OPEN ACCESS
Vietnam Journal of Computer Science
Vol. 10, No. 2 (2023) 197–216

DOI: 10.1142/S2196888822500336

World Scientific
www.worldscientific.com

A Graph-Based Topic Modeling Approach to Detection of Irrelevant Citations

Phu Pham*, Hieu Le† and Nguyen Thanh Tam‡

Faculty of Information Technology
HUTECH University, Ho Chi Minh City, Vietnam
**pta.phu@hutech.edu.vn*
†*lt.hieu@hutech.edu.vn*
‡*nt.tam88@hutech.edu.vn*

Quang-Dieu Tran

Ho Chi Minh National Academy of Politics
Hanoi, Vietnam
dieutq@hcma.edu.vn

Received 3 June 2022
Revised 27 June 2022
Accepted 28 June 2022
Published 5 October 2022

In the recent years, the academic paper influence analysis has been widely studied due to its potential applications in the multiple areas of science information metric and retrieval. By identifying the academic influence of papers, authors, etc., we can directly support researchers to easily reach academic papers. These recommended candidate papers are not only highly relevant with their desired research topics but also highly-attended by the research community within these topics. For very recent years, the rapid developments of academic networks, like Google Scholar, Research Gate, CiteSeerX, etc., have significantly boosted the number of new published papers annually. It also helps to strengthen the borderless cooperation between researchers who are interested on the same research topics. However, these current academic networks still lack the capabilities of provisioning researchers deeper into most-influenced papers. They also largely ignore quite/irrelevant papers, which are not fully related with their current interest topics. Moreover, the distributions of topics within these academic papers are considered as varying and it is difficult to extract the main concentrated topics in these papers. Thus, it leads to challenges for researchers to find their appropriated/high-qualified reference resources while doing researches. To overcome this limitation, in this paper, we proposed a novel approach of paper influence analysis through their content-based and citation relationship-based analyses within the biographical network. In order to effectively extract the topic-based relevance from papers, we apply the integrated graph-based citation relationship analysis with topic modeling approach to automatically learn the distributions of keyword-based labeled topics in forms of unsupervised learning approach, named as TopCite. Then, we base on the constructed graph-based paper–topic structure to identify their relevancy levels. Upon the

identified relevancy levels between papers, we can support for improving the accuracy performance of other bibliographic network mining tasks, such as paper similarity measurement, recommendation, etc. Extensive experiments in real-world AMiner bibliographic dataset demonstrate the effectiveness of our proposed ideas in this paper.

Keywords: Citation network analysis; topic modeling; irrelevancy citation classification.

1. Introduction

In the recent years, with the tremendous development of Internet in multiple platforms, the number of online/e-printed journals has dramatically increased. It has led to significant increases in the number of annual publications. These developments also make it faster and easier for research communities to reach new research findings,[1,2] international collaborations[3] in multiple disciplines. In general, within the supports of online journal submission/publishing platforms, the academic communities have published millions of research papers annually and the publication rates are still continuously increasing over the time. The increase in academic publication amount has provided multiple advances to researchers to not only obtain new knowledge but also be rapidly notified with the state-of-the-art findings in their recent research domains. However, with the rapid increases of publications recently, the literature reviews/studies while doing research are becoming more challenging than ever and considered as the most time-consuming task. In fact, at the beginning/starting phases of academic/research projects, researchers have to spend a lot of time to read as much as possible the recent publications, which are related to their project's problems. In order to reach suitable publications for literature reviews, a common way that most of the researchers have used is keyword-based finding/searching through Internet-based search engines, like Google, Bing, etc. To effectively support the researchers to effectively filter and hit their desired papers, several academic search engine/bibliographic network-based platforms have been constructed, such as Google Scholar,[a] CiteSeerX,[b] Research Gate,[c] etc. Besides the basic search/filtering mechanism for finding proper papers based on the set of input keywords, these scholar-aided platforms are also equipped with the advanced features related for paper/citation/collaboration recommendation.[4] For example, in Google Scholar and Research Gate, these platforms have been designed to look for references of relevant publications that researchers have already found and read. They will also be able to identify experts in researchers' interesting domains as well as suggest high-quality papers from these experts to them. Moreover, by evaluating the citation relationships between papers that researchers have read, these academia-aided platforms also combine the citation-based and keyword-based correlations between papers to provide more correct/meaningful search results for researchers. However, the reliance of keyword-based and citation relation-based features to

