks [13]. Faced with text data containing rich information, the unstructured text is represented as a vector by modeling the text, which contains information about the semantic structure of the text expression, and then the text vector is used as the basis for subsequent text processing work [14]. Therefore, the quality of the vectorized representation of text can directly affect the accuracy and efficiency of the algorithm for subsequent tasks. In order to obtain vector representations that can adequately express text information and provide better support for higher-level text computing tasks, research on text representation models is extremely necessary and has a direct role in promoting research and development in the field of natural language processing.

In this paper, based on the analysis of domestic and international researches on text representation models and similarity calculation algorithms, word and sentence vector models and similarity calculation algorithms are studied in depth. The research work of the paper has two parts: the sentence vector representation model based on feature contribution degree and the text similarity calculation algorithm with multimodel weighted fusion. Based on the research of existing SIF sentence vector model, an improved model of sentence vector representation based on feature contribution degree is constructed. Since the original model only introduces generic word frequency information in the calculation of feature word weights and transports all words in the text to participate in the calculation of sentence vectors without any difference in the process of generating sentence vectors, without considering the screening of feature words with low contribution to the task, the obtained sentence vector representation has unfocused semantic information and poor task relevance, which affects the accuracy of subsequent text calculation tasks. To address the above shortcomings, firstly, this model improves the information gain calculation formula by introducing intraclass words frequency and intraclass and interclass differentiation factors to further enhance the effect of text feature selection. Then, a feature contribution factor that can portray the contribution of features to the task is constructed by combining the generic word frequency factor. Finally, this factor is used to remove the feature words with low contribution to the task, and the remaining strong feature words are involved in the subsequent calculation of the sentence vector, which can obtain a sentence vector representation with concentrated semantic information and strong task focus and improve the computational efficiency of the model to a certain extent. The experiments show that the improved model achieves better experimental results in the two basic tasks of text classification and similarity calculation.

2. Related Work

The knowledge ontology base is a "question-answer" base of user questions, and the answers to these questions are stored in a database. When a user asks a question, the system usually first searches the Frequently Asked Question (FAQ)

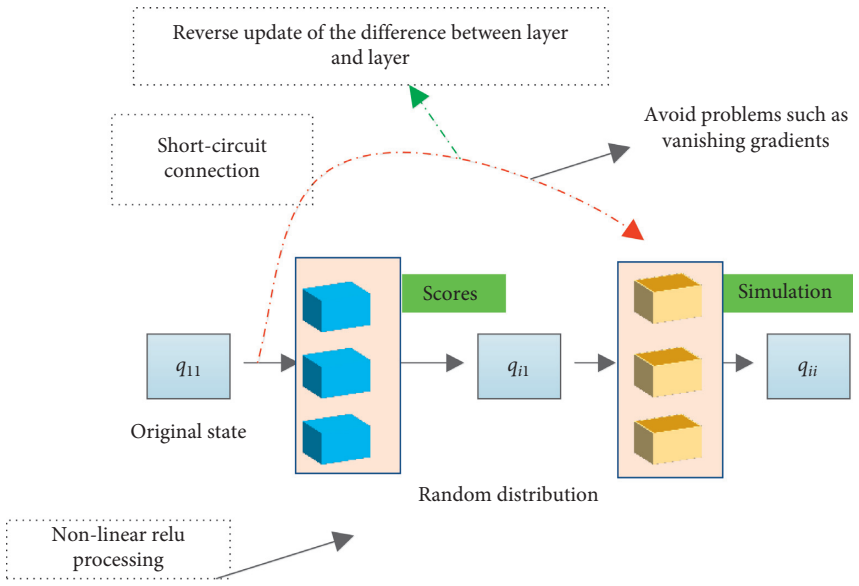

Figure 1: Illustration of the algorithm for retrieving the relevance of text content to the database.

set and then searches for the most similar answer to the question and returns it to the user [15]. If the system cannot find a satisfactory answer in the FAQ database, the system will automatically switch to the knowledge ontology database for retrieval. Among them, the similarity algorithm is the key technology in the intelligent Q&A system, which is used to realize the finding of the most similar question in the knowledge ontology database. By calculating the similarity between the questions asked by the user and the answers stored in the knowledge ontology database, the answer with the highest similarity is selected as the answer to the question asked by the user, and the corresponding answer information is returned to the user [16].

Currently, sentence similarity algorithms have a wide range of applications in reality, and their research status affects the research progress in other related fields. The sentence similarity algorithm is one of the key techniques in question-answer systems. Based on this, this paper designs and implements an English intelligent quiz system by applying sentence similarity algorithm to an elementary school English intelligent quiz system [17]. After the examination, the answer and the recordings are automatically submitted. The invigilator is opened to view the report status. The sentence similarity algorithm not only improves the operation efficiency of the English intelligent quiz system, but also accurately answers the English questions of elementary school students by considering the characteristics of English question words and semantics in various aspects. The system can also meet the functional requirements of English intelligent quiz learning for elementary school students and better help them achieve diversified English learning.

