Journal of Organizational and End User Computing
Volume 35 • Issue 1

Similarity Discriminating Algorithm for Scientific Research Projects

Chong Li, Computer Network Information Center, Chinese Academy of Sciences, China
Jinjie Zhang, University of Chinese Academy of Sciences, China
Anyu Wang, University of Chinese Academy of Sciences, China
Xuemin Liu, Computer Network Information Center, Chinese Academy of Sciences, China*
Yunchsun Sun, Beijing Normal University, China
https://orcid.org/0000-0001-6064-3380
Shibo Zhang, Computer Network Information Center, Chinese Academy of Sciences, China
Zhixia Ji, Computer Network Information Center, Chinese Academy of Sciences, China
Justin Z. Zhang, University of North Florida, USA
https://orcid.org/0000-0002-4074-9505

ABSTRACT

An enormous challenge for project management is to identify similar research projects accurately and efficiently among numerous proposals. To address this challenge, this paper proposes an algorithm to calculate the similarity between research projects using an improved generating method for fused word order sentence vectors based on USIF (unsupervised random walk sentence embeddings). The experimental results show that the proposed algorithm is about 15.8% more accurate than the existing approaches. The authors also propose a pre-checking algorithm by introducing a complex research cooperation graph to enhance query efficiency. The results show the pre-checking method reduces the query time cost by 96% on average.

KEYWORDS

Entity Embedding, Scientific Research Cooperation Network, Sentence Vector Generation Algorithm, Similarity Discrimination Algorithm

INTRODUCTION

Checking the similarity of a scientific research project to other projects is the first step in determining whether it is worthy of funding. According to statistics, the duplication rate of research projects in China is 40%(Zhang et al., 2011). The repeated scientific research projects have caused a waste of scientific research resources and affected the national scientific and technological layout.

DOI: 10.4018/JOEUC.332008
*Corresponding Author

Research project similarity discrimination algorithm is a comprehensive technology involving natural language processing, knowledge graphs, information retrieval, and other fields. Combining multi-domain knowledge and research project data helps screen existing research projects similar to those in applications (including similar research contents, research objects, and research objectives), providing a reference for reviewers and funding agencies. Current research in scientific project similarity discrimination mostly focuses on keyword extraction, text similarity calculation, and project clustering, ignoring the correlation relationships embedded in the data. There are still some deficiencies in the model design, accuracy, and query efficiency of the algorithm.

To address the above problems, this paper conducts research based on the data of completed projects and project results of the National Natural Science Foundation of China. To improve the accuracy of scientific research project similarity discrimination, we propose a method for generating fused word order sentence vectors (IUFWO) based on improved Unsupervised Random Walk Sentence Embeddings (USIF). This method can improve the semantic characterization ability of USIF by introducing part-of-speech weight and position weight and integrating word order features into sentence vectors. Based on IUFWO, this paper designs a new research project similarity calculation method. This method judges the similarity of scientific research projects by the weighted sum of cosine similarity between the project name, abstract, keywords, and the conclusion summary of scientific research projects and improves the accuracy of the similarity.

Projects submitted by scholars with close cooperation are usually more likely to be similar or duplicates. From the perspective of the query efficiency of degree discrimination, the project cooperation relationship information between scholars and entities is extracted to construct a scientific research cooperation network, which is the basis for the scientific research project similarity discrimination algorithm. This algorithm prioritizes checking for duplication of projects where a collaborative relationship exists between participants. The experimental results show that the improved sentence vector generation method is about 16% higher than the TF-IDF weighted method so that the sentence vector can more accurately express the semantics of the text. The similarity calculation method of scientific research projects makes the similarity judgment results more discriminative. Compared with the calculation method of the average similarity of each content item, it is improved by about 15.8%. The similarity discrimination algorithm of scientific research projects based on a scientific research cooperation network makes the detection process more targeted. When there are repeated projects among related scholars, the troubleshooting time is shortened by 96% on average, which improves the efficiency of large-scale checking.

RELATED WORK

Research Project Similarity Calculation

Similarity calculation of scientific research projects is a very important process in science and technology management. The calculation of scientific research project similarity is to confirm whether there is a project that is highly similar to or the same as the research content of the project to be checked through text mining, information retrieval, and other technologies from the existing project data.

There are few studies on similarity algorithms for research projects out of China. In countries which starts early in deep learning, natural language processing, data mining, and other fields, a lot of research and exploration has been conducted for duplication detection in academic papers, patents, and code. For instance, Burrows et al. (2007) proposed a code plagiarism detection algorithm based on text similarity measurement and local alignment to solve the problem that students often use other people's code without authorization. This method is highly scalable and has a similar effectiveness with the popular JPlag and MOSS. Kharat et al. (2013) built a plagiarism detection system for research papers based on TF-IDF and LSI to detect plagiarism semantically. In order to detect cross-lingual plagiarism, Ehsan et al. (2016) designed a cross-language plagiarism detection

algorithm based on topic keywords. This algorithm uses a topic-based segmentation method to convert suspicious documents into a group of related paragraphs and uses a proximity-based model to retrieve documents with the best-matching paragraphs. Park et al. (2012) developed a patent technology similarity detection algorithm based on SAO (Subject-Action-Object), aiming at the defect that keyword vectors cannot reflect specific technology discovery and the structural relationship between technology components. They described the structural relationship between technology components in patents through the SAO, used WordNet to calculate the semantic similarity between them, and applied multi-dimensional scaling to map the relationship between patents to two-dimensional space. Arabi et al. (2022) proposes two methods to identify Extrinsic plagiarism, using pre-trained network technique of words embedding FastText and TF-IDF weighting technique to form two structure and semantic matrices and calculate two similarity values.

