Scholar Publishing Group
Zoology and Animal Physiology
https://doi.org/10.38007/ZAP.2021.020201
ISSN 2789-8369 Vol. 2, Issue 2: 1-12

English Text Similarity Detection Algorithm Based on Animal Algorithm

Shihui Xiang*

Panyapiwat Institute of Management, Nonthaburi, Thailand

**corresponding author*

Keywords: Text Similarity, Similarity Detection, Animal Algorithm, Semantic Unit Division, Part-Of-Speech Space Definition

Abstract: The text similarity detection algorithm has a wide range of applications in the processing of massive natural language text information. Unlike simple and complete repetitive search, the complexity of natural language has caused great difficulties in the calculation of text semantic similarity. The Simhash algorithm does not involve the semantic information of the text, and cannot support the semantic problems of natural language processing such as synonyms and polysemes. Therefore, using the "dimensionality reduction" advantage of animal algorithms in English text processing and the efficiency of the retrieval process, aiming at its inability to recognize semantically similar text content, this paper studies the English text similarity detection algorithm of animal algorithms. Aiming at the shortcomings of simhash in the semantic similarity of text, this paper proposes a semantic code design based on the synonym word forest and context through the study of existing synonym expansion schemes. Based on the comprehensive improvement scheme, a semantic fingerprint generation algorithm incorporating synonym information is proposed, which solves the problem of similar texts that cannot identify replacement synonyms. The experiments in this paper show that after testing the sample data of the algorithm in this paper, under the condition of using k = 3 parameter determination, the accuracy and recall of correct identification are over 77%. In contrast, the two indexes of the traditional simhash algorithm and the word frequency statistical algorithm are only about 70%. This proves that the improved algorithm proposed in this paper has achieved relatively good results for identifying multiple similar modification situations, especially the problem of synonym replacement.

1. Introduction

With the rapid development of the Internet era and the rapid progress of science and technology, more and more information is flooded on network platforms. People use the network to send and receive emails, query information, and post Weibo. The Internet has undoubtedly become an essential part of people's daily communication. According to statistics, 80% of the information on the network platform is in the form of text. Therefore, deep mining and research of text information has very important practical significance. At present, text mining has been intelligent retrieval and

natural language processing. The field has a very wide range of applications, so research on text mining has a high economic value.

Text mining involves text modeling and representation, text word weight calculation, text similarity calculation, text information extraction, application research of potential information in text, etc [1,2]. Among them, the text similarity algorithm is a crucial one in text mining Algorithms are the link between basic research such as text modeling and representation and upper-level application research of text latent information [3]. Cosine similarity lacks in-depth consideration of text length information when measuring the degree of similarity between texts. Cosine similarity is uniformly used in the calculation of similarity between texts of different lengths. It is completely ignored, which greatly restricts the accuracy of similarity calculation [4-5]. Text classification, text clustering problems in information retrieval, related question answering systems, and intelligent retrieval in search engines all require text similarity algorithms to support them. Therefore, the systematic research on text similarity algorithms is perfected. The text characteristics and language characteristics are studied by different text similarity algorithms, which is undoubtedly of great practical significance and application value. As we all know, search engines and question answering systems are currently very popular information retrieval tools. A very important algorithm in search engines and question answering systems is the text similarity algorithm. However, according to the relevant research report on the search engine industry. 33.1% of users clearly stated that it is becoming increasingly difficult to find target information. Users' dissatisfaction with wireless search is as high as 31.3%, and 39.7% of users indicate that they are only marginally satisfied with wireless search [6]. 39.5% of users do not use wireless search. The reason for satisfaction is that the validity of the returned results is too low, and 20.3% of the users are because the returned results are too poor. According to incomplete statistics, when users use the information retrieval system, about one-third of the time is used to screen resources. Therefore, a systematic study of text similarity algorithms is conducted to improve the text similarity calculation. Accuracy has very important economic value and social significance.

