Scientometrics (2022) 127:3343–3366
https://doi.org/10.1007/s11192-022-04389-4

Reviewer recommendation method for scientific research proposals: a case for NSFC

Xiaoyu Liu[1] · Xuefeng Wang[2] · Donghua Zhu[2]

Received: 23 September 2021 / Accepted: 19 April 2022 / Published online: 14 May 2022

Abstract

Peer review is one of the important procedures to determine which research proposals are to be funded and to evaluate the quality of scientific research. How to find suitable reviewers for scientific research proposals is an important task for funding agencies. Traditional methods for reviewer recommendation focus on the relevance of the proposal and knowledge of candidate reviewers by mainly matching the keywords or disciplines. However, the sparsity of keyword space and the broadness of disciplines lead to inaccurate reviewer recommendations. To overcome these limitations, this paper introduces a reviewer recommendation method (RRM) for scientific research proposals. This research applies word embedding to construct vector representation for terms, which provides a semantic and syntactic measurement. Further, we develop representation models for reviewers' knowledge and proposals, and recommend reviewers by matching two representation models incorporating ranking fusions. The proposed method is implemented and tested by recommending reviewers for scientific research proposals of the National Natural Science Foundation of China. This research invites reviewers to provide feedback, which works as the benchmark for evaluation. We construct three evaluation metrics, Precision, Strict-precision, and Recall. The results show that the proposed reviewer recommendation method highly improves the accuracy. Research results can provide feasible options for the decision-making of the committee, and improve the efficiency of funding agencies.

Keywords Reviewer recommendation · Knowledge representation · Word embedding · Scientific research proposal selection · Peer review

JEL Classification O32

Mathematics Subject Classification 68U15

Xiaoyu Liu
Xiaoyu.liu2019@foxmail.com

[1] Department of Management, Beijing Electronic Science & Technology Institute, Beijing 100070, China

[2] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

Introduction

In many countries, the government implemented various S&T policies and strategies to support basic research, promote the development of S&T, and advance the application of S&T. The Chinese government funding supports an increasing number of scientific research projects in recent years. In 2020, the National Natural Science Foundation of China (NSFC) received a total of 276,617 project proposals, 45,656 proposals of which got funding. The huge volume of proposals challenges the efficiency of the reviewer recommendation system. Despite the number of proposals, the quality of peer review would heavily influence the efficiency of fund use. Thus, how to recommend suitable reviewers for each proposal with efficiency and accuracy is a vital task for the funding agencies.

The main task for reviewer recommendation is to provide a list of reviewers who provide a high level of expertise and valuable professional judgments on given proposals (Henriksen & Traynor, 1999; Wang et al., 2005). The candidate reviewers would generate from the principal investigators who have shouldered similar research projects, the researcher who have published related scientific publications, or the nomination by the committee of the funding agency. Their educational background, research interest, publications, and projects are collected to depict their knowledge. The proposal can be represented by the disciplines and its content. The reviewer assignment would rely on manual selection from the academic committee, as well as automatic recommendation. The huge demand for peer review would challenge the efficiency of manual selection, thus, the automatic recommendation becomes a competitive way for reviewer recommendation. Various automatic methods are implemented to recommend the reviewers, including the generative probabilistic model (Balog et al., 2006, 2012), network-based model(Serdyukov et al., 2008; Silva et al., 2013; Xu et al., 2012), voting model (Macdonald & Ounis, 2006, 2008). Machine learning, text mining, and complex networks supply the advanced techniques for reviewer recommendation.

Two key factors influencing the reviewer recommendation accuracy are the representation model of reviewers' knowledge and proposals, as well as the matching between corresponding representation models (Karimzadehgan et al., 2008). Disciplines, research topics, and keywords are considered the main elements for the representation of reviewers' knowledge and proposals (Abdoul et al., 2012; Mirzaei et al., 2019). However, the disciplines may be broad and varying, the claimed research topics are irregular, and the keywords space is sparse. These limitations lead to the inaccurate representation of reviewers' knowledge and proposals. On the other hand, the matching is usually considered as a similarity calculation between representation models, while ignoring the ranking fusion of similarities from multi-systems. The state-of-the-art reviewer recommendation methods would utilize multi-source data to depict the knowledge of reviewers, especially for reviewer recommendations for scientific research proposals(Silva et al., 2013). The applicability of the ranking fusion methods is to be discussed.

To improve the efficiency and accuracy of the reviewer recommendation, this research proposes a reviewer recommendation method (RRM) for scientific research proposals incorporating semantic relationships and ranking fusion. We introduce word embedding techniques to represent terms as continuous low-dimensional vectors. The similarities between different terms can be calculated by the cosine distance, which is a syntactic and semantic measurement. Besides, this model could avoid the sparsity issue when using the vector space model (VSM), and it also has other advantages, such as simplicity, robustness, high accuracy, and low computational cost (Galke et al., 2017; Hu et al., 2018). Based on the word embedding techniques, this research proposes two representation models: the reviewer knowledge representation model integrating their publications and projects and the proposal representation model incorporating the context. Note that the reviewer knowledge representation model contains the scientific publications as well as the projects, the similarities between the above representation models should be calculated separately, i.e., the similarity between the reviewer's scientific publications and the proposal, and the similarity between the reviewer's projects and the proposal. The similarities from different systems are integrated by comparison of three methods, Average, SumRank, and CombRKP. Then, we could generate the ranking list of reviewers for each proposal. The top reviewers are recommended for the proposals, which assist the panel chairs or division managers with the reviewer assignment.

To demonstrate the feasibility and effectiveness of the proposed method, we take the application proposals of NSFC as a case study. NSFC is the largest funding agency in China, and it receives a great number of proposals each year. In 2018, the NSFC proposed the construction of intelligent peer review as one reform plan, aiming to strengthen support for future proposal reviews. We take the proposals of the Science and Technology Management and Policy division (G0404) as a case study. The evaluation of the proposed method is implemented by questionnaires to the reviewers. We take their judgment as the benchmark and propose three evaluation metrics, Precision, Strict-precision, and Recall. Our proposed reviewer recommendation method attempts to provide timely, efficient, and accurate recommended reviewers for each proposal and contribute to the reviewer assignment progress.

The remainder of the paper is organized as follows. The "Literature review" section discusses the related works of expert recommendation, including focuses and methods. The "Proposed RRM" section presents the proposed reviewer recommendation approach for scientific research proposals. The "Case study" section provides the experiment, evaluation metrics, and results for the proposed approach. The "Conclusions" section summarizes the contributions of this research and discusses limitations and future work.

Literature review

Reviewer recommendation is a subtopic of the expert recommendation, while it has distinguishing application scenario. So, we would step back to review the research on the expert recommendation and then focus on the reviewer recommendation for scientific research proposals. The literature review consists of the focuses of expert recommendation, the approaches for the expert recommendation, and the reviewer recommendation for scientific research proposals.

Focuses of expert recommendation

There are mainly five research questions in expert recommendation: (1) what is the demand for expert recommendation; (2) how is one determined to be an expert; (3) how is the expertise model generated; (4) what matching techniques are implemented; (5) how is the performance of the expert recommendation (Lin et al., 2017). This section would discuss the five questions mentioned above.

The demand for expert recommendation varies from reviewer recommendation for scientific research projects, researcher recommendation for collaborations, the expert recommendation for online Q&A communities, supervisor recommendation for postgraduates, enterprise expert recommendation, and so on(Protasiewicz et al., 2016; Wang et al., 2017; Xu et al., 2012; Zhang et al., 2016). The specificity and independence of the domain also vary, corresponding to the demands of expert recommendation. Some expert recommendation approaches consider a wide spectrum of disciplines, e.g., the collaborator recommendation for the online scientific community (Wang et al., 2013; Yang et al., 2015). Other approaches focus on one specific domain or organization, e.g., supervisor recommendation and enterprise expert recommendation (Balog et al., 2006; Campbell et al., 2003; Zhang et al., 2016). Domains within high-level specificity tend to have systematic and detailed information, which helps to map the knowledge of experts and measure the relevance of knowledge demands and candidate experts.

The identification of experts deals with the issue of what kind of capability should an expert own for a specific demand. Stankovic et al. (2010) review the expertise hypotheses for expert search on the web. They classify them into four categories, i.e., “Authorship of High-quality Content”, “Online Activity”, “Real Life Activities and Achievements” and “Reputation and Authority.” However, it is still difficult to nominate candidates for the difficulty in collecting all the information and the trade-offs in choosing the expertise hypotheses (Mockus & Herbsleb, 2002). When the scenario comes into researcher recommendation, some attributes would be taken into consideration, e.g., professional knowledge, innovation performance, practical experience, availability, quality, and quantity of publications. The committees, panel chairs, and managers would nominate some candidates. Besides, outsourcing information would also provide insights into the candidates.