In China, there have been some research works on similarity algorithms for scientific research projects, and all of them have achieved certain results. For example, Zhao et al. (2015) proposed a research project similarity calculation method based on the semantic understanding of HowNet and TF-IDF, which calculates the semantic similarity between words to calculate the similarity between texts by HowNet. Lin (2013) suggested a similarity calculation method based on an item knowledge representation model to calculate the similarity between scientific research items by word weighting and character matching. Zuo (2010) developed a similarity calculation method for research projects based on non-Segmentation techniques. He constructed a vector space model for research projects by using a suffix tree, mining the frequent closed item sets in the tree, and calculating the similarity of project applications based on this vector space model. Liu (2015) created a method to identify unlisted words based on the frequency-directed network of word sequences to extract professional terms in project applications, constructed a vector space and graph model based on the extracted unlisted words as project feature words to represent project applications, and calculated the similarity of project applications based on this model.

Scientific Cooperation Network

Katz and Martin (1997) define scientific research cooperation as a process in which researchers work together to generate new scientific knowledge. However, in cooperation, only some of the actual contributions can be quantified, while many other aspects cannot. Researchers in the signature of the scientific research project and coauthors in the paper produced by the project are usually regarded as the cooperators of the project. Kretschmer (1997) researched the cooperative relationship network and proposed the concept of the co-authorship network. He defined the cooperation network as a network formed by scholars through cooperation or co-authorship, taking papers, patents, projects, and other scientific research achievements as carriers. Using computer databases of scientific papers in physics, biomedical research, and computer science, Newman (2001) constructed networks of cooperation between scientists in each of these disciplines. He used vertices to represent scientists and edges to represent their cooperation. He also studied a variety of statistical properties of the cooperation network. Based on the above network, Newman (2004) proposed a measure of this strength based on the number of papers coauthored by pairs of scientists and the number of other scientists with whom they coauthored those papers to solve the problem that simple networks cannot capture the variation in the strength of collaborative ties. Through the combination of this method and the cooperative network, he proposed a variety of ways to answer the question, "Who is the best-connected scientist?". Analyzing a cooperative network constructed from scientific databases helps us to track the dynamic evolution of the network, and the results and methodologies developed in the cooperative network may help us to systematically study other complex evolving networks (Barabasi et al., 2002). The analysis of cooperation networks provides a good reference for the analysis of cooperation patterns and has been widely used in many fields (Glanzel & Schubert, 2004).

RESEARCH PROJECT SIMILARITY DISCRIMINATION ALGORITHM

This section introduces the similarity discrimination algorithm of scientific research projects based on the cooperation network proposed in this paper. This algorithm focuses on the accuracy and query efficiency of similarity discrimination. First, the sentence vector generation method is improved based on USIF (Unsupervised Random Walk Sentence Embeddings) (Ethayarajh, 2018) algorithm. On this basis, a similarity calculation method for scientific research projects is designed to improve the accuracy of similarity discrimination. Second, the scientific research cooperation relationship between scholars is extracted from the project data and publications linked data to build a scientific research cooperation network, and the graph embedding algorithm is used to generate the scholar entity representation vector, so as to pre-check the projects participated by scholars with high cosine similarity with the current project leader's embedded vector, which improves the query efficiency of similarity discrimination.

Sentence Vector Generation Method Based on Improved USIF Fusion Word Order

Usually, sentence vectors are generated by combining the word vectors in a sentence, and the combination method determines the semantic representational power of the sentence vectors. The common unsupervised sentence vector generation methods, such as direct sum-averaging of word vectors, TF-IDF weighting, cannot gauge the syntactic and semantic roles played by each word in the sentence.

USIF method is an unsupervised sentence vector generation method proposed by scholar Kawin Ethayarajh in 2018, which was awarded the best paper at the ACL conference Repl4NLP workshop and has outperformed some supervised models, such as RNN and LSTM (Ethayarajh, 2018). USIF modifies the random walk model of text generation (Arora et al., 2017) proposed by Arora and other scholars through "smoothing", deduces the probability formula of sentence vector generation, and reduces it to a weighted formula. The USIF approach improves the semantic expression ability of sentences by subtracting common expressions from the set of training sentences on the basis of the word weights based on word frequency.

Using the excellent semantic characterization ability of USIF, this paper proposes a Sentence Vector Generation Method Based on Improved USIF Fusion Word Order (IUFWO). By introducing part-of-speech weight and positional weight and using Doc2Vec to integrate word order features into the sentence vectors, the semantics of the sentences can be more accurately characterized. The specific improvements of the algorithm are as follows.

Incorporate Part of Speech Weight

In a sentence, notional words are the main components of the sentence, which can act as grammatical components independently and imply the basic meaning of the sentence; while function words are words that cannot act as grammatical components independently and have a weaker representation of the meaning of the sentence. The notional words in Chinese mainly contain six types of nouns, verbs, adjectives, pronouns, quantifiers, and numerals. Statistics are made on the distribution of the above six categories of notional words in the text of scientific research projects, and the results are shown in table 1. According to the grammatical rules of Chinese and the distribution of six types of notional words in the text of scientific research projects, this paper sets up the part of speech weights shown in table 2.

Table 1. Distribution of six types of notional words

	Nouns	Verbs	Adjectives	Pronouns	Quantifiers	Numerals
Quantity	174103	112572	24547	178	2347	3329
Percentage	54.91%	35.50%	7.74%	0.06%	0.74%	1.05%

Table 2. Part of speech weight distribution

	Notional Words						Function Word
Part of Speech	Nouns	Verbs	Adjectives	Pronouns	Quantifiers	Numerals	/
Weights	0.4	0.3	0.2	0.1	0.1	0.1	0.0

Incorporate Part of Position Weights

The different positions of words in the paragraph have different contributions to the meaning expression of the text. The research results of American scholar P.E. Baxendale show that the probability that the first sentence or the last sentence of a paragraph is the central sentence of a paragraph is 92%, that is, the words appearing at the beginning or the end of a paragraph can better reflect the meaning of the text (Tian et al., 2021). Therefore, this paper sets the position weight of the word w appearing in the first or last sentences of a paragraph as ps(w)=0.2, and the position weight of word w in other positions as ps(w)=0.1.

Calculation of Word Weights

In combination with the word frequency weight calculated by USIF, the calculation formula of word weight is as follows:

$$wg\left(w\right) = tf\left(w\right) * \left(1 + \ tg\left(w\right) + ps\left(w\right)\right) \tag{1}$$

where tf(w) is the word frequency weight calculated by USIF of the word w, tg(w) is the part of speech weight of the word w, and ps(w) is the position weight of the word w.