The vectorized representation of text reflects important features embedded in the textual language, such as semantic, syntactic, and text structure information. The study of traditional text representation models originated from the Boolean model-based text representation proposed by Lancaster et al. in 1973, which uses binary values 0 and 1 to represent text. As a primitive text representation, the Boolean model is simple and easy to implement, especially for handling large-scale data tasks more efficiently [18]. However, this binary representation structure lacks the semantic information in the text. To solve the problem of missing information in this model, researchers proposed the classical Vector Space Model (VSM). VSM represents text as vectors, using vectors corresponding to points in the feature space with the same number of vector dimensions as the number of feature words. Although this model can express certain semantic information of text, it cannot portray the semantic and syntactic structure information among text features due to the independence of each dimension in the feature space, which may cause ambiguity of multiple meanings of words. In addition, the expansion of the corpus leads to an increase in the dimensionality of the text vectors in the model, and the vectors used to represent a text will be extremely sparse and not utilized for subsequent processing. The traditional text representation model solves the problem of text representation in a certain sense, but the represented text vector contains only the shallow semantic information of the text, and the representation vector is high-dimensional and sparse, which directly affects the complexity of the computational process and the accuracy of the subsequent tasks [19]. The main purpose of system testing is to conduct a most comprehensive test on the software system, so that this system project can meet the product demands and conform to the overall and detailed system design.

To improve the capability and accuracy of text vector representation, some researchers have proposed a series of topic generation models starting from mining the latent features of text feature space. The latent semantic indexing model was proposed in 1990, and its core idea is to use singular value decomposition to decompose the text-word matrix into three matrices: document-topic, topic-word, and word-word, to realize the dimensionality reduction representation of text features. However, the model lacks a

statistical basis, and the representation results obtained by the model have low interpretability. On this basis, the latent Dirichlet allocation model has been proposed. The LDA model gives the hidden topic probability distribution of each text in the corpus and maps the text to the hidden topic space. The model uses a limited number of hidden topics to achieve a low-dimensional representation of the text. Since the topic model can obtain a low-dimensional vector representation of text and has a solid theoretical foundation, it has been highly valued by researchers, and many improved models have been proposed, such as hierarchical LDA, hierarchical Dirichlet process, and associative topic model. These models are good at mining the semantic information of text from the topic space using potential topic features, but there are still problems of long training time and unsatisfactory processing of short text.

3. English Text Similarity Calculation and Interaction

3.1. Algorithm Based on Vector Space Model. Algorithms based on vector space models act directly on the sequence or combination of feature words of the original text and use the matching degree or distance of feature words of two texts to determine similarity. This type of model is to represent the text into vectors and then measure the similarity of the text by the spatial distance between the vectors. The distance between vectors is mainly calculated using algorithms such as matching coefficient, cosine similarity, and Euclidean distance. The vectors here are mainly in the form of binarized one-hot vector. The role of this vector is mainly to transform the unstructured text information into computable vector form, the vector itself contains minimal semantic information, and the vector is high-dimensional and sparse. The vector space-based model only considers the surface words and does not take into account the semantic and structural information contained in the sentences, which leads to poor coverage of text features and low accuracy, resulting in inaccurate text similarity calculation results. Due to its limitations, the vector space-based model only uses the simple word frequency information of words in the text to transform the unstructured text into vectors, ignoring the semantic and contextual relationships of the text, resulting in a complex and inaccurate text similarity calculation process.

The Vector Space Model (VSM) considers that a text consists of several independent words. These independent words constitute the feature set of the text. Each feature item is given a different weight by combining the word frequency information of the text, and the spatial vector of the text is formed by taking the weights of all the feature items as components. Finally, we calculate the semantic distance based on the spatial vectors of the two texts to obtain the text similarity calculation results. Given a text T, with t_i denoting a feature item in the text and W_i denoting the weight value of t_i in the text T, then, in the multidimensional vector space, the weights of all feature items in the text T form the vector text are the value of the vector text in a certain dimension.

After the necessary preprocessing of the text, the weighting of the feature items in the vector space model is a crucial step, and the TF-IDF is usually used to calculate the weights of the feature items in existing studies. The TF-IDF method considers both whether a single feature item can express the information of a single text and whether the feature item can distinguish the text from other texts. After calculating the weights, the vector space representation of the text is obtained. Following that, the similarity between texts can be calculated. Currently, Euclidean distance is generally used to calculate the degree of similarity between two texts. The feature vectors of texts k_i and k_j are as follows:

$$\begin{aligned} V_{ki} &= \{w_1, w_2, \ldots, w_i\}, \\ V_{kj} &= \{w_1, w_2, \ldots, w_j\}. \end{aligned} \tag{1}$$

The text similarity is

$$V_{\max(kj)} = \oiint_{i=0} \sum_{i=1}^{ki} \operatorname{Max}(ki + kj) \frac{\sqrt{V_{ki} V_{kj}}}{ki + kj}. \tag{2}$$