[a] Google Scholar: https://scholar.google.com.vn/.
[b] CiteSeerX: https://citeseerx.ist.psu.edu/.
[c] Research Gate: https://www.researchgate.net/.

identify relevant papers sometimes can be considered as inefficient and time-consuming due to the fact that large computational efforts must be spent for the processes of paper's content indexing and paper relationship analyzing. Moreover, the keyword-based indexing/searching mechanism is also considered as ineffective for specific conditions in which multiple complex academic concept and topic relevancy are taken into consideration by the users/researchers. Specifically, the keyword-based search mechanism might be unable to account for identifying academic concepts, which are formed in different names/mathematic notation forms. In addition, the usage of keyword sets to represent for the papers while indexing is also considered as insufficient to cover the distributions of topics within papers. Thus, it fails to compute the topic-based relevancies between papers to provide accurate recommendations for researchers. Since the search results are insufficient or incorrect due to the ambiguity of researchers' input keywords, they might lead the researcher backward in time or go with wrong direction. Therefore, further researches in paper similarity and citation relationship analysis are currently widely attended by many researchers to overcome these existing challenges.

1.1. *Recent achievements in paper similarity analysis and challenges*

From the past, most of scholar-aided search engines/recommendation systems are designed upon the content-based indexing/searching approach, a.k.a., bag-of-words (BOWs) approach, in which the proposed models[5–8] attempt to measure the similarity between the set of keywords within users' queries and indexed document corpus to find relevant papers. In general, the content-based paper searching techniques highly relied on the process of accessing the contents of the papers, such as metadata, title, abstract, body, etc. To leverage the accuracy of return papers, different word weighting mechanisms in different content blocks of the papers are proposed. However, these traditional content-based filtering techniques still suffered several limitations related to the sparse textual representations of BOWs approach as well as the capability of capturing various concept/topic distributions from the given text corpora. Moreover, the content-based paper similarity analysis techniques also suffered another issue related to the failure in identifying academic concepts with different forms, which normally occurred in multi-disciplined scientific papers. Thus, the search engines might tend to return irrelevant results, which covered the different topics/fields. On the other hand, there are several attempts in paper similarity measurement and recommendation by applying the collaborative filtering paradigm in which the researchers' references/interactions within the given online bibliographic platforms are carefully taken in consideration to provide appropriated paper recommendations.

In the recent times, there are several studies[9–11] that demonstrated the effectiveness of evaluating the user's interest levels/ratings on specific papers in order to efficiently rank and suggest similar ones. However, the collaborative filtering techniques also suffered several challenges related to the cold-start issue as well as the

practical implementations in which the authors/researchers are unwilling to reveal their interests/ratings on the papers they have read. Thus, this approach is considered as quite limited and incapable for applying in real-world search engine/academia-aided systems. Moreover, within this approach, the content-based/topic-based features of academic papers are largely ignored. Thus, they might not be involved much in the paper relevancy evaluation process. Within the relation-based analysis approach,[12,13] the citation relationships between papers are utilized to identify the relatedness between papers. The similarity levels between scientific papers are identified by using different aspects, such as co-citation, citation proximity, bibliographic coupling, etc. By extracting these relation-based features, the given academia-aided systems can find relevant papers upon researchers' input papers, which might be slightly similar with the input ones in both content-based and topic-based aspects. However, this approach sometimes still suffered limitations due to the irreverent citations within the evaluated papers, especially with multi-disciplined papers. Incorporating the citation relation-based approach, there are other approaches, which mainly depend on the graph-structured features of bibliographic networks,[14,15] such as topic taxonomy,[16] general domain knowledge graph,[17,18] etc., to identify the similarity between scientific papers. However, the proposed techniques within the citation relationship and graph-based approaches still suffered challenges related to the thorough evaluations on the topic distributions within multi-disciplined papers, which can highly influence on the paper similarity measurement processes. In fact, citation relationships are considered as most important features, which can be utilized to identify the relatedness between source cited papers and the citing ones. However, realistic analyses have shown that a set of citations in a paper are not equally important and some citations are more relevant than the others. For example, in a paper, less-important papers might be cited once, whereas other more-important papers are cited multiple times. This fact indicates that a list of references in a paper should have different levels of importance as well as content-based relevancies.