Incorporation of Word Order Features

Although the improved USIF can eliminate the meaningless components caused by the weighted average of word vectors to a certain extent, it cannot solve the problem that Word2Vec word vectors do not contain word order features. This paper uses the same corpus set to train Doc2Vec model. The sentence vector generated by Doc2Vec is weighted average with the sentence vector weighted by the improved USIF to integrate the word order features into the sentence vectors.

$$\widehat{c_s} = \frac{\varphi * \widetilde{c_s} + \ \phi * \overset{\vee}{\mathrm{c}}_s}{\varphi + \ \phi} \tag{2}$$

where φ, ϕ are the weighting coefficients, $\overset{\vee}{c_s}$ is the Doc2Vec trained sentence vector of sentence s, $\widehat{c_s}$ is the final sentence vector of the sentence s.

A text is composed of several sentences, so the representation vector of the text is obtained by summing and averaging the sentence vectors of all sentences in the text, as shown in (3).

$$e_t = \frac{\sum_{i=1}^{n} \widehat{c_s}}{n} \tag{3}$$

where s is the sentence in the text t, $\widehat{c_s}$ is the sentence vector of sentence s, n is the number of sentences in text t.

Research Project Similarity Calculation Method

According to the project data, the content items with semantic analysis value in the research projects include the project name, keywords, abstract, and conclusion abstract, so the similarity of the research projects is obtained by using the weighted average of the cosine similarity of the above four content vectors, which is calculated in (4). Since the newly declared project does not have a conclusion abstract, the similarity is calculated by the abstract and the conclusion abstract of the closed project in the database.

$$proj_{sim(i,j)} = \alpha * n_{sim(i,j)} + \beta * k_{sim(i,j)} + *a_{sim(i,j)} + \delta * c_{sim(i,j)} \tag{4}$$

where $n\_sim(i,j)$, $k\_sim(i,j)$, $a\_sim(i,j)$ and $c\_sim(i,j)$ respectively represents the cosine similarity between the project name, keyword, abstract and conclusion abstract of project i and project j, α, β, γ, and δ are the weights of the corresponding items.

To obtain the reasonable weight value of each content item, multiple groups of experiments were carried out based on the completed project data, and similar project data were constructed by means of multi-language translation, synonym replacement, reversing word order, deleting, and adding some content.

The improved sentence vector generation method is used to vectorize the text, and the cosine similarity is used to calculate the similarity between each content item. The similarity of each item pair is defined by the empirical value. Some experimental data are shown in Table 3. Thus, a group of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are obtained, and each data point $x_i = (n\_sim_i, k\_sim_i, a\_sim_i, c\_sim_i)$, y_i are the empirical values of project similarity. The least squares method was used to calculate the weights of each content item. The least squares method is the most commonly used method to solve curve fitting problems, which

Table 3. Similarity of each content item and empirical value of item similarity

Project Pairs	Name Similarity	Keyword Similarity	Abstract Similarity	Conclusion Abstract Similarity	Similarity Empirical Value of Projects
1	0.56	0.53	0.46	0.46	**0.50**
2	0.45	0.49	0.50	0.49	**0.46**
3	0.58	0.55	0.62	0.62	**0.56**
4	0.44	0.52	0.49	0.51	**0.46**
5	0.45	0.51	0.53	0.52	**0.48**
6	0.98	0.95	0.99	0.98	**0.98**
7	0.90	0.98	0.89	0.99	**0.92**
8	1.0	0.91	0.92	0.90	**0.91**
9	0.79	0.90	0.93	0.92	**0.85**
10	0.96	0.94	0.76	0.97	**0.88**

minimizes the error sum of squares by $\sum\left(z_i - y_i\right)^2$ to obtain the function coefficients $\alpha = 0.35$, $\beta = 0.25$, $\gamma = 0.25$ and $\delta = 0.15$, where z_i are the fitted values of the functions.

Research Cooperation Network Construction and Entity Embedding

A research cooperation network is an objective description of the project cooperation relationship between scholars in the field and is the basis for discerning whether there is an association relationship between scholars. In the research project data and project achievement data, it is possible to obtain not only the information of scholars' entities but also the association relationship between scholars and projects. First, the author's information on the paper in the project achievement data can be used to get the coauthor scholars of the paper. Second, according to the project number associated with the paper, we can get the scholars of the joint projects. Finally, by associating the coauthors in the project results with the research projects, we can get the coauthors of each project, and the network structure is shown in Figure 1. To improve the accuracy and query efficiency of the subsequent graph embedding work, this paper attempts to simplify the heterogeneous network into a homogeneous network, only retain the scholar entity, and add an edge between the scholars who jointly participate in the project. The simplified network structure is shown in Figure 2.

There are many complex relationships in the cooperation network of scientific research projects, and if only the cooperation relationships in the graph are used to query the associated scholars, it will not only require massive computational resources, but also consume too much time, and at the same time, the degree of association among scholars is difficult to measure effectively. Therefore, it is necessary to vectorize the scholar entities in the graph and ensure that the embedding vectors of scholars can well represent the network topology, node connections, and node neighborhood features.

This paper uses the Node2Vec (Grover and Leskovec, 2016) algorithm for scholar entity embedding. In the experiment, the walk is more inclined to depth-first, and the embedded vectors are more representative of the "homogeneity" of the network, even if the embedded vectors of scholars who often cooperate are more similar. In addition, in the scientific research project cooperation network, scholars with close cooperation relationships usually have multiple edges. During a random walk, the existence of multiple edges between two nodes is equivalent to weighting the walk probability, so edge weights are not set. The core parameter settings of Node2Vec are shown in table 4.