Each text in the LF-LDA model is represented by some topic probability vector obeying the Dirichlet distribution. Since the text representation vectors generated by the LF-LDA model are hidden topic vectors obeying the probability distribution, the distance function Jensen-Shannon (JS), which is more suitable for measuring the probability distribution, is used. Then, the generated text topic vector is as follows:

$$k = \{k_1, k_2, \ldots, k_i\}. \tag{3}$$

Its similarity calculation can be based on

$$\operatorname{sim}(k_1, k_2, \ldots, k_i) = DJS \frac{1}{2} \left[\oint_{i=o} \frac{\sqrt{\sum_{i=1}^{n} k_i^2}}{k_1 + \cdots + k_i} \right]. \tag{4}$$

Before calculating the bulldozer distance, it is necessary to define the distance d_{ij} between the eigenvolume in P and any of the eigenvolumes in Q. The Euclidean distance is generally used when the two eigenvariables are vectors, and the KL distance is generally used when the two eigenvariables obey a certain probability distribution. The classical application of bulldozer distance is to find the optimal solution of linear programming in transportation problems. Suppose that the same batch of identical goods is to be transported from n factories to m warehouses. In this practical problem, P is a set of multiple factories, P_1 to P_n represent n factories, and factory P_i has goods with weight PW_i. Q is a set of multiple warehouses, Q_1 to Q_m represent m warehouses, and warehouse Q_j has the maximum capacity QW_j. The distance to transport goods with weight ijf from P_i to Q_j is defined as d_{ij}, then this problem. The optimization objective is to minimize the following cost function:

$$Q = \sum_{i=0}^{m} \left[d_{ij} + f_{ij} \right]. \tag{5}$$

Subject to

$$f_{ij} \le 1, \quad i, j \in [0, 9],$$

$$\sum_{i=0}^{q} f_{ij} \le W_{F_i}, \tag{6}$$

$$\sum_{i=0}^{m} f_{ij} = \frac{\oint_{i=1} \left[f_{ij} + q_{ij} \right]}{W_{Q_p}}.$$

Applying the bulldozer distance to the text similarity calculation yields the word-shift distance algorithm. The word-shift distance measures the difference between two documents, where the difference refers to the minimum cost required to transfer words from one document to another, i.e., the minimum amount of distance. Then, applying the bulldozer algorithm to the text similarity calculation problem requires solving two problems: (i) how to represent the unstructured text as a vector or obey a certain probability distribution and (ii) constructing a cost function for word transfer between texts.

3.2. Sentence Vector Model for Feature Contribution Degree. Although the smoothed inverse frequency model can generate the sentence vector of the text better, only the word frequency information generated using the common dataset is introduced in the calculation of the weights of each word in the sentence vector. In the process of generating the sentence vector, all words in the text are transported to participate in the calculation of the sentence vector without any difference, resulting in the fact that the sentence vector obtained by the model will have high dimensionality and unfocused semantic information. In addition, the feature words are not optimized when oriented to specific tasks, which affects the accuracy of text computation tasks. This section introduces feature selection methods to the model for the above deficiencies to obtain a better sentence vector representation.

The latent variable generation model is a model proposed in 2016 to portray the dynamic generation process of sentences. The model assumes that the t_{th} word is generated at the t_{th} moment until the complete sentence is produced, as shown in Figure 2. The discourse vector is a latent variable that represents the direction of word generation in the sentence, so the discourse vector of the sentence is denoted as C_t. The random wandering of the discourse vector C_t drives the whole process of sentence dynamic generation. Each word w in the sentence is a 2-dimensional vector.

To obtain the relationship between a word and its sentence, the model uses the inner product of the word vector v_w of word w and the discourse vector t_c at the current moment to represent the relationship. The model assumes that the probability of word w occurring in sentence t_c at time t is a log-linear relationship between their inner products:

$$Q(w) = C_t\{t_1, t_2, \ldots, t_i\}. \tag{7}$$

Based on the idea of latent variable generation model, the random wandering model aims to obtain the relationship between the sentence vector and the word vector of a sentence. To simplify the formulation, the model represents the discourse vector at all moments as a deterministic vector s_c since t_c changes extremely little during the generation of sentences. Since some words in the corpus occur outside the specified context and some word occurrences (stop words) affect the discourse vector generation, the SIF model adds two smoothing terms to this model.

(1) It generates a cumulative term in the model, where α is a fixed hyperparameter. $Q(w)$ is the probability of word w occurring in the text corpus. With the introduction of this term, although the inner product of a word and the discourse vector are small, the word also has a certain probability to participate in the subsequent calculation, which solves the problem of unregistered words to some extent.

(2) The common discourse vector *cd*, a smoothing term representing the "central idea" of the sentence, is the most important component of the sentence. The model finds that the longer the projection of word w along the c_0 vector is, the more the smoothing term will increase the probability of word w.