As a core process of the expert recommendation, the expert knowledge representation model would influence the recommendation results decisively. Expert representation modeling starts with the selection of expertise resources. There are mainly two ways for resource selection (Vivacqua et al., 2009). Traditional expert recommendation relies on an expert database, also known as a knowledge catalog or knowledge map. The initial establishment of the database relied on experts to give a detailed description of their expertise. For example, Microsoft’s SPUD system, Hewlett-Packard’s CONNEX system, and SAGE’s People-Finder system are all built in this way. However, the expert database has certain shortcomings. (1) The establishment of the system is labor-intensive. (2) The description of expert expertise is usually incomplete and general, while the knowledge demands are usually specific and detailed, which leads to the inaccuracy of the matching process. (3) The specific expertise and skills of experts are constantly changing, and the update is resource-consuming and time-costing.

In addition to relying on manual work, expert search and recommendation systems can also discover expert expertise through indirect or secondary resources. Many researchers use personal websites to track the expertise of experts, while other researchers take the web data, scientific publications, patent data, project proposals, and other data containing the information of expertise as the sources for collecting and identifying the expertise of candidate experts (Guy et al., 2013; Serdyukov & Hiemstra, 2008a, 2008b; Sriramoju, 2015). This method has certain advantages compared to expert search based on expert databases, but the volume of obtained data may be huge. Thus, the recommendation is essential to reduce the negative effects of information overload (Sriramoju, 2015). The expert knowledge representation models have plenty of formats by introducing various techniques and tools, such as ontology, graph theory, VSM, social network, and text mining (Ma et al., 2012; Manning et al., 2008; Turney & Pantel, 2010). Daud and his colleagues come up with time topic modeling to represent experts’ knowledge (Daud, 2012; Daud et al., 2010).

The matching technique aims at measuring the similarities between the demand and the supply, giving clues about "who knows what." Corresponding to different knowledge representation models of experts, there are plenty of matching techniques. These techniques would extract some features or standards as the criteria for capturing the degree of matching. The method is repeatable and efficient. If the matching process is reasonable, the matching results are also transparent and predictable. But we also notice that it could be unable to handle extreme situations. We would like to expand the discussion on the matching techniques in the next section. Despite the research on independent reviewer recommendation, there are some researches concerned with the group-based reviewer recommendation. The reviewer group recommendation aims at recommending a set of reviewers who can thoroughly evaluate all aspects of the proposal's content to each proposal. At the same time, it also addresses the balance of the workload of the reviewers (Hoang et al., 2021; Mirzaei et al., 2019).

How to evaluate the performance of the expert recommendation is a tough problem. The expert recommendations usually have different scenarios, and different methods also lead to un-comparable results. Wildly used test datasets in the expert finding task are obtained from different sources such as third parties (CSRIO Enterprise Research Collection, CERC), colleagues (W3C), and candidates themselves (UvT) (Balog & de Rijke, 2008). Recommendation results are comparable for these datasets, while how to evaluate the results remains a problem for other specific cases. To the best of our knowledge, a possible way is to evaluate the satisfaction scored by users or authentic experts.

Approaches for expert recommendation

There are plenty of expert recommendation approaches, and this research would review three main categories: content-based approaches, network-based approaches, and hybrid approaches.

Content-based expert recommendation approaches

The content-based expert recommendation method focuses on calculating similarities between proposals and experts. The models include Boolean models, VSM (Biswas & Hasan, 2007; Yukawa et al., 2001), and probability models (Manning et al., 2008).

In VSM, a document-term matrix is constructed: each line represents a document vector and each column represents a term. The element of the matrix could be a binary variable representing whether the term occurs in the document, or a metric variable representing the term frequency in the document. Likewise, the relationship between experts and documents could be depicted by an expert-document matrix, in which each line represents the document that is related to (i.e., authored by, verified by) the expert. By aggregating the expert-document matrix and document-term matrix, the expert knowledge could be represented by the vector space of terms. Correspondingly, the knowledge demand could also be represented as a demand-term matrix, and the similarity between a demand and an expert could be measured by the cosine similarity of the two vectors. Shon et al., (2017) develop an automatic matching system using keywords with fuzzy weights. The system is developed based on the MapReduce framework created by Hadoop. The evaluation of the system focuses on the execution time while ignoring the accuracy.

What we should mention here is that the above method assumes the terms are statistically independent of each other, which is often not the case in the real world. To overcome this shortcoming, some tools are introduced to measure the semantic relationship of terms. For example, WordNet is a large-scale electronic lexical database, where words are connected via semantic and lexical similarities (Fellbaum, 1998). ConceptNet 5 is a large-scale, multilingual, domain-general knowledge graph (Speer & Havasi, 2013). The words and phrases of natural language are connected with labeled, weighted edges. These tools are wildly used in the task of text clustering (Vo et al., 2015; Wei et al., 2015), and word sense disambiguation (Sachdeva et al., 2014).

However, these methods have some limitations, such as the connotation inconsistencies, incorrect translations of polysemous literals, etc. (Erjavec & Fišer, 2006). Specifically, the semantic relationships of terms are still inadequate for most languages, and the databases focus on general knowledge, which leads to inaccuracy or gaps for knowledge within a specific S&T domain. Rather than measuring the semantic relationships of terms by the above methods, this research would introduce a word embedding method, the Word2vec model. It represents each term as a low-dimensional continuous vector, each dimension of which is an independent feature from the other. Besides, it could learn the semantic relationships of terms by training the datasets of the specific language, as well as the specific domain. In this way, the term, document, and expert knowledge representation could be represented as a vector, and each dimension is independent. The semantic measurement is introduced to reviewer recommendation research. For example, Zhao et al. (2018) construct a Word Mover's Distance–Constructive Covering Algorithm to solve the reviewer recommendation problem as a classification issue. Tan et al. (2021) propose a word and semantic-based iterative model (WSIM) to assign reviewers. Shen et al. (2021) develop the deep variational matrix factorization with knowledge embedding for the recommendation system.

Network-based expert recommendation approaches

Different from the content-based expert recommendation approaches which focus heavily on textual information, the network-based expert recommendation approaches make use of linkages. There are mainly two ways to build the network. First, experts are taken as vertexes, and two vertexes are connected once they have relationships. Different kinds of networks could be built corresponding to different relationships, such as the collaboration network (Huang et al., 2019). Second, the network is built by representing experts, documents, words, and other attributes as the vertexes, and their relationships are taken as edges. It could be viewed as a multi-layer network or a heterogeneous network. In this kind of network, the edge would represent different relationships: it could be the collaboration between two experts, or the authorship of a document by an expert (Xu et al., 2012).

There are lots of tools commonly used in network-based expert recommendation approaches. Social network analysis is applied to detect the community of experts, find experts with similar interests, activities, or experiences, and infer possible connections in the future. The recommendation could be made by applying link prediction (Liben-Nowell & Kleinberg, 2007) or vertex similarity analysis (Leicht et al., 2006). These approaches are applied to recommend potential collaborators (Xu et al., 2012), find suitable committee candidates for academic conferences (Han et al., 2013), and recommend participants for scientific research projects (Wang et al., 2017).

Hybrid expert recommendation approaches

Hybrid approaches cover the amalgamation of the above approaches. Content-based expert recommendation approaches are implemented to build profiles of experts, identify expertise/research topics, and represent the knowledge, while network-based approaches are applied to depict the network of researchers, detect the communities, and find authoritative experts within a community. For example, Davoodi et al. (2013) take advantage of content-based profiles of experts enriching them with a semantic kernel, then, detecting the social communities of experts by considering their experience, background, personal preferences, and knowledge. The recommendation is made by seeking the representatives of the most relevant community. Xu et al. (2012) combine social network and semantic concept analysis for the personalized academic researcher recommendation. Pradhan et al. (2021) utilize the content-based approach and the network-based approach to form a decision support system for reviewer recommendation, which integrates the topic network, the citation network, and the reviewer network into a reviewer recommender system.

The network-based approach has some limitations. Firstly, the relationship-based approach would lack the understanding of experts' knowledge. Secondly, the approach may suffer from the "cold start" problem, which may ignore the newcomers. Thirdly, the expandability of the approach is not ensured. Compared to the network-based approach, the content-based approach has a better understanding of the knowledge, which is most important in the reviewer recommendation. Compared to the hybrid approach, the content-based approach is easy to apply. Thus, this research would develop a reviewer recommendation method by introducing the content-based approach.