Figure 1. Research project cooperation network before optimization

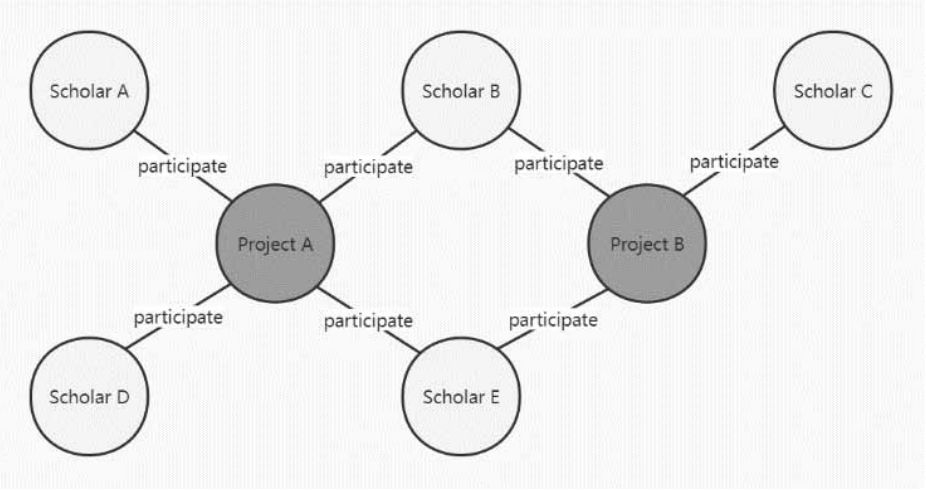


Figure 2. Optimized research project cooperation network structure

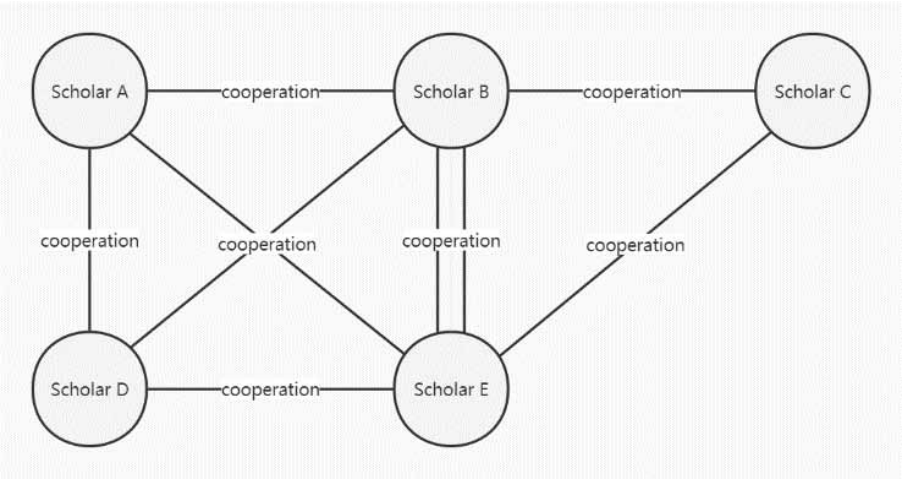


Table 4. Node2vec algorithm core parameter settings

embeddingDimension	inOutFactor	returnFactor	Iterations
50	0.1	0.8	30

Design and Optimization of Discriminant Algorithm Process

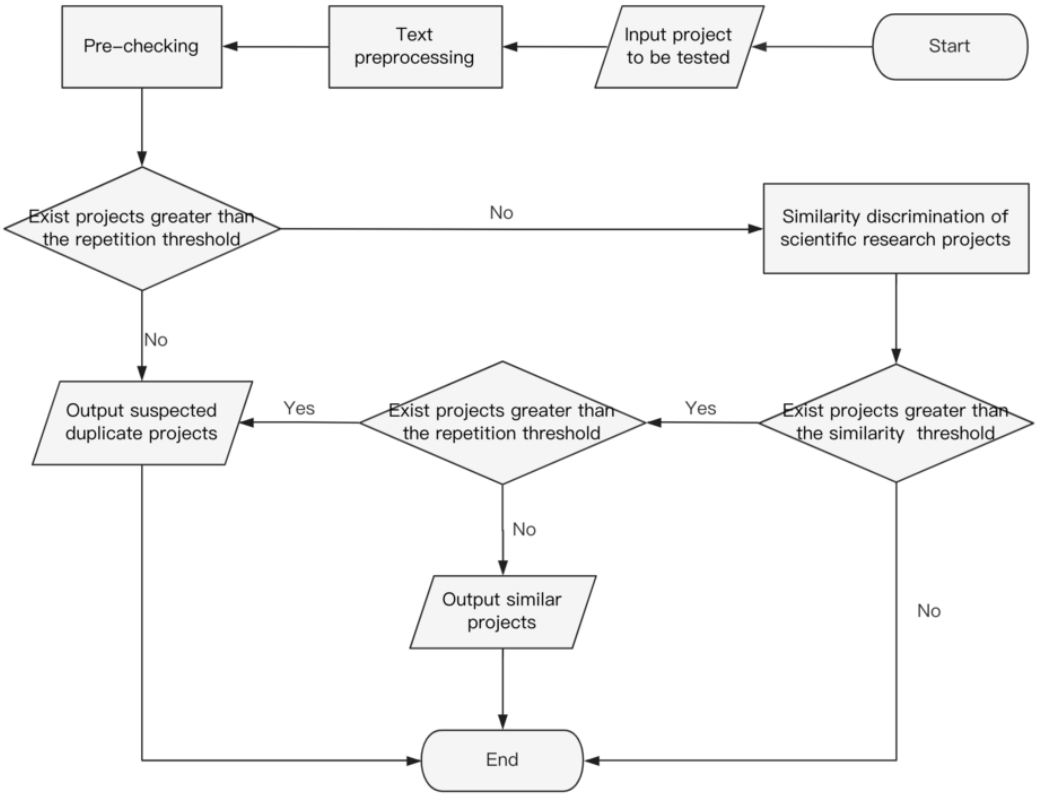
In general, scholars with close project cooperation have similar research fields, and there are often certain connections in the real world (such as the same units or laboratory). The project submitted by them is more likely to have strong correlations and similarities.