After adding the smoothing term, based on a fixed discourse vector c_s, the probability of a certain word w occurring in sentence s is

$$Q(w) = C_t\{t_1, t_2, \ldots, t_i\} + (1 + \alpha)\frac{q[c_s, w]}{t}. \tag{8}$$

To obtain the sentence vector representation, the model assumes that the word vector representation vw for each word in the corpus obeys approximately uniform distribution over the text vector space. Thus, for any c_s, the z-values are the same. Under this assumption, the c_s vectors are estimated using the maximum likelihood method, which yields the following likelihood function:

$$P(w) = \prod C_t\{t_1, t_2, \ldots, t_i\} + (1 + \alpha)\frac{q[c_s, w]}{t}. \tag{9}$$

The relationship between the sentence vectors and the word vectors is thus obtained, and the representation of the sentence vectors can be obtained by using a weighted average of the word vectors. Finally, the model is optimized by subtracting the vector c_0 of the sentence vector in the direction of its first principal component. The final sentence vector representation results are obtained. The smoothed inverse frequency model, though, better portrays the sentence generation model under the statistical laws of the general corpus. However, it only takes into account the word frequency information of sentences and does not further consider the influence of different types of text corpus on sentence vector generation in specific tasks. To maintain the generality of the smoothed inverse frequency model while improving its accuracy in

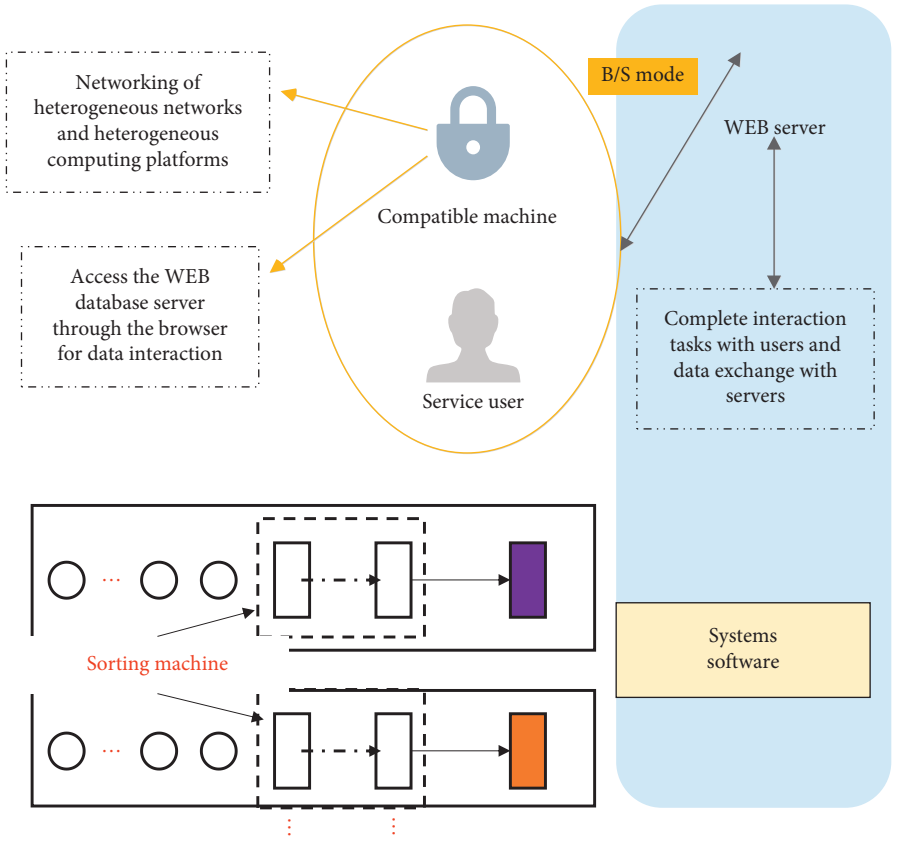

Figure 2: Latent variable generation model process diagram.

downstream tasks, this section proposes a sentence vector representation model based on feature contribution degree. An improved information gain method is used to preemptively remove the features with low contribution before the model computes the sentence vectors, thereby improving the differentiation of the sentence vectors among different categories and thus obtaining higher accuracy in computational tasks.

4. The Realization of English Intelligent Communication

4.1. System Model. Based on the above similarity calculation, we designed an English intelligent quiz system, as shown in Figure 3. The English Intelligent Quiz System is a more advanced information retrieval technology in the field of nature, which aims to enable elementary school students to ask questions about basic English sentences in daily life and then get an accurate answer by using this system. Simply put, the English Smart Quiz system is designed to analyze the English questions asked by students and understand the meaning of the questions and then return the answers to the students. The first module is the question analysis module, which consists of five parts: word reduction, lexical annotation, question type analysis, question-answer type analysis, and keyword extraction. The second module is the similarity calculation module, which is the core content of the Q&A system, and this part mainly includes word similarity calculation and sentence similarity calculation. The third module is answer extraction, which mainly focuses on similarity sorting, filtering answers, and outputting answers. The last module is knowledge pushing; this part is mainly to present answers to users in diversified forms by understanding different users' knowledge level levels, related knowledge difficulty, and users' preference for resource types.

The question analysis module is an indispensable part of the intelligent question-and-answer system. Its goal is to enable the computer to understand the semantics of the user's query and prepare for the subsequent work in the answer extraction module. Accurate question analysis helps the system use the appropriate answer extraction methods and strategies for different categories of questions during the answer extraction module. When the user enters an English question, the question analysis module will analyze and process it, and the processing procedures include the following:

(1) Word reduction changes all words in the question-answer sentences back to their prototypes. For example, the verbs "had" and "has" become have, the plural word "sports" becomes sport, and the verb "becomes" is and "was" becomes be.

