Scientometrics (2022) 127:6787–6808
https://doi.org/10.1007/s11192-022-04555-8

Academic collaborations: a recommender framework spanning research interests and network topology

Xiaowen Xi[1] · Jiaqi Wei[2] · Ying Guo[2] · Weiyu Duan[2]

Received: 16 May 2021 / Accepted: 5 October 2022 / Published online: 17 October 2022

Abstract
Fruitful academic collaborations have become increasingly more important for solving scientific problems, participating in research projects, and improving productivity. As such, frameworks for recommending suitable collaborators are attracting extensive attention from scholars. In an effort to improve on the current solutions, we have developed an approach that produces recommendations with better precision, recall, and accuracy. Our strategy is to comprehensively consider the similarity of both scholars' research interests and their collaboration network topologies, leveraging the benefits of these two common similarity indicators into one unified collaborator recommendation framework. A Word2Vec model creates word embeddings of research interests, which solves the problem of calculating similarity solely based on co-occurrence, not context, while a Node2Vec model automatically extracts and learns the topological features of a co-authorship network, moving beyond just local features to capture global network features as well. Then the CombMNZ method is used to fuse the results of the two similarity measures. A ranked collaborator list is then generated to recommend potential collaborators to the target scholars. The workings of the framework and its benefits are demonstrated through a case study on academics in the field of intelligent driving and a comparison with the three baselines: Random Walk with Restart (RWR), Latent Dirichlet Allocation (LDA), and Researcher's Interest Variation with Time (RIVT). Our framework should be of benefit to academics, research centers, and private-enterprise R&D managers who are seeking partners. We hope that, through the framework's recommendations, collaborators will form strong partnerships and be able to achieve the ultimate goal of completing research projects, solving scientific problems, and promoting discipline development and progress.

Keywords Academic collaborator recommendation · Research interest · Network topology · Word embedding · Network embedding

✉ Ying Guo
guoying_cupl@126.com

[1] Archives of Chinese Academy of Sciences, Beijing, China

[2] China University of Political Science and Law, Beijing, China

Introduction

As scientific endeavors become more complex and more comprehensive, academic collaborations have gradually become the primary means of conducting scientific research. Thus, the scale and scope of academic collaborations has burgeoned with the rapid development of scientific activity (Merton, 1973; Earlier work by Ziman, 1994). Academic collaboration means establishing a joint research team of scholars to exchange knowledge, share resources, and, hopefully, use the power of new and different perspectives to generate thinking that is greater than the sum of its parts. The ultimate goal, of course, is to successfully complete research projects, find high-quality solutions to scientific problems with greater efficiency, and to contribute to the development and progress of an entire field. As Guns and Rousseau (2014) state, academic collaboration helps to improve the efficiency of scientific inquiry and research output (see also Abramo et al., 2009; Tang, 2013). In this vein, scholars and scientists, just like the professionals in any sector, typically aspire to collaborate with researchers of the highest-level possible in their field.

In the late 1990s, Melin (2000) conducted a survey to determine the main factors influencing the decision to collaborate with one scholar over another. What the results show is that possessing special skills and owning unique data or equipment are primary considerations for scholars in academic collaboration. Some of the recent studies have considered the research interests of scholars. For instance, Gang et al. (2015) point out that having similar research interests is an indispensable and intrinsic motivation promoting the collaboration of scholars. Cooperation with high-level scholars can effectively reduce the cost of research and improve the efficiency and quality of the research that is done (Lab & Tollison, 2000).

From the perspective of research results, the co-authorship network is one of the most significant manifestations of academic collaboration. In a co-authorship network, the nodes represent authors and the links represent co-author relationships. As such, co-authorship networks impart a great deal of information about scientific collaboration, and so they have been widely used in associated thematic fields. For example, they have been used to analyze author collaboration structures, discover scientific research communities, and discuss interdisciplinary relations. They have also been used in systems to recommend potential academic collaborators (Abramo et al., 2012; Guns & Rousseau, 2014). Therefore, the recommendation framework presented in this article mainly focuses on the similarities between scholars' research interests and the topologies of the co-authorship networks they belong to.

Existing systems generally fall into two categories: collaborator recommendations based on similar research interests (Kong et al., 2017) and collaborator recommendations based on the structure of co-authorship networks (Dong et al., 2013; Pham et al., 2011). Frameworks based on similar research interests mostly use scholars' published papers as a data source with a term frequency-inverse document frequency (TF-IDF) topic model as the method used to extract the scholars' research interests in the form of keywords and subject terms. Additionally, all these methods use a bag-of-words model to represent the thesis of the document. Yet a bag-of-words cannot and does not reflect the order, semantics, and syntactic relations of a document. Another problem with this approach is that the content of the document is represented with high-dimensional sparse vectors, so problems with computational complexity arise as the amount of data increases (Li et al., 2018a, 2018b; Wang et al., 2016). The frameworks based on the structure of co-authorship networks generally calculate similarity through the network's topological features using indicators like

the common neighbor index, the restart random walk index, the local path index, and so on. However, these indicators mostly represent the network using a discrete adjacency matrix. Yet the high dimensionality and sparseness of the matrix means that issues with computational efficiency can arise. Further, existing studies only consider the first-order adjacency relationships between nodes; they ignore the network's more complicated higher-order structural relationships, such as pathways and frequent substructures (Li et al., 2017).

To address these issues, our framework takes both the scholars' research interests and their co-authorship network structures into account. The framework, which combines both word and network embeddings, produces collaborator recommendations through three main steps: (1) the research interests of scholars are extracted from a corpus of articles with the Word2Vec model, then the similarity between scholars' interests is calculated in terms of cosine distance; (2) a co-authorship network is constructed, then the similarities between topological features are extracted and calculated with the Node2Vec model; and (3) the results of both similarity measures are integrated using the CombMNZ method and sorted according to a final similarity score to produce a ranked list of recommendations.

To verify the effectiveness and efficiency of our framework, we conducted an empirical analysis on the field of intelligent driving and compared the results with three single-feature models for finding potential collaborators. The results show our recommendations are more accurate, have a recall rate and have a higher F1 score. Notably, the approach can be applied to a range of fields/sectors/industries with little to no modifications, and the results can provide useful insights and recommendations to academics, research centers, and private sector R&D managers.

The remainder of this paper is organized as follows. Section 2 contains a review of work related to the recommendation for academic collaborators. The details of our proposed model are presented in Sect. 3. The empirical study and results are given in Sect. 4. Section 5 offers our conclusions, the limitations of this research, and avenues for future study.

Related works

Recommending collaborators with similar research interests

Researchers have explored many different ways of improving the accuracy of collaboration recommendations. Among these methods, recommendations based on similar research interests is the mainstream. These studies usually revolve around text mining techniques, using titles, abstracts, and keywords to extract authors' research interests. Text matching is also used to recommend scholars with similar research interests.

In early studies on general recommender systems, a scholar's research interests were mainly captured by extracting salient terms and phrases from a dataset. Balabanovic and Shoham (1997), for example, extracted 100 important keywords from web pages and used them as recommendation results for those searching for similar content. The next main advancement came with feature weighting through techniques such as TF-IDF, i.e., determining the importance of particular words and weighting them accordingly (Taşcı & Güngör, 2013; Ping & De-Gen, 2016). However, due to the ambiguity of natural language (synonyms, polysemy, etc.), comparing scholars based on keywords does not always accurately reflect the actual similarity of the researchers' academic interests. Topic models are credited with solving some of these ambiguity problems and also for raising general interest in feature extraction. One representative approach to topic modeling is LDA, which

involves constructing a three-layer Bayesian model to discover hidden topic structures in large document collections (Blei et al., 2008). Due to its advantages, researchers have widely employed LDA models to mine and extract the features of scholars' research interests. This approach substantially improves the accuracy of the recommendations (Kong et al., 2017; Weng et al., 2010). Along these lines, Mimno and McCallum (2007) proposed a novel topic model called Author-Persona-Topic (APT). In this model, each author can write under one or more "personas", which are represented as independent distributions over hidden topics. Kawamae (2010) presented the Latent Interest Topic model (LIT), which introduces a latent variable into each document and each author layer in a coherent generative model. Rosen-Zvi et al. (2010) proposed a new unsupervised learning technique for extracting information about authors and topics from large text collections. Considering that a scholar's research interests will generally change over time, Xu et al. (2014) devised the Author-Topic over Time (AToT) model. Pradhan et al. (2020) proposed the Researcher's Interest Variation with Time, which gives more weight to recently published papers in order to capture the current areas of interest of the researcher. However, as mentioned, all these models treat the documents as a bag-of-words and assume that words occur independently without considering the contextual semantics of the document, so the resulting recommendations cannot be completely reliable.