Based on the above reasonable assumptions, this paper proposes a scientific research project similarity discrimination algorithm based on the scientific research project cooperation network. When judging the similarity of scientific research projects, the similarity threshold is set to two levels:

1. Repetition threshold
   It is used to screen projects that are suspected of repetition for more rigorous duplicate checking. If a project whose similarity is higher than the repetition threshold is found in the database, it is considered that the target project and the project in the database have repetition in the content of the research object, research method, etc., and may have "repeated declaration" and other problems.
2. Similarity threshold
   It is used to filter projects with similar contents, and to check projects that are related to the target projects in terms of research objects, research methods, etc. By screening similar projects, the project reviewers can be provided with richer and more effective reference information, such as previous funded projects, research progress in related fields, scientific research layout, etc., so as to assist the reviewers in making decisions.

Based on the above two similarity thresholds, the discrimination process is divided into "pre-checking" and "full checking", as shown in Figure 3.

Figure 3. Similarity discriminative process of scientific research projects



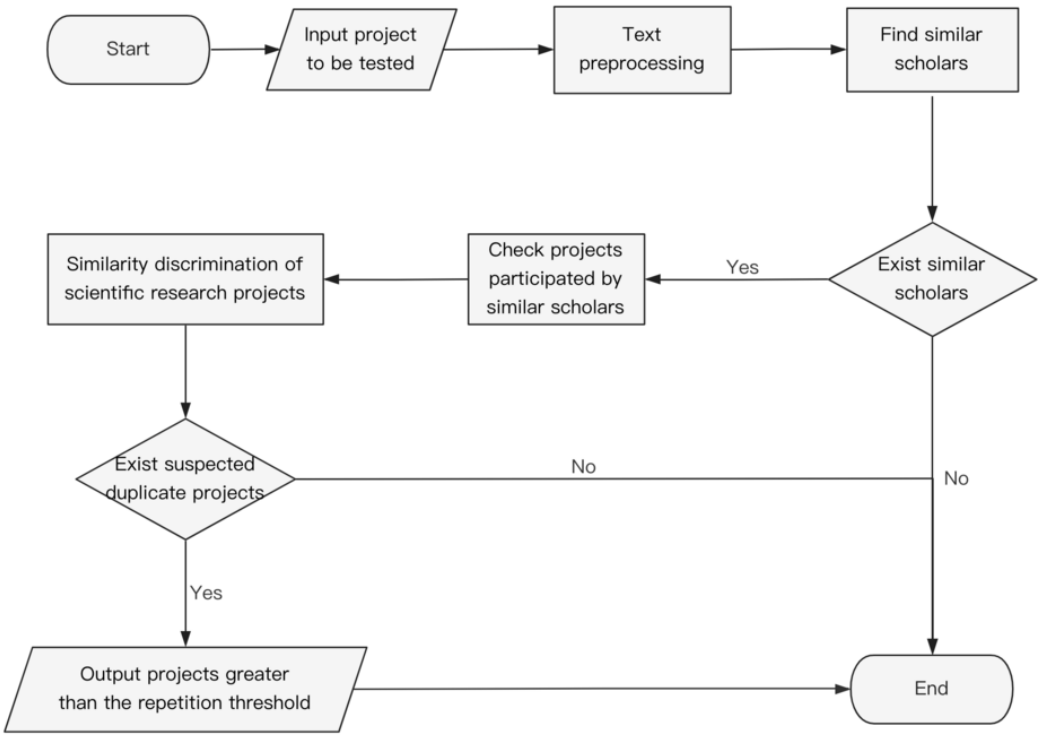
1. Pre-checking
   Calling the pre-checking algorithm based on the cooperation network of scientific research projects to give priority to the screening of projects participated by associated scholars. If there are projects greater than the repetition threshold, you can output suspected duplicate projects directly and end the detection without full checking. The pre-checking process is shown in Figure 4. If there are no suspected duplicate items, enter the full checking process.
2. Full checking
   Eliminating the detected projects during full checking. If there are projects greater than the repetition threshold in the screening, directly output the possible duplicate projects and end the checking. If there are no suspected duplicate projects, output all projects greater than the similar threshold and end the checking.

EXPERIMENT VERIFICATION AND ANALYSIS

Data Sets and Pre-Processing

In the experiment, this paper selects the closed projects and their scientific achievements from 2010-2019 in the Department of Information Science of the National Natural Science Foundation of China as the experimental data, with a total of 29,498 projects and 663,175 related scientific research achievements Project data has multiple column attributes, including project name, type, head of the project, supporting unit, keywords, abstract and conclusion abstract. The project achievements include

Figure 4. Pre-checking algorithm process



papers, monographs, awards and patents, among which papers account for 87%. The attributes of paper data mainly include the title, author, unit, abstract and keywords.

To reduce the ambiguity of duplicate scholar names in the data, this paper uses the DBSCAN (Ester et al., 1996) clustering algorithm based on the co-authorship network to disambiguate the authors with duplicate scholar names, so as to make the author reference items correspond to the author entities one by one. After the name disambiguation, this paper uses the neo4j (Webber, 2012) platform to build a scientific research cooperation network.

In addition, this paper takes the project name, keywords, abstract, conclusion abstract, and the title and abstract of the paper achievements as the corpus. After word segmentation and removal of stop words, Word2Vec (Mikolov et al., 2013) is used for training, and the word vectors obtained from the training are saved as the corpus for subsequent research.

Verification of the Semantic Characterization Capability of Sentence Vector Generation Methods

To verify that the IUFWO sentence vector generation method has certain advantages over other sentence vector generation methods in terms of text semantic characterization ability, multiple sets of experiments are conducted on the project abstract texts selected from the test dataset. The experiment compares four sentence vector generation methods: TF-IDF weighted sentence vector generation method (TW), USIF weighted sentence vector generation method (UW), TF-IDF weighted sentence vector generation method fusion word order (TFWO) and USIF weighted sentence vector generation method fusion word order (UFWO). Some experimental data are shown in table 5.

As shown in table 5, the IUWFO sentence vector generation method generally makes the similarity between dissimilar abstracts lower and the similarity between similar abstracts higher, and there is

about 16% improvement compared to the TF-IDF weighted sentence vector generation method and 9.5% improvement compared to the USIF weighted sentence vector generation method, which can calculate the similarity between items more accurately.