Qi Zhang believed that various hashing methods have been proposed to capture the similarity between text, visual, and cross-media information. However, most existing works use the word wrap method to represent textual information. Since different forms of words may have similar meanings, these methods cannot handle semantic similarity of text well [7]. To solve these problems, he proposed a new method, semantic cross-media hashing (SCMH), which uses continuous word representations to capture text similarity at the semantic level and uses depth Deep belief network (DBN) to build the correlation between different models. In order to prove the effectiveness of the proposed method, he evaluated the proposed method on three commonly used cross-media datasets. His experimental results show that the performance of his method is significantly better than existing methods. In addition, the efficiency of this method is comparable to or better than other hash methods [8]. Li believed that string-based similarity measures can take advantage of basic facts in a scalable manner and can operate at an abstraction level, which is difficult to counter at the code level. He introduced ITect, a scalable malware similarity detection method based on information theory. ITect detects the entropy mode of the target file in different ways to achieve 100% accuracy and 90% accuracy, but it can also detect 100% recall. Its accuracy and accuracy on the malware of the combination of Kaggle and VirusShare exceed VirusTotal [9]. C. Wang believed that saliency detection refers to automatically finding the most important target based on human visual characteristics in an unknown scene. In order to improve the accuracy of saliency detection, he proposed a saliency detection algorithm based on robust foreground seeds. First, he used the Harris angle and boundary connectivity algorithm to obtain two different convex hulls. The original target area is determined by the intersection of the convex hulls. Secondly, the similarity detection between the superpixels in the convex hull and the outer edge of the convex hull is performed.

When superpixels are similar to most of the outer edges, remove them to get more accurate foreground seeds. A new graph structure is constructed by using anchor graph to express the relationship between data nodes. Then the foreground seeds and the background seeds are used to sort the manifold, and two different significant results are obtained. Finally, a saliency map is obtained by optimizing a new cost function. His experimental results show that the algorithm he proposed further improves the performance of recall and precision [10].

This paper analyzes and studies the differences between Chinese and English in terms of word meaning expressions. For English texts with different length characteristics, the study of text mainly relies on keywords, word frequency co-occurrence relations, and part-of-speech relations to convey semantic information in thematic expressions. The semantic similarity algorithm of the semantic dictionary is constrained by the capacity of the dictionary during the calculation of text similarity. It proposes to segment the text with the part-of-speech as the identifier, and merge the same-sex keywords into a part-of-speech vector. The weight of each dimension in the keyword space is calculated by the semantic dictionary through the semantic dictionary. In this way, the weights in each dimension in the part of speech space include the semantic relevance of the keywords, and the similarity calculation between texts can be performed. It is reflected by the similarity of part-of-speech space, which is mainly used to compare similarities between short texts.

2. Proposed Method

2.1. Text Similarity Detection Algorithm

(1) Text similarity detection problems and evaluation criteria

The text similarity detection problem is mainly to solve the problem of determining whether there is copying or plagiarism between texts. Generally, the text content is quantified by a certain text processing method, such as generating a text vector or a digital fingerprint, and then according to the corresponding similarity calculation method, the similarity of the text content is reflected by quantifying the similarity between the models [11]. Generally, a percentage result is generated based on the repetition rate of similar content, that is, the text copy ratio. The higher the proportion of copying, the more serious the plagiarism of the article is, and the lower the original possibility is [12-13].

The separation or LCS algorithm calculates the text similarity, that is, the magnitude of the text modification [14-15]. However, changes at the semantic level are much more complicated. The simplest modification is to replace the keywords of the text with synonyms, for example: “Research on large-scale text similarity detection technology” is replaced with “Research on mass file check and detection methods”. Obviously, the contents of the two sentences are almost the same in terms of semantic meaning, but due to the replacement of some of the keywords, it is not possible to directly determine that there is plagiarism in the two sentences. Further, by rephrasing the semantic content of the text and expressing the same content with other grammatical structures, this situation cannot be identified by means of string matching. However, due to different expression habits and the development of machine translation is not yet perfect, there is currently no effective method to detect and identify such plagiarism in interpreting foreign languages [16].

(2) Text similarity detection algorithm based on word frequency

In CDSDG, CHECK, SCAM, and many other systems, related methods based on word frequency statistics are used as the core algorithm for development. The reason is that such algorithms have great advantages in model understanding, processing methods, and improvement scope. Algorithms based on the category of word frequency statistics usually combine the vector space model of the data. As the name implies, the keywords of the text are organized into the model unit vector of the text. By treating the occurrence frequency of each keyword as the weight of the

corresponding unit vector, finally Generate space vector for text. The basic idea is that the frequency of the keywords of the document has different effects on the semantic content of the document. Except for meaningless stop words, the higher the frequency of keywords, the greater the relevance of the main content of the article and the keywords. The more the word can express the semantics of the document [17]. The research focus of this type of algorithm is generally divided into the following two aspects: the construction of text models and the measurement of model similarity.

1) Selection of text model

The process of transforming text from a natural language expression model into a model with a certain feature composition is the first step in text similarity detection. Only by establishing a suitable text representation model, which clearly and accurately reflects the expression of the text, can the content overlap of the text be reflected through the subsequent model similarity calculation [18-19]. In general research, text data feature models mainly include the three classic models of vector space model, Boolean model, and probability model, as well as other extended models derived from these three types of models, such as the extended Boolean model based on set theory. Latent semantic indexing model of vectors, as well as probability-based reasoning network and trust network models.