The Word2Vec model excels at resolving the abovementioned issues. It is an efficient word embedding technique that can learn the semantics of terms in a context and can produce a dense low-dimensional vector for each word. It has been proven to capture precise syntactic and semantic word relationships and form high-quality word vectors from massive unstructured text-based datasets (Mikolov et al., 2013a, 2013b). The power of the Word2Vec model has made it a hugely successful solution to many document classification, topic extraction, and clustering problems. For example, Li et al., (2018a, 2018b) proposed a novel clustering model combining LDA and Word2Vec that yields highly accurate results when clustering article abstracts. Zhang et al. (2018) combined Word2Vec with k-kernel clustering to produce a new method of topic extraction. Lilleberg et al. (2015) proposed a novel text classification model that integrates Word2Vec and semantic features with support vector machine (SVM). These applications verify that the Word2Vec model has more advantages than topic models, TF-IDF, and so on.

Word2Vec also paves the way for applications that can recommend academic collaborators. On these grounds, we applied the Word2Vec model to extract more finely-grained features representing the research interests of scholars in our framework to improve the accuracy of the recommendations.

Recommending collaborators with similar network topology

Among the network analysis approaches to collaborator recommendation, co-authorship prediction is an important line of work. In this field, many different similarity indicators have been explored to analyze and predict which candidate collaborations are likely to have the most potential for success (Zhang et al., 2015). The types fall into two main categories: one based on node attributes, the other on network topology.

Node indicators reflect characteristics like the researcher's affiliation, geographic location, etc. Recommendations are made by calculating the similarity of these features between pairs of scholars (Gollapalli et al., 2012; Shibata et al., 2012). For example, Liben-Nowell & Kleinberg (2007) constructed a co-authorship network in the field of physics, and then used the paper title, institution, and geographic location features to recommend

collaborators. However, like all information science methods, these approaches have some limitations. First, node attribute information is often difficult to obtain and, second, once obtained, relevant attributes need to be filtered from irrelevant attributes. Determining which attributes are relevant is one part of this problem, and how to filter them is the other.

With the development of complex networks, link prediction methods have been widely employed to forecast links between authors that might appear in future—these future links being potential academic partners (Lv et al., 2010). For example, Kong et al. (2017) used a RWR indicator to measure the academic impact of researchers in a co-authorship network. Xia et al. (2014), also using a RWR model, considered three academic factors to determine the importance of each link in a social network: co-author order, most recent collaboration date, and the number of joint papers published. Pradhan and Pal (2020) introduced a multi-level fusion-based model for collaborator recommendation and employed an RWR model to recommend the top N collaborators. These studies have greatly improved the accuracy of recommendations; however, the efficacy of all these indicators mainly depends on manual design and selection. Yet, practically speaking, extracting topological features should be an automatic process. In addition, these indicators can only extract the local structural features around a node; accurately extracting global structures is still a complicated task.

Like Word2Vec does with phrases and sentences, Node2Vec is able to transform network structures into a low-dimensional vector space and then calculate the semantic connections between nodes while effectively preserving the global network structure (Peng et al., 2018). Also, like Word2Vec, Node2Vec's clear advantages over other models has seen it applied to a wide variety of tasks. For example, Deepika et al. (2018) introduced Node2Vec as a model for predicting the contraindications between drugs, proving it performed better than a basic classifier. Hu et al. (2019) used Node2Vec to detect communities in a complex network, demonstrating the advantages of the model in learning the topological features of nodes, while Kazemi and Abhari (2020) applied Node2Vec to feature extraction from the scientific literature. In the related research of recommender system, Chen et al. (2017) put forward an improved spectral clustering-based collaborative filtering framework based on the Node2Vec algorithm. The framework overcomes the sparsity and efficiency challenges encountered by traditional recommendation methods. Liu et al. (2020) proposed a deep learning-enhanced framework for implicit feedback recommendation. They also learned new distributed representations of users and items via Node2Vec to improve their negative sampling strategy. These applications not only verify that the Node2Vec model has distinct advantages in network analysis but also that it could be used to build a solid application for recommending academic collaborators. Thus, we selected Node2Vec as our method of automatically extracting the topological features from co-authorship networks for subsequent similarity calculations.

Methodology

The recommendation framework for academic collaborators proposed in this paper is based on word embedding and a network embedding model. It comprises four main steps: (1) data acquisition and preprocessing; (2) using Word2Vec to train word vectors and measure the similarity of research interests between scholars; (3) using Node2Vec to train the node vectors and measure the similarity of network topological features between scholars; and (4) using the CombMNZ model to integrate the above results to form an appropriate academic collaborator recommendation list. The validity of the model in this paper is verified

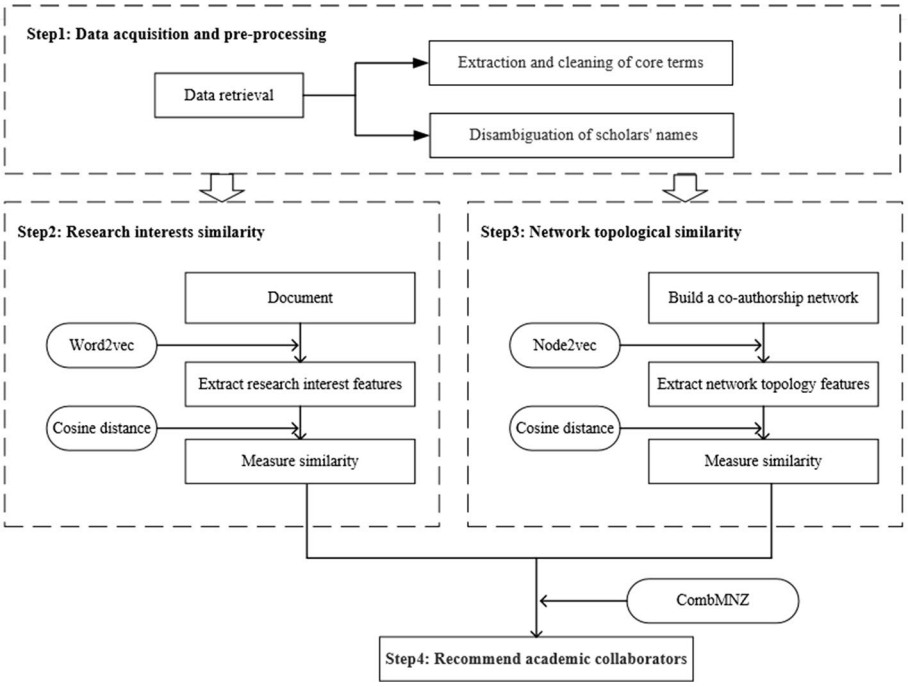

Fig. 1 Analytical framework

by comparative analysis with other models. An overview of the framework is provided in Fig. 1.

Data acquisition and preprocessing

This step involves retrieving and downloading academic articles from the Web of Science database and performing data preprocessing on academic information contained in the articles, such as the scholar's research interests and co-authorship relationships. We use professional desktop text mining software VantagePoint[1] to extract key features such as the author, year of publication, title, and abstract. With a raw dataset of terms assembled, the subsequent data preprocessing procedure cleans the terms and disambiguates the author names in two separate steps.