Validation of the Effect of the Similarity Calculation Method for Scientific Research Projects

To verify the effect of the weight setting of the similarity calculation method of scientific research projects on improving the accuracy of similarity discrimination of scientific research projects, this paper uses the above test dataset and compares the calculation methods of weighted average and sum average of each content item based on the IUWFO sentence vector generation method. Some experimental data are shown in table 6. The similarity calculation method of scientific research projects proposed in this paper has better project differentiation, resulting in lower similarity between dissimilar projects and higher similarity between similar projects, which is about 15.8% higher than the sum average similarity calculation method of each content item and can effectively improve the accuracy of similarity discrimination of scientific research projects.

Validation of Pre-Checking Algorithm

A pre-checking algorithm based on scientific research project cooperation network to improve the query efficiency of scientific research project similarity discrimination. In order to verify the effectiveness of the algorithm, this paper obtained the impact of whether to use pre-checking on time consumption through repeated experiments, as shown in table 7. It can be seen from the table that under the premise of linear checking if there are duplicate projects among related scholars, the checking time will be reduced by 96% on average when using pre-checking compared with not using pre-checking, and the larger the data scale in the database, the more obvious the efficiency improvement of pre-checking, which can effectively improve the screening efficiency. When there are no duplicate items among related scholars, or when there are no duplicate items in the database that match the input items, it is necessary to compare the input items with each item in the database to prevent misjudgment or omission. Therefore, the difference between the two methods of duplicate checking is not significant in terms of the maximum duplicate checking time.

Table 5. Similarity of research project based on different sentence vector generation methods

Algorithm	Not Similar Abstract					Similar Abstract				
TW	0.54	0.46	0.63	0.52	0.56	0.94	0.94	0.88	0.87	0.90
UW	0.53	0.51	0.60	0.55	0.53	0.97	0.96	0.91	0.89	0.91
TFWO	0.53	0.57	0.60	0.47	0.52	0.96	0.90	0.88	0.88	0.93
UFWO	0.51	0.53	0.59	0.51	0.51	0.96	0.93	0.90	0.90	0.92
IUWFO	0.50	0.48	0.59	0.49	0.50	0.97	0.95	0.93	0.90	0.92

Table 6. Comparison of the effects of setting weights for similarity calculation methods of scientific research projects

Weight Setting	Not Similar Projects					Similar Projects				
Weighted Average	0.46	0.53	0.49	0.42	0.51	0.87	0.93	0.95	0.87	0.91
Sum Average	0.51	0.54	0.52	0.41	0.57	0.82	0.90	0.93	0.87	0.86

Table 7. Impact of pre-checking on time

Average Time Spent	Find Related Scholars	Duplicate Projects Exist Among Related Scholars	No Duplicate Projects Among Related Scholars
Using Pre-checking	3 seconds	540 seconds	13301 seconds
Not Using Pre-checking	0 seconds	13274 seconds	13274 seconds

Validation of Similarity Discrimination Algorithm for Scientific Research Projects

To verify that the research project similarity discrimination algorithm based on the research cooperation network can accurately and efficiently screen suspected duplicate and similar projects, several experiments were conducted in this paper.

In the experiment, the project repetition threshold was first set to 0.85 based on empirical values, similarity checking was performed in the database, and no duplicate projects were found. To further verify the effectiveness of the discriminative process, duplicate project data were constructed and entered the database using methods such as mutual translation between Chinese and English, synonym replacement, reversing the order of words, and deleting and adding parts of the content. The repetition threshold is adjusted and verified through many experiments. When the project repetition threshold is set to 0.90, all artificially constructed repetition projects can be accurately and completely checked out without other irrelevant projects. Taking project A in table 8 as the test project, the results show that the discrimination algorithm can check out all the artificially constructed repeated items, and some of the checked repeated items are shown in table 8 (project B).

Second, the similarity threshold is set to 0.75 according to the empirical value, and then the similarity threshold is adjusted and verified through many experiments. When the project similarity threshold is set to 0.70, the items associated with the target project can be checked out more accurately, and there are no other irrelevant projects. Similarly, taking project A in table 8 to be tested as an example, the artificially constructed duplicate projects are deleted and checked in the database. Some similar projects found are shown in table 8 (project C and project D). Both project C and project D are related to research on "Underwater Acoustic Communication", involving the application of technologies such as "Underwater Acoustic Channel" and "Doppler", which are related to project A in terms of research objects and research methods. Project C and Project D are not completely identical to Project A, and the algorithm here only determines the possibility of their duplication. Although there is some overlap in research methods, content, and other aspects between them, whether they are duplicated still needs to be determined by project reviewers. Through these related project contents, project reviewers can more accurately and efficiently understand the research progress and scientific research layout related to "Underwater Acoustic Communication", to make more scientific decisions.

CONCLUSION

This paper proposes a similarity discrimination algorithm for research projects based on the research project cooperation network using the data of closed projects and project achievements from 2010 to 2019 in the Department of Information Science of the National Natural Science Foundation of China. To improve the accuracy of similarity discrimination, a sentence vector generation method based on improved USIF fusion word order is proposed, and on this basis, a similarity calculation method for scientific research projects is designed. To improve the query efficiency of similarity discrimination, this paper constructs a research cooperation network based on the cooperation relationship between scholar entities through the rule that projects submitted by scholars with close cooperation are usually more likely to be similar or duplicates. On this basis, a research project discrimination algorithm based on the research cooperation network is proposed. This algorithm prioritizes checking for duplication

