Information Retrieval Journal (2021) 24:175–204
https://doi.org/10.1007/s10791-021-09390-8

Improved reviewer assignment based on both word and semantic features

Shicheng Tan[1,2,3] · Zhen Duan[1,2,3] · Shu Zhao[1,2,3] · Jie Chen[1,2,3] · Yanping Zhang[1,2,3]

Received: 26 May 2020 / Accepted: 26 February 2021 / Published online: 2 April 2021

Abstract

Assigning appropriate reviewers to a manuscript from a pool of candidate reviewers is a common challenge in the academic community. Current word- and semantic-based approaches treat the reviewer assignment problem (RAP) as an information retrieval problem but do not take into account two constraints of the RAP: incompleteness of the reviewer data and interference from nonmanuscript-related papers. In this paper, a word and semantic-based iterative model (WSIM) is proposed to account for the constraints of the RAP by improving the similarity calculations between reviewers and manuscripts. First, we use the improved language model and topic model to extract word features and semantic features to represent reviewers and manuscripts. Second, we use a similarity metric based on the normalized discounted cumulative gain (NDCG) to measure semantic similarity. This metric ignores the probability value (quantitative exact value) of the topic and considers only the ranking (qualitative relevance), thus reducing overfitting to incomplete reviewer data. Finally, we use an iterative model to reduce the interference from nonmanuscript-related papers in the reviewer data. This approach considers the similarity between the manuscript and each of the reviewer's papers. We evaluate the proposed WSIM on two real datasets and compare its performance to that of seven existing methods. The experimental results show that the WSIM improves the recommendation accuracy by at least 2.5% on the top 20.

Keywords Reviewer assignment · Semantic-based model · Word-based model

Mathematics Subject Classification 68T99

✉ Shu Zhao
zhaoshuzs2002@hotmail.com

✉ Yanping Zhang
zhangyp2@gmail.com

[1] Key Laboratory of Intelligent Computing and Signal Processing, Anhui University, Ministry of Education, Hefei, Anhui Province 230601, China

[2] School of Computer Science and Technology, Anhui University, Hefei 230601, China

[3] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei, Anhui Province 230601, China

1 Introduction

Paper-reviewer recommendation refers to the automated process for selecting candidates to perform peer review. This enables journal and conference committees to match papers quickly and accurately to reviewers (McGlinchey et al. 2019). Manually conducting this pairing is labor intensive. Furthermore, it is difficult for a nonprofessional chair to assign suitable matches. Many reviewer assignment systems exist to automate this process (e.g., the Toronto paper matching system (Charlin and Zemel 2013), SubSift (Flach et al. 2010), the Microsoft conference management toolkit[1], the global review assignment processing engine (GRAPE) (Di Mauro et al. 2005), Erie (Li and Hou 2016), advanced reviewer assignment system (Kou et al. 2015b), and decision support system (Hoang et al. 2019)). These systems are completely automated and have been used for many real conferences (e.g., NIPS, ICML, CVPR, ICCV, and ACML).

The problem of paper-reviewer recommendation is known as the reviewer assignment problem (RAP) (Tayal et al. 2014). Dumais and Nielsen (1992) was the first paper to address this problem. The author treated the RAP as an information retrieval issue and used the latent semantic indexing (LSI) model, to establish the relationship between the reviewer and the paper. With the development of the topic model, Mimno and McCallum (2007) used the more advanced latent Dirichlet allocation (LDA) model and author-topic (AT) model and proposed an author-persona-topic (APT) model to better represent the topics covered by a reviewer. These methods are based on semantic information. To further mine the features of the reviewers and papers, some people used word-based information. Peng et al. (2017) used the term frequency-inverse document frequency (TF-IDF) to mine the statistical characteristics of reviewers and papers. They combined this approach with the topic model to propose the time-aware and topic-based (TATB) model. However, these methods neglect the constraints of the RAP: incompleteness of the reviewer data and interference from nonmanuscript-related papers in the reviewer data. We present these two challenges and their corresponding solutions below.

1.1 Incompleteness of the reviewer data

It is not practical to obtain accurate and latest full-text papers of all reviewers because data collection and processing are difficult, and there are even multilingual data. We usually only take the titles and abstracts of the reviewers' papers as reviewer data. Incomplete reviewer data have difficulty accurately and quantitatively reflecting the field (topic) of the reviewer's expertise. To resolve this problem, we use a ranking-based approach to turn the topic distribution into an ordered sequence so that the quantitative probability value of the topic can be ignored, thereby reducing the influence of inaccurate probability values of topics and reducing overfitting to incomplete reviewer data. We first obtain the reviewer and target manuscript topics using the topic model. For the ranking-based approach, we use the normalized discounted cumulative gain (NDCG) as a similarity metric to compute the semantic similarity between the reviewer and the manuscript.

[1] https://cmt3.research.microsoft.com/Content/CMT.html

1.2 The interference from nonmanuscript-related papers

When assigning reviewers, we focus on the authors (reviewers) of papers that are highly similar to the manuscript but do not focus on whether this author has published many papers that are not related to the manuscript. In contrast, when calculating full-text similarity, we focus on documents with paragraphs that are highly similar to the query term but also on documents with many other nonsimilar paragraphs. Therefore, when calculating the similarity between the reviewers' papers as a whole and the manuscript, a large number of irrelevant papers may excessively reduce the similarity between the reviewers and the manuscripts. To resolve this problem, we calculate the similarity between each reviewer's paper and the manuscript, thus highlighting the importance of papers that are highly similar to the manuscript. We also measure the impact of low-similarity papers by calculating the similarity of the manuscript and all papers of the reviewer. This is because it is difficult to directly weigh the impact of low-similarity papers, e.g., how many low-similarity papers can be equivalent to one high-similarity paper? Finally, we combine these two factors in an iterative way.

Our contributions in this paper are summarized as follows.

(1) We propose a word and semantic-based iterative model (WSIM) that considers for the first time the constraints of the reviewer assignment problem by improving the metrics between reviewers and manuscripts.
(2) We use the NDCG as the similarity metric to compute the semantic similarity of the topic. This approach ignores the probability value (quantitative exact value) of the topic and considers only the ranking (qualitative relevance), thus reducing overfitting to incomplete reviewer data.
(3) We use an iterative model to reduce the interference in the assignment from nonmanuscript-related papers in the reviewer data. This approach considers the similarity between the manuscript and each of the reviewer's papers, thus reducing the importance of nonmanuscript-related papers in the reviewer data.
(4) We perform experiments through two datasets, six metrics, and seven comparison algorithms to show that our model is effective in overcoming the challenges.

This paper is organized as follows. Section 2 describes the related research. Section 3 provides the problem formulation, explains our proposed model, describes the model learning algorithm, and introduces the applications of the model. Section 4 describes the experimental setup, the comparison methods, and the performance results. Finally, Sect. 5 concludes this paper.

2 Related work

The authors of Dumais and Nielsen (1992) were the first to discuss automated reviewer recommendations, acknowledging the importance of this task for journal editors as well as the drawbacks of manual assignment. This problem has many names including the conference paper assignment problem (CPAP) (Goldsmith and Sloan 2007), RAP (Wang et al. 2008), paper-reviewer assignment (PRA) (Long et al. 2013), and reviewer assignment (RA) (Wang et al. 2013). Dumais and Nielsen (1992) divided the problem into two processes: selecting the most suitable reviewers for a manuscript and determining the most suitable reviewers for many manuscripts in the case of restrictions. The former is termed retrieval-based RAP (RRAP) (Kou et al. 2015a), while the latter can be termed constrained multiaspect committee review assignment (CMACRA) (Karimzadehgan and Zhai 2009), assignment-based RAP (ARAP) (Kou et al. 2015a) or the multiagent resource allocation problem (MARA) (Lian 2018).

We use the terms RRAP and ARAP as in (Kou et al. 2015a). The ARAP focuses more on optimization issues (Yeşilçimen and Yıldırım 2019) (e.g., how many relevant manuscripts need to be assigned to each reviewer to achieve global optimization?). This paper focuses on RRAP, which can be divided into three categories of related methods: based on semantic information, based on word information, and based on other information. The semantic information corresponds to some relationship between words, typically described by the topic. Word information is used to define the relationship between the reviewer and the manuscript through statistical word frequency and other information. In addition to semantic information and word information, nontextual information can be used to calculate similarity, including classification information (Zhang et al. 2020a; Liu et al. 2016) pertaining to the paper and information provided by the reviewers (Rigaux 2004; Di Mauro et al. 2005). A rule-based (Di Mauro et al. 2005), collaborative filtering (Rigaux 2004) or network-based (Fair and accurate reviewer assignment in peer review 2019; Xu et al. 2019; Anaya et al. 2019) method is often used for this type of information. We focus on methods based on semantics, words, and a combination of these types of information.

2.1 Semantic-based approach

Dumais and Nielsen (1992) transformed RRAP into a retrieval problem using latent semantic indexing (LSI) to extract semantic information and used cosine similarity to calculate the similarity between the reviewer and the manuscript. LSI is a common method for extracting topic information, and Ferilli et al. (2006) and Li and Hou (2016) also used this method. pLSA (Karimzadehgan and Zhai 2012, 2009) and LDA (including variants) (Charlin and Zemel 2013; Misale and Vanwari 2017; Kim and Lee 2018) are improved methods for extracting topic information. Karimzadehgan et al. (2008) first used the pLSA model to obtain the topic and calculate the similarity between the reviewer and manuscript. Mimno and McCallum (2007) first used the LDA model to extract semantic information and proposed the APT model to improve LDA with respect to describing textual information from reviewers and manuscripts. Li and Watanabe (2013) combined the APT model with a time factor to measure the degree of expertise of reviewers. Based on this, Peng et al. (2017) employed TF-IDF to consider word information. Kou et al. (2015a) used the topic weighted coverage calculation based on the topic feature of LDA and proposed the branch-and-bound algorithm (BBA) to find reviewers in the fastest time. In addition to the topic model, Ogunleye et al. (2017) used word2vec to calculate similarity. Zhao et al. (2018) transformed the RRAP into a classification problem and used the word mover's distance (WMD) method to calculate similarity and then used the constructive covering algorithm (CCA) to simultaneously classify reviewers and manuscripts. In (Zhang et al. 2020b), RRAP was also cast as a multilabel classification task in which the reviewers were assigned according to multiple predicted labels.