(2) The lexical annotation is done for each word after reduction. These include words with verbs, words

with nouns, words with adjectives, and words with adverbs.

(3) Question types are analyzed where some common English question types proposed by users are analyzed, such as "What kind of". . .. . .How do you like ", " Which is. . .. . ." and so on.

(4) Question-answer type analysis determines to which type of thing the question asked by the user belongs. There are various types of question answer, such as answer and contextual dialogue. Each type contains four types of media presentation: audio, video, picture book, and picture. There are three levels: difficulty, medium, and easy. In addition, users can enter a word individually to query the result.

The types of word answers include answers, original sentences from the text, situational passages, and extended example sentences. The division and determination of words, as well as question-answer types, not only enrich the form of English knowledge presentation, but also improve elementary school students' motivation and interest in learning English to a certain extent.

4.2. Similarity Calculation Module. The sentence similarity calculation method is mainly used to select a suitable sentence by calculating the similarity between two utterances. The similarity of the utterances mainly includes word form, syntax, and semantics. The larger the similarity value derived from the calculation results, the closer the information of the two sentences in terms of word form, syntax, and semantics. In this study, the distance-based similarity algorithm is used to calculate the similarity of English interrogative sentences, which is based on the Word Net conceptual-semantic classification dictionary, and the conceptual word similarity calculation method is used to calculate the similarity between English words, and then the semantic similarity between English utterances is obtained according to the calculation method of the pinched cosine similarity.

The answer extraction module is to analyze the alternative answer interrogatives obtained from the information retrieval module in terms of lexical, syntactic, and semantic aspects, and the answers need to be sorted. In addition, the system also needs to set a threshold value, and only when the similarity of statements is greater than the set threshold value, the retrieved results are output, and the retrieved results are filtered through the mandatory keyword table to remove the contents that are irrelevant to the retrieved results, and then the answers are refined according to the categories to which the query question sentences belong, and the most appropriate answers to the questions are returned to the users in a way that is consistent with their knowledge level. The module mainly obtains the most similar alternative answers from the knowledge ontology database through question-sentence similarity calculation and then carries out answer extraction, sorts the alternative answers with similarity value greater than 0.8 according to the similarity value, and pushes the alternative answers to the knowledge push module according to the question types obtained by the question analysis module.

The knowledge push module is an important part of the system and is unique in that it takes into account the individual interests of users, each of whom has his or her own unique learning style and is able to push different knowledge based on the user's browsing information. The intelligent Q&A system provides a variety of answers to users based on their prior knowledge level, interests, and learning styles, and different users will get answers to questions that match their knowledge level. In the knowledge pushing module, the content of question-answer presentation is divided into three levels according to the difficulty level: easy, medium, and difficult. The answers to questions with different difficulty levels are recommended by the background of different users' knowledge levels (as shown in Figure 4).

The knowledge ontology library of the system is divided into three parts: knowledge ontology, textbook organization ontology, and resource library. The knowledge ontology includes word ontology and sentence ontology with different attributes; each attribute consists of resources such as pictures, audio, video, and picture books to meet the individual needs of different learners. Since the algorithm of the similarity calculation module is to find the sentence with the highest similarity among the same question words, which will miss the sentences with high semantic similarity among different question words, a special sentence ontology library is established to link such sentences to improve the finding accuracy. Both the word textbook organization ontology and the sentence textbook organization ontology inherit the textbook organization ontology and contain all the attributes of textbook features. Distance-based similarity calculation is one of the most used algorithms, and its basic idea is to measure the semantic distance between two conceptual words in the semantic dictionary tree by obtaining their path lengths, and a negative correlation is shown between semantic similarity and semantic distance. If the semantic distance between two words is larger, then the similarity is lower; on the contrary, if the semantic distance between two words is smaller, then the similarity is higher, as shown in Figure 5.

5. Results and Discussion

In this paper, we use the vector space model to represent English utterances when measuring the similarity between utterances, and the vector space model is one of the better methods to represent text. The basic idea of the vector space model (VSM) is to partition the smallest units containing semantic meaning such as words and phrases in a text and then use their corresponding similarity values as the elements in a vector. This vector space model can accurately and objectively represent the semantic information of English text. After vectorized representation of two English question-and-answer sentences, the semantic similarity between English utterances is obtained using the vector similarity measure, the angle cosine. Since the multifeature fusion-based word-shift distance algorithm (MMF) involves the selection of text feature words, i.e., the word selection