Table 8. Examples of repetitive research projects and similar research projects

	Project A	Project B	Project C	Project D
Name	Data acquisition and reliable transmission of underwater sensor mobile node (Original Chinese: 水下传感移动节点的数据获取与可靠传输)	Research on data acquisition and transmission of underwater aircraft sensor (Original Chinese: 水下航行器传感数据采集与传输研究)	Research on underwater acoustic communication equalization based on sparse characteristics of channel (Original Chinese: 基于信道稀疏特性的水声通信均衡研究)	Modeling and parameter estimation of broadband time-varying underwater acoustic channel (Original Chinese: 宽带时变水声信道的建模与参数估计)
Keywords	Underwater acoustic communication; underwater acoustic system; underwater acoustic channel; channel estimation; channel equalization (Original Chinese: 水声通信;水下声系统;水声信道;信道估计;信道均衡)	Channel estimation; channel equalization; underwater acoustic communication; underwater acoustic system (Original Chinese: 信道估计;信道均衡;水声通信;水下声系统)	Underwater acoustic communication; sparse channel; channel equalization (Original Chinese:水声通信;稀疏信道;信道均衡)	Underwater acoustic communication; underwater acoustic channel; channel estimation; broadband nature; time-varying (Original Chinese: 水声通信;水声信道;信道估计;宽带本质;时变)
Abstract	In this research, basic theories and key technologies for underwater acoustic communication of underwater mobile node were studied for important applications of underwater light aircraft sensing and monitoring in order to solve the technical problems of data acquisition and transmission of underwater sensor mobile node. In view of random time-space-frequency variable parameter, multipath, high attenuation rate, low sound velocity and strict bandwidth of underwater acoustic channel as well as small size, high maneuverability and strong self-noise of underwater aircraft platform, the author carried out channel analysis and modeling, channel estimation and equalization, communication signal processing, and communication system building and estimation in order to solve the problems of underwater acoustic mobile channel modeling and its influence mechanism on communication, and multipath-Doppler double extension dynamic channel multi-parameter joint estimation based on……(Abridged) (Original Chinese: 本项目面向水下轻型航行器传感监测的重要应用,通过研究水下移动节点水声通信基础理论与关键技术,解决水下传感移动节点数据获取与传输中面临的技术难题。针对水声信道随机时-空-频变参、强多途、大衰减、低声速和严带宽等特性,以及水下航行器平台小尺度、高机动性、强自噪声等特点,开展信道分析与建模、信道估计与均衡、通信信号处理、通信系统构建与评估等方面的研究工作,解决水声移动通信动态信道建模及其对通信的影响机理、多途-多普勒双扩展动态信道多参数联合估计、基于…)	In this research, key technologies for underwater acoustic communication of underwater aircraft were studied for important applications of underwater aircraft sensing and monitoring in order to solve the technical problems of data acquisition and communication of underwater aircraft sensor. In view of random time-space-frequency variable parameter, high attenuation rate and low sound velocity of underwater acoustic channel as well as small volume, high maneuverability and strong self-noise of aircraft, the author carried out analysis and modeling, communication signal processing, channel estimation and equalization, and communication system building and estimation in order to solve…… (Abridged) (Original Chinese: 该项目是面向水下航行器传感与监测的重要应用,通过研究水下航行器水声通信的关键技术,解决了水下航行器传感数据采集与通信面临的技术难题。针对水声信道时空频变参数随机、衰减大、声速低等特点,以及航行器体积小、机动性强、自噪声强等特点,进行分析建模、通信信号处理、信道估计与均衡、通信系统构建与评估等研究工作,解决…)	On account of urgent need for national defense and civil use and high complexity and variability of underwater acoustic channel, underwater acoustic communication becomes a research focus of the communication field. Underwater acoustic channel equalization technology is a key technology for improving underwater acoustic communication reliability and accuracy. In this research, the author studies equalization technologies suitable for time-varying underwater acoustic channels with ultra-long response and sparse characteristics combined with features of underwater acoustic channel and research results of signal processing and wireless communication. This research will focus on the theories and methods for ultra-long response and sparse characteristics of underwater acoustic channel, channel equalization and channel shortening equalization technologies against multipath delay and Doppler shift, and underwater acoustic MIMO channel equalization technology. Research result……(Abridged) (Original Chinese: 由于国防和民用的迫切需要,以及水声信道极其复杂多变的特性,水声通信成为通信领域的前沿热点研究课题之一。水声信道均衡技术是水声通信中提高通信可靠性和准确性的关键性技术。本项目从水声信道的特征出发、结合声信号处理和无线通信的研究成果,研究适合于具有超长响应和稀疏特性的时变水声信道的均衡技术。本项目将重点研究水声信道超长响应和稀疏特性建模的理论与方法、抗多途时延和多普勒频移的信道均衡和信道缩短均衡技术、水声MIMO信道均衡技术。研究成果…)	Different from onshore time-varying channel, underwater acoustic channel is a broadband time-varying channel. Time-varying Doppler effect cannot be simply expressed and compensated using frequency shift, and its frequency shift will change with frequency and time, which should be described using Doppler scale expansion. It is such Doppler scale expansion that severely restricts performance of underwater acoustic communication receiver. Presently, physical layer technologies for underwater acoustic communication are mainly studied to elevate the capacity and transmission reliability of such time-frequency bi-dispersion broadband channel. This research……(Abridged) (Original Chinese: 与陆上时变信道不同,水声信道是一个宽带时变信道。时变引起的多谱效应不能简单地用频率偏移来表示和补偿,其频率偏移量会随频率大小和时间而变化,这一特点需要使用多普勒尺度扩展来描述。正是这种多谱勒尺度扩展严重限制了水声通信接收机的性能。当前,水声通信物理层技术的研究主要集中在提高这种时频双色散宽带信道的容量和传输可靠性。本课题从…)

continued on following page

Table 8. Continued

	Project A	Project B	Project C	Project D
Conclusion Abstract	(No Conclusion Abstract of the declared project)	Underwater wireless transmission restricts the development Chinese marine information technology and technological breakthroughs of this field are of great significance for advancing Chinese ocean strategy.…… (Abridged) (Original Chinese: 水下无线传输制约了我国海洋信息技术的发展,在此领域实现技术突破对于推进我国的海洋战略具有极其重要的意义。)	(No Conclusion Abstract of the declared project)	To build a maritime power and implement the smart ocean strategy, there is an extensive demand for high-speed high-reliability underwater acoustic communication. Presently, two critical issues urgently need to be addressed for underwater acoustic communication: First, how to conduct high speed communication under low signal-to-noise ratio and limited bandwidth,…… (Abridged) (Original Chinese: 海洋强国与智慧海洋战略的开启,对高速可靠水声通信存在广泛需求。当前水声通信亟待解决两个关键问题:一是如何在低信噪比与有限带宽下实现高速通信…)
Similarity		0.913	0.784	0.743

of projects where a collaborative relationship exists between participants. Experiments show that the improved sentence vector generation improves by about 16% compared with the TF-IDF weighted method and improves by about 9.5% compared with the USIF weighted sentence vector generation method, which can characterize the sentence semantics more accurately. According to the experiment, the research project discrimination algorithm based on the research cooperation network shortens the checking time by 96% on average when there are duplicate projects among the associated scholars, which can screen duplicate projects and similar projects more accurately and efficiently, thus more effectively assisting project reviewers in decision making.