2.2 Word-based approach

The most commonly used word-based methods are keyword matching (Sidiropoulos and Tsakonas 2015; Protasiewicz et al. 2016; Shon et al. 2017; Dung et al. 2017), TF-IDF (Hettich and Pazzani 2006; Flach et al. 2010; Peng et al. 2017), and the language model (LM) (Mimno and McCallum 2007; Tang et al. 2010; Charlin et al. 2012). Tang and Zhang (2008) calculated the similarity between reviewers and manuscripts by constructing a keyword network and using cosine similarity for keyword matching. Protasiewicz (2014) added publication time information to calculate keyword weights. Dung et al. (2017) improved the keyword matching results by improving the Knuth-Morris-Pratt (KMP) algorithm. Yarowsky and Florian (1999) first used TF-IDF and cosine similarity to calculate the similarity between the reviewer and manuscript. Basu et al. (2001) used a TF-IDF-based information integration system (WHIRL) combined with collaborative filtering. They obtained the recommendation source matrix using the scores retrieved by WHIRL. Biswas and Humayun (2007) mapped keywords to topics based on TF-IDF, which combines ontology-driven inferences. Protasiewicz et al. (2016) directly retrieved relevant reviewers using a full-text index based on TF-IDF. Charlin and Zemel (2013) used LM as the similarity calculation method for the Toronto paper matching system.

2.3 Approach combining semantic and word information

Few existing methods simultaneously consider the semantic and word information of reviewers and manuscripts to capture the semantic and word similarity between a reviewer and a manuscript. Tang et al. (2010, 2012) were the first to combine the language model and LDA to calculate the similarity between reviewers and manuscripts. Peng et al. (2017) used term frequency-inverse document frequency (TF-IDF) to mine the word information of reviewers and papers. They combined this approach with the topic model to propose the time-aware and topic-based (TATB) model.

These semantic-based or word-based approaches treat the reviewer assignment problem as an information retrieval problem but do not take into account the constraints of the reviewer assignment problem. Hence, we propose a WSIM based on LDA and LM to account for the constraints of the reviewer assignment problem by improving the similarity calculations between reviewers and manuscripts.

3 Proposed model

In this section, we first formulate the reviewer assignment problem and notation used in this paper. Then, we describe the word and semantic information extraction. Finally, we detail the ranking-based approach and iterative model for considering the constraints of the reviewer assignment problem.

3.1 Problem definition and notation

First, we define our terms in a formal way. We define a set of reviewer papers $\mathbf{D} = \{d_1, d_2, ..., d_{|\mathbf{D}|}\}$ and a set of manuscripts $\mathbf{P} = \{p_1, p_2, ..., p_{|\mathbf{P}|}\}$, where d_i and p_i denote

Table 1 Notations

Symbol	Description
r	A reviewer
d	A reviewer's paper
p	A manuscript
$\mathbf{R}$	A set of reviewers, $\{r_1, r_2, ..., r_{\vert\mathbf{R}\vert}\}$
$\mathbf{D}$	A set of reviewer's papers, $\{d_1, d_2, ..., d_{\vert\mathbf{D}\vert}\}$
$\mathbf{P}$	A set of manuscripts, $\{p_1, p_2, ..., p_{\vert\mathbf{P}\vert}\}$
K	Number of topics
V	Number of words
θ_{mat}	The topic distributions to all of reviewers
φ_{mat}	The distribution probability of words to all of topics
ρ_{mat}	The topic distributions to all of reviewer's papers
θ_P	The topic distributions to all of manuscripts
α, β	Dirichlet priors to the distributions θ_{mat} and φ_{mat}
η	The weighting factor
ξ_d	The iterative weight of the reviewer's paper
ξ_r	The iterative weight of the reviewer

the text information (e.g., title, abstract, etc.) of the reviewer's paper and manuscript, respectively. We define a set of reviewers $\mathbf{R} = \{r_1, r_2, ..., r_{|\mathbf{R}|}\}$, where r_i denotes the text information (composed of $d_j \in \mathbf{D}$) of the reviewer.

Then, we define our problem in a formal way. Given three sets $\mathbf{D}, \mathbf{P}, \mathbf{R}$ and *topN* (the number of reviewers required for each manuscript), our goal is to obtain the most suitable *topN* reviewers (a subset of $\mathbf{R}$) for each manuscript $p_i \in \mathbf{P}$.

The definition of the retrieval-based RAP (RRAP) is given above. We solve this problem based on two characteristics of reviewer data. In the next subsection, we will begin to describe the proposed word and semantic-based iterative model (WSIM) for the reviewer assignment problem. Table 1 lists the notation used in the proposed model.

3.2 Feature extraction

To calculate the similarity between the reviewer and the manuscript, we need to obtain the semantic and word features of the reviewer and the manuscript. The semantic feature captures the word cooccurrence information between the topics, and the word feature captures the word cooccurrence information between the documents. These two different levels of information make the similarity calculation more comprehensive.

3.2.1 Semantic features

We use the topic model (LDA) to demonstrate the use of the semantic information corresponding to the reviewer publications and the manuscript text. LDA assumes that the text contains multiple topics, following the unigram hypothesis, and obtains the topics of a text by using Gibbs sampling. We use LDA on each reviewer's textual information to obtain the reviewer-topic distribution θ_{mat}:

$$\begin{aligned}\theta_{mat} &=\{\theta_1, ..., \theta_{|\mathbf{R}|}\}\\ \theta_m &=\{\theta_{m,1}, ..., \theta_{m,K}\}, \quad 1 \leqslant m \leqslant |\mathbf{R}|\\ \theta_{m,i} &=\frac{n_{m,i}+\alpha}{\sum_{j=1}^{K}(n_{m,j}+\alpha)}, \quad 1 \leqslant i \leqslant K\end{aligned} \tag{1}$$

where K denotes the number of topics, $n_{m,i}$ denotes the occurrence of the ith topic within the topics covered by reviewer r_m, as obtained by Gibbs sampling, and α denotes the hyperparameter of the LDA model. After obtaining the reviewer-topic distribution, we can predict the manuscript-topic distribution $\theta_{\mathbf{P}} = \{\theta_{p_1}, ..., \theta_{p_{|\mathbf{P}|}}\}$, where θ_{p_m} denotes the polynomial topic distribution of manuscript p_m. The topic-word distribution $\varphi_{mat} = \{\varphi_, ..., \varphi_K\}$ is similar to θ_{mat}, while m, K, and α are replaced with k, V, and β.

For consistency, the topics of each reviewer are directly represented by the topics of the reviewer's papers. This method requires a separate representation of the reviewer's papers. According to the reviewer-topic model, the paper-topic distribution ρ_{mat} is expressed as Eq. (2):

$$\begin{aligned}\rho_{mat} &= \{\rho_1, ..., \rho_{|\mathbf{D}|}\}\\ \rho_m &= \{\rho_{m,1}, ..., \rho_{m,K}\}, \quad 1 \leqslant m \leqslant |\mathbf{D}|\\ \rho_{m,i} &= \frac{n_{m,i}+\alpha}{\sum_{j=1}^{K}(n_{m,j}+\alpha)}, \quad 1 \leqslant i \leqslant K\end{aligned} \tag{2}$$

where $n_{m,i}$ denotes the occurrence of the ith topic in the mth reviewer's paper $\rho_{m,i}$.

Thus, we represent the semantic features of the textual information using the topic distribution θ_{mat}, $\theta_{\mathbf{P}}$, ρ_{mat}.

3.2.2 Word features

We use the language model to demonstrate the representation of the word information. In the language model, the relevance between a query word w and a paper d_i can be expressed as the probability of generating $P_{LM}(w|d_i)$ or $P_{LM}(w|r_i)$, as follows:

$$\begin{aligned}P_{LM}(w|d_i) &= \frac{N_{d_i}}{N_{d_i}+\lambda} \cdot \frac{tf(w,d_i)}{N_{d_i}} + (1-\frac{N_{d_i}}{N_{d_i}+\lambda}) \cdot \frac{tf(w,\mathbf{D})}{N_{\mathbf{D}}}\\ P_{LM}(w|r_i) &= \frac{N_{r_i}}{N_{r_i}+\lambda} \cdot \frac{tf(w,r_i)}{N_{r_i}} + (1-\frac{N_{r_i}}{N_{r_i}+\lambda}) \cdot \frac{tf(w,\mathbf{R})}{N_{\mathbf{R}}}\end{aligned} \tag{3}$$

where N_{d_i} denotes the length of paper d_i, λ denotes the average length across all of the papers, $tf(w, d_i)$ denotes the number of times word w appears in paper d_i, $tf(w, \mathbf{D})$ denotes the number of times word w appears in all papers $\mathbf{D}$, and $N_{\mathbf{D}}$ denotes the total length of all of the papers. The parameters in $P_{LM}(w|r_i)$ are analogous.

The query term w is derived from any manuscript p_k. To effectively capture the importance of certain low-frequency words and reduce the weight of insignificant high-frequency words, we extract the word collection $\mathbf{p}_k$ without considering the repeated words in the manuscript p_k. Different manuscripts contain different numbers of words, potentially causing an order of magnitude difference in the results for manuscripts of different lengths in the language model.

To solve this problem, we sorted the words in manuscript p_k to obtain the collection of the first t words $\mathbf{W}_{p_k}$, resulting in manuscripts of equal length. This process is described in Eq. (4):

$$\begin{aligned} \underset{\mathbf{W}_{p_k}}{\arg\max} \quad & P_{LM}(w_t|d_i) \\ where \quad & \mathbf{W}_{p_k} = \{w_1, w_2, ..., w_t\} \subseteq \mathbf{p}_k,\ d_i \in \mathbf{D} \\ s.t. \quad & \forall w_j \in \mathbf{W}_{p_k} \\ & \Rightarrow P_{LM}(w_j|d_i) \geqslant P_{LM}(w_{j+1}|d_i) \end{aligned} \tag{4}$$

Finally, we obtain the word-based similarity $LM(p_k, d_i)$ between manuscript p_k and paper d_i:

$$LM(d_i, p_k) = \prod_{w_j \in \mathbf{W}_{p_k}} P_{LM}(w_j|d_i) \tag{5}$$

Thus, we represent the word features of the textual information using an improved language model.

3.3 Ranking-based approach and iterative model

After obtaining the features of the reviewers and the manuscript, we detail the ranking-based approach and iterative model for considering the constraints of the reviewer assignment problem.

3.3.1 Ranking-based approach

To reduce the influence of inaccurate probability values of topics, we use the NDCG as the similarity metric to turn the topic distribution into an ordered sequence so that the quantitative probability value of the topic can be ignored, thereby reducing the influence of inaccurate probability values of topics. This approach ignores the probability value (quantitative exact value) of the topic and considers only the ranking (qualitative relevance), thus reducing overfitting to incomplete reviewer data.