Reviewer recommendation for scientific research projects

The reviewer recommendation for the scientific research project is a sub-topic of expert recommendation, whose main responsibility is recommending capable scientific researchers for proposals. It has characteristics as follows. First, the data source of reviewer recommendation is S&T-based, including projects, publications, patents, awards, and nominations (Biswas & Hasan, 2007; Cook et al., 2005; Liu et al., 2016; Protasiewicz et al., 2016). The data is heterogeneous, which leads to the consideration of how to provide a reviewer list integrating different kinds of data. The discussion is still to be explored. This research would compare three ranking fusion methods for reviewer recommendation.

Second, there are lots of factors influencing the performance of the reviewer recommendation. The relevance of reviewers and proposals is a certain important factor, while the level of expertise also counts. Except for the quantity and quality of the research, other factors are taken into consideration, e.g. the responsiveness of reviewers, and the objectivity and impartiality of the review (Alhosan et al., 2014). Yong et al. (2021) develop a framework for reviewer recommendation based on a knowledge graph and rules matching. Pradhan et al. (2021) propose a proactive decision support system for reviewer recommendations in academia. They explore various aspects, including relevance between reviewer candidates and submission, authority, expertise, diversity, and conflict of interest, and then integrate them into the proposed framework.

Last but not the least, how to evaluate the performance of the reviewer recommendation approaches is difficult. Since the real reviewers of proposals usually remain confidential, the recommended results are usually evaluated by experts. This is somehow an indirect evaluation. Some researchers also use time efficiency, effect on avoiding the interest conflict, balance, and rationality as the evaluation criteria (Liu et al., 2016). For example, Shon et al. (2017) develop an automatic matching system using keywords with fuzzy weights. The system is developed based on the MapReduce framework created by Hadoop. The evaluation of the system focuses on the execution time while ignoring the accuracy. This research proposes an evaluation method through questionnaires, which is a first-hand evaluation. The detailed evaluation metrics are available in "Evaluation".

Proposed RRM

This research proposes a reviewer recommendation method for the scientific research proposal by introducing tools from text mining, bibliometrics, and machine learning. As discussed in "Introduction", two key factors influencing the reviewer recommendation accuracy are the representation model of reviewers' knowledge and proposals, as well as the matching between corresponding representation models (Karimzadehgan et al., 2008). Previous studies mainly represent reviewers' knowledge by disciplines, research topics, and keywords, which suffers from the broadness of disciplines, the irregularity of research topics, and the sparseness of keyword space. To perform a solid foundation for matching, the representation model of reviewers' knowledge and proposals should be as accurate as possible. The Word2vec model gives a solution for the representation of reviewers' knowledge and proposals, for its semantic and synthetic measurement. On the other hand, the matching between candidate reviewers and proposals should take advantage of the representation model, and deal with similarities from multi-systems. That's why this research proposes a two-step ranking: calculating the semantic similarities of representation models and performing a recommended reviewer list by ranking fusion methods.

Figure 1 shows the framework of the proposed reviewer recommendation method. This section is organized as follows. The "Profiling" section provides the data collection strategy for the profiling purpose, including collecting scientific research proposals from the funding agency and retrieving scientific publications and the research background of candidate reviewers. The "Training" section discusses the training of the Word2vec model. The "Representing" section represents the proposals and the knowledge of candidate reviewers as low-dimensional continuous vectors. The "Ranking" section proposes and compares three ranking fusion methods to measure the similarities between scientific research proposals and candidate reviewers. The "Evaluation" section discusses three evaluation metrics to test the accuracy of the proposed reviewer recommendation method. Note that this research would take proposals to be reviewed as "proposals" and the projects shouldered by candidate reviewers as "projects."

Profiling

Profiling is the process of selecting expertise resources, collecting related data, and identifying attributes to characterize a given object. In this research, the profiling process aims at gathering fundamental information to develop comprehensive representation models for the knowledge of candidate reviewers and proposals.

For proposals, it could be profiled by the context of each proposal. The candidate reviewers are identified by the disciplines and keywords of the proposals. For candidate reviewers, data should be retrieved by developing an information retrieval strategy. As

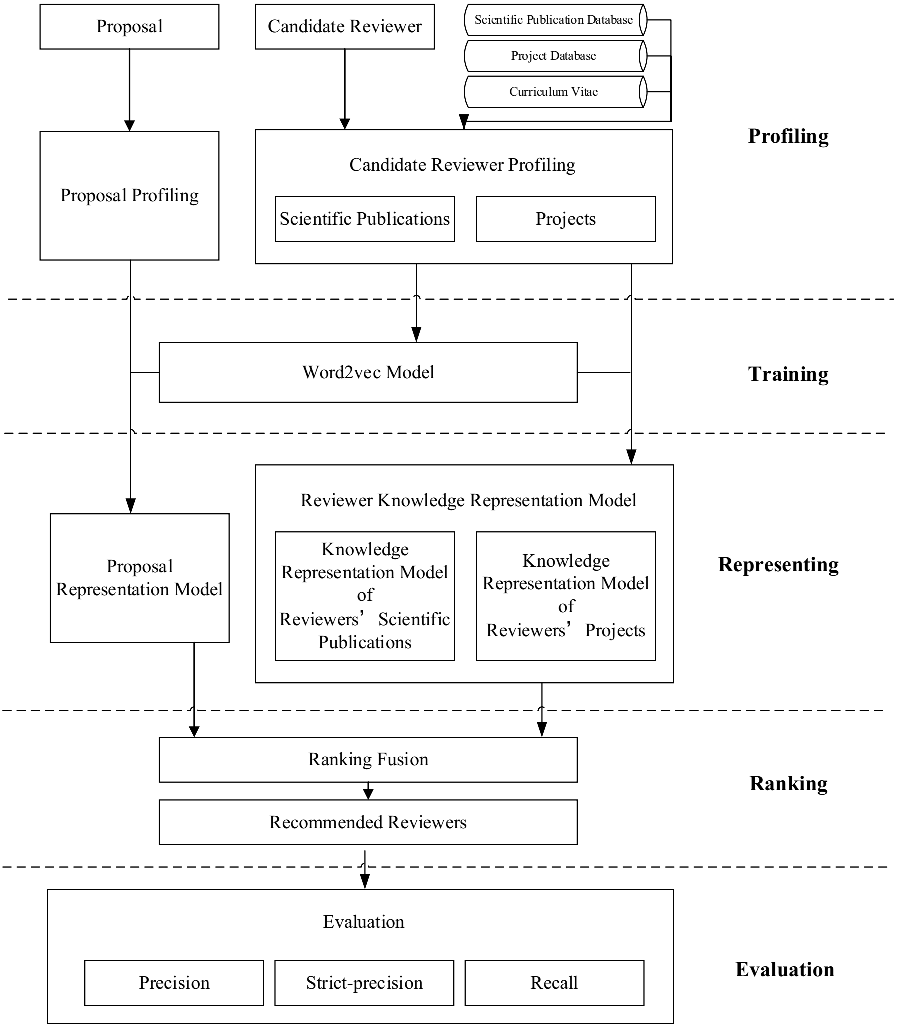
Proposal
Candidate Reviewer
Scientific Publication Database
Project Database
Curriculum Vitae
Proposal Profiling
Candidate Reviewer Profiling
Scientific Publications
Projects
Profiling
Word2vec Model
Training
Proposal Representation Model
Reviewer Knowledge Representation Model
Knowledge Representation Model of Reviewers' Scientific Publications
Knowledge Representation Model of Reviewers' Projects
Representing
Ranking Fusion
Recommended Reviewers
Ranking
Evaluation
Precision
Strict-precision
Recall
Evaluation

Fig. 1 The framework of proposed reviewer recommendation method

discussed in the literature review, traditional profiling relying on self-claimed information lacks efficiency, accuracy, and timeliness. On the contrary, open data sources have the potential to build accurate profiles automatically and objectively (Xu et al., 2016). In research communities, publishing scientific papers and applying for scientific research projects are two main activities of researchers. Thus, publications and projects can be considered as two kinds of important evidence to represent reviewers' expertise. In this research, we trace the scientific research projects and scientific publications of each candidate reviewer to profile the experts' knowledge.

Training

For content-based reviewer recommendation systems, the reviewers' knowledge is usually represented by a collection of keywords. By introducing the VSM model, the reviewer-term matrix is built to represent the reviewers' knowledge, in which each column represents a reviewer and each row represents a term. However, the VSM model has some shortcomings. On the one hand, the high-dimensional reviewer-term matrix would lead to a sparse problem, which could bring a significant decrease in the quality of the matching between the knowledge of reviewers and proposals. On the other hand, the model assumes that each row/ each term is independent, which is not in line with the real world, the semantic relationship between the terms is ignored. To overcome the above shortcomings, this research uses the Word2vec model to represent the reviewers' knowledge.