The purpose of cleaning the terms is to extract the terms that most intuitively express the academic interests of the authors; these are the "core" terms. The procedure is as follows. First, the title and abstract fields are merged, and VantagePoint performs word segmentation. Noise is then removed, and synonyms are merged with a term clumping process based on a fuzzy semantic matching algorithm developed by Zhang et al. (2014). We extracted the terms and phrases that appeared more than six times for further analysis, and

[1] https://www.thevantagepoint.com/.

deleted the general and irrelevant terms, such as development, methods, significant, etc. The final culled list forms the vocabulary of core terms.

Non-standardized scholar names, such as abbreviated initials and full first names can reduce the reliability and accuracy of the final recommendations given. Therefore, cleaning and disambiguating the scholar names was essential. Considering that the probability of two persons with the same name working at the same institution is much smaller than the probability two people with the same name at different institutions, we cleaned the names of scholars by institution. First, we matched scholars by research institution and then constructed a correspondence matrix between the scholars and the research institutions. The fuzzy matching algorithm in VantagePoint then merged similar names followed by a manual judgment process to make the final merges and selections.

Research interest similarity

Step two is a two-stage process. First, the research interests of the scholars need to be extracted from the corpus of articles, then the cosine similarity of interests between pairwise scholars needs to be calculated.

Interest mining with Word2Vec

Word2Vec focuses on sequential combinations of words in a corpus and exploits the idea of neural networks to train a language model that maps each word to a vector. Word2Vec includes two model options for updating parameters to suit different situations (Mikolov et al., 2013a). The Skip-gram model is based on a sliding window scheme, where each target term is used to predict the surrounding words. Conversely, the CBOW model uses the surrounding words to predict the target word. For our purposes, the training procedure in Skip-gram produces a more accurate result.

Following Mikolov et al. (2013b), the main objective of Skip-gram model is to maximize the average logarithmic conditional probability Pt.

$$\mathrm{Pt} = \frac{1}{m}\sum_{j=1}^{m}\sum_{-i\le d\le i}\log \Pr\left(w_{j+d}|w_j\right) \tag{1}$$

where i is the size of the sliding window, w_{j-d} and w_{j+d} are the first and last words of the target word w_j, and m represents the total number of words in a given corpus. A softmax function is used to formulate the probability $\Pr\left(w_{j+d}|w_j\right)$ as follows:

$$\Pr\left(w_{j+d}|w_j\right) = \frac{\exp\left(v_{j+d}^{T} * v_j\right)}{\sum_{w=1}^{N}\exp\left(v_w^{T}\cdot v_j\right)} \tag{2}$$

where v_w^T represents the transpose of each word vector in the corpus, and N is the total number of words in the corpus.

To reduce the cost of calculating the word vectors and to accelerate vector learning, Mikolov et al. (2013b) use a negative sampling method. The basic idea is to maximize the joint probability of the positive and negative samples, which means the average logarithmic conditional probability becomes:

$$\mathrm{Pt} = \frac{1}{m}\sum_{j=1}^{m}\sum_{-i\le d\le i}\log\delta\left(x_{j+d}\cdot x_j\right) + k\cdot N\left(w' \sim w_{j+d}\right)\cdot\log\delta\left(x'\cdot x_j\right) \tag{3}$$

where $i.$ is the number of negative samples, and $N\left(w' \sim w_{j+d}\right)$ denotes the negative sample collection of context word w_{j+d}. The optimization is performed via stochastic gradient descent, and the gradients are calculated using backpropagation neural networks.

The next step is to produce accurate eigenvectors othe scholars' research interests. Specifically, given a series of documents $\mathrm{D} = \left\{d_1, d_2, \ldots, d_n\right\}$ with a vocabulary of N words $\left\{w_1, w_2, \ldots, w_n\right\}$, the Word2Vec model maps each word in the vocabulary to a fixed-length vector $\left\{v(w_1), v\left(w_2\right), \ldots, v\left(w_n\right)\right\}$ based on the co-occurrence relationship between the documents and words. The document vector $\mathrm{v}\left(d_i\right)$ is then calculated by adding each word vector as follows:

$$v\left(d_i\right) = \sum_{n=1}^{m}\mathrm{v}\left(w_n\right) \tag{4}$$

where m represents the number of words in the document.

The author vector $\mathrm{v}\left(c_i\right)$ is then computed by adding each document vector according to the co-occurrence relationships between documents and authors as follows:

$$v\left(c_i\right) = \sum_{i=1}^{n} v\left(d_i\right) \tag{5}$$

where n is the number of documents written by the author.

Calculating cosine similarity

With the fixed-dimension feature vectors of the research interests generated, the next step is to calculate the similarity of interests between researchers. Of the many methods of measuring similarity, we chose the popular and widely-used cosine similarity index, which is identified to be one of the best metrics for similarity calculations and more suitable for processing high-dimensional data (Dehak et al., 2011), formulated as

$$\mathrm{sim}(A,B) = \cos(A,B) = \frac{A\cdot B}{AB} = \frac{\sum_{j=1}^{m} a_j b_j}{\sqrt{\sum_{j=1}^{m} a_j^2} * \sqrt{\sum_{j=1}^{m} b_j^2}} \tag{6}$$

where the vector of author A is (a_1, a_2, a_3, a_4,…, a_m), and the vector of author B is (b_1, b_2, b_3, b_4,…, b_m).

Topological similarity

Calculating topological similarity follows a similar procedure: generate vectors and construct a co-authorship network through the Node2Vec model, then calculate the topological similarities between scholars in terms of cosine distance.

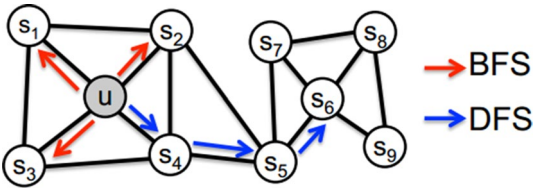

Fig. 2 The different search strategies of BFS and DFS from the initial node *u* (Grover & Leskovec, 2016)

Topology mining with Node2Vec

Building a co-authorship network is a simple process of mapping the links between the co-authors of the articles in the dataset. But mining network structure for author recommendations involves predicting whether a pair of nodes should be linked based on one or more similarity indicators. Common indicators include the common neighbor index, Random Walk with Restart, and the local path index. However, as mentioned in the literature review, the features of these similarity indicators need to be designed and defined manually, and they are probably not applicable to a random situation. Further, they can only measure local structures. Our solution to these problems is to replace the traditional indicators with a network embedding technique, i.e., Node2Vec, which maps the co-authors (nodes) to a low-dimensional feature space. This method has been proven to maximize the likelihood of preserving network neighborhoods (Cui et al., 2018).

The Node2Vec model adopts a second-order random walk strategy to sample the neighborhood nodes—an approach that smoothly interpolates between breadth-first sampling and depth-first sampling. To capture the adjacency attributes of node neighborhoods, breadth-first sampling generally takes samples from around the initial nodes, while depth-first sampling tends to sample nodes with similar network structures that are farther away (Grover & Leskovec, 2016). Therefore, Node2Vec is able to learn the node embedding representations within the same network community and node embedding representations with similar structural features. As shown in Fig. 2, if the number of nodes sampled is set to 3, then breadth-first sampling will sample the node sequence s_1, s_2, s_3, and depth-first sampling will sample the node sequence s_4, s_5, s_6.

To use an analogy with Word2Vec, the Node2Vec model regards a node as a word and regards the random paths it generates as a sentence. Thus, the neighbor(hood) of a node can be thought of as its context. The Skip-gram model is motivated by this idea, and, like Word2Vec, Node2Vec also uses Skip-gram to learn the node representations but with an architecture extended to suit networks (Grover & Leskovec, 2016).

The main objective of Node2Vec is to maximize the log-probability of the neighbors of node n in the node sequence:

$$\mathrm{Pt} = \sum_{n \in V} \log P\big(N_m(n) | f(n)\big) \tag{7}$$

where $f(n)$ is the current representation of n, and $N_m(n)$ is a network neighborhood of node n generated through the neighborhood sampling strategy m.