Among the above models, the vector space model is the most classic and most popular among researchers. The vector space model has been applied in many existing text processing systems, such as SMART text retrieval system, N. Shivakumar and other SCAM systems developed on the basis of relative frequency improvement. The automatic text classification system of the Institute of Computing Technology of the Chinese Academy of Sciences, TREC filtering processing. The model conforms to the semantic logic of text and uses words to embody semantics. By filtering and refining the vocabulary of the document to obtain a keyword set, and then organizing the keywords to obtain a unit vector in a multidimensional space, the unit direction of each dimension represents the information of a keyword in the text, and the direction of extension in this dimension Determined by related weights. The formula is expressed as formula (1), where V(f) represents the space vector of text f, and wn represents the feature vector representation of the nth keyword. In the database, the vector dimension n of the model should take the same value. And ω_n represents the feature weight corresponding to the nth word. Finally, the correlation between the models is used to calculate the similarity between the models, so as to obtain the text similarity. Therefore, the generation of the vector space model mainly depends on the determination of two parts, the filtering of text keywords and the calculation of feature weights.

$$V(f)=w_1(f)\cdot\omega_1+w_2(f)\cdot\omega_2+\cdots+w_n(f)\cdot\omega_n \tag{1}$$

2) Feature keyword screening

Choosing keywords that can reflect the semantics of the text body plays a vital role in the effect of the model expression. Generally due to the difference in text length, the vocabulary of the document is also very relevant, especially for texts with tens of thousands of words. If the entire vocabulary is extracted as the feature vector, the dimension of the model will be very high. Some irrelevant words may also have opposite effects on the theme of the article, or they may be meaningless. According to research by related scholars, 2% -5% of words in general text content are the most suitable keywords for text features. Therefore, the filtering of text feature keywords can effectively avoid “dimensional disaster”, reduce noise, and improve the expression effect of the model.

2.2. Word Segmentation Method

Computers need to recognize and process the English text they want to “read”. The first step is to

split the text into the smallest elements (words), and then proceed to the next step based on the set of words. This technique of recognizing and dividing text into words is called word segmentation. Because English articles are composed of single letters, punctuation, and spaces, the segmentation of the entire article is much less difficult than that of Chinese articles, but there are still some special cases that will affect the word segmentation results of English articles. For example, many words in English are connected by hyphens, and new words formed have completely different meanings (for example, green-house refers to a greenhouse, not a green house). If a word is split according to a hyphen, it will inevitably affect its original meaning, and even affect the understanding of the entire article, causing serious errors. So it is necessary to confirm the correctness of the segmentation in advance.

2.3. Stop Word Filtering Method

In rule-based languages such as Chinese and English, there is a class of words that appear very frequently, but they do not make any contribution to adding semantic information to the text. Such words are collectively called stopwords. For example, “the, to, for, and, on, of” in English, etc. They appear to make the grammatical structure of sentences more accurate, but they are not actually Contains any valid semantic information. These words will appear in any article, and if you count the word frequency of the text, these words will always come out on top. Therefore, it is very necessary to remove such words before data processing. This process of elimination is called stopword filtering. Stopword filtering is usually performed by matching and filtering from an existing stopword list. The whole process is a simple query process, that is, for each word in the word set, query the stop word list to see if there is the same word corresponding to it, and if so, delete the word from the word set. Among them, the query step can be optimized the most. Generally, a hash table, a binary search tree, and the like can be used for query operations. Of course, the lowest time complexity is to use a hash table for searching.

3. Experiments

3.1. Experimental Environment

(1) Hardware environment:
(a) CPU: Lenovo Xcon (R) E3-1231 v3@3.4GHZ;
(b) Memory space: 16.00GB;
(2) Software environment:
(a) Operating system: Microsoft Windows 10 Professional;
(b) Development platform: My Eclipse Professional 2014;

3.2. Experimental Data and Comparative Experimental Design

This paper proposes an English text similarity detection algorithm based on animal algorithms. The following experiments are used to verify its efficiency and performance.