Word2vec model is a word representation model, using a one-layer neural network to map one-hot sparse word vectors into an n-dimensional dense vector (Mikolov et al., 2013). The basic assumption of the language model is that the meaning of a word can be inferred from its context. Word2Vec takes a single word as input and outputs a single vector representation of that word, essentially 1–1 mappings between words and their respective vectors. It would be very useful for the reviewer recommendation since we want to train word vectors by the datasets and apply them to the proposals. Since BERT and GPT-2 generate contextual embeddings, it takes input as a sequence (usually a sentence) rather than a single word. With BERT and GPT-2, the vector representations of words will vary based on the specific sequences inputting. The output is a fixed-length vector representation of the input sentence. It provides sub-word embeddings and sentence representations. For some words, there may be a single sub-word while, for others, the word may be decomposed into multiple sub-words. The word representation is dynamic. For this instance, we would choose the Word2vec model in our scenarios. It is an effective technique to elicit knowledge from large text corpora in an unsupervised manner, with advantages of high accuracy and low computational cost. The Word2vec model has been widely adopted in many similarity measurements related studies, and its good performance is proven (Galke et al., 2017; Hu et al., 2018; Zhang et al., 2018). There are two important models of Word2vec, the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. The CBOW calculates the probability distribution of a certain word based on consecutive words before and after. The Skip-gram model predicts the words surrounding a target word. The former has better performance dealing with short texts.

Before forming the term vector by the Word2vec model, natural language processing (NLP) and term frequency-inverse document frequency (TFIDF) are implemented to segment the documents and identify the terms. Since the Word2vec model only deals with single terms, representing them as a vector. Some of the identified terms are multi-terms, and we connect each word by underscores. In this way, the word embedding vectors for each term could be trained. The data obtained by the above preprocessing can be taken as the input of the Word2vec model to train the term representation vector.

Note that the term representation model is the one trained in the reviewer knowledge representation. Though the training dataset does not include the proposals, the vector space is reasonable for proposal representation. There are mainly three reasons. (1) The recommendation is set in the situation in which the reviewer and the proposal share a similar knowledge base since the candidate reviewers are nominated according to the proposal. (2) The proposals and reviewers' documents share similar terms. (3) The previous research wildly-use the pre-trained term representation model to analyze the similarity of terms in new datasets. For example, the Global Vector (GloVe) word representation is trained by the data from Wikipedia2014 and Gigaword5, released in 2014 by the computer science department at Stanford University (Pennington et al., 2014). Various researchers apply this pre-trained representation for sentiment analysis, automatic answers, disease detection, etc. (Magooda et al., 2016; Sharma et al., 2017; Yancheva & Rudzicz, 2016). The pre-trained dataset does not include the experiment data, but it is still available for similarity calculation using the pre-trained vector. This research utilizes the pre-trained Word2vec vectors to represent proposals.

Representing

By adopting the Word2vec model, we could represent the term in project proposals and scientific publications as a continuous vector. Specifically, each term could be represented as an n-dimensional vector, i.e., $\text{term}_i = \left(p_{i1}, p_{i2}, \ldots, p_{in}\right)$, p_{im} indicates the probability of term i on the m-th dimension, $m \in [1, n], i \in [1, s]$. Based on this, a term representation matrix could be built. Each column represents a dimension of the Word2vec model, and each row represents a term. Specifically, the term representation matrix could be marked as P.

Further, we employ the TFIDF and SVM to build a document-term matrix, in which each cell takes the TFIDF value of the term if the document contains the term, and 0 otherwise. We denote $tfidf_{hi}$ as the TFIDF value of term i in document h, $i \in [1, s], h \in \left[1, j\right]$. Previous works (Arora et al., 2016; Blacoe & Lapata, 2012; Mitchell & Lapata, 2008, 2010) show that computing phrase or sentence embeddings by composing word embeddings using operations on vectors and matrices performs very well. Thus, we build a document representation model in this way. A TFIDF-weighted document representation matrix could be built by multiplying the document-term matrix and the term representation matrix, as $\text{DOC} = \text{TFIDF} * P$.

The reviewer knowledge representation model could be generated by integrating his/her documents. The scientific publications and project proposals are two main outcomes of researchers, while they have different emphases. Thus, for proposals and scientific publications, we would build two knowledge representation models respectively. For each reviewer, the reviewer knowledge representation model contains two components: the knowledge representation models for projects and publications. The $\text{exp}_k^{\text{Grant}}$ represents the knowledge of the k-th expert's project proposals, while $\text{exp}_k^{\text{Pub}}$ represents the knowledge of the k-th expert's scientific publications. They could be calculated as the mean of corresponding document representation vectors.

By extracting information from proposals, we could get terms and a document-term matrix. We inherit the term representation model trained before and calculate the representation model of proposals by multiplying the TFIDF-weighted document-term matrix and the term representation model. Similar to the reviewer knowledge representation model, we get the proposal representation model as $\text{DEMAND} = \text{TFIDF}_{\text{DEMAND}} * P$.

Ranking

Since the employment of multi-source data, the ranking process also has two steps. First, the relevance between the proposals and reviewers' publications/projects is calculated by cosine distance separately, marked as m_{jk}^{Pub} and m_{jk}^{Grant}. Second, a list of reviewers is generated by introducing three methods: (1) Average: Reviewers rank by the average matching score from high to low; (2) SumRank: Reviewers rank by the sum of rankings; (3) CombRKP: Reviewers rank by the reciprocal sum of rankings. The research will compare the construction ideas and applicability of the three methods mentioned above.

Average

The Average method ranks the reviewers by the average matching score from high to low, and the ranking list is marked as r_{jk}^{Average}. This is a common method to deal with scores from different systems. However, it inherits the shortcomings of "average", such as the sensitivity with the distribution, especially, the extremum. For example, a reviewer does not undertake a project, so the matching score between the reviewer's project and the proposal is 0. In this case, even if the reviewer's scientific publications match the proposal very well, it is still difficult for such a reviewer to have a higher overall ranking. This deficiency of the average index will significantly affect the accuracy of the reviewer recommendation.

To solve this shortcoming, this research proposes two ranking methods. The ranking method considers two hypotheses. First, reviewers recommended through multiple rankings may have a higher relevance. Second, among the recommended reviewers, the reviewer with the higher ranking has higher relevance to the proposal. On these bases, this research introduces two reviewer ranking methods, SumRank and CombRKP.

SumRank

The SumRank method sums the ranking order of reviewers in multiple rankings and then ranks the sum of the rankings in ascending order. By sorting the matching degree between the reviewer's grant and the proposal from high to low, the reviewer ranking based on projects can be obtained, marked as r_{jk}^{Grant}. In the same way, the reviewer ranking based on scientific publications is marked as r_{jk}^{Pub}. The sum of the ranking is calculated by $r_{jk}^{\text{Sum}} = r_{jk}^{\text{Grant}} + r_{jk}^{\text{Pub}}$. By ranking r_{jk}^{Sum} from low to high, the ranking of the knowledge matching degree between reviewer k and the proposal j is $1_{\text{score }(j,k)\in\{1,0.5,0\}}$. This method could avoid the shortcomings that bring by "average". In the previous example, the reviewer could be selected since the high ranking of the reviewer's scientific publications with the proposal.

CombRKP

Nuray and Can proposed the CombRKP method of ranking fusion (Nuray & Can, 2006), and the calculation formula of CombRKP is: CombRKP $\left(d_j\right) = \sum_{i=1}^{p} \frac{1}{\text{rank}_i(d_j)}$. The p demonstrates the total number of the retrieval systems. The d_j denotes document j, while $\text{rank}_i(d_j)$ denotes the ranking of document j retrieved in the i-th retrieval system. The CombRKP is widely used in the research on reviewer systems, and the final recommendation result list is formed by fusing the results returned by multiple retrievals. Cormack et al. (2009) tested on the TREC data set and found that the CombRKP performed better than the Condorcet Fuse method. Chen et al. (2010) tested on multiple data sets and found that the CombRKP is significantly better than other data fusion methods. Zainal et al. (2013) compared the CombRKP with other data fusion methods based on some data in the field of bioinformatics and concluded that the CombRKP is significantly better than other methods.