To ensure the optimization problem is tractable, Grover and Leskovec (2016) defined a formula for Pt based on the assumptions of conditional independence and a symmetrical feature space:

$$\mathrm{Pt} = \sum_{n \in V} \left[-\log E_u + \sum_{u \in N_m(n)} f(u) * f(n) \right] \tag{8}$$

where $E_u = \sum_{n \in V} exp(f(n) * f(v))$. To reduce the complexity of model training, negative sampling technique and stochastic gradient ascent method are used for model optimization (Mikolov et al., 2013b).

The process of acquiring the features of scholars' network topology begins by taking a co-authorship network $G = (V, E)$ as input. After running the Node2Vec model, we have an $n * \beta$matrixN_{raw} representing a focus author's network topology features, where n denotes the number of nodes, β is the parameter that determines the dimension of the node's vector representation, and the final output is $N_{raw} = \{v_1, v_2, \ldots, v_n\}^T$.

Calculating cosine similarity

Calculating the cosine similarity between the network topological features of each scholar follows the same basic principles as described in Sect. 3.2.2. Using the authors M and N as an example, where the vector of author M is (m_1, m_2, m_3, m_4,…, m_t) and the vector of author N is (n_1, n_2, n_3, n_4,…, n_t), the equation for calculating the cosine similarity between the authors is

$$\mathrm{sim}(M, N) = \cos(M, N) = \frac{M \cdot N}{MN} = \frac{\sum_{j=1}^{t} m_j n_j}{\sqrt{\sum_{j=1}^{t} n_j^2} * \sqrt{\sum_{j=1}^{t} n_j^2}} \tag{9}$$

Recommendations with CombMNZ

The goal in this stage is to integrate the two similarities and rank the candidate collaborators from high to low according to their similarity. Typical data fusion algorithms are either score-based or rank-based (Donald & Saari, 1999; Edward & Joseph, 1994). We opted for a score-based algorithm, more specifically CombMNZ (Macdonala & Ounis, 2008) because it is the most widely used in the field of recommendation.

To fuse the similarity results with CombMNZ in a fair way, the dimensions of each similarity measure first need to be standardized, as shown in Eq. (10):

$$\mathrm{Score}_{\mathrm{normal}} = \frac{\mathrm{score} - \mathrm{score}_{\mathrm{min}}}{\mathrm{score}_{\mathrm{max}} - \mathrm{score}_{\mathrm{min}}} \tag{10}$$

where $\mathrm{score}_{\mathrm{min}}$ denotes the lowest number of dimensions across the two values, and $\mathrm{score}_{\mathrm{max}}$ is the highest. Following Eunice et al. (2016), the CombMNZ calculation is then

$$\mathrm{Score}_{\mathrm{combmnz}} = n(j, w) * \sum_{n=1}^{N} m_n * \mathrm{score}_{\mathrm{normal}}(j, w_n) \tag{11}$$

where $n(j, w)$ denotes the number of times scholar j appears in the score w of each dimension, $\mathrm{score}_{\mathrm{normal}}(j, w_n)$ denotes the standardized score of the scholar in the w_n item ($w_n \leq 2$), and m_n denotes the weight of each dimension derived with a greedy strategy. On this basis

and according to the rankings on the recommendation list, scholars who have not cooperated before are recommended.

The assessment metrics

To verify the performance of our recommendation framework, we conducted extensive experiments following the measurement method proposed by Xia et al. (2014). In summary, we used precision, recall, and F1 as our evaluation metrics to make for an easy comparison with other contemporary methods. The calculation formulas for each evaluation index were:

$$\text{Precision} = \frac{A}{A+B} \tag{12}$$

$$\text{Recall} = \frac{A}{A+C} \tag{13}$$

$$F1 = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{14}$$

Given a recommendation from the testing set: A is the number of recommended scholars that have collaborated.B is the number of recommended scholars that have not collaborated. C is the number of scholars who have not been recommended but have collaborated.

Case study

To verify the efficacy of our framework and to showcase its benefits, we selected the field of intelligent driving as a case study. This is an area of research that is highly anticipated to be one of the next hot development trends in automotive engineering.

Sections 4.1 through 4.4 outline the experimental settings for each of the four main steps of the framework, along with the observations made at each step to validate the intermediate results. Section 4.5 provides a comparison of our overall recommendations with two of the most commonly used methods and a novel method: RWR, LDA and RIVT.

Data collection and preprocessing

To assemble our corpus, we retrieved papers published between 2010 and 2018 from the Web of Science database using a search strategy drawn from Kwon et al. (2019) as follows:

TS = (((Self-driving or autonomous or driverless) near/4 (transport* or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane))) or TS = (((drone near/2 autonomous) or (uav near/4 autonomous))) or TS = ((robot* near/1 (transport* or mobile or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane)) AND (autonomous or self-driving or driverless)) or TS = ("autonomous driv*") or TS = (((robot* near/1 (transport* or mobile or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane)) OR (drone or uav)) AND (path or planning or planner or plan)) or TS = (((robot* near/1 (transport* or mobile or car or motorcar or vehicle or automobile or

aircraft or airplane or aeroplane)) OR (drone or uav)) AND (2D or 2-D or 3D or 3-D or map or localization or tracking or navigat* or obstacle or avoid*)).

The search returned 36,433 records. NLP preprocessing with VantagePoint yielded 8,637 core terms and phrases that appeared in six or more records. Outstanding scientists were defined as those who had published N or more papers—a criteria put forward by Price (1963). Formally, the calculation is $N = 0.749\left(\beta_{\max}\right)^{1/2}$, where $\beta_{\max}$ is the highest number of papers published by any author in the dataset. Thus, we selected 813 researchers with five or more publications for future analysis.

We then selected data from 2010 to 2013 to complete the remaining three main steps of the recommendation framework (Sects. 4.2–4.4), along with the observations made at each step to validate the results. To further verify the quality of recommendations (Sect. 4.5), we first divided the publications into a training set and a test set by year, while trying to ensure the ratios chosen were integer ratios if possible. When taking 2013 as the cut-off point, we found that the ratio of training set and test set is almost 3:7. Therefore, we divided the 2010–2013 data into a training set containing 10,042 papers. 813 outstanding scholars were selected as the research objects. Hence, the constructed co-authorship network contained 813 nodes and 2,939 edges. The data from 2014 to 2018 was used as the test set. It contained 26,404 papers with co-authorship network containing 3,515 edges, of which 3,393 edges were newly formed collaboration relations and 122 edges reflected previous collaborations.

With this set, we compared the three baselines: RWR, LDA, and RIVT. RWR is coupled with a topological model. LDA discovers hidden topic structures in large document collections, which greatly improves the accuracy of the recommendations and has become one of the representative methods for collaborator recommendation (Blei et al., 2008; Kong et al., 2017; Weng et al., 2010). RIVT is a novel model for analyzing the distribution of authors' research interests. It adds a weighted vector to the topic distribution after LDA is performed (see Pradhan et al., 2020, for a more detailed description of the algorithm).

Constructing the network of research interests with Word2Vec

As mentioned, we selected data from 2010 to 2013 to construct the training network of research interests with Word2Vec. The model's parameters were all default parameters. We then generated the research interest vectors for all scholars based on the co-occurrence of core terms and documents, and the co-occurrence of documents and scholars. Equation (6) subsequently gave us an 813×813 symmetrical similarity matrix of research interests. Lopez-Franco Michel and Anonymous had the least similar at 0.881717. The mean value, median, and standard deviation values were 0.988698, 0.990406 and 0.00757, respectively.

We plotted the author similarity network using the ITGInsight tool (Wang et al., 2021). Figure 3 shows part of the author similarity network. Here, the size of the node represents the number of published papers for each scholar, and the thickness of the lines indicates the degree of similarity between the scholars' research interests.