Although the discrimination algorithm proposed in this paper achieves the improvement of accuracy and query efficiency, some shortcomings of the work were found during the experimental process. For the research on the algorithm and model of scientific project similarity discrimination, future work can start from the following aspects, such as further enriching the data used in scientific project similarity discrimination and designing a scientific project similarity discrimination model that supports incremental data updates.

REFERENCES

Arabi, H., & Akbari, M. (2022). Improving plagiarism detection in text document using hybrid weighted similarity. *Expert Systems with Applications*, *207*, 118034. doi:10.1016/j.eswa.2022.118034

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. *International conference on learning representations*.

Barabasi, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*(3-4), 590–614. doi:10.1016/S0378-4371(02)00736-7

Burrows, S., Tahaghoghi, S. M., & Zobel, J. (2007). Efficient plagiarism detection for large code repositories. *Software, Practice & Experience*, *37*(2), 151–175. doi:10.1002/spe.750

Ehsan, N., & Shakery, A. (2016). Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. *Information Processing & Management*, *52*(6), 1004–1017. doi:10.1016/j.ipm.2016.04.006

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD (pp. 226–231). Academic Press.

Ethayarajh, K. (2018). Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP* (pp. 91–100). doi:10.18653/v1/W18-3012

Gao, Q., Wang, L., & Wang, Z. (2021). Research on least square curve fitting and optimization algorithm. *Industrial Control Computer*, *34*, 100–101.

Glanzel, W., & Schubert, A. (2004). Analysing scientific networks through coauthorship. In *Handbook of quantitative science and technology research* (pp. 257–276). Springer.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855–864). doi:10.1145/2939672.2939754

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. doi:10.1016/S0048-7333(96)00917-1

Kharat, R., Chavan, P. M., Jadhav, V., & Rakibe, K. (2013). Semantically detecting plagiarism for research papers. *International Journal of Engineering Research and Applications*, 3.

Kretschmer, H. (1997). Patterns of behaviour in coauthorship networks of invisible colleges. *Scientometrics*, *40*(3), 579–591. doi:10.1007/BF02459302

Lin, J. H. (2013). *Research and application of similarity calculation in science and technology project management system* [Ph.D. thesis]. Hangzhou Dianzi University. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201402&filename=1014229060.nh

Liu, H. N. (2015). *Research on similarity calculation of scientific and technological projects based on Hadoop* [Master's thesis]. Hebei University of Technology. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201601&filename=1015998385.nh

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

Newman, M. E. (2001). Scientific collaboration networks. i. network construction and fundamental results. *Physical Review. E*, *64*(1), 016131. doi:10.1103/PhysRevE.64.016131 PMID:11461355

Newman, M. E. (2004). Who is the best connected scientist? a study of scientific coauthorship networks. In *Complex networks* (pp. 337–370). Springer. doi:10.1007/978-3-540-44485-5_16

Park, H., Yoon, J., & Kim, K. (2012). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics*, *90*(2), 515–529. doi:10.1007/s11192-011-0522-7

Tian, H. P., Ma, B., & Feng, J. (2021). Text similarity calculation based on multi-model weighted fusion. *Computer Engineering and Design*, *042*, 3239–3245.

Webber, J. (2012). A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on systems, programming, and applications: software for humanity* (pp. 217–218). doi:10.1145/2384716.2384777

Zhang, J. L., Huang, C., & Chen, R. H. (2011). Deepening scientific and technological novelty search and expanding social services. *Library Forum*, 130–132, 145.

Zhao, S. J., & Chen, Q. (2015). Item similarity calculation method based on semantics and tf-idf. *Computer Times*, 1–3.

Zuo, C. (2010). *Research and implementation of duplication-checking of scientific and technological projects based on non-word segmentation technology* [Ph.D. thesis]. Chongqing University. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2011&filename=2010216146.nh

Chong Li, Ph.D., senior engineer, senior member of the Chinese Computer Federation. His main research interests include advanced software architecture, recommendation technology, and big data management.

Jinjie Zhang, postgraduate student. His main research interests are big data management, recommendation technology, advanced software architecture, etc.

Anyu Wang, postgraduate student. His main research interests are big data management, natural language processing, advanced software architecture, etc.

Xuemin Liu, M.S., senior engineer. His main research interests include research on key technologies of large-scale information system and application of data intelligence.

Yunchuan Sun is a professor and the director of the Institute of Big data in Finance, Business School, Beijing Normal University. He is an IEEE senior member and acts as the associate Editor of Personal and Ubiquitous Computing, founder of IIKI series events (from 2012-2022). He received his Ph.D. degree from the Institute of Computing Technology, CAS in 2009. His research interests include Fintech, IoT, semantic technologies, and information security.

Shibo Zhang, MA.Sc,software engineer. His main research interests are advanced software architecture, data governance, data analysis, etc.

Zhixia Ji, M.S., senior engineer. Her main research interests include research on key technologies of large-scale information system and application of data intelligence.

Justin Zhang is a faculty member in the Department of Management at Coggin College of Business in University of North Florida. He received his Ph.D. in Business Administration with a concentration on Management Science and Information Systems from Pennsylvania State University, University Park. His research interests include economics of information systems, knowledge management, electronic business, business process management, information security, and social networking. He has published research articles in various scholarly journals, books, and conference proceedings. He is the editor-in-chief of the Journal of Global Information Management. He also serves as an associate editor and an editorial board member for several other journals.