Since there is no authoritatively recognized similarity retrieval data set in Chinese studies, the data of this experiment was obtained by processing SOGOU-T from Sogou Corpus web page data. At the same time, considering the inefficiency of making full-text similar samples and the imbalance of the corpus, this paper mainly verifies the recognition based on sentence-level fingerprints. Therefore, the following processing is performed on the data set: all 17,910 Chinese short texts are divided into text blocks, and 108,154 Chinese sentences are obtained as all experimental corpus samples. Then 150 samples were randomly selected to manually add or delete

part of the content, synonym replacement, complete copying, etc. to do the plagiarism treatment, as the target sample data was mixed into the overall sample. Based on the accuracy of part-of-speech weighted fingerprints and word frequency-based weighted fingerprint algorithms in some randomly selected sample data, the semantic animal algorithm and simhash fingerprint algorithm in this paper have PR curves under different threshold conditions, and the same simhash algorithm and tf-idf The algorithm has three sets of controlled experiments in terms of accuracy and efficiency, to investigate and analyze the actual effect of the algorithm improvement work done in this paper.

3.3. Measurement Criteria

The performance evaluation standard of the detection algorithm is generally described by the PR curve, where P refers to the accuracy rate, and in this experiment, it refers to the ratio of the system to determine the correct result: R refers to the recall rate, which is a measure of the proportion of the system's correct results covering all positive samples , So it's also called recall.

4. Discussion

4.1. Contrastive Analysis of Part-of-Speech Weighted Semantic Fingerprint and Part-Frequency Weighted Semantic Fingerprint Algorithm

In order to illustrate the impact of adjusting fingerprint feature weights on the accuracy of fingerprint matching, in addition to the experimental processing steps involved in the algorithm, the rest will be controlled by a unified processing flow to ensure the control of the variable environment. Text preprocessing,

Under the same conditions of word segmentation tools and statistical calculation formulas, compare the differences in experimental results in certain aspects. This experiment mainly examines the adjustment of feature weights in the semantic fingerprint generation algorithm, instead of common word frequency parameters, and the introduction of keyword part-of-speech as a measurement index. Whether this improvement can improve the similarity detection of semantic fingerprints. Under different Hamming distance judgment thresholds, the actual accuracy results based on part-of-speech weights and word-frequency weights are shown in Table 1, and Figure 1 is drawn based on the data in the table.

Table 1. Comparison of the accuracy of judgment of fingerprints at different thresholds k based on the two weights of part-of-speech and word-frequency (%)

Hamming distance judgment threshold(k)	Part-of-speech weights	Word frequency weight
1	39.5	37.5
2	61.4	59.4
3	78.6	76.6
4	85.9	83.9
5	88.1	86.1
6	89.3	87.3
7	90.4	88.4
8	91.8	89.8
9	92.4	90.4
10	92.6	90.6

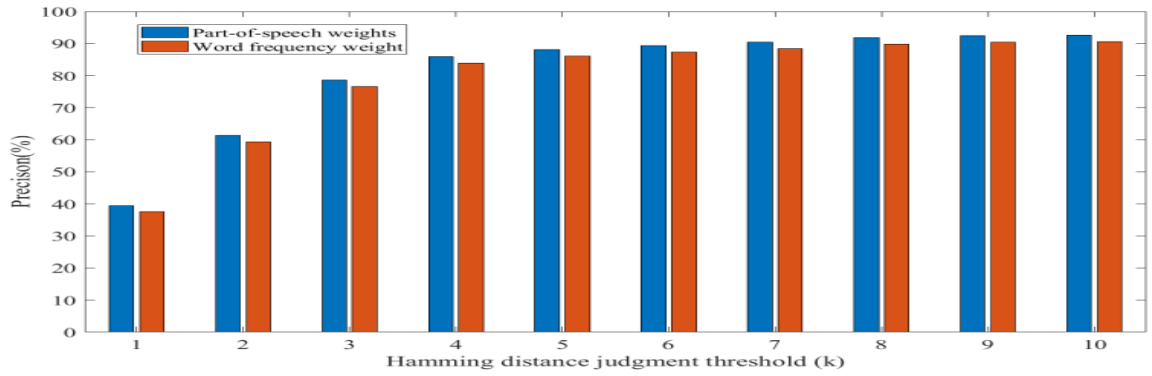

Figure 1. Accuracy curves based on two weights of word frequency and part of speech at different values of k

As shown in Figure 1, the abscissa corresponds to the distribution of the Hamming distance threshold k in the table, and the ordinate measures the accuracy. The black curve indicates the recognition accuracy of semantic fingerprints based on weights under different thresholds. It can be intuitively found that the use of part-of-speech as the semantic fingerprint weighting has a certain improvement in accuracy compared with the semantic frequency recognition based on word frequency weights, but because of the stop word processing in the sentence, it is easy to remove other parts of speech. The remaining features are mostly nouns and verbs, which are not very distinguishable. In the same way, due to the limitation of the length of the sentence, the frequency discrimination between keywords is not obvious. Therefore, the improvement of the part-of-speech weights has a certain effect, but it is not significant.