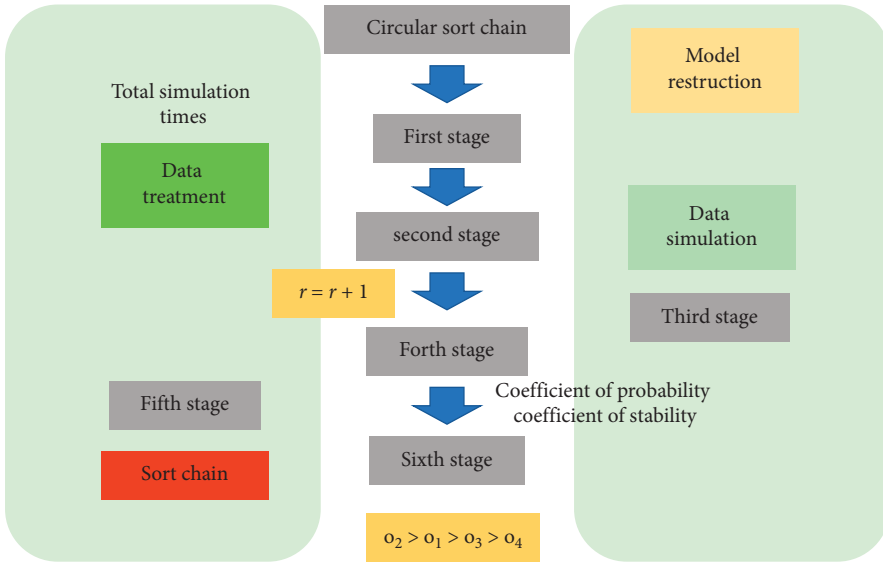

Figure 3: English intelligent question-and-answer system model.

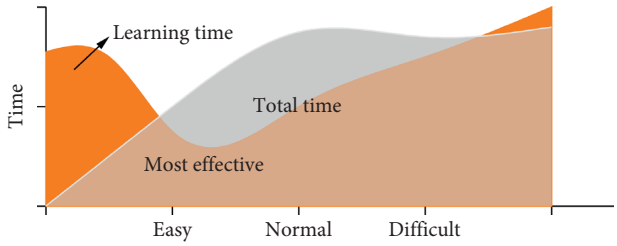

Figure 4: Question-answer difficulty level.

step, different selection ratios have a great impact on the similarity calculation results. If the proportion of selected text feature words is too small, the model does not contain enough information about the text, resulting in unsatisfactory algorithm results. If the proportion is too large, the model contains too much redundant information, which may affect the accuracy of the algorithm, and the computational complexity of the algorithm will increase.

Clustering is the most common type of information mining model, which requires neither training nor prelabeled document categories. Experiment 1 uses two clustering algorithms, K-mean, and DBSCAN, to determine under what proportion of text feature words are selected to better represent the text information. Since the dataset II is also a common clustered text dataset, it has a high recognition in related researches. Therefore, in this experiment, dataset II is selected as the experimental dataset, and the normalized mutual information index (NMI) is chosen to evaluate the clustering results. The larger the NMI value is, the more similar the clustering cluster structure obtained by the algorithm is to the real clustering cluster structure, and the more information the text contains. In this section, the K-means clustering algorithm is run under the scikit-learn package, and the DBSCAN algorithm is implemented by the source code. The experimental results are shown in Figure 6 to demonstrate the effect of different percentages of text feature words on the clustering effect.

As can be seen from Figure 7, good clustering results can be achieved when about 60% of the key text feature items of the target text are selected. If the percentage is lower than this, the number of key feature items is small and cannot express the information contained in the text, resulting in inaccurate results. On the contrary, it will add too much redundant information and reduce the independence between texts, resulting in unsatisfactory results. As known from the experimental results, increasing the weight value of the word-shift distance algorithm with multifeature fusion increases the recall rate significantly, which firstly affirms the effectiveness of this distance calculation algorithm for the task of text similarity calculation, and secondly fusing multiple key text features is an extremely necessary and feasible idea. The weight value of the IIGSIF-hire Sim algorithm is not sensitive to the recall rate, mainly because the hierarchical pooling IIGSIF model retains a certain degree of word order and spatial information, but it still lacks the more critical lexical and location information in the text; on the other hand, it is also because the IIGSIF model eliminates the textual. This may also affect the results of text similarity calculation.

From the data of the comparative experimental results of the five algorithms on the four types of datasets, the experimental results of the text similarity calculation algorithm with multimodel weighted fusion (MMFSim) proposed in this section are better than the other control algorithms in the three evaluation indexes of accuracy, recall, and F1 value, more generally. This is due to the fact that the MMFSim algorithm not only utilizes the word frequency information in the text, but also fuses the lexical, semantic, sentence position, and text structure information of the text and constructs multifeature fusion weights with this, which makes full use of the key information contained in the text and calculates the transfer distance between words more