Taking Sukhatme Gaurav S as an example (center far right), we find that the three authors with the most similar research interest to his are Hover Franz, Hollinger Geoffrey A, and Smith Ryan N, in that order. To verify the validity of these similarity scores, we conducted a manual analysis under the guidance of domain experts and found overlaps in several areas of their research, including underwater inspection, sensor networks, and ocean monitoring. In robotic sensor networks, underwater inspection, and underwater

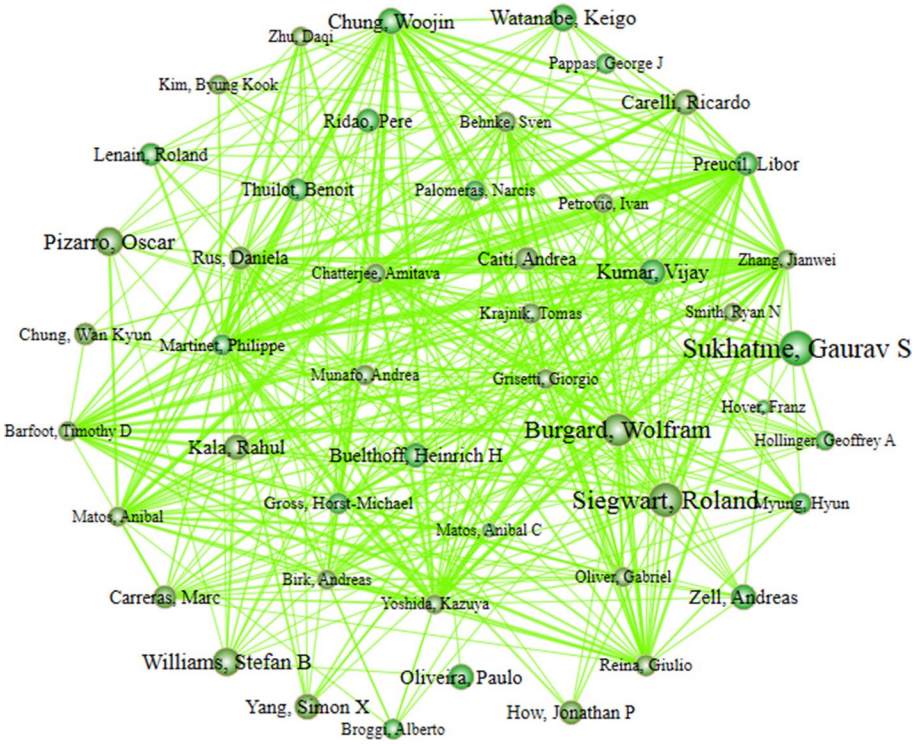

Fig. 3 The similarity network of research interests

vehicles, Sukhatme and Hollinger have jointly published about nine papers. Smith and Sukhatme have collaborated on underwater vehicles, particularly using ocean models, monitoring, prediction and tracking (Smith et al., 2011). Further, Hover mainly focus on robotic sensor networks, underwater data collection, and underwater inspections (Hollinger et al., 2011). Notably, Sukhatme, Hover and Hollinger have co-authored several papers on data collection using robotic sensor networks. Thus, there is no question that these authors share highly similar research interests (Hollinger et al., 2012) confirming the accuracy of the similarity calculations.

Constructing the co-authorship network with Node2Vec

In this section, we selected data from 2010 to 2013 to construct the co-authorship network with Node2Vec. The parameter settings for the Node2Vec model were all default settings. Equation (9) yielded the 813×813 topology matrix of cosine similarity between scholars, which was again symmetrical. Anvar Amir and Lu Tien-Fu shared the greatest similarity (0.999954), and Paley Derek A and Ramos Fabio had the least (0.000181). The mean, median, and standard deviation values were 0.485076, 0.455575, and 0.174119, respectively.

Figure 4 shows part of the author similarity network. The size of nodes indicates the number of collaborators associated with that scholar. The thickness of the lines represents the degree of similarity between the two connected scholars.

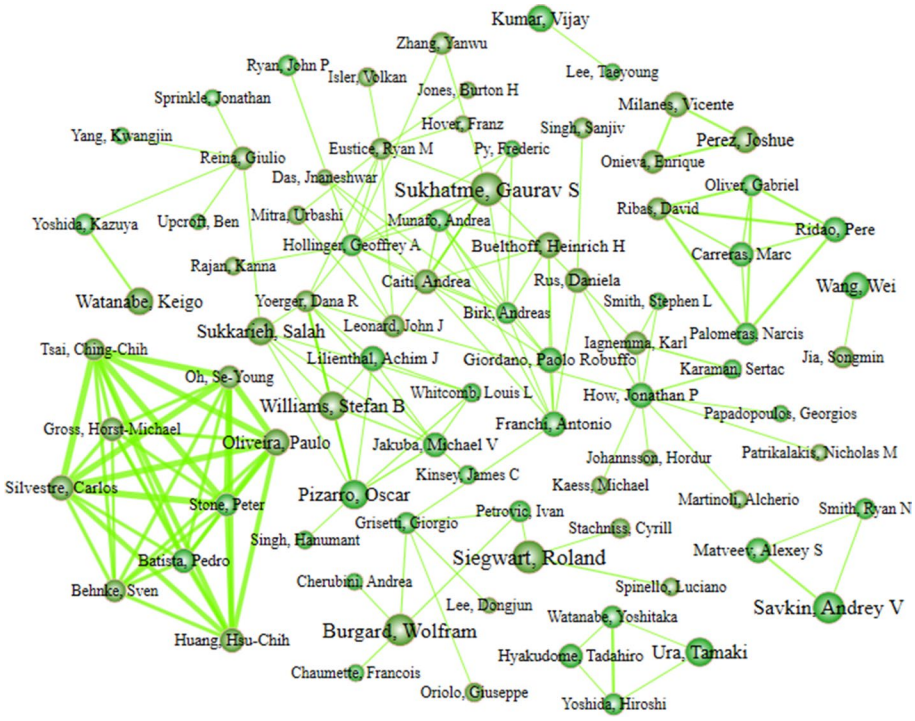

Fig. 4 The co-authorship network showing the topological similarities between scholars

With the help of domain experts, we found three kinds of positional relationships between researchers in this co-authorship network. Take Burgard Wolfram's cohort (lower left) as an example, and particularly Grisetti Giorgio, Stachniss Cyrill, and Petrovic Ivan, who are the three scholars with the highest similarity to Burgard in Fig. 4. Burgard, Petrovic and Grisetti are directly linked, which means they have collaborated on at least one paper. This is the first kind of positional relationship. Second, Burgard and Stachniss belong to different clusters in the network. Third, Burgard and Stachniss are connected through a common neighbor. These three positional relationships, especially the second, confirm that the network embedding method is able to capture global network topological factors, making up for the limitation of only considering common neighbors or paths.

Ranking the candidate collaborators with CombMNZ

As a preliminary assessment of the framework's ability to make appropriate recommendations, we randomly selected Caiti, Andrea as the target scholar and generated a list of the 20 scholars with the most similar research interests to hers, as shown in Table 1. Likewise, Table 2 shows the 20 scholars with the greatest topological similarity.

Using the CombMNZ method to integrate the two similarity indexes, we generated a final list of recommendations. The top-10 ranked candidates are shown in Table 3.