4.2. Comparative Experiments with Word Frequency Simhash Algorithm and Tf-Idf Algorithm

The improvement of animal algorithms in this paper is mainly based on the design of the calculation process of semantic coding, in addition to the previously verified effect of adjusting the part-of-speech weights. Therefore, the control variable verification experiment is also performed here to compare with the word frequency simhash algorithm based on the conventional algorithm. Since the tf-idf algorithm does not involve the calculation of the Hamming distance, the first step of the experiment is only compared with the simhash algorithm. At the same time, find the most reasonable threshold setting, use this threshold parameter, and compare it with the tf-df algorithm. The experimental results are shown in Table 2 and Figure 2.

As shown in Figure 2, the meaning of the coordinates is the same as in the previous experiment, where SF represents the semantic fingerprint proposed in this paper. Through this comparison, the fingerprints proposed in this paper have greatly improved the recognition accuracy and recall compared with the original simhash algorithm. It shows that the improvement of synonym expansion coding in this paper is relatively effective in improving the effect of digital fingerprint detection. And it can be clearly seen that when the threshold is set to 3, both algorithms reach a higher level on the two indicators of P and R. Therefore, in the following comparison with the tf-df algorithm, the algorithm in this paper, The simhash algorithms all use 3 as the threshold parameter to compare the effects.

As shown in Table 3, the comparison results of the three algorithms under the same experimental environment and relatively optimal parameter selection. After testing the sample data, the algorithm in this paper uses the parameter determination condition of k = 3, and the accuracy and recall rate of identification are over 77%. In contrast, the two indexes of the traditional simhash algorithm and

the word frequency statistical algorithm are only about 70%. This proves that the improved algorithm proposed in this paper has achieved relatively good results in identifying multiple similar modification situations, especially the problem of synonym replacement, and to a certain extent makes up for the lack of recognition ability of simhash algorithm for similar processing. However, due to the deviation between the manually modified target samples and the corpus information, there is still a lot of room for improvement in the algorithm in improving the accuracy and recall.

Table 2. Comparison of the accuracy and recall of the algorithm and word frequency simhash algorithm under different thresholds (%)

Hamming distance judgment threshold(k)	Text similarity detection algorithm in this paper		Word frequency simahash algorithm	
	Accuracy	Recall rate	Accuracy	Recall rate
1	39.5	85.7	36.8	82.1
2	61.4	81.7	56.2	77.3
3	78.6	77.6	71.8	71.2
4	85.9	73.5	79.6	65.7
5	88.1	65.2	84.1	59.4
6	89.3	57.4	87.1	51.1
7	90.4	52.1	88.8	47.4
8	92.1	47.6	89.6	44.3
9	92.4	43.8	90.5	41.7
10	92.6	41.1	91.2	39.5

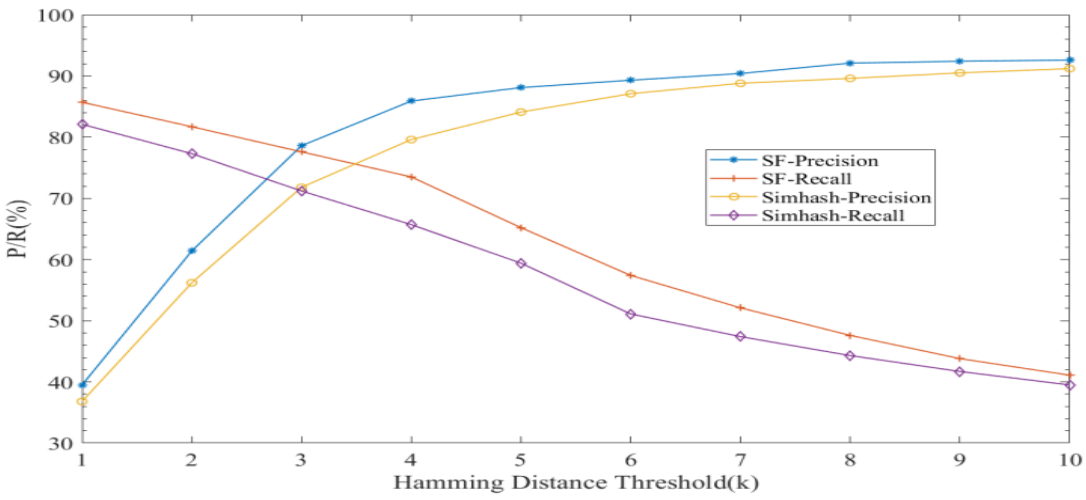

Figure 2. PR curve of semantic fingerprint and word frequency simhash in this paper