1.2. *Our motivations and contributions in this paper*

Graph-based topic modeling approach for irrelevant citation classification. Mainly inspired from the recent achievements in both content-based and citation relation-based paper similarity measurement and recommendation techniques, in this paper, we proposed a novel hybrid paper similarity measurement technique, which combines the citation relation and topic feature analysis techniques to identify the similarity scores between papers as well as handling irrelevant citation classification task, called as TopCite (as illustrated in Fig. 1). Our proposed TopCite model is designed as a graph-based topic modeling approach in which the paper–topic relationships are extracted from papers through the labeled topic modeling approach. The labeled topic taxonomy of the given paper set is constructed by using defined main keywords of papers within the given bibliographic network. In more specifics, within

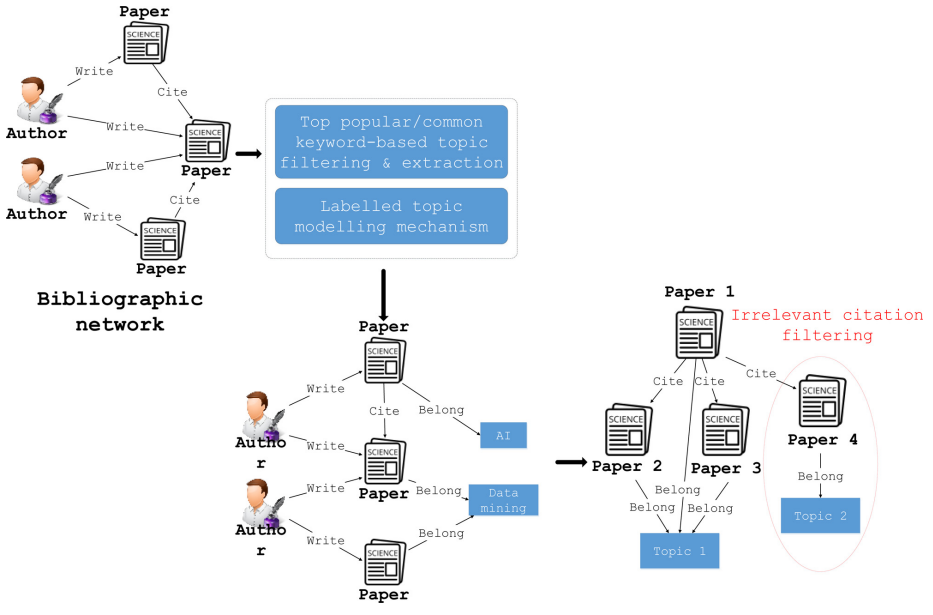


Fig. 1. The architecture of our proposed TopCite model.

the approach of our proposed TopCite model, in order to effectively achieve the topic distributions from the given scientific paper resources, we mainly applied the topic modeling model.[19,20] By doing this, we learn the distributions of labeled topics in papers, which are identified as the set of main keywords with the given paper set. The sets of main keywords of the given paper corpus can be assumed as the general topic taxonomy, which is similar to the ACM computing classification system (CCS).[d]

In order to eliminate the unpopular topics/keywords within the given paper corpus, we only filtered the top common keyword-based topics, which are the most frequently occurred/identified in the corpus. Then, the LLDA model supports to produce the probabilistic distributions of labeled common topics across the given paper set. The latent labelled topic distributions are used as the representations of these papers, which are later used to identify the similarity between citation-link papers through the out-of-shelf similarity metrics (e.g. Euclidean distance, cosine similarity) in the form of topic relatedness analysis process. Finally, we applied the topic-based similarity measurement to identify the relatedness levels of references in each paper. It supports to effectively deal with the irrelevant/less-relevant citation identification problem, which directly assists for leveraging the performance of similar paper retrieval as well as recommendation. In general, our contributions in this paper can be summarized as three-folds, including

- With the given paper corpus, we apply the main keyword-based topic filtering mechanism to achieve a set of top common topics, which are used as the general

[d]ACM Computing Classification System (CCS): https://dl.acm.org/ccs.

taxonomy for the given paper set. Then, the LLDA model is applied to achieve the probabilistic labeled topic distributions across all papers. These labelled topic distributions are considered as the rich-semantic representations of these papers, which are later utilized to identify the paper–topic relationships in forms of graph-based topic modelling approach.