The CombRKP calculates the total ranking based on the sum of the reciprocal of the knowledge matching degree ranking. The method believes that reviewers with high rankings should be given higher weights, and reviewers with low rankings should be given lower weights. The formula for calculating the reciprocal summation of the matching degree ranking is $r_{jk}^{\mathrm{RKP}} = 1/r_{jk}^{\mathrm{Grant}} + 1/r_{jk}^{\mathrm{Pub}}$. By ranking r_{jk}^{RKP} from low to high, the ranking of the knowledge matching degree between reviewer k and the proposal j, marked as $r_{jk}^{\mathrm{CombRKP}}$, could get.

By incorporating different data fusion methods, this research develops three reviewer recommendation methods: (1) W2V-Average: using the Average method to integrate similarities between the reviewer knowledge representation and the proposal representation; (2) W2V-SumRank: using the SumRank method to integrate similarities between the reviewer knowledge representation and the proposal representation; (3) RRM: using the CombRKP method to integrate similarities between the reviewer knowledge representation and the proposal representation.

Evaluation

This research uses a questionnaire to verify the rationality of the recommendation results. The questionnaire lists the titles and abstracts of proposals in detail. We invite reviewers to rate the familiarity of the proposals according to their knowledge. One point means the respondent is familiar with the proposal and suitability for recommendation; 0.5 means the respondent is relatively familiar with the proposal and reasonableness for recommendation; 0 means the respondent is unfamiliar and unsuitable for the recommendation. We distribute the questionnaire through the Internet.

The research compares and analyzes the results of the reviewer recommendation with the questionnaire to verify the effectiveness of the proposed method. To evaluate the performance of the proposed method, the research proposes three evaluation indicators for the recommendation results: Precision, Strict-precision, and Recall.

Rank_j^w is defined as the set of reviewers whose matching degree with the proposal j is in the top w position; Rec_k^w is defined as the set of proposals whose matching degree with the reviewer k is in the top w position. There is the following equivalence relationship:$k \in \mathrm{Rank}_j^w \Leftrightarrow j \in \mathrm{Rec}_k^w$.

Precision is the most important evaluation indicator for reviewer recommendation. The previous study find that 59.9% of the reviewers reject to review an NSFC proposal because they were unfamiliar with the content (Li et al., 2014). It shows that recommending unsuitable reviewers to the proposals would lead to low efficiency. The paper puts forward a Precision indicator to describe the percentage of reviewers who can review the proposal (scores of 1 and 0.5) in the top w ranking. The Precision indicator is calculated as follows:

$$\mathrm{Precision}_j = \frac{\sum_{k \in \mathrm{Rank}_j^w} 1_{\mathrm{score}\ (j,k) \in \{1,0.5\}}}{\left|\mathrm{Rank}_j^w\right|} \tag{1}$$

J represents the j-th proposal, k represents the k-th reviewer, $j \in [1, m], k \in [1, n]$. score (j, k) represents the score of the proposal j given by the reviewer k, the value range is $\{1, 0.5, 0\}$. $1_{\mathrm{score}\ (j,k) \in \{1,0.5\}}$ means that when the score of the proposal j given by reviewer k belongs to $\{1, 0.5\}$, the value is 1.

The formula (1) calculates the Precision of the proposal j. By averaging all items, we get the overall Precision, as follows:

$$\text{Precision} = \frac{1}{m}\sum_{j=1}^{m}\text{Precision}_j \tag{2}$$

On this basis, this research constructs a Strict-precision index, which characterizes the proportion of reviewers suitable to review the proposals (score of 1) in the top w reviewers. The Strict-precision index is calculated as follows:

$$\text{Strict} - \text{precision}_j = \frac{\sum_{k\in\text{Rank}_j^w} 1_{\text{score }(j,k)\in\{1\}}}{\left|\text{Rank}_j^w\right|} \tag{3}$$

For all projects, the calculation formula for Strict-precision is as follows:

$$\text{Strict} - \text{precision} = \frac{1}{m}\sum_{j=1}^{m}\text{Strict} - \text{precision}_j \tag{4}$$

In addition to the Precision and Strict-precision, this research constructed a Recall index to characterize the percentage of proposals that reviewers are suitable for review in recommended proposals. Its calculation formula is as follows:

$$\text{Recall} = \frac{1}{n}\sum_{k=1}^{n}\frac{\sum_{j\in\text{Rec}_k^{\text{top}}} 1_{\text{score }(j,k)\in\{1\}}}{\sum_j 1_{\text{score }(j,k)\in\{1\}}} \tag{5}$$

This research takes two models as the baseline methods, i.e., the social network-empowered research analytics framework (RAF) proposed by Silva, Guo, and Ma (2013), and the Apache Lucene. The RAF takes Scholarmate.com as the platform to build researcher profiles from three dimensions: relevance, productivity, and connectivity. The RAF calculates the average matching scores between the proposal and the reviewers' self-claimed keywords, projects, scientific publications, and social tags. The calculation of relevance in RAF is $r_{ij} = \alpha\text{Self}_{ij} + \beta\text{Grant}_{ij} + \gamma\text{Pub}_{ij} + \delta\text{Social}_{ij}$. Since the self-claimed keywords and social tags are not publicly available, and the projects and the scientific publications are good data sources to depict reviewers, this research applies a simplified formula. In the RAF, the projects and scientific publications share the same weight. Thus, the calculation of relevance is $r'_{ij} = 0.5 * \text{Grant}_{ij} + 0.5 * \text{Pub}_{ij}$. The RAF calculates the matching score between the reviewer's projects/ scientific publications and the proposal by the corresponding VSM vectors. The difference between RAF and W2V-Average is whether they consider the semantic relationships of the data. The comparison between RAF and W2V-Average can illustrate the changes brought by the Word2vec model.

This research introduces Apache Lucene, a typical full-text search engine, as the baseline method. Apache Lucene project (https://lucene.apache.org/) develops open-source search software, providing powerful indexing and search features, as well as spellchecking, hit highlighting, and advanced analysis/tokenization capabilities. Thus, this research uses the Apache Lucene to find reviewers for proposals. Since the Apache Lucene would return a list of reviewers for projects and publications respectively, we would combine the Apache Lucene and the ranking methods mentioned above. The Lucene-SumRank would integrate the two recommended lists by the SumRank, while the Lucene-CombRKP would take advantage of CombRKP to combine the two lists.

Case study

Experiment

NSFC has comprehensively introduced and implemented a rigorous and objective merit-review system to fulfill its mission of supporting basic research, fostering talented researchers, developing international cooperation, and promoting socio-economic development. The funding agency comprises one general office, five bureaus, and nine departments, and these departments are partitioned into divisions focusing on specific research domains. Though there are nine departments and subdivided divisions, it is still difficult for the committee to assign suitable reviewers for each proposal, considering the enormous demand for peer review. Besides, each division could manage multiple related research topics. For example, the Science and Technology Management and Policy division (G0404) includes projects on intelligence property, scientometrics, technology management practice. Thus, it is too difficult for the committee to be familiar with all the research topics. An automatic or half-automatic recommendation system could make it more efficient and economical. Besides, reasonable assignment of reviewers could benefit the objective evaluation and promote the efficiency of scientific funding. Thus, how to assign suitable reviewers to proposals is critical for NSFC. This research takes the reviewer recommendation for NSFC as a case study.

To implement our reviewer recommendation method, we take the proposals within one specific division as the proposals to be reviewed, while taking the principal investigators who have shouldered projects within this division as the candidate reviewers. Furthermore, we would like to retrieve the publications for each candidate reviewer, which works for the reviewer knowledge representation. The website of NSFC (http://output.nsfc.gov.cn/) announces the information of the funding projects from 1986 to 2015. The project information includes the project approval number, title, project type, application code, principal investigators, the title of the principal investigator, the affiliation of the principal investigator, research period, funding amount, abstract of proposal, and conclusions.

This study takes the reviewer recommendation under the Science and Technology Management and Policy division (G0404) as a case study. As of the data collection of this research, the latest proposals announced on the website are 12 proposals granted in 2015. We take these 12 proposals as the test dataset. Second, we would establish a candidate reviewer list. Generally, candidates are researchers who have been the principal investigator of the NSFC project. But some principal investigators could not participate in the peer-review process due to retirement or physical reasons, so this research would take the principal investigator who received funding projects from 2005 to 2014 as candidate reviewers. From 2005 to 2014, 136 researchers shouldered 167 projects in the G0404 Science and Technology Management and Policy division, including 4 key projects, 114 general projects, and 49 youth science fund projects. These researchers are selected as candidate reviewers. Third, we would retrieve the data to represent the knowledge of each candidate. As discussed before, we consider two data sources: publications and projects. Web of Science database is generally accepted as the most recognized database of scientific publications covering all disciplines. It also provides a high standard of collecting peer-reviewed journals. By Web of Science database, scientific publications of these candidate reviewers are retrieved and collected to establish the knowledge base for each candidate reviewer. Based on the data of NSFC, we collect the projects of these 136 researchers. Further, the organization information of the above researchers is collected. By applying the unique identifier of researchers, this research builds a dataset for each researcher, including the organizations, education background, scientific publications, and grants. Finally, this research collects 982 publications and 167 projects of 136 candidate reviewers.