Table 1 The 20 scholars with the most similar research interests

Rank	Scholar	Similarity	Rank	Scholar	Similarity
1	Munafo, Andrea	0.999905	11	Anvar, Amir	0.997159
2	Calabro, Vincenzo	0.999804	12	Chitre, Mandar	0.997148
3	Casalino, Giuseppe	0.997896	13	Sousa, Joao	0.997059
4	Cruz, Nuno A	0.997686	14	Matos, Anibal C	0.997055
5	Kremer, Ulrich	0.997479	15	Ridao, P	0.996889
6	Sukhatme, Gaurav S	0.997464	16	Smith, Ryan N	0.996855
7	Sousa, J B	0.997415	17	Singh, Hanumant	0.996695
8	Woithe, Hans Christian	0.997280	18	Hover, Franz	0.996684
9	Ament, Christoph	0.997275	19	Seto, Mae	0.996661
10	Hover, Franz S	0.997253	20	Wietfeld, Christian	0.996630

Table 2 The 20 scholars with greatest topological similarity

Rank	Scholar	Similarity	Rank	Scholar	Similarity
1	Munafo, Andrea	0.995798	11	Sukhatme, Gaurav S	0.744630
2	Casalino, Giuseppe	0.995695	12	Lee, Dongjun	0.736304
3	Calabro, Vincenzo	0.995351	13	Giordano, Paolo Robuffo	0.726276
4	Birk, Andreas	0.990512	14	Mitra, Urbashi	0.714087
5	Aguiar, A Pedro	0.939023	15	Isler, Volkan	0.713870
6	Antonelli, Gianluca	0.923837	16	Grabe, Volker	0.712004
7	Chiaverini, Stefano	0.880891	17	Jones, Burton H	0.705787
8	Arrichiello, Filippo	0.873056	18	Oriolo, Giuseppe	0.700646
9	Buelthoff, Heinrich H	0.750107	19	Hollinger, Geoffrey A	0.699834
10	Franchi, Antonio	0.746333	20	Chao, Yi	0.690461

Table 3 The top 10 recommended collaborators for Roland Siegwart based on the 2010–2013 dataset

No	Recommended collaborator
1	Munafo, Andrea
2	Calabro, Vincenzo
3	Casalino, Giuseppe
4	Birk, Andreas
5	Antonelli, Gianluca
6	Arrichiello, Filippo
7	Aguiar, A Pedro
8	Chiaverini, Stefano
9	Sukhatme, Gaurav S
10	Jones, Burton H

An in-depth manual review of Caiti's academic background shows these recommendations to be appropriate. For example, Munafo and Caiti have overlapping interests in autonomous underwater vehicles, underwater acoustic network, mobile sensor networks, distributed algorithms, and more, and have both published many influential papers. In addition,

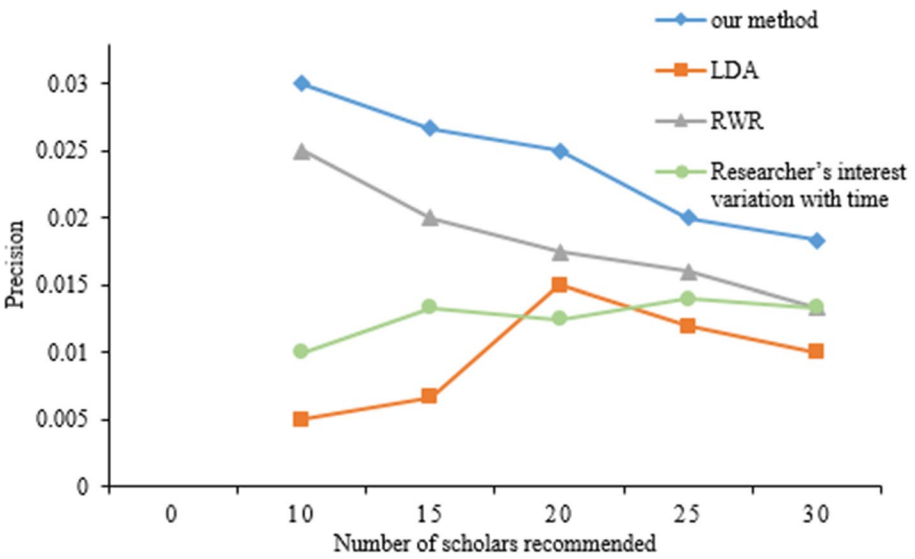

Fig. 5 Precision

both scholars often attend the IEEE OCEANS Conference. Similarly, the research interests of Arrichiello mainly include cooperative caging, autonomous aquatic surface vehicles, multi-robot systems, underwater robots, mobile robots, etc., which closely match Caiti's. They both work at universities in Italy and have published papers together on the mobile underwater sonar technology in the following years. Based on this analysis, it is reasonable to conclude that the framework can recommend realistic and fruitful collaborations.

Comparative evaluations

From the literature, we found three methods for collaborator recommendations, among which the LDA model makes recommendations based on the features of scholars' similar research interests, the RIVT is a novel model for analyzing the distribution of authors' research interests, and the RWR model makes recommendations based on the features of network topology. As described above, to compare the quality of recommendations produced by these one-feature approaches with those of our framework, we use precision, recall, and F1 scores as the three assessment metrics. These are outlined next.

To assess the effectiveness of the framework as a collaborator recommendation tool, we randomly selected 20 target scholars in the test set and explored scholars who may collaborate with them in the future, i.e., potential authors they have never cooperated with. The results are given in Figs. 5, 6, and 7.

From Fig. 5, we can see that the precision of our method and RWR show a similar downwards trend as the recommendation list increases, while RIVT and LDA show a small upward trend. Our method had the highest precision, followed by RWR, then RIVT, and finally LDA. Because precision reflects how many of the recommendations to the target scholar are already collaborators, it is clear that our method does provide superior recommendations.

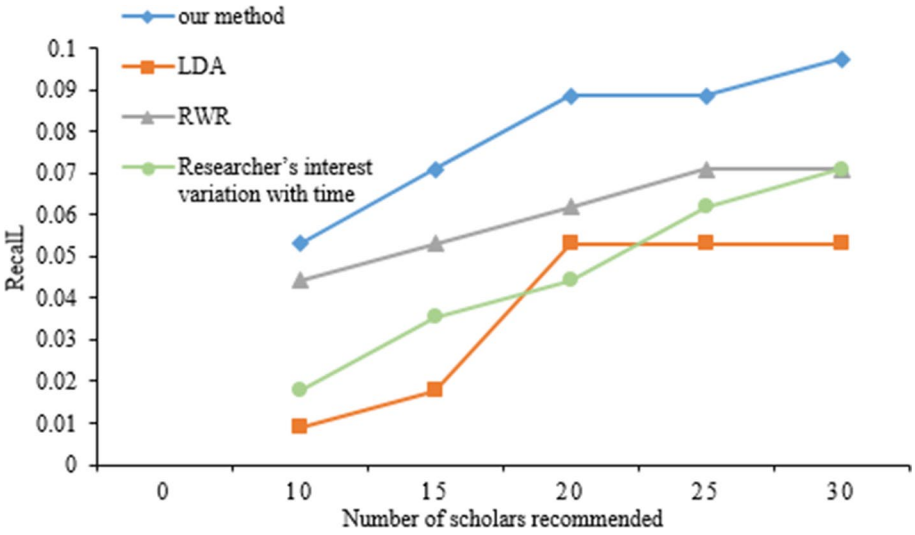

Fig. 6 Recall

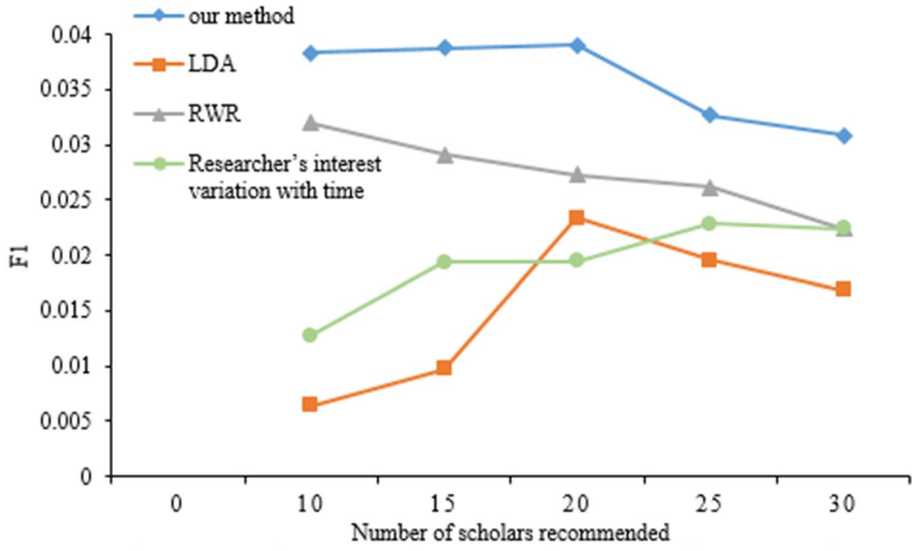

Fig. 7 F1 Score

The recall curves in Fig. 6 for the four models all show an upward trend as the list of recommendations increases. As with precision, our framework had the highest recall rate, followed by RWR, then RIVT, and finally LDA. Recall is a measure of accuracy, essentially reflecting the probability of an efficient recommendation. Hence, these results confirm that our framework generates more accurate recommendations than the other two methods.