- From the extracted latent labeled topics in previous step as well citation relationships between papers, we constructed a topic–paper-based graph structure. Within this graph-based structure, we have two types of relationships, including the paper–paper citations and paper–topic relatedness. In order to identify the proper paper–topic relationships, we mainly selected the highest distributed latent topic of each paper upon the extract topic distributions, which have been achieved in previous step. To effectively store and extract the information from the given constructed topic–paper graph structure, we store them into a graph database in order to ensure the capability of efficient and scalable data retrieval/processing. In our implementation, we choose Neo4J as the main graph-based database platform to store our data.
- Finally, upon the constructed topic–paper graph, we conduct the irrelevant citation classification task, which depends on the common belonged topics of a given evaluated paper and its citing paper set. This task is formally formulated as a binary classification problem in which a citing paper/reference is categorized as relevant/irrelevant citation. To demonstrate the effectiveness of our approach in handling irrelevant citation classification problem, we conducted extensive experiments in the real-world AMiner dataset.

The left contents of our papers are organized into three sections. In the next section, we briefly reviewed about recent works, which are related to scientific paper similarity measurement and citation analysis domain. Then, in the next section, we formally present the methodology and detailed implementations of our proposed TopCite model for dealing with the irrelevant citation classification problem. In the fourth section, we present extensive experiments to demonstrate the effectiveness of our proposed ideas in this paper as well as further information on how we organize the AMiner bibliographic network into graph-based database to improve the data storage scalability and query performance. Finally, we conclude our achievements in this paper in the last section and highlight some potential improvements for future works.

2. Related Works

For many years, research communities in multiple disciplines have been widely engaged in finding better ways for dealing with scientific paper similarity measurement as well as citation relationship analysis problems. A significant number of researchers have published in this domain with worthy contributions to our communities for the purpose of faster and more accurate research works. There existed different paper

relevancy, and citation analysis methods have been proposed recently, which can be categorized into three main approaches, the textual/content-based, collaborative filtering and relational/graph-structural analysis approach.

Textual/content-based approach. Within this approach, in the forms of primitive textual analysis and representation learning problem, researchers have tried to extract distinctive features from scientific papers to obtain the unique representations of scientific papers, which are later served for the purposes of indexing and similarity searching.[6–8] The BOW-based techniques and their family (e.g. TF-IDF, n-grams analysis, word information gain) have been widely adapted in the past to achieve the nonlinear transformed embeddings of papers, which mainly represented as important weights of their inside occurred keywords/phases. The recent works[6,7] have shown that the BOW-based textual analysis and learning techniques, such as TF-IDF, can still be considered as affordable for handling the keyword-varied representation learning objectives and can achieve reasonable accuracy results for the scientific paper similarity measurement task. However, the BOW-based techniques still suffered major challenges related to the simplicity and sparse representations of textual documents. Moreover, while important keywords/phases in scientific papers are independently evaluated, it can't carry the sequential relationship information between continuous words. Thus, they can't achieve better quality representations of the papers as well as provide more accurate similarity measurement results. Last but not the least, the BOW-based textual representation approach is also considered as unable to achieve the topic-varied distributions between papers, especially in multi-disciplined papers.

Collaborative filtering approach. Majorly inspired from the collaborative filtering paradigm (e.g. matrix factorization, factorization machine) in the recommendation domain, researchers have tried to formulate the scientific paper representation learning as the low-ranged latent feature factorization process. In this approach,[9–11] the shared researchers' preferences, which indicate their interests on specific papers, are utilized to achieve the representations. Such as, well-known works of Liu *et al.,*[9] in adopting the association mining technique[21] to achieve the rich-structural representations of papers, which are based on their citation relationships and users' interactions. Then, the achieved representations are used to identify the similarity scores between papers for conducting relevant search/recommendation tasks. Similar to that representation learning paradigms in the works of Murali *et al.*[10] and Sakib *et al.*,[11] the rich-structural features within the multi-typed relationships of papers and users/researchers have been utilized to facilitate the paper representation learning process through different latent feature extraction/factorization methods within the collaborative filtering paradigm. Even, proposed methods in this approach have demonstrated significant improvements in both scientific paper representation learning/similarity measurement and recommendation, they still suffered several limitations related to the realistic application implementation in case that users/researchers are unwilling to share their interests/ratings on recent interacted papers. Moreover, proposed models within this approach also

encountered the drawbacks of largely ignoring the textual/topic-varied features inside scientific papers to deliver richer-semantic paper search results.