In this case, the projects are written in Chinese while the publications use English. Thus, this research trains two Word2vec models. First, a Word2vec model for publications is trained by the WOS publications of all the candidate reviewers. We extract the title and abstract of the scientific publications for the training. We clean (consolidate) the data, following the clean-up process of Liu and Porter (2020). We identify 269 terms by the TFIDF, and we connect each word in multi-terms by underscores. In this way, the representation vector for multi-terms can be calculated. The consolidated data are the input of the training of the Word2vec model. By training the Word2vec model, we get a continuous vector for each term. Second, a Word2vec model for the project is trained by collecting all the proposals of the G0404 division from 1986 to 2014. Similarly, we get 305 Chinese terms and their representation vector.

Further, we could build the publication-term matrix and the project-term matrix. By aggregating the terms representation matrix and the publication-term matrix/ the project-term matrix, we get the scientific publication representation vectors/ the project representation vectors. The knowledge representation vectors of reviewers' scientific publications and the knowledge representation vectors of reviewers' projects can be calculated. On the other hand, the titles and the abstracts of proposals are available in both Chinese and English. Thus, we could represent the Chinese knowledge representation vectors for proposals and the English knowledge representation vectors for proposals.

The relevance between reviewers and proposals is constructed by two parts: the similarity between reviewers' scientific publications and proposals, marked as m_{jk}^{Pub} and the similarity between reviewers' projects and proposals, marked as m_{jk}^{Grant}. This research uses the ranking fusion methods introduced in the "Proposed RRM" section to recommend reviewers for each proposal. Specifically, we recommend top 10, top 15, and top 20 reviewers for each proposal, by the W2V-Average, the W2V-SumRank, and the RRM.

In our case, the proposed reviewer recommendation method is valid on the 12 proposals of the G0404 division. The actual reviewers of these proposals are not publicly available. Thus, the research uses a questionnaire to verify the rationality of the recommendation results. The specific steps are introduced in "Evaluation". See supplementary materials for the questionnaire. A total of 15 valid questionnaires were collected, and the reviewers participating in the survey came from 11 institutions and 7 cities. There are a total of 136 candidates in the case study, and the number of questionnaires returned accounts for 11.03% of the total number of candidates. There are 102 reviewers in the top 20 rankings, and the other 34 reviewers are not in the top 20 rankings. The number of questionnaires returned accounted for 14.71% of the 102 reviewers. Although the number of questionnaires returned is limited, it is representative in terms of the proportion. The research compares and analyzes the results of the reviewer recommendation with the questionnaire to verify the effectiveness of the proposed method.

Results

This research recommends top 10, top 15, and top 20 reviewers for each proposal, by six methods: the Lucene-SumRank, the Lucene-ComRKP, the RAF, the W2V-Average, the W2V-SumRank, and the RRM. Table 1 lists the evaluation results. The average Precisions of the Lucene-SumRank, the Lucene-ComRKP, the RAF, the W2V-Average,

Table 1 Evaluation of recommended reviewers

Top	Method	Precision (%)	Strict-precision (%)	Recall (%)
10	Lucene-SumRank	58.33	37.50	13.54
	Lucene-CombRKP	58.33	37.50	12.50
	RAF	66.67	25.00	12.20
	W2V-Average	72.73	37.88	26.85
	W2V-SumRank	86.36	51.06	21.43
	RRM	86.11	61.11	24.35
15	Lucene-SumRank	75.00	50.00	17.16
	Lucene-CombRKP	75.00	51.39	18.20
	RAF	66.67	36.25	18.27
	W2V-Average	82.58	55.45	41.37
	W2V-SumRank	84.09	52.58	41.49
	RRM	84.72	61.11	27.74
20	Lucene-SumRank	75.00	48.61	22.72
	Lucene-CombRKP	75.00	48.61	22.72
	RAF	70.83	38.75	21.55
	W2V-Average	82.27	46.71	45.83
	W2V-SumRank	84.09	51.49	48.63
	RRM	87.22	53.06	37.08

the W2V-SumRank, and the RRM are 69.44%, 69.44%, 68.06%, 79.19%, 84.85%, and 86.02%. Compared with the Lucene-SumRank and the Lucene-ComRKP, the Precisions of the W2V-Average, the W2V-SumRank, and the RRM increase by 9.75%, 15.41%, and 16.58%. Compared with the RAF, the Precisions of the W2V-Average, the W2V-SumRank, and the RRM increase by 11.14%, 16.79%, and 17.96%. Using the RAF, reviewers have a 32% probability of being unfamiliar with the proposal and unwilling to review it. The RRM reduces this value to 14%. In the case of recommending top 10 reviewers, the Precision of the RRM is slightly lower than that of the W2V-SumRank, but it is still at a high level.

The average Strict-precisions of the Lucene-SumRank, the Lucene-ComRKP, the RAF, the W2V-Average, the W2V-SumRank, and the RRM are 45.37%, 45.37%, 33.33%, 46.68%, 51.71%, and 58.43% correspondingly. Compared with the Lucene-SumRank and the Lucene-ComRKP, the Strict-precisions of the W2V-SumRank and the RRM increase by 1.31% and 6.34%. Compared to RAF, the Strict-precisions of the W2V-Average, the W2V-SumRank, and the RRM increased by 13.35%, 18.38%, and 25.09%. In the above three cases, the Strict-precision of the RRM is the highest. It shows a 58.43% probability that the reviewers recommended through the RRM show a high matching degree with the proposal to be reviewed.

The average Recalls of the Lucene-SumRank, the Lucene-ComRKP, the RAF, the W2V-Average, the W2V-SumRank, and the RRM are 17.81%, 17.46%, 17.34%, 38.02%, 37.18%, and 29.72%. Compared with the RAF, the Recalls of the W2V-Average, the W2V-SumRank, and the RRM increase by 20.67%, 19.84%, and 12.38%.

Discussion

Regardless of the top 10 reviewers, the top 15 reviewers, or the top 20 reviewers recommended, the Precision, Strict-precision, and Recall of RRM are significantly higher than those of the baseline methods. It shows that reviewers obtained through the proposed method are more likely to be familiar with the research content of the corresponding proposals and willing to participate in the review. The results show that the three reviewer ranking methods based on matching degree fusion proposed in this research effectively improve the accuracy between reviewers and proposals. The effectiveness of the RRM is verified.

Second, the Precision and Strict-precision of the W2V-Average are higher than those of the RAF, which illustrates the improvement by introducing the Word2vec model. The W2V-Average and the RAF rank the reviewers by the average matching score between the reviewers' publications/ projects and proposals. The difference is that the former uses the proposed reviewer knowledge representation model and proposal representation model incorporating the Word2vec model, and the latter uses the simple keyword-based method, incorporating the VSM model. As discussed before, the proposed reviewer knowledge representation model based on the Word2vec model considers the semantic relationships between the terms, while the VSM ignores them. The consideration of semantic relationships leads to improved accuracy, which is one of the contributions of this research.

Third, the Precision and Strict-precision of the W2V-SumRank, and the RRM are overall higher than those of W2V-Average, which addresses the improvement by the ranking fusion methods. Three methods incorporate different ranking fusion methods, building on the reviewer knowledge representation model and the proposed representation model. The SumRank and the CombRKP would use the rankings from multi-systems, rather than matching scores, to build the overall ranking. It overcomes the disadvantages of scores: once the score from one system is quite low, the overall ranking would be affected heavily. In the scenario of reviewer recommendation, either the reviewer's publications or the reviewer's projects are highly relevant to the proposal, the reviewer could be a good choice for the recommendation. The ranking fusion methods could meet the requirement very well. The high Precisions of the W2V-SumRank and the RRM illustrate the improvement of introducing ranking fusion methods.

Fourth, this research also compares the RRM method with the subject-based methods, which shows the RRM has better performance. Our case study uses proposals from one division of the NSFC (G0404 the Science and Technology Management and Policy division) as a case study. To some extent, it is a within-subject reviewer recommendation. Due to the results from our questionnaire, about 30% of the respondents reports that they are unfamiliar and unsuitable for the recommendation. That is to say, recommending reviewers by subject would lead to an accuracy of around 70%. We believe the data shows the low accuracy of the subject-based method.