The NDCG similarity between reviewer r and manuscript p is expressed as $\mathrm{NDCG}_K(r, p)$ using θ_{mat} and $\theta_{\mathbf{P}}$. $\mathrm{NDCG}_K(r, p)$ must be normalized to calculate the topic similarity. A topic's NDCG (tNDCG) similarity is expressed as Eq. (6), and $\mathrm{tNDCG}_K(r, d)$ is analogous.

$$\begin{aligned} \mathrm{tNDCG}_K(r,p) &= \frac{\mathrm{NDCG}_K(r,p) - \frac{\mathrm{bDCG}_K}{\mathrm{iDCG}_K}}{1 - \frac{\mathrm{bDCG}_K}{\mathrm{iDCG}_K}} \\ \mathrm{NDCG}_K(r,p) &= \frac{\mathrm{DCG}_K(r,p)}{\mathrm{iDCG}_K} \end{aligned} \tag{6}$$

where iDCG_K, bDCG_K, and $\mathrm{DCG}_K(r, p)$ are further defined as:

$$
\begin{aligned}
\mathrm{iDCG}_K &= \sum_{i=1}^{K} \frac{y(i)}{\log_2(i+1)} \\
\mathrm{bDCG}_K &= \sum_{i=1}^{K} \frac{K-i+1}{\log_2(i+1)} \\
\mathrm{DCG}_K(r,p) &= \sum_{i=1}^{K} \frac{y\big(rank[x(\theta_r), i, x(\theta_p)]\big)}{\log_2(i+1)} \\
where \quad & x(\theta_r) = \{k_1, \dots, k_K\}, \theta_r \in \theta_{mat}, \theta_p \in \theta_{\mathbf{P}} \\
s.t. \quad & \forall i \in [1, K-1] \Rightarrow \theta_{r,k_i} \geqslant \theta_{r,k_{i+1}},\ \theta_{r,k_i} \in \theta_r
\end{aligned}
\tag{7}
$$

where $x(\theta_r)$ denotes the probability ranking order of the topics (in reverse order). $rank[x(\theta_r), i, x(\theta_p)]$ represents the ranking of topic k_i of $x(\theta_p)$ in $x(\theta_r)$. The function y denotes the rank value function and $y(i) = i^{-\frac{1}{2}}$. The bDCG (bad DCG) denotes the lower bound of $\mathrm{DCG}_K(r,p)$. The role of bDCG is to achieve the normalization of $\mathrm{NDCG}_K(r,p)$.

3.4 Iterative model

To reduce the interference in the assignment from nonmanuscript-related papers in the reviewer data, we calculate the similarity between each reviewer's paper and the manuscript, thus highlighting the importance of papers that are highly similar to the manuscript. Then, we measure the impact of low-similarity papers by calculating the similarity of the manuscript and all papers of the reviewer. This is because it is difficult to directly weigh the impact of low-similarity papers, e.g., how many low-similarity papers can be equivalent to one high-similarity paper? Finally, we combine these two factors in an iterative way.

When we combine these two factors using an iterative model, the similarity of one reviewer to the manuscript is influenced by the similarity of the manuscript and each paper for that reviewer, and the similarity of one reviewer's paper to the manuscript is influenced by the similarity of each author (reviewer) to the manuscript. We can describe this with the following formula, Eq. (8):

$$
\begin{aligned}
&\begin{cases} \gamma^0[r] = \mathrm{tNDCG}_K(r,p) \cdot LM(r,p) \\ \gamma^k[r] = (1-\xi_d)\gamma^{k-1}[r] + \xi_d \cdot \pi^k[f_{rd}(r)] \end{cases} \\
&\begin{cases} \gamma^0[d] = \mathrm{tNDCG}_K(d,p) \cdot LM(d,p) \\ \gamma^k[d] = (1-\xi_r)\gamma^{k-1}[d] + \xi_r \cdot \pi^k[f_{dr}(d)] \end{cases}
\end{aligned}
\tag{8}
$$

where $\gamma^k[r]$ denotes the relevance of reviewer r to the manuscript at the kth iteration and $\gamma^k[d]$ denotes the relevance of the reviewer's paper d to the manuscript at the kth iteration. Further, ξ_d denotes the iterative weight of the reviewer's paper, ξ_r denotes the iterative weight of the reviewer, $f_{rd}(r)$ denotes all of the papers of reviewer r, and $f_{dr}(d)$ denotes all of the reviewers of the reviewer's paper d.

In the above formula, $\pi^k[f_{rd}(r)]$ is essential. It denotes the relevance of the reviewer's r paper to the manuscript. Because nonmanuscript-related papers overshadow manuscript-related papers, we highlight the importance of papers that are highly similar to the manuscript by the function π. By formulating the relevance of reviewer r's collection of papers as $f_{rd}(r)$, different weights can be assigned to reviewers' papers with different influences that can distinguish reviewers' papers of different levels of importance.

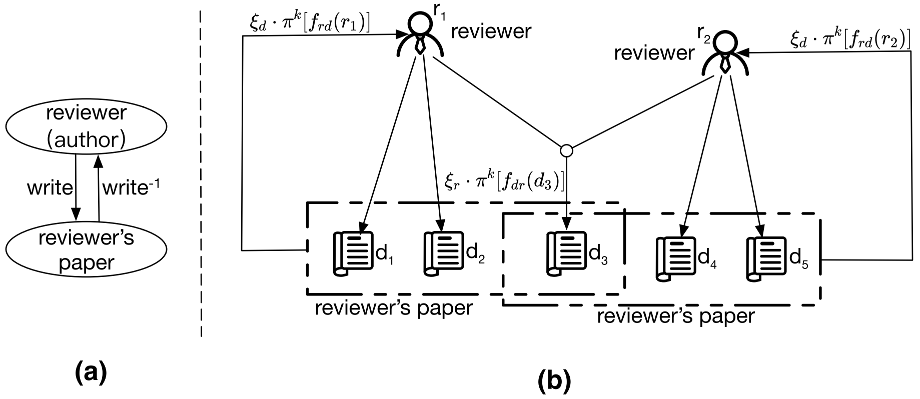

Fig. 1 An example of WSIM reducing the interference in the assignment from nonmanuscript-related papers in the reviewer data. **a** The relationship schema between the reviewer and the reviewer's paper; **b** the iterative process based on the relationship schema

We determine the ranking μ_r^k of the relevance between all of the reviewers' manuscripts and the target manuscripts for reviewer r: (in the kth iteration)

$$\begin{aligned} & \mu_r^k = \{\gamma^{k-1}[d_1], \dots, \gamma^{k-1}[d_h]\} \\ where\quad & d_i \in f_{rd}(r),\ h = |f_{rd}(r)| \\ s.t.\quad & \forall i \in [1, h-1] \Rightarrow \gamma^{k-1}[d_i] \geqslant \gamma^{k-1}[d_{i+1}] \end{aligned} \tag{9}$$

Similarly, the ranking μ_d^k of the relevance between all of the authors of paper d (reviewers) and manuscripts is represented by Eq. (10):

$$\begin{aligned} & \mu_d^k = \{\gamma^{k-1}[r_1], \dots, \gamma^{k-1}[r_l]\} \\ where\quad & r_i \in f_{dr}(d),\ l = |f_{dr}(d)| \\ s.t.\quad & \forall i \in [1, l-1] \Rightarrow \gamma^{k-1}[r_i] \geqslant \gamma^{k-1}[r_{i+1}] \end{aligned} \tag{10}$$

$$\overbrace{\underset{\mu^k_{r,1}}{\overbrace{-\ -\ -\ -\ -\ -\ -\ -}^{\eta}}\ \mathrlap{\underbrace{\phantom{\underset{\mu^k_{r,2}}{\overbrace{-\ -\ -\ -}^{(1-\eta)\eta}}\cdots\underset{\mu^k_{r,i+1}}{\overbrace{-\ -}^{(1-\eta)^i\eta}}\cdots\underset{\mu^k_{r,h}}{\overbrace{-}^{(1-\eta)^{h-1}\eta}}\ \overbrace{-}^{(1-\eta)^h}}}_{1-\eta}}\underset{\mu^k_{r,2}}{\overbrace{-\ -\ -\ -}^{(1-\eta)\eta}}\cdots\underset{\mu^k_{r,i+1}}{\overbrace{-\ -}^{(1-\eta)^i\eta}}\cdots\underset{\mu^k_{r,h}}{\overbrace{-}^{(1-\eta)^{h-1}\eta}}}^{1-(1-\eta)^h}\ \overbrace{-}^{(1-\eta)^h} \tag{11}$$

To ensure the stability of the iteration, we must normalize the accumulation of the relevance of all of the reviewer's papers. The most relevant reviewer paper is assigned a weight of η, and all of the reviewer's remaining papers are assigned a weight of $1 - \eta - (1 - \eta)^h$, which is weighted in a recursive method. Equation (11) shows how the function π assigns weights to the papers that are most relevant to the target manuscript. In the kth iteration, the relevance $\pi^k[f_{rd}(r)]$ of the reviewer's r paper to the manuscript and the relevance $\pi^k[f_{dr}(d)]$ of the reviewer's paper d to the manuscript are expressed as Eq. (12):

$$\pi^k[f_{rd}(r)] = \sum_{i=0}^{h-1} \frac{(1-\eta)^i \eta}{1-(1-\eta)^h} \mu^k_{r,i+1}$$
$$\pi^k[f_{dr}(d)] = \sum_{i=0}^{l-1} \frac{(1-\eta)^i \eta}{1-(1-\eta)^l} \mu^k_{d,i+1} \tag{12}$$

where $h = |f_{rd}(r)|$ denotes the number of papers authored by reviewer r, η denotes the weighting factor, and $l = |f_{dr}(d)|$ denotes the number of authors (reviewers) of the reviewer's paper d. Further, $\mu^k_{r,i+1}$ denotes the $(i+1)$-ranked relevance score in μ^k_r.