Figure 7 shows that the curve of RWR starts to decline with the increase of the number of recommended authors, while the curve of RIVT rises slowly. When the number of recommended authors is greater than 20, the curve of our model and the curve of LDA start to drop. F1 scores integrate Precision and Recall into one metric as a reflection of efficiency and accuracy. These results again emphasize the advantages of our approach.

Overall, we find that our framework makes higher quality recommendations than the three current benchmark solutions. More specifically, we drew the following insights from this analysis:

(1) Fusing similarity indicators based on research interests and topological structures significantly increased the quality of the recommendations.
(2) The Word2Vec method solved the problems of a lack of context and the scalability issues associated with traditional text mining technologies.
(3) The Node2Vec method removed the need to manually design and define indicators, saving on manpower and producing recommendations based on global network features.

Conclusion

The ability of a single researcher to predict the new and unknown grows increasingly difficult as the complexity of research activities and the junction of disciplines becomes more apparent. Thus, in today's academic world, collaborations have become part and parcel of solving technical problems, increasing research efficiency, and improving output quality. Therefore, recommender systems for academic collaborations have become an important topic of interest. To date, two main strategies have been developed: one based on similar research interests, the other on network topological similarity. However, with the advent of the academic big data era, recommendation strategies become increasingly difficult (George et al., 2014). The growing number and extent of authors, publications, journals and other scholarly entities, compounded by the changing scholar relationships, challenge the ability to consider enough information to produce high-quality recommendations.

To produce higher quality recommendations, we developed a unified framework for recommending academic collaborators that combines similarity based on research interests and similarity based on topology, leveraging both types of indicators through word embedding and network embedding. The Word2Vec model (Le & Mikolov, 2014; Mikolov et al., 2013a, 2013b) calculates the similarity of research interests, which solves the problem of semantic term extraction. While the Node2Vec model calculates topological similarity, which solves the problem of automatically extracting global network topology features. Notably, both improve the accuracy of similarity calculations. The CombMNZ method then fuses the two similarity values together and produces a ranked list of top-K collaborators as recommendations.

Notably, isolated researchers and researchers with few co-authors get an equal chance of inclusion in the final recommendation. That said, for the purposes of our analysis, we considered that collaboration with high-level authors can effectively reduce research costs and improve research efficiency. Hence, we took the number of published papers as a threshold and focused on authors whose publications exceeded that threshold. The results demonstrate the effectiveness of the framework. We also compared the recommendations produced by our framework against LDA, RWR and Researcher's interest variation with

time. The results show that our method performs better in terms of precision, recall, and F1 score, confirming an overall improvement in recommendation quality from our strategy.

Overall, the main innovation of our paper is to develop a novel framework for recommending academic collaborators with similar research interests and network topology features. By integrating both word embedding and network embedding, the framework produces more accurate recommendations than existing methods. As demonstrated, this system can help researchers and private-enterprise R&D managers to provide valuable references for finding potential partners. At the same time, this framework can also serve as a basis for further improvement or to inspire related research.

The limitations of our current research offer opportunities for future inquiry. These are summarized as follows. (1) Word embedding and network embedding techniques both contain some parameters; however, methods of training these parameters for optimal benefit is a task that falls into the field of machine learning. (2) We have based our recommendations on only two criteria: the similarity of research interests and the co-authorship features. However, other factors can also indicate the likelihood of a good collaboration, such as citations, or institutional ties. In future, we will consider adding more of these factors into our framework. (3) Although we verified the recommendation results on one dataset from one field. It would be beneficial to test the framework with more datasets and other fields of inquiry. The same goes for the comparisons. A broader assessment of current indicators and strategies could either refute or lend further support to our results.

Acknowledgements This paper was supported by the National Natural Science Foundation of China (NSFC) (Grant Nos. 72274219, 71874013 and 71810107004) and Program for Qian Duansheng Excellent Researcher in China University of Political Science and Law. The previous version of this work is published on Artificial Intelligence + Informetrics (AII) 2021 Workshop (Xi et al., 2021). The authors are very grateful for the valuable comments and suggestions from reviewers, which significantly improved the quality of the paper.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [XX], [JW] and [WD]. Formulation or evolution of overarching research goals and oversight and leadership responsibility for the research activity planning and execution was performed by [YG]. The first draft of the manuscript was written by [XX] and all authors commented on previous versions of the manuscript.

Declarations

Conflict of interest The authors declare that they have not conflict of interest, and actual or potential financial interests also.

References

Abramo, G., D'Angelo, C. A., & Costa, F. D. (2009). Research collaboration and productivity: Is there correlation? *Higher Education, 57*(2), 155–171.

Abramo, G., D'Angelo, C. A., & Costa, F. (2012). Identifying interdisciplinary through the disciplinary classification of coauthors of scientific publications. *Journal of the American Society for Information Science and Technology, 63*(11), 2206–2222.

Ahsan, N., Williams, S. B., Jakuba, M., Pizarro, O., & Radford, B. (2010). Predictive habitat models from AUV-based multibeam and optical imagery. In *OCEANS 2010 MTS/IEEE SEATTLE* (pp.1–10).

Balabanovic, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Communication of the ACM, 40*(3), 66–72.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of International AAAI Conference on Web and Social Media* (pp. 361–362).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2008). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.
Cai, D., He, X., & Han, J. (2008). Training linear discriminant analysis in linear time. In *2008 IEEE 24th International Conference on Data Engineering* (pp. 209–217).
Chen, J. Y., Wu, Y. Y., Fan, L., Lin, X., Zheng, H. B., Yu, S. Q., & Xuan, Q. (2017). Improved spectral clustering collaborative filtering with node2vec technology. In *2017 IEEE 14th International Workshop on Complex Systems and Networks (IWCSN)* (pp. 330–334).
Cui, P., Wang, X., Pei, J., & Zhu, W. W. (2019). A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering., 31*(5), 833–852.
Deepika, S. S., & Geetha, T. V. (2018). A meta-learning framework using representation learning to predict drug-drug interaction. *Journal of Biomedical Informatics, 84*, 136–147.
Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio Speech and Language Processing., 19*(4), 788–798.
Saari, D. G. (1999). Explaining all three-alternative voting outcomes. *Journal of Economic Theory, 87*(2), 313–355.
Dong, Y., Tang, J., Wu, S., Tian, J. L., Chawla, N. V., Rao, J.H., & Cao, H. H. (2013). Link prediction and recommendation across heterogeneous social networks. In *2012 IEEE 12th International Conference on Data Mining* (pp. 181–190).
Edward, A. F., & Joseph, A. S. (1994). Combination of multiple searches. *NIST SPECIAL PUBLICATION*, 243–243.
Eunice, T., Iris, S., Humphrey, L., & Yiu-Kai, N. (2016). Making personalized movie recommendations for children. In *Proceedings of 18th International Conference on Information Integration & Web-based Applications & Services* (pp. 96–105).
Faleiros, T. D. P., & Lopes, A. D. A. (2015). Bipartite graph for topic extraction. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (pp. 4361–4362).
Gang, L., Li, L., Jin, M., & Ye, G. (2015). Empirical research on similarity of research interests in co-authorship network. *Library and Information Service, 59*(2), 75.
George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal, 57*(2), 321–326.
Glänzel, W., & Czerwon, H. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics, 37*(2), 195–221.
Gollapalli, S., Mitra, P., & Giles, C. (2012). Similar researcher search in academic environments. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 167–170). https://doi.org/10.1145/2232817.2232849.
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).
Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics, 101*, 1461–1473.
Hildrun, K. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics, 60*(3), 409–420.
Hollinger, G. A., Choudhary, S., Qarabaqi, P., Murphy, C., Mitra, U., Sukhatme, G. S., Stojanovic, M., Singh, H., & Hover, F. (2011). Communication protocols for underwater data collection using a robotic sensor network. In *Proceedings of the IEEE GLOBECOM Workshops* (pp.1308–1313).
Hu, F., Liu, J., Li, L. H., & Liang, J. (2019). Community detection in complex networks using Node2vec with spectral clustering. *Physica A: Statistical Mechanics and Its Applications, 545*(1), 123633.
Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1
Kawamae, N. (2010). Latent interest-topic model: finding the causal relationships behind dyadic data. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 649–658).
Kazemi, B., & Abhari, A. (2020). Content-based Node2Vec for representation of papers in the scientific literature. *Data & Knowledge Engineering, 127*(5), 101794.
Kong, X., Jiang, H., Wang, W., Bekele, T. M., Xu, Z. Z., & Wang, M. (2017). Exploring dynamic research interest and academic influence for scientific collaborator recommendation. *Scientometrics, 113*(1), 369–385.
Krishnamurthy, B., Puri, N., & Goel, R. (2016). Learning Vector-space representations of items for recommendations using word embedding models. *Procedia Computer Science, 80*, 2205–2210.
Kwon, S., Liu, X., Porter, A. L., & Youtie, J. (2019). Research addressing emerging technological ideas has greater scientific impact. *Research Policy, 48*(9), 103834.