The Precision measures whether reviewers are suitable to review the recommended proposals; the Recall measures whether proposals that reviewers are suitable to review are recommended to them. Regarding the evaluation of the reviewer recommendation, it is more important to measure whether the recommended reviewers are suitable for the proposal, which directly affects the efficiency of peer review. It makes the significance of reviewer recommendation methods with high precision. As for the Recall, it is not as important as the precisions. For example, 100 researchers are suitable for a proposal and

5 of them are recommended. It would lead to a low recall, while it is a reasonable recommendation. The precision and recall are trade-off relationships, and the reviewer recommendation method would emphasize more on precisions. The result shows that the RRM owns the highest precision overall and it is selected as the ranking fusion method of the proposed reviewer recommendation method for scientific research proposals.

Conclusions

Since the implementation of the peer review, evaluation from reviewers has played a vital role in deciding the approval of scientific research projects and the qualification of scientific achievements. Finding suitable reviewers for scientific research proposals is fundamental to ensuring the fairness, justice, and efficiency of the peer review. Therefore, it is of great theoretical and practical significance to develop the reviewer recommendation for scientific research proposals.

This research proposes a reviewer recommendation method (RRM) for scientific research proposals, which significantly improves the accuracy of reviewer recommendation: the Precision increases by about 17%, and the Strict-precision increases by about 15%, compared to baseline methods. This research contributes to the scholarly endeavor toward the automated reviewer recommendation.

First, this research constructs a reviewer knowledge representation model and a proposal representation model considering semantic relationships, which significantly improve the accuracy of the reviewer recommendation. The common reviewer recommendations usually take terms or disciplines as the basic units to depict reviewers and proposals and ignore the semantic relationships between them. This research introduces the word embedding technique to represent reviewers and proposals as low-dimensional continuous vectors. In this way, we can calculate the similarity semantically. This model also could avoid the sparse problem led by VSM. The results show that proposed representation models bring a significant increase in the quality of the reviewer recommendation.

Second, this research introduces the ranking fusion method for reviewer recommendations, which highly improves the accuracy. This research takes advantage of multi-source data aiming at depicting the reviewer comprehensively. For single-source data, we can take the reviewer recommendation as calculating the cosine distance between the reviewer knowledge representation model and the proposal representation model. As for multi-source data, previous methods usually take the "average" of the similarities in multi-systems for the recommendation. This research discusses the shortcomings of previous methods and compares the W2V-SumRank, and RRM with the W2V-Average. The results show that the W2V-SumRank and RRM would improve the accuracy in Precision and Strict-precision.

Third, this research validates the results of the reviewer recommendation by a questionnaire to the candidate reviewer, which provides a first-hand evaluation. This research invites the candidate reviewers to rate suitable proposals, through a questionnaire. The questionnaire results can verify the accuracy of the proposed reviewer recommendation method. The evaluation of the reviewer recommendation is sticky since the actual reviewer is not publicly available. It is still a lack of standard recommendation datasets, especially for the scientific research proposal of NSFC. The common practice would invite some reviewers in the field to evaluate the results. This research improves the evaluation by introducing first-hand evaluation from reviewers.

Last but not the least, this research takes the reviewer recommendation for scientific research proposals of the National Natural Science Foundation of China as a case study, which verified the feasibility and effectiveness of the proposed method. Research results can provide feasible options for the decision-making of the committee, and improve the efficiency of scientific funding agencies.

There are some considerations to the present study. Firstly, validation datasets for reviewer recommendation of the funding agency need to be continuously constructed. The wildly-used datasets for reviewer recommendation are usually set in the scenario of recommending reviewers for a paper, which utilizes the single-source data. In the case of recommending reviewers for proposals, the nature of proposals and papers is different. The multi-source data, e.g., proposals and papers, should be utilized. We would like to build a validation dataset for proposal reviewer recommendations in the future. Indeed, both Chinese and English papers are of great value for the expert recommendation, which is also the limitation of this research. We do appreciate combining the Chinese and English papers as data sources for the reviewer recommendation in the future.

Secondly, how to evaluate the accuracy of the reviewer recommendation for funding agencies remains a sticky problem. Previous research on proposal reviewer recommendations for a national research management institute mainly focuses on the feasibility and time cost. For example, Shon and his colleagues (2017) come up with a proposal reviewer recommendation system based on big data for the Korean research management institute, and the efficiency test focuses on the execution time. Silva et al. (2013) propose a social network-empowered research analytics framework for project selection for NSFC, and the evaluation emphasis on the computing efficiency and the feedback from review panels. As far as we know, the accuracy of proposal reviewer recommendations is still lacking quantitative analysis. This research uses a questionnaire to validate the reviewer recommendation results for NSFC. It gives a way to quantify the accuracy of proposal reviewer recommendations. We would like to expand the distribution of the questionnaire and improve the response rate, aiming at providing a solid base for reviewer recommendations.

Thirdly, how to define the candidates for peer review is one of the important factors affecting the reviewer recommendation results. In this research, reviewers who have undertaken projects of a certain discipline are taken as candidates for the reviewer recommendation, but these candidates are often limited. When faced with a large number of scientific research proposals, only relying on reviewers who have undertaken projects before may cause a shortage of options. For reviewers who have not undertaken projects before, especially for young scholars, a selection mechanism should be established. It can not only expand the range of candidates but also increase the diversity of reviewers.

Finally, the research takes the reviewer recommendation as ranking a list of reviewers with the highest similarities, which does not consider the diversity of the reviewer group. The diversity of the reviewer group would influence the peer review to some extent. Future research could address how to construct a reviewer group with diversity to achieve better peer review.

The reviewer recommendation method for scientific research proposals is still in the era of combining manual assignment and automatic recommendation. Rather, the proposed reviewer recommendation method can provide feasible options for the decision-making of the committee. Along with the development of techniques from machine learning and natural language processing, the reviewer recommendation method will advance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-022-04389-4.

Acknowledgements This work was undertaken with support from the National Natural Science Foundation of China (Award #72104013, #71673024). The findings and observations contained in this work are those of the authors and do not necessarily reflect the views of the National Natural Science Foundation of China.

References

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., Durand-Zaleski, I., & Alberti, C. (2012). Peer review of grant applications: Criteria used and qualitative study of reviewer practices. *PLoS ONE, 7*, e46054.

Alhosan, N., Fayyoumi, A., & Faris, H. (2014). Shaping an experts locator system: Recommending the right expert. *Journal of Theoretical & Applied Information Technology, 66*, 645–651.

Arora, S., Liang, Y., & Ma, T. (2016). A simple but tough-to-beat baseline for sentence embeddings. *International conference on learning representations.*

Balog, K., Azzopardi, L., & De Rijke, M. (2006). Formal models for expert finding in enterprise corpora. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 43–50). ACM.

Balog, K., & de Rijke, M. (2008). Associating people and documents. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. W. White (Eds.), *Advances in information retrieval* (pp. 296–308). Springer.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise retrieval. *Foundations and Trends® in Information Retrieval, 6*, 127–256.

Biswas, H. K., & Hasan, M. M. (2007). Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment. *2007 International Conference on Information and Communication Technology* (pp. 82–86). IEEE.

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 546–556). Association for Computational Linguistics.

Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). Expertise identification using email communications. *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 528–531).

Chen, B., Mueller, C., & Willett, P. (2010). Combination rules for group fusion in similarity-based virtual screening. *Molecular Informatics, 29*, 533–541.

Cook, W. D., Golany, B., Kress, M., Penn, M., & Raviv, T. (2005). Optimal allocation of proposals to reviewers to facilitate effective ranking. *Management Science, 51*, 655–661.

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *International ACM SIGIR conference on research and development in information retrieval* (pp. 758–759).

Daud, A. (2012). Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems, 26*, 154–163.

Daud, A., Li, J. Z., Zhou, L. Z., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems, 23*, 615–625.

Davoodi, E., Kianmehr, K., & Afsharchi, M. (2013). A semantic social network-based expert recommender system. *Applied Intelligence, 39*, 1–13.

Erjavec, T., & Fišer, D. (2006). Building the Slovene Wordnet: First steps, first problems. *Proceedings of the third international WordNet Conference—GWC.*

Fellbaum, C. (1998). WordNet: An electronic lexical database. *The encyclopedia of applied linguistics.* MIT Press.

Galke, L., Saleh, A., & Scherp, A. (2017). Word embeddings for practical information retrieval. *INFORMATIK 2017.*

Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., & Ronen, I. (2013). Mining expertise and interests from social media. *Proceedings of the 22nd international conference on World Wide Web* (pp. 515–526). ACM.

Han, S., Jiang, J., Yue, Z., & He, D. (2013). Recommending program committee candidates for academic conferences. *Proceedings of the 2013 workshop on Computational scientometrics: theory & applications* (pp. 1–6).