Figure 1a depicts the relationship schema between the reviewer and the reviewer's paper, resulting in $f_{rd}(r)$ and $f_{dr}(d)$. Figure 1b depicts an example of the iterative process based on the relationship schema. In this example, the reviewer's papers $\{d_1, d_2, d_3\}$ influence reviewer r_1 through $\xi_d \cdot \pi^k[f_{rd}(r_1)]$, and the reviewers $\{r_1, r_2\}$ influence the reviewer's paper d_3 through $\xi_r \cdot \pi^k[f_{dr}(d_3)]$, all of which together form a coupled random walk.

```
Algorithm 1: word and semantic-based iterative model
Input: Set of reviewers and reviewer's papers: $\mathbf{R}, \mathbf{D}$; Manuscript: $p$; Number of
       reviewers assigned to a manuscript: $topN$; Average number of papers by
       reviewer: $\tau$; Number of iterations: $k'$
Output: Reviewers most relevant to the manuscript $p$
Calculate all of topic distributions of $\mathbf{R}$ and $\mathbf{D}$
for for each reviewer $r \in \mathbf{R}$ do
    $\gamma^0_{\mathbf{R}} \leftarrow \gamma^0[r] = \mathrm{tNDCG}_K(r,p) \cdot LM(r,p)$
end
for for each reviewer's paper $d \in \mathbf{D}$ do
    $\gamma^0_{\mathbf{D}} \leftarrow \gamma^0[d] = \mathrm{tNDCG}_K(d,p) \cdot LM(d,p)$
end
$\gamma^0_{\mathbf{D}} \leftarrow$Reverse sorting $\gamma^0_{\mathbf{D}}$
Number of papers per level: $i = topN \cdot \tau$
for get $i$ scores from $\gamma^0_{\mathbf{D}}$ in order do
    $\gamma^0_{\mathbf{D}} \overset{i\ times}{\longleftarrow}$ Calculate the average of $i$ relevance scores
end
for for the $k$-th iteration, $k \in [1, k']$ do
    $\gamma^k[r] = (1-\xi_d)\gamma^{k-1}[r] + \xi_d \cdot \pi^k[f_{rd}(r)]$
    $\gamma^k[d] = (1-\xi_r)\gamma^{k-1}[d] + \xi_r \cdot \pi^k[f_{dr}(d)]$
end
Return: The first $topN$ reviewers in $\gamma^{k'}_{\mathbf{R}}$
```

Reviewers of the same manuscript do not consider rankings. We use an averaging process for the relevance of the reviewer's papers before iteration so that the papers of the *topN* reviewers do not consider the ranking. First, we calculate the sorting of γ^0_d for each paper. Then, we evaluate all γ^0_d according to the reviewer's average number of papers and the number of reviewers to be assigned to the target manuscript, and we average the relevance of the reviewer's papers. In Algorithm 1 outlined below, we describe the main process of WSIM.

Thus, we use a ranking-based approach and iterative model to consider the constraints of the reviewer assignment problem and to obtain the most suitable *topN* reviewers for each manuscript.

Table 2 First dataset

	Reviewers	Manuscripts
Number	400	100
Range of years	1996–2014	2015
Number of papers included in *r* or *p*	50–150	1
Number of fields included in all *r* or *p*	262	81
Range of number of words included in the paper	100–733	151–313

4 Experiments

In this section, we evaluate the effectiveness of our WSIM method. We construct experiments using closed-world settings (Price and Flach 2017) with a fixed predetermined pool of reviewers to conduct a comparison with seven existing methods.

4.1 Dataset

Typically, journals do not disclose their specific manuscript review process because of fairness and privacy, so it is difficult to obtain a real manuscript review process. This problem makes it difficult to use current real datasets. For example, in paper (Karimzadehgan et al. 2008), their dataset[2] is too small and lacks time. Another example is in paper (Tang et al. 2012), whose dataset[3] lacks reviewer paper information and assignment results. There is also a paper (Kou et al. 2015a), whose datasets[4] lack the allocation results that can be used for evaluation. The paper (Mimno and McCallum 2007) provides a manually assigned dataset for NIPS2006, but it is not publicly available. Therefore, we used two data sources to construct a real dataset. All datasets are released on GitHub[5].

4.1.1 First dataset

Table 2 describes the dataset in detail. This dataset consists of reviewer profiles, which comprise their publications (including titles, abstracts, and years) and labels. The label indicates the peer review relationship between the target manuscript and the reviewer. It uses a binary value to describe whether the reviewer can review the target manuscript.

We apply a rule to obtain the labels for the field (classification) of reviewers and manuscripts: a reviewer who has published at least 10 papers in a field corresponding to the manuscript is eligible to review that manuscript. In this setup, each reviewer has at least one field that corresponds to at least 10 papers published by that reviewer (whether or not consistent with the target manuscript), which forms the qualification for becoming a candidate reviewer.

[2] http://sifaka.cs.uiuc.edu/ir/data/review.html

[3] https://www.aminer.cn/expertisematching

[4] http://degroup.cis.umac.mo/reviewerassignment/

[5] https://github.com/aitsc/WSIM/tree/main/datasets

Table 3 Second dataset

	Reviewers	Manuscripts
Number	1885	685
Range of years	1992–2015	2016
Number of papers included in *r* or *p*	50–123	1
Number of fields included in all *r* or *p*	175	46
Range of number of words included in the paper	160–300	160–291

Table 4 Validation dataset

	Reviewers	Manuscripts
Number	337	80
Range of years	1996–2014	2015
Number of papers included in *r* or *p*	50–150	1
Number of fields included in all *r* or *p*	260	57
Range of number of words included in the paper	100–918	102–412

4.1.2 Second dataset

This dataset comes from the public data source of arXiv[6], which contains a total of 1,180,081 papers. All papers contain titles, abstracts, authors, publication time, subject, and 1,031,734 papers without MSC classification. We use the subject as the field. As with the processing of the first dataset, we constrain the information of reviewers and manuscripts during preprocessing. The difference is that reviewers who have published at least 20 papers in a field corresponding to the manuscript are eligible to review that manuscript. Table 3 describes the details of the dataset, and finally, we obtain 1885 reviewers and 685 manuscripts from the second dataset, which simulates a medium-sized conference.

4.1.3 Validation dataset

To find a common set of hyperparameters, we constructed a validation dataset using the same methods and data sources as the first dataset. Table 4 describes the dataset in detail.

4.2 Comparison methods

We compare our WSIM with the following seven methods, which include classic algorithms and state-of-the-art algorithms: LDA (equivalent to the author-topic model) (Mimno and McCallum 2007), LM (Charlin and Zemel 2013), LDA-LM (Tang et al. 2010), TATB (time-aware and topic-based model) (Peng et al. 2017), KCS (keyword cosine similarity) (Protasiewicz et al. 2016), BBA (Kou et al. 2015a), and WMD (Kusner et al. 2015).

[6] ftp://3lib.org//oai_dc/arxiv

LDA This method calculates the cosine similarity of the topic distribution probability between the reviewer and the manuscript to determine the appropriate reviewers for the manuscript.

LM The field of the manuscript is regarded as a query term; the method calculates the probability that the query term is present in the reviewer's information to obtain the appropriate reviewers for the manuscript (see Eq. (3)).

LDA-LM This approach combines the results of LDA and LM to determine the appropriate reviewers for the manuscript based on the total score.

TATB Based on LDA, the papers published by reviewers are assigned different weights over time and multiplied by the results of TF-IDF to determine the appropriate reviewers based on the resulting scores.

KCS This method uses the Kea algorithm to extract the keywords of the reviewers and target manuscripts, assigns weights to the keywords with respect to the publication time of the paper in which the keyword is located and calculates the cosine similarity between the reviewer and the target manuscript.

BBA This approach uses LDA to obtain the topic distribution of the reviewers and the target manuscripts. The topic distribution of all of the reviewers for a target manuscript is considered as a whole (a group of reviewers), and the branch-and-bound method is used to quickly determine the appropriate reviewers.

WMD This approach uses word2vec to calculate the word embedding of the reviewers and the target manuscripts and then uses earth mover's distance to calculate the similarity between the text excerpts.

4.2.1 Hyperparameters

We perform a random search (Bergstra and Bengio 2012) in the hyperparameter space using the validation dataset, and the result is as follows. The hyperparameters in the LDA model of the WSIM and comparison methods include the number of fields (topics) K, the hyperparameter α, the hyperparameter β, and the number of iterations, which are set to 50, 0.5, 0.1, and 3000, respectively. The hyperparameters in the WSIM include t, η, ξ_d, and ξ_r, which are set to 80, 0.25, 0.05, and 0.05, respectively. WMD uses 300-dimensional word embedding. The hyperparameters used for the other comparison methods are consistent with the respective original papers. Our implementations are available on GitHub[7].

4.3 Evaluation metrics

We use the methods to find the *topN* reviewers for each manuscript and compare the result of each method with the labels in the dataset. We use the precision, recall, and F1 score as evaluation metrics. We also employ several popular information retrieval measures (Büttcher and Clarke 2016) including mean averaged precision (MAP), normalized discounted cumulative gain (NDCG), and bpref (Buckley and Voorhees 2004). The metrics are defined below:

[7] https://github.com/aitsc/WSIM

Table 5 Method performance comparison for the first dataset

topN	Methods	P	R	F$_1$	MAP	NDCG	bpref
10	LDA	0.3500	0.1058	0.1625	0.2408	0.3557	0.6466
	LM	0.5280	0.1634	0.2496	0.4107	0.5536	0.7565
	LDA-LM	0.5290	0.1638	0.2501	0.4115	0.5540	0.7570
	TATB	0.3520	0.1071	0.1643	0.2524	0.3647	0.6524
	KCS	0.1530	0.0497	0.0750	0.0832	0.1624	0.5398
	BBA	0.1630	0.0492	0.0756	0.0815	0.1851	0.5512
	WMD	0.4990	0.1526	0.2337	0.3812	0.5302	0.7461
	WSIM	**0.5690**	**0.1725**	**0.2647**	**0.4559**	**0.5932**	**0.7796**
20	LDA	0.3330	0.2005	0.2503	0.2036	0.3417	0.6555
	LM	0.4680	0.2828	0.3526	0.3281	0.5027	0.7422
	LDA-LM	0.4670	0.2823	0.3519	0.3271	0.5019	0.7419
	TATB	0.3365	0.2029	0.2531	0.2103	0.3492	0.6578
	KCS	0.1325	0.0836	0.1025	0.0611	0.1449	0.5522
	BBA	0.1245	0.0715	0.0908	0.0497	0.1502	0.5535
	WMD	0.4570	0.2733	0.3420	0.3101	0.4897	0.7319
	WSIM	**0.5025**	**0.2955**	**0.3722**	**0.3614**	**0.5367**	**0.7601**
30	LDA	0.3110	0.2730	0.2908	0.1772	0.3246	0.6545
	LM	0.4180	0.3719	0.3936	0.2756	0.4601	0.7260
	LDA-LM	0.4180	0.3717	0.3935	0.2759	0.4599	0.7257
	TATB	0.3140	0.2768	0.2942	0.1822	0.3307	0.6560
	KCS	0.1183	0.1092	0.1136	0.0478	0.1322	0.5526
	BBA	0.1147	0.0974	0.1053	0.0393	0.1377	0.5521
	WMD	0.4203	0.3660	0.3913	0.2682	0.4568	0.7206
	WSIM	**0.4500**	**0.3897**	**0.4177**	**0.3084**	**0.4941**	**0.7443**
50	LDA	0.2910	0.4126	0.3413	0.1510	0.3063	0.6488
	LM	0.3602	0.5124	0.4230	0.2189	0.4064	0.7020
	LDA-LM	0.3580	0.5101	0.4207	0.2182	0.4047	0.7019
	TATB	0.2928	0.4184	0.3445	0.1540	0.3106	0.6498
	KCS	0.0978	0.1426	0.1160	0.0337	0.1132	0.5486
	BBA	0.1166	0.1619	0.1356	0.0310	0.1329	0.5544
	WMD	0.3646	0.5145	0.4268	0.2151	0.4062	0.7013
	WSIM	**0.3760**	**0.5312**	**0.4403**	**0.2353**	**0.4272**	**0.7133**