On the other hand, in terms of time efficiency, this algorithm has certain advantages over other algorithms. The matching time of the digital fingerprint algorithm is shorter than the vector cosine calculation process of the tfidf algorithm. For the first two fingerprint algorithms, the fingerprint processing time is not much different, and the difference in overall processing time is mainly reflected in the fingerprint matching process. Compared with the simhash algorithm, the algorithm in this paper is further segmented, and the position information is added to reduce the part of the redundant comparison calculation process. The actual running time of the experiment is reduced, which is consistent with the expected results of the theory.

Table 3. Comparison of the results of this paper's algorithm, simhash algorithm, and tf-idf algorithm

Detection algorithm	Accuracy(%)	Recall rate(%)	F1 value(%)	Detection time/ms
Algorithm	78.6	77.6	78.1	1573
Simhash algorithm	71.8	71.2	71.5	1804
Tf-idf algorithm	70.8	69.5	70.1	2156

4.3. Analysis of Word Similarity Experiment Results

(1) Analysis of Yihara similarity calculation experiment

According to Hownet, words are expressed by one or several concepts, and concepts are represented by a series of meanings. Therefore, Yihara similarity is the basis of word similarity calculation, so this article first verifies the Yihara similarity algorithm. The root node is Yoshihara's "Entity", and other Yoshiharas are descendants of "Entity". This article defines the depth of the root node "Entity" as 0, and the depth of Yoshihara "All Things" and "Space" is 1. Analogy; at the same time, the density of Yihara "Entity" is 0, then the density of Yihara "All Things" and "Space" is 1.

Each parameter used in the Yihara similarity calculation formula should have been obtained through a large number of investigations and repeated experiments, but in order to save time and cost, the experimental parameter values were used in this experiment. The results of Yihara's similarity experiment are shown in Table 4.

Table 4. "Yihara" similarity experiment results

Serial number	Yoshihara	Method one	Method two
1	"Everything"and "Space"	0.444	0.322
2	"Beast"and "Human"	0.444	0.444
3	"Tree" and "Flower"	0.444	0.500
4	"Beast" and "Beast"	0.615	0.653
5	"Substance" and "Everything"	0.615	0.477

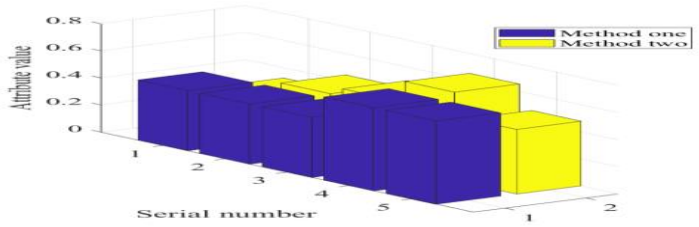

Figure 3. "Yihara" similarity experiment comparison chart

As shown in Table 4 and Figure 3, the first three groups of Yihara similarity calculated by method 1 are the same, but it is obvious that the first three groups of Yihara similarity are different. Because the method takes into account the relative distance between Yoshihara, and does not consider the influence of Yoshihara density and depth on Yoshihara's similarity. The similarity of the second group of Yihara calculated by method two is greater than the first group of Yihara similarity. According to the characteristics of Chinese vocabulary, the similarity of the second group of Yihara is indeed greater than that of the first group of Yihara. Because the depth of the second group of Yihara is greater than the depth of the first group of Yihara, the deeper the depth of Yihara in the Yihara hierarchy tree, the more specific the concept it expresses.

(2) Experimental analysis of word similarity calculation

This article uses four different word similarity algorithms to calculate word similarity, and

conducts experimental comparison:

Method 1: Use the first independent meaning similarity as the word similarity;

Method 2: Word similarity algorithm;

Method 3: Hownet's word similarity algorithm;

Method 4: The text similarity algorithm in this paper.