Lab, D. N., & Tollison, R. D. (2000). Intellectual collaboration. *Journal of Political Economy, 108*(3), 632–661.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (pp. 1188–1196).

Lee, S., & Bozeman, B. (2003). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673–702.

Li, C., Guo, J., Lu, Y., Wu, J., & Liu, P. (2018a). LDA meets Word2Vec: A novel model for academic abstract clustering. *Companion of the The Web Conference* (pp. 1699–1706).

Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. In *IEEE International Conference on Cognitive Informatics & Cognitive Computing* (pp. 136–140).

Lopes, G. R., Moro, M. M., Wives, L. K., & de Oliveira, J. P. M. (2010). Collaboration recommendation on academic social networks. In *Proceedings of the 29th International Conference on Conceptual Modeling* (pp. 190–+).

Li, C. Z., Lu, Y., Wu, J. F., Zhang, Y. R., Xia, Z. Z., Wang, T. C., Yu, D. T., Chen, X. R., Liu, P. D., & Guo, J. Y. (2018b). LDA Meets Word2Vec: A novel model for academic abstract clustering. In *Proceedings of the 27th World Wide Web (WWW) Conference* (pp. 1699–1706).

Li, L., Wang, W., Yu, S., Wan, L., Xu, Z., & Kong, X. (2017). A modified Node2vec method for disappearing link prediction. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress* (pp. 1232–1235).

Liu, Y. Z., Tian, Z. Q., Sun, J. S., Jiang, Y. C., & Zhang, X. (2020). Distributed representation learning via node2vec for implicit feedback recommendation. *Neural Computing & Applications, 32*(9), 4335–4345.

Lv, L., & Zhou, T. (2010). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications, 390*, 1150–1170.

Macdonald, C., & Ounis, I. (2008). Voting techniques for expert search. *Knowledge & Information Systems, 16*(3), 259–280.

Man, T., Shen, H., Liu, S., Jin, X., & Cheng, X. (2016). Predict anchor links across social networks via an embedding approach. In *International Joint Conference on Artificial Intelligence* (pp. 1823–1829).

Matveev, A. S., Wang, C., & Savkin, A. V. (2012). Real-time navigation of mobile robots in problems of border patrolling and avoiding collisions with moving and deforming obstacles. *Robotics & Autonomous Systems, 60*(6), 769–788.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy, 29*(1), 31–40.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *Computer Science, 2*(12), 27–35.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 500–509).

Pham, M. C., Cao, Y., Klamma, R., & Jarke, M. (2011). A clustering approach for collaborative filtering recommendation using social network analysis. *Journal of Universal Computer Science, 17*(4), 583–604.

Ping, N., & De-Gen, H. (2016). TF-IDF and rules based automatic extraction of Chinese keywords. *Journal of Chinese Computer Systems, 37*(4), 711–715.

Pradhan, T., Sahoo, S., Singh, U., & Pal, S. (2020). A proactive decision support system for reviewer recommendation in academia. *Expert Systems with Applications, 169*, 114331.

Pradhan, T., & Pal, S. (2020). A multi-level fusion based decision support system for academic collaborator recommendation. *Knowledge-Based Systems, 197*, 1–23.

Price, D. (1963). *Little science, big science*. Columbia University Press.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS), 28*(1), 1–38.

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the American Society for Information Science and Technology, 63*(1), 78–85.
Smith, R. N., Cazzaro, D., Invernizzi, L., Marani, G., Choi, S. K., & Chyba, M. (2011). A geometric approach to trajectory design for an autonomous underwater vehicle: Surveying the bulbous bow of a ship. *Acta Applicandae Mathematicae, 115*(2), 209–232.
Sooho, L., & Barry, B. (2005). Scientific collaboration‖the impact of research collaboration on scientific productivity. *Social Studies of Science*, 35(5), 673–702.
Tang, L. (2013). Does "birds of a feather flock together"matter-Evidence from a longitudinal study on US–China scientific collaboration. *Journal of Informetrics, 7*(2), 330–344.
Taşcı, Ş, & Güngör, T. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications, 40*(12), 4871–4886.
Wang, X. F., Zhang, S., & Liu, Y. Q. (2021). ITGInsight-discovering and visualizing research fronts in the scientific literature. *Scientometrics*. https://doi.org/10.1007/s11192-021-04190-9
Wang, Z., Long, M., & Zhang, Y. (2016). A hybrid document feature extraction method using latent Dirichlet allocation and Word2Vec. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)* (pp. 98–103). IEEE.
Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank:finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (pp. 261–270).
Widyotriatmo, A., & Hong, K. S. (2011). Navigation function-based control of multiple wheeled vehicles. *IEEE Transactions on Industrial Electronics, 58*(5), 1896–1906.
Williams, S. B., Pizarro, O., Webster, J. M., Beaman, R. J., Mahon, I., Johnson-Roberson, M., & Bridge, T. C. L. (2010). Autonomous underwater vehicle–assisted surveying of drowned reefs on the shelf edge of the great barrier reef, australia. *Journal of Field Robotics, 27*(5), 675–697.
Xi, X. W., Guo, Y., & Duan, W. Y. (2021). Recommendation of academic collaborators: a methodology incorporating word embedding and network embedding. In *Proceedings of the 1st Workshop on AI + Informetrics (AII2021) co-located with the iConference 2021* (pp. 47–57).
Xia, F., Chen, Z., Wang, W., Li, J., & Yang, L. T. (2014). Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. *IEEE Transactions on Emerging Topics in Computing, 2*(3), 364–375.
Xu, S., Shi, Q., Qiao, X., Zhu, L., Jung, H., Lee, S., & Choi, S. P. (2014). Author-Topic over Time (AToT): a dynamic users' interest model. In *Mobile, ubiquitous, and intelligent computing* (pp. 239–245). Springer.
Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics, 8*(2), 295–309.
Zhang, Q., Xu, X., Zhu, Y., & Zhou, T. (2015). Measuring multiple evolution mechanisms of complex networks. *Scientific Reports, 5*, 10350.
Zhang, J. (2017). Research collaboration prediction and recommendation based on network embedding in co-authorship networks. *Proceedings of the Association for Information Science & Technology, 54*(1), 847–849.
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*(4), 1099–1117.
Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change., 85*, 26–39.
Ziman, J. M. (1994). *Prometheus bound*. Cambridge University Press.
Zuckerman, H. A. (1968). Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and Its Ambiguity. *American Journal of Sociology, 74*(3), 276–291.