Table 6 Method performance comparison for the second dataset

topN	Methods	P	R	F_1	MAP	NDCG	bpref
10	LDA	0.3893	0.1441	0.2104	0.2686	0.4324	0.6957
	LM	0.6340	0.2354	0.3433	0.5517	0.6937	0.8402
	LDA-LM	0.6349	0.2357	0.3438	0.5529	0.6945	0.8407
	TATB	0.4254	0.1572	0.2296	0.3171	0.4744	0.7232
	KCS	0.0342	0.0126	0.0184	0.0140	0.0393	0.4718
	BBA	0.0648	0.0237	0.0347	0.0521	0.1183	0.5059
	WMD	0.6121	0.2275	0.3317	0.5160	0.6662	0.8252
	WSIM	**0.6742**	**0.2510**	**0.3658**	**0.5925**	**0.7258**	**0.8568**
20	LDA	0.3153	0.2331	0.2681	0.1847	0.3654	0.6695
	LM	0.4734	0.3516	0.4035	0.3718	0.5600	0.7784
	LDA-LM	0.4746	0.3525	0.4045	0.3729	0.5611	0.7789
	TATB	0.3306	0.2445	0.2811	0.2089	0.3902	0.6838
	KCS	0.0293	0.0218	0.0250	0.0082	0.0340	0.4919
	BBA	0.0389	0.0286	0.0329	0.0268	0.0810	0.5072
	WMD	0.4482	0.3327	0.3819	0.3416	0.5320	0.7647
	WSIM	**0.4996**	**0.3716**	**0.4262**	**0.4024**	**0.5854**	**0.7963**
30	LDA	0.2594	0.2869	0.2724	0.1397	0.3151	0.6495
	LM	0.3716	0.4137	0.3915	0.2732	0.4694	0.7344
	LDA-LM	0.3723	0.4145	0.3922	0.2739	0.4701	0.7348
	TATB	0.2698	0.2986	0.2835	0.1567	0.3344	0.6593
	KCS	0.0298	0.0331	0.0314	0.0061	0.0332	0.4992
	BBA	0.0303	0.0334	0.0318	0.0183	0.0655	0.5077
	WMD	0.3529	0.3928	0.3718	0.2512	0.4465	0.7222
	WSIM	**0.3849**	**0.4290**	**0.4057**	**0.2915**	**0.4851**	**0.7467**
50	LDA	0.1918	0.3538	0.2487	0.0929	0.2501	0.6202
	LM	0.2613	0.4837	0.3394	0.1769	0.3614	0.6785
	LDA-LM	0.2619	0.4847	0.3400	0.1773	0.3620	0.6788
	TATB	0.1971	0.3635	0.2556	0.1033	0.2632	0.6261
	KCS	0.0306	0.0568	0.0398	0.0044	0.0328	0.5057
	BBA	0.0159	0.0287	0.0204	0.0107	0.0444	0.5044
	WMD	0.2502	0.4638	0.3251	0.1628	0.3453	0.6695
	WSIM	**0.2666**	**0.4939**	**0.3463**	**0.1787**	**0.3656**	**0.6804**

Table 7 Experimental results of the original methods and their improved versions (first dataset)

Methods	top10	top20	top30	top50
LM	0.5280	0.4680	0.4180	0.3602
I-LM	*0.5600*	*0.4960*	*0.4403*	*0.3754*
LDA	0.3500	0.3330	0.3110	0.2910
LDA-ED	0.3180	0.2980	0.2713	0.2372
LDA-JS	0.3220	0.3175	0.2973	0.2678
LDA-NDCG	*0.3520*	*0.3445*	*0.3267*	*0.2968*
LDA-LM	0.5290	0.4670	0.4180	0.3580
LDA-NDCG+I-LM	0.5610	0.4955	0.4423	**0.3760**
WSIM	**0.5690**	**0.5025**	**0.4500**	**0.3760**

Table 8 Experimental results of the original methods and their improved versions (second dataset)

Methods	top10	top20	top30	top50
LM	*0.6340*	*0.4734*	0.3716	0.2613
I-LM	0.6320	0.4726	*0.3747*	*0.2651*
LDA	0.3893	0.3153	0.2594	0.1918
LDA-ED	0.3234	0.2364	0.1827	0.1241
LDA-JS	0.3028	0.2531	0.2165	0.1658
LDA-NDCG	*0.4292*	*0.3566*	*0.3029*	*0.2333*
LDA-LM	0.6349	0.4746	0.3723	0.2619
LDA-NDCG+I-LM	0.6336	0.4750	0.3762	**0.2666**
WSIM	**0.6742**	**0.4996**	**0.3849**	**0.2666**

$$\text{Precision: P} = \frac{1}{N}\sum_{1}^{N}\frac{TP}{TP+FP}$$

$$\text{Recall: R} = \frac{1}{N}\sum_{1}^{N}\frac{TP}{TP+FN}$$

$$\text{Macro-F1 score: F}_1 = \frac{2\cdot\text{P}\cdot\text{R}}{\text{P}+\text{R}}$$

$$\text{Mean Average Precision: MAP} = \frac{1}{N}\sum_{1}^{N}\frac{1}{R_n}\sum_{i=1}^{R_n}(\text{P@}i\cdot R(c_i))$$

$$\text{NDCG} = \frac{1}{N}\sum_{1}^{N}\frac{\sum_{i=1}^{n}\frac{R(c_i)}{\log_2(i+1)}}{\sum_{i=1}^{n}\frac{1}{\log_2(i+1)}}$$

$$\text{Binary preference: bpref} = \frac{1}{N}\sum_{1}^{N}\frac{1}{R_n}\sum_{r=1}^{R_n}(1-\frac{\sum_{i=1}^{r}(1-R(c_i))}{R_n})$$

where $N = |\mathbf{P}|$, TP denotes the number of true positives, FP denotes the number of false positives, and FN denotes the number of false negatives. In addition, $n = topN$, R_n is the number of reviewers who are eligible to review the target manuscript, and $R(c_i) = 1$ if the i-th retrieved candidate is relevant to the target manuscript and $R(c_i) = 0$ otherwise.

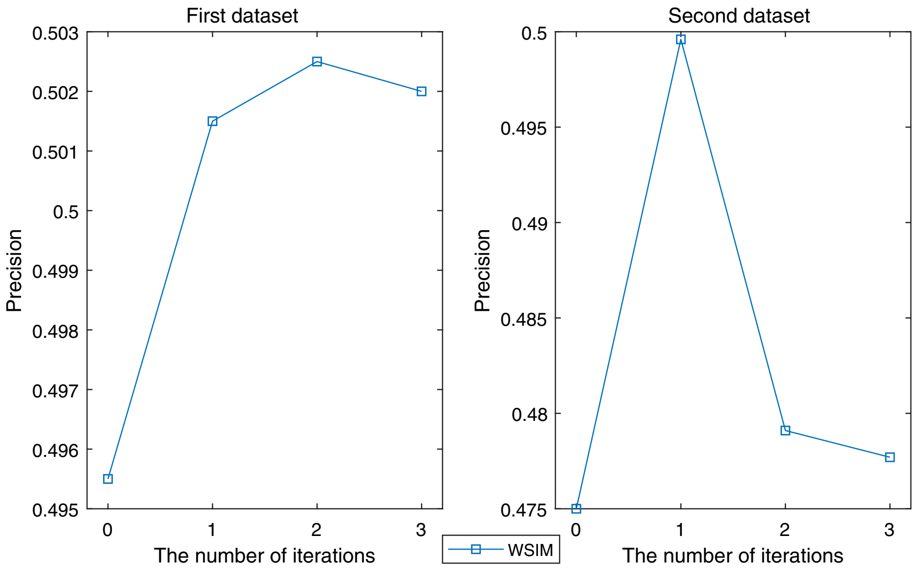

Fig. 2 Precision achieved with a varied number of iterations

4.4 Experimental results

We examined the performance of the WSIM and comparison methods when *topN* was 10, 20, 30, and 50. Tables 5 and 6 show the results of the WSIM and the seven comparison methods with respect to the precision, recall, F1 score, MAP, NDCG, and bpref metrics on the first dataset and second dataset. The results indicate that our proposed WSIM is superior to all of the comparison methods, including the latest RRAP method (TATB). This proves our effectiveness in overcoming challenges. The WSIM is better than the three types of methods it is based on, namely, LM (word-based), LDA (semantic-based), and LDA-LM (word and semantic-based). The WSIM also outperforms other methods: KCS (word-based), BBA (semantic-based), TATB (word and semantic-based), and WMD (word embedding or semantic-based). This is because we consider the constraints of RAP, not just RAP, as an information retrieval problem.

Here, some analyses of comparative methods are presented: (1) The performance of LDA is weaker than LM, which is better suited for short texts (title and abstract). (2) The direct combination of LDA and LM does not result in an improvement in performance (Table 5) because this combination is not complementary and may result in incorrect results being adversely affected. (3) TATB uses the TF-IDF method but does not suitably represent the word information. (4) KCS uses keyword weight calculations but does not represent semantic information. (5) BBA uses topic-based coverage to calculate relevance, without considering word information, and this coverage only finds the appropriate reviewer group for the target manuscript and does not ensure that each reviewer is appropriate for the target manuscript. (6) WMD uses semantic information but does not consider the constraints of RAP.