Table 5. Word similarity experiment results

Serial number	Words 1	Words 2	Method one	Method two	Method three	Method four
1	Classmate	Friend	1.000	0.834	0.800	0.579
2	Man	Woman	1.000	0.668	0.861	0.806
3	Treasure	Gem	1.000	0.623	0.130	0.383
4	Analysis	Study	0.802	0.753	0.444	0.735
5	Invention	Create	0.900	0.873	0.615	0.397
6	Beautiful	Ugly	1.000	0.914	0.815	0.741
7	Doctors	Patient	1.000	0.879	0.665	0.652
8	Chicken	Duck	1.000	1.000	1.000	1.000
9	Apple	Table	0.211	0.209	0.111	0.107

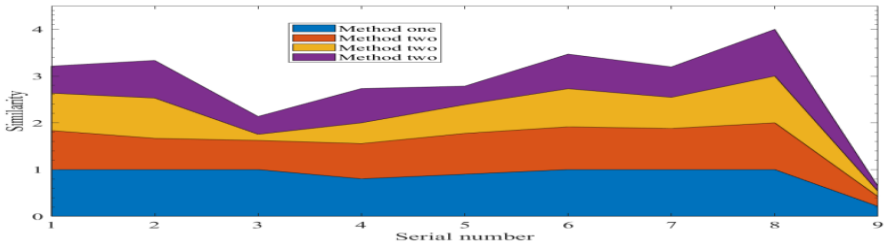

Figure 4. Comparison of experimental results of word similarity

As shown in Table 5 and Figure 4, Method 1 simply uses the similarity of the first independent meaning as the word similarity. The algorithm is too rough. As long as the first independent meaning is the same, the two words are the same, that is, the similarity is 1, this does not meet the characteristics of Chinese vocabulary. Method 2 is a bit more detailed than the calculation algorithm in Method 1. The relevant index coefficients are added to the word similarity calculation. However, the similarity of the sixth group of words is too large, which is not consistent with human understanding of Chinese words, and the calculation results are not accurate enough. Method three is more accurate than method two, because method three takes into account the similarity of four types of meanings and the similarity of concepts in the calculation of word similarity. The calculation result of method 4 (word similarity algorithm in this paper) is more consistent with human understanding of Chinese words, and the calculation result is more accurate than the other three methods. Because the word similarity algorithm in this paper incorporates semantic originality density and semantic depth, the influence of semantic original primary and secondary restraint relationship on word similarity is considered.

5. Conclusion

This article proposed a prototype system based on the core of animal algorithms, and carried out comparative experiments with other related algorithms on the accuracy, recall and performance of the improved algorithm through the system processing flow and the processing of the corpus. According to the analysis of comparative experimental results, it is found that the improved algorithm proposed in this paper can not only achieve rapid matching and localization of similar

texts, but also perform better than other algorithms in terms of accuracy and recall. It is fully proved that the semantic extension coding and weight adjustment based on animal algorithms, and the improvement of the multi-level segmentation index based on location information proposed on the segmentation index are reasonable and effective.

In this paper, the semantic density factor and semantic depth factor of Yihara are added in the Yihara similarity calculation. Summarizing the word similarity experiments shows that the improved word similarity algorithm calculates results more accurately and is more semantically consistent with people's understanding of words. The word similarity algorithm in this article is based on the Hownet dictionary. However, a large number of network words and new terms are not included in the dictionary Hownet. Therefore, the similarity calculation of this part of the word in this article cannot be performed temporarily. However, considering the rigor of academic papers, in general, such network vocabulary rarely appears in academic papers, so the English text similarity algorithm in this paper can be used for similarity detection of papers.

The paragraph similarity in this paper is obtained by taking the average value of the sequence combination of the maximum sentence similarity. According to the paragraph similarity experiment, it can be known that the calculation result of the algorithm is accurate and reliable. The paragraph similarity depends on the similarity of each sentence that composes it, which is consistent with human interpretation of English paragraphs. Therefore, the English text similarity algorithm in this paper can be used for text similarity detection. The similarity of the English text in this paper is obtained by taking the average value of the sequence combination of the maximum paragraph similarity after incorporating the position weight. According to the text similarity experiment, we can see that the algorithm works well. According to the fixed structure of academic papers, different weights are given to different paragraph positions, similar to the overall grasp of English papers by humans. Therefore, the English text similarity algorithm in this paper is more accurate and reasonable, and can be applied to the similarity detection of the paper.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] *Sara Muslih Mishal , Murtadha M. (2022). Hamad, Text Classification Using Convolutional Neural Networks, Fusion: Practice and Applications, 7(1), pp. 53- 65 https://doi.org/10.54216/FPA.070105*

[2] *Cao, J., van Veen, E. M., Peek, N., Renehan, A. G., & Ananiadou, S. (2021). EPICURE: Ensemble Pretrained Models for Extracting Cancer Mutations from Literature. In 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS) pp. 461-467. https://doi.org/10.1109/CBMS52027.2021.00054*

[3] *M. I. Schlesinger, E. V. Vodolazskiy, & V. M. Yakovenko.(2016). "Fréchet Similarity of Closed*