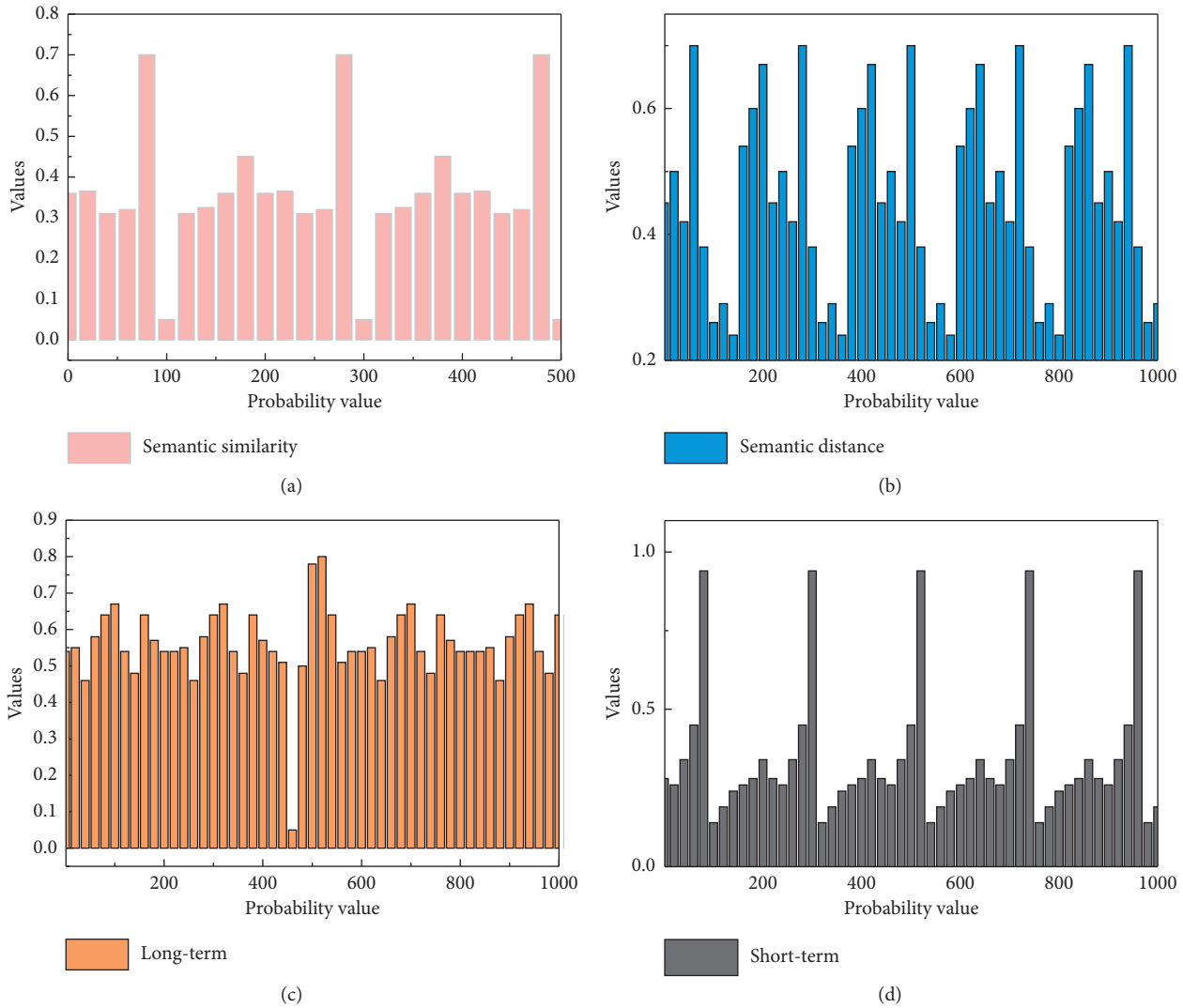

Figure 5: Relationship between semantic similarity and semantic distance.

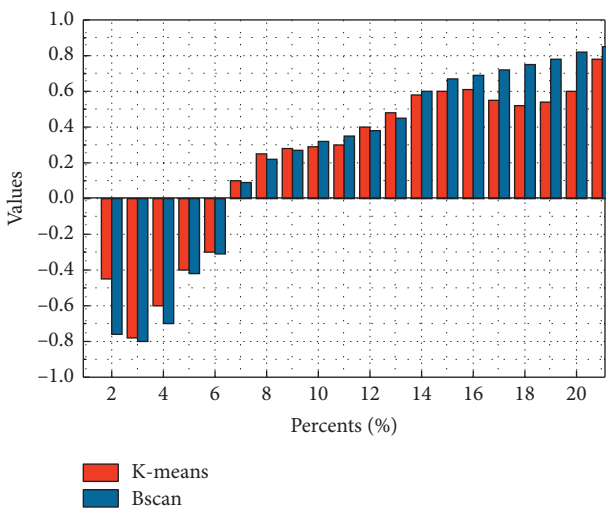

Figure 6: Effect of different proportions of text feature words on the results of clustering experiments.

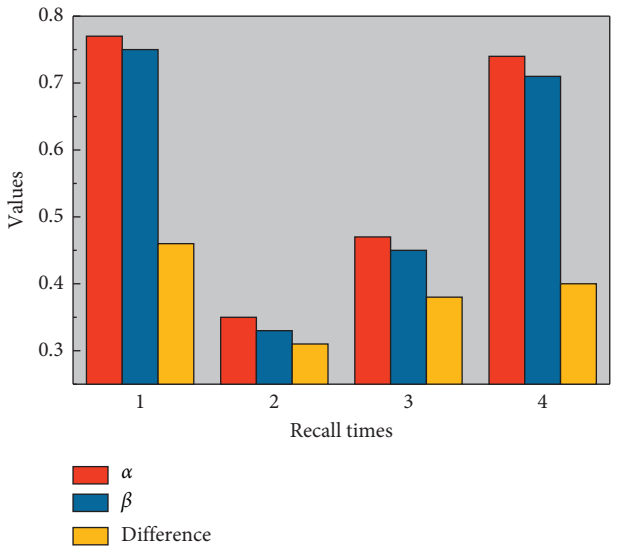

Figure 7: Recall of experimental results with different values of α and β.

accurately and reasonably. In addition, the text structure information is also considered using the hierarchical pooling operation of sentence vectors. Therefore, the experimental structure on the four datasets in this experiment is better than other text similarity calculation algorithms, with higher computational accuracy and better model performance on the whole dataset. Since the algorithm also sets the optimal ratio of key text feature terms to improve the efficiency of the algorithm operation while holding as much text information as possible.