Table 9 Two-sided paired t-test results between the WSIM and other comparison methods

Methods	First dataset			Second dataset		
	Mean	*t*-value	*p*-value	Mean	*t*-value	*p*-value
WSIM	0.5025	NaN	1.0000	0.4998	NaN	1.0000
LDA+LM	0.4670	−2.7947	0.0209	0.4747	−4.5561	0.0014
LM	0.4680	−2.7889	0.0211	0.4735	−4.9981	0.0007
WMD	0.4570	−2.873	0.0184	0.4561	−11.4248	0.0000
TATB	0.3365	−8.431	0.0000	0.3306	−25.3586	0.0000
LDA	0.3330	−8.5863	0.0000	0.3154	−25.3241	0.0000
KCS	0.1325	−12.457	0.0000	0.0294	−82.3878	0.0000
BBA	0.1275	−14.5907	0.0000	0.0388	−84.5804	0.0000

4.5 Ablation analysis

We conduct an ablation analysis on the WSIM to examine the effectiveness of each component, including the improved LM, ranking-based approach and iterative model. First, for improved LM, we list the existing method LM and its improved method improved LM (I-LM). Second, for the ranking-based approach, we list the existing LDA method and its improved LDA-NDCG method. The original LDA method uses cosine similarity, and we further compare Euclidean distance (LDA-ED) and Jensen-Shannon divergence (LDA-JS). Finally, for the iterative model, we list the existing LDA-LM and improved WSIM methods, including the LDA-NDCG+improved LM (LDA-NDCG+I-LM), which is a zero-iterative WSIM. Tables 7 and 8 show the precision of these methods on the first dataset and second dataset, respectively, with *topN* values of 10, 20, 30, and 50. The underline indicates the best result in the current component. The bold font is the best result in the current column. We have the following observations and analysis:

(1) Iterative models are helpful for performance improvement. The performance of the WSIM exceeds all LDA-IM and 75% of LDA-NDCG+I-LM. This is because the iterative model reduces the interference in the assignment from nonmanuscript-related papers in the reviewer data. (2) The ranking-based approach is helpful for performance improvement. The performance of LDA-NDCG exceeds that of LDA, LDA-ED, and LDA-JS. This is because the ranking-based approach reduces the influence of inaccurate probability values of topics. (3) Improved LM is helpful for performance improvement. I-LM outperforms LM in 75% of the results. This is mainly because I-LM alleviates the problem caused by the inconsistent text length in the LM method.

We explore the influence of the number of iterations on the algorithm performance. Figure 2 shows the precision (*topN*=20) for different numbers of iterations. Performing zero iterations denotes that the interference in the assignment from nonmanuscript-related papers is not considered. On the first dataset, increasing the number of iterations from one to two results in improved performance because the importance of papers that are highly similar to the manuscript is considered with more iterations. On the second dataset, increasing the number of iterations to one results in the best improved performance. As the number of iterations increases, the proportion of $\pi^k[f_{rd}(r)]$ in $\gamma^k[r]$ increases at the same time. $\pi^k[f_{rd}(r)]$ has highlighted the importance of papers that are highly similar to the manuscript, $\gamma^0[r]$ has highlighted the importance of all papers by reviewers, and both are indispensable. Therefore, continuing to increase the number of iterations can overstate

Table 10 Two-sided paired t-test results between all methods

		Methods	WSIM	LDA+LM	LM	WMD	TATB	LDA	BBA	KCS
The results of six metrics	First dataset	WSIM	–	101111	101111	111111	111111	111111	111111	111111
		LDA+LM	000000	–	000000	000000	111111	111111	111111	111111
		LM	000000	000000	–	000000	111111	111111	111111	111111
		WMD	000000	000000	000000	–	111111	111111	111111	111111
		TATB	000000	000000	000000	000000	–	000000	111111	111111
		LDA	000000	000000	000000	000000	000000	–	111111	111111
		BBA	000000	000000	000000	000000	000000	000000	–	000000
		KCS	000000	000000	000000	000000	000000	000000	000000	–
	Second dataset	WSIM	–	111111	111111	111111	111111	111111	111111	111111
		LDA+LM	000000	–	100101	111111	111111	111111	111111	111111
		LM	000000	000000	–	111111	111111	111111	111111	111111
		WMD	000000	000000	000000	–	111111	111111	111111	111111
		TATB	000000	000000	000000	000000	–	111111	111111	111111
		LDA	000000	000000	000000	000000	000000	–	111111	111111
		BBA	000000	000000	000000	000000	000000	000000	–	111111
		KCS	000000	000000	000000	000000	000000	000000	000000	–
Lowest confidence level	First dataset	WSIM	–	0.9517	0.9552	0.9818	0.9997	0.9998	1.0	1.0
		LDA+LM	0	–	0	0.4743	0.9998	0.9998	1.0	1.0
		LM	0	0.2237	–	0.7170	0.9998	0.9998	1.0	1.0
		WMD	0	0	0	–	0.9993	0.9993	1.0	1.0
		TATB	0	0	0	0	–	0.5758	0.9999	0.9996
		LDA	0	0	0	0	0	–	0.9999	0.9999
		BBA	0	0	0	0	0	0	–	0
		KCS	0	0	0	0	0	0	0	–

Table 10 (continued)

	Methods	WSIM	LDA+LM	LM	WMD	TATB	LDA	BBA	KCS
Second dataset	WSIM	–	0.9987	0.9989	1.0	1.0	1.0	1.0	1.0
	LDA+LM	0	–	0.9546	0.9999	1.0	1.0	1.0	1.0
	LM	0	0	–	0.9986	1.0	1.0	1.0	1.0
	WMD	0	0	0	–	1.0	1.0	1.0	1.0
	TATB	0	0	0	0	–	1.0	1.0	1.0
	LDA	0	0	0	0	0	–	1.0	1.0
	BBA	0	0	0	0	0	0	–	0.9995
	KCS	0	0	0	0	0	0	0	–

Table 11 Bias, variance, generalization error (GE) of the WSIM and comparison methods on two datasets

Methods	First dataset			Second dataset		
	Bias	Variance	GE	Bias	Variance	GE
WSIM	**0.4975**	0.0414	**0.2889**	**0.5004**	0.0194	**0.2699**
LDA+LM	0.5330	0.0409	0.3250	0.5254	0.0192	0.2952
LM	0.5320	0.0415	0.3245	0.5266	0.0193	0.2966
WMD	0.5430	0.0387	0.3335	0.5440	0.0169	0.3128
TATB	0.6635	0.0582	0.4985	0.6694	0.0156	0.4637
LDA	0.6670	0.0589	0.5038	0.6847	0.0155	0.4843
KCS	0.8675	0.0225	0.7750	0.9707	0.0015	0.9437
BBA	0.8725	**0.0062**	0.7674	0.9612	**0.0013**	0.9253

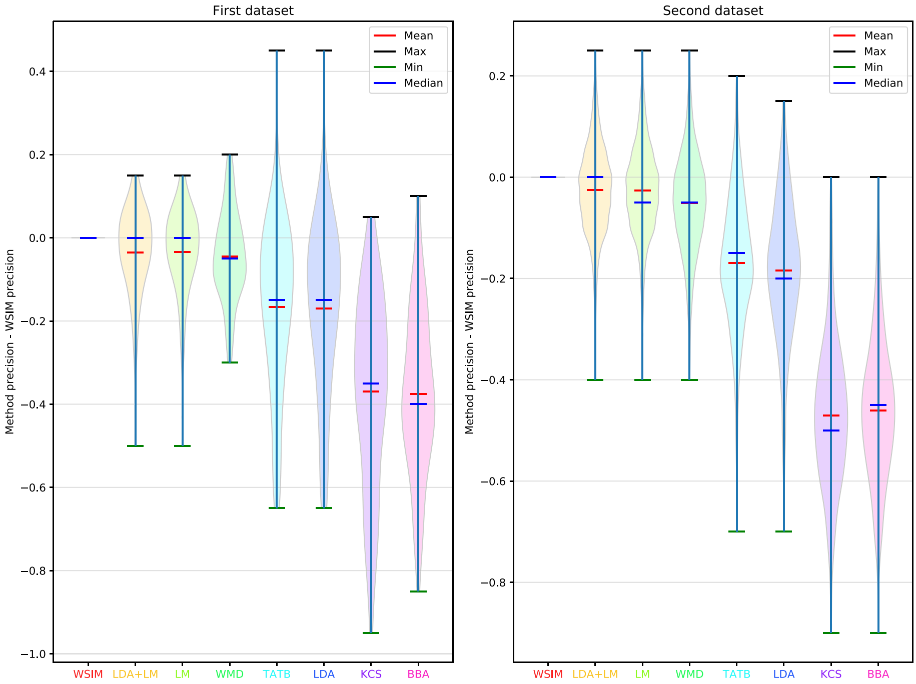

Fig. 3 Precision bias between all methods and the WSIM on each manuscript

the importance of papers that are highly similar to the manuscript and degrade the final performance.

4.6 Significance test

In this subsection, we analyze the statistical significance of method performance improvement through a significance test. We randomly divide all manuscripts into ten through tenfold cross-validation and then compare the precision of the WSIM and all comparison methods through the two-sided paired t-test (Smucker et al. 2007). Table 9 shows the mean,

Table 12 The values of four different hyperparameters in the WSIM

Hyperparameters	Value					
ξ_d	0.01	**0.05**	0.1	0.15	0.2	0.25
ξ_r	0.01	**0.05**	0.1	0.15	0.2	0.25
η	0.05	0.15	**0.25**	0.35	0.45	0.55
t	20	50	**80**	110	140	170

t-value, and *p*-value of precision (*topN*=20) on the two datasets. The first row is a paired *t*-test for WSIM and WSIM, and they do not differ. The confidence level of the WSIM over the other methods is at least 97.5% in both datasets. This proves that the performance improvement of the WSIM is statistically significant.

For a more comprehensive analysis of the statistical significance between the performance of all methods, we extended the results presented in Table 9 to all methods and all metrics. Table 10 shows the results. The upper part of Table 10 shows whether each method (column) passes the significance test of outperforming each method (row) on six metrics (in the order of P, R, F_1, MAP, NDCG, and bpref). We set the confidence level at 97.5% and record 1 if it passes the significance test at this confidence level; otherwise, we record 0. For example, "101111" means that only the recall R does not pass the significance test. The bottom half of Table 10 shows the lowest value of the confidence level among the six metrics. For example, in the first dataset, the lowest confidence level at which WSIM outperforms LDA+LM is 0.9517, which is the confidence level of recall R, as seen in the upper part of Table 10. We can obtain the performance ranking between different methods: WSIM > LDA+LM ≥ LM ≥ WMD > TATB ≥ LDA > BBA ≥ KCS. From Table 10, we can see that this ranking's confidence level is at least 95%.