Polygonal Curves", International Journal of Computational Geometry & Applications, 26(01), pp.53-66. https://doi.org/10.1142/S0218195916500035

[4] Daniel Lamprecht, Kristina Lerman, Denis Helic, & Markus Strohmaier.(2017). "How the Structure of Wikipedia Articles Influences User Navigation", New Review of Hypermedia & Multimedia, 23(1), pp.29-50. https://doi.org/10.1080/13614568.2016.1179798

[5] Henning Pohl, Christian Domin, & Michael Rohs.(2017). "Beyond just Text: Semantic Emoji Similarity Modeling to Support Expressive Communication", Acm Transactions on Computer Human Interaction, 24(1), pp.1-42. https://doi.org/10.1145/3039685

[6] Ward Peeters, John Linnegar, Marilize Pretorius, & Marina Vulovic. (2017). "Review of Weideman, Albert (2017) Responsible Design in Applied Linguistics: Theory and Practice", English Text Construction, 10(1), pp.179-185. https://doi.org/10.1075/etc.10.1.10pee

[7] Ahmed A. Elngar , Mohamed Arafa , Amar Fathy , Basma Moustafa , Omar Mahmoud , Mohamed Shaban , (2021). Nehal Fawzy, Image Classification Based On CNN: A Survey, Journal of Cybersecurity and Information Management, 6(1), pp. 18-50 https://doi.org/10.54216/JCIM.060102

[8] Qi Zhang, Yang Wang, Jin Qian, & Xuanjing Huang.(2016). "A Mixed Generative-discriminative Based Hashing Method", IEEE Transactions on Knowledge & Data Engineering, 28(4), pp.845-857. https://doi.org/10.1109/TKDE.2015.2507127

[9] Li, X.(2016). "Method for Semantic Similarity Detection in English Based on Ontology", Journal of Computational & Theoretical Nanoscience, 13(12), pp.9464-9468. https://doi.org/10.1166/jctn.2016.5866

[10] C. Wang, Y. Fan, & B. Li.(2017). "Saliency Detection Based on Robust Foreground Selection", Journal of Electronics & Information Technology, 39(11), pp.2644-2651.

[11] Abhijit Saha , Arnab Paul, (2019). Generalized Weighted Exponential Similarity Measures of Single Valued Neutrosophic Sets, International Journal of Neutrosophic Science, 2019(II), pp. 57-66. https://doi.org/10.54216/IJNS.000201

[12] Alexandros Belesiotis, Dimitrios Skoutas, Christodoulos Efstathiades, Vassilis Kaffes, & Dieter Pfoser.(2018). "Spatio-textual User Matching and Clustering Based on Set Similarity Joins", Vldb Journal, 27(10), pp.1-24. https://doi.org/10.1007/s00778-018-0498-5

[13] Lu-Fang Lin.(2016). "The Impact of Video-based Materials on Chinese-speaking Learners' English Text Comprehension", English Language Teaching, 9(10), pp.1. https://doi.org/10.5539/elt.v9n10p1

[14] Lee, Y. J. (2017). "First Steps Toward Critical Literacy: Interactions with an English Narrative Text Among Three English as a Foreign Language Readers in South Korea", Journal of Early Childhood Literacy, 17(2), pp.45-55. https://doi.org/10.1177/1468798415599048

[15] O'Grady, Gerard.(2016). "Given/new: What do the Terms Refer to?: a First (small) Step", English Text Construction, 9(1), pp.9-32. https://doi.org/10.1075/etc.9.1.02ogr

[16] Margaret Berry.(2016). "Dynamism in Exchange Structure", English Text Construction, 9(1), pp.33-55. https://doi.org/10.1075/etc.9.1.03ber

[17] Isa Abdullahi Baba, & Evren Hincal.(2017). "Global Stability Analysis of Two-strain Epidemic Model with Bilinear and Non-monotone Incidence Rates", European Physical Journal Plus, 132(5), pp.208. https://doi.org/10.1140/epjp/i2017-11476-x

[18] Xiunan Wang, & Xiao-Qiang Zhao.(2017). "A Climate-based Malaria Model with the Use of Bed Nets", Journal of Mathematical Biology, 77(30), pp.1-25. https://doi.org/10.1007/s00285-017-1183-9

[19] Bin-Guo Wang, Wan-Tong Li, & Zhi-Cheng Wang.(2016). "A Reaction–diffusion Sis Epidemic Model in an Almost Periodic Environment", Zeitschrift Für Angewandte Mathematik Und Physik Zamp, 66(6), pp.3085-3108. https://doi.org/10.1007/s00033-015-0585-z