Henriksen, A. D., & Traynor, A. J. (1999). A practical R&D project-selection scoring tool. *IEEE Transactions on Engineering Management, 46*, 158–170.

Hoang, D. T., Nguyen, N. T., Collins, B., & Hwang, D. (2021). Decision support system for solving reviewer assignment problem. *Cybernetics and Systems, 52*, 379–397.

Hu, K., Wu, H. Y., Qi, K. L., Yu, J. M., Yang, S. L., Yu, T. X., Zheng, J., & Liu, B. (2018). A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word-2Vec model. *Scientometrics, 114*, 1031–1068.

Huang, Y., Porter, A., Zhang, Y., & Barrangou, R. (2019). Collaborative networks in gene editing. *Nature Biotechnology, 37*, 1107–1109.

Karimzadehgan, M., Zhai, C., & Belford, G. (2008). Multi-aspect expertise matching for review assignment. *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1113–1122).

Leicht, E. A., Holme, P., & Newman, M. E. (2006). Vertex similarity in networks. *Physical Review E, 73*, 026120.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*, 1019–1031.

Li, D., Hao, Y., & He, X. (2014). The review and reconstruction for the peer review expert information database of science foundation. *Bulletin of National Natural Science Foundation of China, 2014*, 209–213.

Lin, S., Hong, W., Wang, D., & Li, T. (2017). A survey on expert finding techniques. *Journal of Intelligent Information Systems, 49*, 255–279.

Liu, O., Wang, J., Ma, J., & Sun, Y. (2016). An intelligent decision support approach for reviewer assignment in R&D project selection. *Computers in Industry, 76*, 1–10.

Liu, X., & Porter, A. L. (2020). A 3-dimensional analysis for evaluating technology emergence indicators. *Scientometrics, 124*, 27–55.

Ma, J., Xu, W., Sun, Y. H., Turban, E., Wang, S., & Liu, O. (2012). An ontology-based text-mining method to cluster proposals for research project selection. *IEEE Transactions on Systems, Man, and Cybernetics Part a: Systems and Humans, 42*, 784–790.

Macdonald, C., & Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task. *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 387–396). ACM.

Macdonald, C., & Ounis, I. (2008). Voting techniques for expert search. *Knowledge and Information Systems, 16*, 259–280.

Magooda, A. E., Zahran, M., Rashwan, M., Raafat, H., & Fayek, M. (2016). Vector based techniques for short answer grading. *The Twenty-Ninth International Flairs Conference.*

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

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

Mirzaei, M., Sander, J., & Stroulia, E. (2019). Multi-aspect review-team assignment using latent research areas. *Information Processing & Management, 56*, 858–878.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. *Proceedings of ACL-08: HLT* (pp. 236–244).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*, 1388–1429.

Mockus, A., & Herbsleb, J. D. (2002). Expertise browser: A quantitative approach to identifying expertise. *Proceedings of the 24th international conference on software engineering. ICSE 2002* (pp. 503–512). IEEE.

Nuray, R., & Can, F. (2006). Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management, 42*, 595–614.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Pradhan, T., Sahoo, S., Singh, U., & Pal, S. (2021). A proactive decision support system for reviewer recommendation in academia. *Expert Systems with Applications, 169*, 114331.

Protasiewicz, J., Pedrycz, W., Kozłowski, M., Dadas, S., Stanisławek, T., Kopacz, A., & Gałężewska, M. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems, 106*, 164–178.

Sachdeva, P., Verma, S., & Singh, S. K. (2014). An improved approach to word sense disambiguation. *2014 IEEE international symposium on signal processing and information technology (ISSPIT)* (pp. 235–240). IEEE.

Serdyukov, P., & Hiemstra, D. (2008a). Being omnipresent to be almighty: The importance of the global web evidence for organizational expert finding. *Proceedings of the SIGIR 2008a Workshop on Future Challenges in Expertise Retrieval (fCHER)* (pp. 17–24).

Serdyukov, P., & Hiemstra, D. (2008). Modeling documents as mixtures of persons for expert finding. *European conference on information retrieval* (pp. 309–320). Springer.

Serdyukov, P., Rode, H., & Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1133–1142). ACM.

Sharma, Y., Agrawal, G., Jain, P., & Kumar, T. (2017). Vector representation of words for sentiment analysis using GloVe. *2017 International conference on intelligent communication and computational techniques (ICCT)* (pp. 279–284). IEEE.

Shen, X. X., Yi, B. L., Liu, H., Zhang, W., Zhang, Z. L., Liu, S. Y. Y., & Xiong, N. X. (2021). Deep variational matrix factorization with knowledge embedding for recommendation system. *IEEE Transactions on Knowledge and Data Engineering, 33*, 1906–1918.

Shon, H. S., Han, S. H., Kim, K. A., Cha, E. J., & Ryu, K. H. (2017). Proposal reviewer recommendation system based on big data for a national research management institute. *Journal of Information Science, 43*, 147–158.

Silva, T., Guo, Z., Ma, J., Jiang, H., & Chen, H. (2013). A social network-empowered research analytics framework for project selection. *Decision Support Systems, 55*, 957–968.

Speer, R., & Havasi, C. (2013). ConceptNet 5: A large semantic network for relational knowledge. *The people's web meets NLP* (pp. 161–176). Springer.

Sriramoju, S. B. (2015). A framework for keyword based query and response system for web based expert search. *International Journal of Science and Research, 6*, 391.

Stankovic, M., Wagner, C., Jovanovic, J., & Laublet, P. (2010). *Looking for Experts? What can Linked Data do for you?* LDOW.

Tan, S., Duan, Z., Zhao, S., Chen, J., & Zhang, Y. (2021). Improved reviewer assignment based on both word and semantic features. *Information Retrieval Journal, 24*, 175–204.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141–188.

Vivacqua, A. S., Oliveira, J., & De Souza, J. M. (2009). i-ProSE: Inferring user profiles in a scientific context. *The Computer Journal, 52*, 789–798.

Vo, D.-T., Hai, V. T., & Ock, C.-Y. (2015). Exploiting language models to classify events from Twitter. *Computational Intelligence and Neuroscience, 2015*, 4.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W. G., & Zhang, Z. J. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems, 54*, 1442–1451.

Wang, K., Wang, C. K., & Hu, C. (2005). Analytic hierarchy process with fuzzy scoring in evaluating multidisciplinary R&D projects in China. *IEEE Transactions on Engineering Management, 52*, 119–129.

Wang, Q., Ma, J., Liao, X., & Du, W. (2017). A context-aware researcher recommendation system for university-industry collaboration on R&D projects. *Decision Support Systems, 103*, 46–57.

Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X. (2015). A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications, 42*, 2264–2275.

Xu, W., Sun, J., Ma, J., & Du, W. (2016). A personalized information recommendation system for R&D project opportunity finding in big data contexts. *Journal of Network and Computer Applications, 59*, 362–369.

Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y. K., & Xu, W. (2012). Combining social network and semantic concept analysis for personalized academic researcher recommendation. *Decision Support Systems, 54*, 564–573.

Yancheva, M., & Rudzicz, F. (2016) Vector-space topic models for detecting Alzheimer's disease. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 2337–2346) Long Papers.

Yang, C., Ma, J., Silva, T., Liu, X. Y., & Hua, Z. S. (2015). A multilevel information mining approach for expert recommendation in online scientific communities. *Computer Journal, 58*, 1921–1936.

Yong, Y., Yao, Z., & Zhao, Y. (2021) A framework for reviewer recommendation based on knowledge graph and rules matching. *IEEE International Conference on Information Communication and Software Engineering (ICICSE)* (pp. 199–203). Sichuan Inst Elect.

Yukawa, T., Kasahara, K., Kato, T., & Kita, T. (2001). An expert recommendation system using concept-based relevance discernment. *Proceedings 13th IEEE International Conference on Tools with Artificial Intelligence. ICTAI 2001* (pp. 257–264). IEEE.

Zainal, A. A., Yusri, N., Malim, N., & Arif, S. M. (2013). The Influence of similarity measures and fusion rules toward turbo similarity searching. *Procedia Technology, 11*, 823–833.

Zhang, M., Ma, J., Liu, Z., Sun, J., & Silva, T. (2016). A research analytics framework-supported recommendation approach for supervisor selection. *British Journal of Educational Technology, 47*, 403–420.

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H. S., & Zhang, G. Q. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics, 12*, 1099–1117.

Zhao, S., Zhang, D., Duan, Z., Chen, J., Zhang, Y. P., & Tang, J. (2018). A novel classification method for paper-reviewer recommendation. *Scientometrics, 115*, 1293–1313.