4.7 Bias-variance decomposition

In this subsection, we analyze the generalizability of all methods, and we perform a bias-variance decomposition on the precision of each manuscript. We use precision=1.0 as the true output and calculate the bias between it and each method's precision. The generalization error is equal to the square of the bias plus the variance. Table 11 shows the bias, variance, and generalization error for each method (*topN*=20) on both datasets. The bias and generalization error of the WSIM are minimal. The variance of the WSIM is almost identical to that of LDA-LM and LM on which it is based. This proves the excellent generalization capability of WSIM, as the WSIM reduces the bias while maintaining the variance.

To further determine whether the performance is helped by a few manuscripts or many manuscripts, we show the precision bias between each method and the WSIM on each manuscript. The violin plot in Fig. 3 shows the distribution of precision bias. Each violin in the figure shows the precision bias of a method on each manuscript and is labeled with the maximum, minimum, mean, and median of the precision (*topN*=20). The wider the violin is, the more the manuscripts that are in that position. From the figure, we can observe that (1) the median value of the precision bias for each comparison method is below zero; and (2) the distribution of the precision bias is close to the normal distribution. This proves that our method improves the performance of most manuscripts, and the improvement is close to the normal distribution.

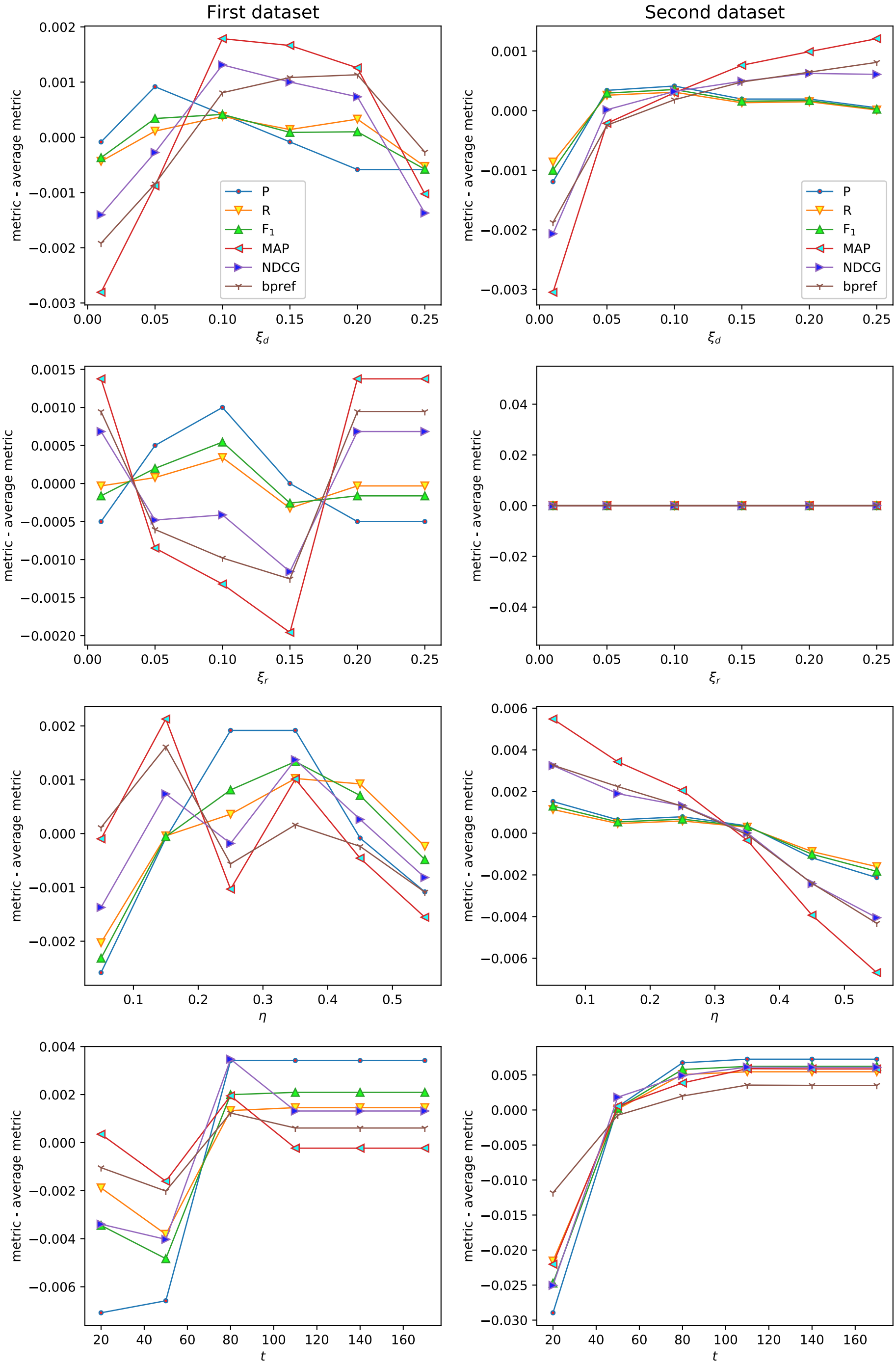

Fig. 4 Performance of the four hyperparameters of the WSIM under different values

Table 13 The matching results for the two manuscripts

	Title	Fields
p_1	RNA sequencing identifies upregulated kyphoscoliosis peptidase and phosphatidic acid signaling pathways in muscle hypertrophy generated by transgenic expression of myostatin propeptide.	f_1,f_2,f_3
r_{11}	Deer antler extract improves fatigue effect through altering the expression of genes related to muscle strength in skeletal muscle of mice.	f_4,f_1
r_{12}	Haplotype-assisted accurate non-invasive fetal whole genome recovery through maternal plasma sequencing.	f_5,f_1
r_{13}	Integrating multiple resources to identify specific transcriptional cooperativity with a Bayesian approach.	f_4,f_3
r_{14}	Highly specific DNA detection from massive background nucleic acids based on rolling circle amplification of target dsDNA.	f_6,f_7
r_{15}	Telomere-associated factor expression in replicative senescence of human embryonic lung fibroblasts.	f_1,f_8
p_2	A new multi-objective particle swarm optimizer using empirical movement and diversified search strategies.	f_9,f_{10},f_3
r_{21}	An evolutionary algorithm based on constraint set partitioning for nurse rostering problems.	f_{11},f_3,f_{11}
r_{22}	Hybridizing invasive weed optimization and simulated annealing algorithm for high-dimensional function optimization.	f_{12},f_3
r_{23}	Multichannel cooperative sensing for cognitive radio with users owning heterogeneous sensing ability.	f_{13},f_{14}
r_{24}	A DC offset adaptive energy detection algorithm.	f_{14},f_3
r_{25}	A tutorial on event-based optimization-a new optimization framework.	f_{14},f_{10}

Table 14 Notions of fields

Symbol	The field
f_1	Molecular biology
f_2	Physical and theoretical chemistry
f_3	Computer science applications
f_4	Biochemistry, genetics and molecular biology(all)
f_5	Biochemistry
f_6	Medicine(all)
f_7	Chemical engineering(all)
f_8	Environmental chemistry
f_9	Control and optimization
f_{10}	Applied mathematics
f_{11}	Computational theory and mathematics
f_{12}	Computer science (miscellaneous)
f_{13}	Computer networks and communications
f_{14}	Control and systems engineering

4.8 Hyperparameter analysis

In this subsection, we show the performance of important hyperparameters of the WSIM at different values to investigate the impact of different hyperparameter values on the performance. We use the method of control variables to analyze the four most important parameters (t, η, ξ_d, ξ_r) in the WSIM. Table 12 shows the six values used for each hyperparameter, and the values given in Section 4.2 are in bold. When the value of a hyperparameter is a variable, the other parameters will use fixed boldface values. Figure 4 shows the experimental results of the WSIM on four hyperparameters, six hyperparameter values, six metrics, and two datasets. The horizontal coordinates show the hyperparameter values. The vertical coordinate shows the performance at the current hyperparameter value minus the average performance of the six hyperparameter values. From the range of values of vertical coordinates, we can obtain the following conclusions: (1) the influence of hyperparameters on performance can be ordered as $t > \eta > \xi_d > \xi_r$; (2) the influence of hyperparameters η, ξ_d, ξ_r on performance is less than 0.7%; and (3) when hyperparameter $t > 110$, its influence on performance tends to be stable.

4.9 Case study

In this subsection, we provide a case study analysis to show the effectiveness of the WSIM with respect to the experimental evaluation and illustrate the practicality of the method.

To illustrate the effectiveness of the WSIM, we show the matching results of two manuscripts (p_1,p_2) in the first dataset. To be reasonable, we chose two test samples with single-sample precision approximating the evaluation results (50.25%). The precision ($topN$=20) of these two manuscripts is 50% and 55%, respectively. Among the reviewers recommended for the manuscript, we focus on five reviewers corresponding to matching errors to show that the WSIM is more effective than the results of the evaluation metric. We use the title and the related fields (classification) of the paper to display the manuscript and reviewer's information. The reviewer's title and related fields are obtained from the most similar reviewer papers with respect to the target manuscript.

Table 13 shows the matching results for the two manuscripts, where the fields use symbolic representations, and Table 14 explains the names of the fields corresponding to the symbols. Among the five reviewers matched to manuscript p_1, reviewers $\{r_{11}, r_{12}, r_{13}, r_{14}\}$ are truly suitable. Among the five reviewers matched to manuscript p_2, reviewers $\{r_{21}, r_{22}\}$ are truly suitable. The reviewers corresponding to matching errors still have many of the appropriate qualifications to review the target manuscript. This is because the groundtruth of the evaluation metrics is strict and using only the label disqualifies some suitable reviewers.

This analysis shows that the WSIM is more effective and practical than the results of evaluation metrics.

5 Conclusions

We proposed an approach named the word and semantic-based iterative model (WSIM) to solve the retrieval-based reviewer assignment problem (RRAP). The WSIM determines the most appropriate reviewers for a target manuscript using a combination of word information and semantic information and considering the constraints of the RAP by improving the similarity calculations between reviewers and manuscripts. We reduce overfitting to incomplete reviewer data and the interference in the assignment from nonmanuscript-related papers in the reviewer data with a ranking-based approach and iterative model. We compare our approach with seven existing methods in closed-world settings, and the experimental results validate the effectiveness of our method.

The RAP includes the retrieval-based RAP, which we address in this paper, and the assignment-based RAP, which requires different strategies for different requirements (O'Dell et al. 2005) and is an interesting problem for future research. In the future, we also plan to provide an efficient system based on our proposed method for use by journals and conferences. In addition, we plan to explore how our methods can be applied to other research topics, such as information retrieval and question-answerers.