6. Conclusion

The most direct manifestation of whether an English quiz system can achieve intelligence is whether it can accurately answer the questions asked by users. This paper introduces the model of English intelligent quiz system and the functions and roles of each module of the model by studying the English sentence similarity algorithm and gives the specific system design and development process, and on this basis, we coded and implemented the English intelligent quiz system for elementary schools based on the similarity algorithm. After practical experience and use, the system can not only answer the questions raised by users intelligently and accurately, but also provide some knowledge answers related to the questions raised by users, and the system is fast to find, and the distance-based similarity algorithm improves the checking efficiency of the English intelligent quiz system. Subsequent research will improve the similarity algorithm in order to more accurately and quickly query the content needed by users and further improve the efficiency of the system. Meanwhile, in the process of the actual use of the system, we will continue to collect the number of times users click on the page resources, browsing time and other content to obtain the user's preference information, and adaptively recommend learning resources to users to achieve personalized learning.

Data Availability

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

Conflicts of Interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. D. Nath, S. Jha, J. M. Meena, and S. P. Syedibrahim, "A survey paper on elastic search similarity algorithm," *Asian Journal of Pharmaceutical and Clinical Research*, vol. 10, no. 13, pp. 361–364, 2017.

[2] Y. M. K. Liu, Y. Ma, Z. Gao, Y. Zang, and J. Teng, "Similarity analysis-based indoor localization algorithm with backscatter information of passive UHF RFID tags," *IEEE Sensors Journal*, vol. 17, no. 1, pp. 185–193, 2017.

[3] S. Ozer, "Similarity domains machine for scale-invariant and sparse shape modeling," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 534–545, 2019.

[4] B. A. Moser, "Similarity recovery from threshold-based sampling under general conditions," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4645–4654, 2017.

[5] P. Wang, H. Cai, and L. Wang, "Design of intelligent English translation algorithms based on a fuzzy semantic network," *Intelligent Automation & Soft Computing*, vol. 26, no. 3, pp. 519–529, 2020.

[6] T. Kouchi, Y. Tanabe, E. J. Smit et al., "Clinical application of four-dimensional noise reduction filtering with a similarity algorithm in dynamic myocardial computed tomography perfusion imaging," *The International Journal of Cardiovascular Imaging*, vol. 36, no. 9, pp. 1781–1789, 2020.

[7] C. Li, "Intelligent system for college English listening and writing training," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 10, pp. 121–133, 2018.

[8] X. Li, "The construction of intelligent English teaching model based on artificial intelligence," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 12, no. 12, pp. 35–44, 2017.

[9] B. Yushau, "Perceptions, challenges and the resources of learning mathematics in English for a "mathematically intelligent" bilingual Arab University student tith weak English background," *International Journal of Mathematics Trends and Technology*, vol. 53, no. 6, pp. 453–463, 2018.

[10] J. Cai and Y. Liu, "Research on English pronunciation training based on intelligent speech recognition," *International Journal of Speech Technology*, vol. 21, no. 3, pp. 633–640, 2018.

[11] S. Bi, "Intelligent system for English translation using automated knowledge base," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 4, pp. 5057–5066, 2020.

[12] S. Dong, "Intelligent English teaching prediction system based on SVM and heterogeneous multimodal target recognition," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 6, pp. 7145–7154, 2020.

[13] W. Gao, L. Zhu, Y. Guo, and K. Wang, "Ontology learning algorithm for similarity measuring and ontology mapping using linear programming," *Journal of Intelligent & Fuzzy Systems*, vol. 33, no. 5, pp. 3153–3163, 2017.

[14] Y. Hai, "Computer-aided teaching mode of oral English intelligent learning based on speech recognition and network assistance," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 4, pp. 5749–5760, 2020.

[15] H. Wen, "Intelligent English translation mobile platform and recognition system based on support vector machine," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 6, pp. 7095–7106, 2020.

[16] M. Yin, "Research and analysis of intelligent English learning system based on improved neural network," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 2, pp. 1721–1731, 2020.

[17] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning," *Nature Methods*, vol. 14, no. 4, pp. 414–416, 2017.

[18] W. Song, Y. Li, P. Yin, and J. Du, "Research on the development and application of English intelligent translation human-machine interface system," *Solid State Technology*, vol. 63, no. 1, pp. 2204–2208, 2020.

[19] C. Zhao and D. Jiang, "Study on schema theory applied in college English esp curriculum based on intelligent optimization algorithm," *Solid State Technology*, vol. 63, no. 1, pp. 1950–1956, 2020.