Acknowledgements This work was partially supported by National High Technology Research and Development Program (Grant # 2017YFB1401903), the National Natural Science Foundation of China (Grants # 61876001, # 61602003 and # 61673020), the Provincial Natural Science Foundation of Anhui Province (Grants # 1708085QF156), and the Recruitment Project of Anhui University for Academic and Technology Leader.

References

Anaya, Antonio R., Luque, Manuel, Letón, Emilio, & Hernández-del-Olmo, Félix. (2019). Automatic assignment of reviewers in an online peer assessment task based on social interactions. *Expert Systems with Applications, 36*(4), e12405. (**ISSN 0266-4720.**).

Basu, Chumki, Hirsh, Haym, Cohen, William W., & Nevill-Manning, Craig. (2001). Technical paper recommendation: A study in combining multiple information sources. *Journal of Artificial Intelligence Research, 14,* 231–252.

Bergstra, James, & Bengio, Yoshua. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2), 281–305.

Biswas, Humayun Kabir & Hasan, Md Maruf. (2007). Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment. In *Information and Communication Technology, 2007. ICICT'07. International Conference on*, pages 82–86. IEEE

Buckley, Chris, & Voorhees, Ellen M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual International ACM SIGIR Conference on research and development in information retrieval*, pages 25–32. ACM

Büttcher, Stefan, Clarke, Charles LA, & Gordon V Cormack. (2016). Information retrieval: Implementing and evaluating search engines. Mit Press.

Charlin, Laurent, & Zemel, Richard. (2013). *The toronto paper matching system: An automated paper-reviewer assignment system* (p. 28). JMLR: W&CP.

Charlin, Laurent, Zemel, Richard S., & Boutilier, Craig. (2012). A framework for optimizing paper matching. *arXiv preprint*arXiv:1202.3706

Di Mauro, Nicola, Basile, Teresa MA, Ferilli, Stefano. (2005). Grape: An expert review assignment component for scientific conference management systems. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 789–798. Springer

Dumais, Susan T., & Nielsen, Jakob. (1992). Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 233–244. ACM

Dung, Nguyen Dinh, Cong, Nguyen Huu, & Anh, Nguyen Tuan. (2017). Algorithm of dynamic programming for paper-reviewer assignment problem. *IRJET, 04*(11)

Ferilli, Stefano, Di Mauro, Nicola, Maria Altomare Basile, Teresa, Esposito, Floriana, & Biba, Marenglen. (2006). Automatic topics identification for reviewer assignment. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 721–730. Springer

Flach, Peter A., Spiegler, Sebastian, Golénia, Bruno, Price, Simon, Guiver, John, Herbrich, Ralf, et al. (2010). Novel tools to streamline the conference review process: Experiences from sigkdd'09. *ACM SIGKDD Explorations Newsletter, 11*(2), 63–67.

Goldsmith, Judy, & Sloan, Robert H. (2007). *The ai conference paper assignment problem* (pp. 53–57). Vancouver. In Proc: AAAI Workshop on Preference Handling for Artificial Intelligence.

Hettich, Seth, & Pazzani, Michael J. (2006). Mining for proposal reviewers: lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 862–871. ACM

Hoang, Dinh Tuye, Nguyen, Ngoc Thanh, & Hwang, Dosam. (2019). Decision support system for assignment of conference papers to reviewers. In *International Conference on Computational Collective Intelligence*, pages 441–450. Springer

Karimzadehgan, Maryam, & Zhai, ChengXiang. (2009). Constrained multi-aspect expertise matching for committee review assignment. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1697–1700. ACM

Karimzadehgan, Maryam, & Zhai, ChengXiang. (2012). Integer linear programming for constrained multi-aspect committee review assignment. *Information Processing and Management, 48*(4), 725–740.

Karimzadehgan, Maryam, Zhai, ChengXiang, & Belford, Geneva. (2008). Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1113–1122. ACM

Kim, Jungil, & Lee, Eunjoo. (2018). Understanding review expertise of developers: A reviewer recommendation approach based on latent dirichlet allocation. *Symmetry, 10*(4), 114.

Kou, Ngai Meng, Hou, U Leong, Mamoulis, Nikos, Gong, Zhiguo. (2015a). Weighted coverage based reviewer assignment. In *Proceedings of the 2015 ACM SIGMOD International Conference on management of data*, pages 2031–2046. ACM

Kou, Ngai Meng, Mamoulis, Nikos, Li, Yuhong, Zhiguo Gong, Ye Li (2015b) et al. A topic-based reviewer assignment system. *Proceedings of the VLDB Endowment*, 8(12): 1852–1855

Kusner, Matt, Sun, Yu, Kolkin, Nicholas, & Weinberger, Kilian. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966

Li, Baochun, & Hou, Y Thomas. (2016). The new automated iEEE infocom review assignment system. *IEEE Network, 30*(5), 18–24.

Li, Xinlian, & Watanabe, Toyohide. (2013). Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers. *Procedia Computer Science, 22,* 633–642.

Lian, Jing Wu. (2018). *Nicholas Mattei, Renee Noble, and Toby Walsh*. The Conference paper assignment problem: Using order weighted averages to assign indivisible goods.

Liu, Ou., Wang, Jun, Ma, Jian, & Sun, Yonghong. (2016). An intelligent decision support approach for reviewer assignment in r&d project selection. *Computers in Industry, 76,* 1–10.

Long, Cheng, Wong, Raymond Chi-Wing, Peng, Yu, & Ye, Liangliang. (2013). On good and fair paper-reviewer assignment. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1145–1150. IEEE

McGlinchey, Noel, Hunter, Tom, Bromley, Jack, Fisher, Ruth, Debiec-Waszak, Anna, & Gaston, Thomas. (2019). Do journal administrators solve the reviewer assignment problem as well as editors? consideration of reviewer rigour and timeliness. *Learned Publishing, 32*(1), 37–46. (**ISSN 0953-1513.**).

Mimno, David, McCallum, Andrew. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 500–509. ACM

Misale, Mohini, & Vanwari, Pankaj. (2017). A survey on recommendation system for technical paper reviewer assignment. In *Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of*, volume 2, pages 329–331. IEEE

O'Dell, Regina, Wattenhofer, Mirjam, & Wattenhofer, Roger. (2005). *The paper assignment problem* (p. 491). Department of Computer Science: Technical report/Swiss Federal Institute of Technology Zurich.

Ogunleye, O., Ifebanjo, T., Abiodun, T., & Adebiyi, AA. (2017). Proposed framework for a paper-reviewer assignment system using word2vec

Peng, Hongwei, Hu, Haojie, Wang, Keqiang, & Wang, Xiaoling. (2017). Time-aware and topic-based reviewer assignment. In *International Conference on Database Systems for Advanced Applications*, pages 145–157. Springer

Price, Simon, & Flach, Peter A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM, 60*(3), 70–79.

Protasiewicz, Jarosław. (2014). A support system for selection of reviewers. In *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pages 3062–3065. IEEE

Protasiewicz, Jaroslaw, Pedrycz, Witold, Kozlowski, Marek, Dadas, Slawomir, Stanislawek, Tomasz, Kopacz, Agata, & Galezewska, Malgorzata. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems, 106,* 164–178.

Rigaux, Philippe. (2004). An iterative rating method: application to web-based conference management. In *Proceedings of the 2004 ACM symposium on Applied computing*, pages 1682–1687. ACM

Shon, Ho Sun, Han, Sang Hun, Kim, Kyung Ah, Cha, Eun Jong, & Ryu, Keun Ho. (2017). Proposal reviewer recommendation system based on big data for a national research management institute. *Journal of Information Science, 43*(2), 147–158.

Sidiropoulos, Nicholas D., & Tsakonas, Efthymios. (2015). Signal processing and optimization tools for conference review and session assignment. *IEEE Signal Processing Magazine, 32*(3), 141–155.

Smucker, Mark D., Allan, James, & Carterette, Ben. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632

Fair and accurate reviewer assignment in peer review. (2019). Ivan Stelmakh, Nihar B Shah, and Aarti Singh. Peerreview4all. *Algorithmic Learning Theory, 98,* 827–855.

Tang, Wenbin, Tang, Jie, Tan, Chenhao. (2010). Expertise matching via constraint-based optimization. In *Web intelligence and intelligent agent technology (wi-iat), 2010 IEEE/wic/acm International Conference on*, volume 1, pages 34–41. IEEE

Tang, Wenbin, Tang, Jie, Lei, Tao, Tan, Chenhao, Gao, Bo., & Li, Tian. (2012). On optimization of expertise matching with various constraints. *Neurocomputing, 76*(1), 71–83.

Tang, Xijin, & Zhang, Zhengwen. (2008). Paper review assignment based on human-knowledge network. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pages 102–107. IEEE

Tayal, Devendra Kumar, Saxena, P. C., Sharma, Ankita, Khanna, Garima, & Gupta, Shubhangi. (2014). New method for solving reviewer assignment problem using type-2 fuzzy sets and fuzzy functions. *Applied Intelligence, 40*(1), 54–73.

Wang, Fan, Chen, Ben, Miao, Zhaowei. (2008). A survey on reviewer assignment problem. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 718–727. Springer

Wang, Fan, Zhou, Shaorui, & Shi, Ning. (2013). Group-to-group reviewer assignment problem. *Computers and Operations Research, 40*(5), 1351–1362.

Xu, Yichong, Zhao, Han, Shi, Xiaofei, & Shah, Nihar B. (2019). On strategyproof Conference peer review. pages 616–622

Yarowsky, David, & Florian, Radu. (1999). Taking the load off the conference chairs-towards a digital paper-routing assistant. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*

Yeşilçimen, Ali, & Yıldırım, E Alper. (2019). An alternative polynomial-sized formulation and an optimization based heuristic for the reviewer assignment problem. *European Journal of Operational Research, 276*(2), 436–450.

Zhang, Dong, Zhao, Shu, Duan, Zhen, Chen, Jie, Zhang, Yanping, & Tang, Jie. (2020a). A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation. *ACM Transactions on Information Systems, 38*(1), 1–20. (**ISSN 1046-8188.**).

Zhang, Dong, Zhao, Shu, Duan, Zhen, Chen, Jie, Zhang, Yanping, & Tang, Jie. (2020b). A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation. *ACM Transactions on Information Systems (TOIS), 38*(1), 1–20.

Zhao, Shu, Zhang, Dong, Duan, Zhen, Chen, Jie, Zhang, Yan-ping, & Tang, Jie. (2018). A novel classification method for paper-reviewer recommendation. *Scientometrics*, pages 1–21