Relational/graph-structural analysis approach. Relying on the paradigm of information network analysis and mining approach, the proposed scientific paper representation learning techniques[14–18] in this approach highly depend on evaluating the graph-structural features that are achieved from multi-typed relationships between papers and their associated entities like authors/researchers, keywords/concepts, topics, etc. Such as, recent survey works of Ali *et al.* listed several significant studies, which demonstrated the effectiveness of incorporating paper–topic relational taxonomy with the given graph-structured bibliographic network to achieve rich representations of scientific papers.[16] Also, on the same research direction, Manrique *et al.*[17] and Tang *et al.*[18] recently proposed the combination between bibliographic network graph-structural analysis with external knowledge graphs to achieve better-quality representations of scientific papers. Then, these rich-structural paper representations are utilized to facilitate the similarity measurement as well as recommendation problems. By mainly relying on the relationships between papers and their associated entities within the given bibliographic network, similar to the collaborative filtering approach, this approach still suffered challenges, which are related to the ignorance of textual/topic analysis inside the paper resources. Thus, they might be unable to achieve better performance in complex/multi-disciplined paper relevancy measurement problems.

3. Methodology and Implementation

In this section, we formally present important background concepts, methodology and detailed implementations of our proposed TopCite model. Our proposed TopCite model supports to achieve the rich-semantic and topic-varied representations of scientific papers within a given bibliographic network. To properly extract the topic distributions from the paper set, we apply the labeled topic modeling approach. In order to optimize the topic inference process through the LLDA model, we apply a custom top common keywords/labeled topics filtering mechanism in which labeled topics are selected based on the occurrence frequency in the given paper set. Then, the extracted topic distributions are utilized to construct the paper–topic graph, which are later applied to facilitate the irrelevant citation classification problem that is mainly focused on this paper. Moreover, the constructed paper–topic network can also further support other integrated topic-based/relation-based similarity search and recommendation.

3.1. *Topic-based scientific paper representation learning*

Topic modeling approach for textual representation learning. Within the natural language processing (NLP) domain, topic modeling is considered as an unsupervised latent topic discovery technique from text corpus. Within the topic modeling approach, Latent Dirichlet Allocation (LDA)[22] is considered as the original

and most well-known latent topic evaluation method, which enables to efficiently learn the probabilistic topic distributions from the given text corpus. In general, within the LDA algorithm, the latent topics are formally defined as a probabilistic distribution over keywords/phases, which occurred inside scientific papers. And, each paper is assumed as a mixture of latent topics, which are distributed with different proportions. In the view of NLP and data representation learning approach, the LDA model supports to generate interpretable fixed (K)-dimensional representations of the scientific papers. Here, (K) represents the number of predefined latent topics at the beginning. Then, these topic-based representations of papers are later utilized to facilitate multiple primitive text mining problems, such as similarity search, clustering, classification, etc. In more specifics, the LDA model is mainly designed upon the assumptions of topic–word/topic–document probabilistic distributions, with a predefined (K) number of latent topics, the set of all latent topics, which are discovered from a paper/document set $(\mathcal{D})$, is represented as (T) and $T = \{t_1, t_2, \ldots, t_K\}$). In addition, within the given paper set $(\mathcal{D}, d \in \mathcal{D})$, we have the associated vocabulary set $(\mathcal{W}, w \in \mathcal{W})$, assuming that the size of $(\mathcal{W})$ is large.

Following the topic–word/topic–document probabilistic assumptions, we have two independent distributions. $P(w|z) = \phi^{(t)}$ is the probabilistic distribution of a specific word (w) over the given (z^{th}) latent topic. $P(z) = \theta^{(d)}$ is the probabilistic distribution of (z^{th}) latent topic in a specific paper (d). Then, in order to achieve the distribution of a specific (i^{th}) keyword/term in any specific scientific paper for all (K) number of latent topics, the distribution is drawn as the following (as shown in Eq. (1)):

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i = j)P(z_i = j), \tag{1}$$

$$P(z_i = j|z_{\neg i}, w_i, d_i, \cdot) \propto \frac{C^{\mathcal{W}T}_{(w_i)j} + \beta}{\sum_{w=1}^{\mathcal{W}} C^{\mathcal{W}T}_{(w_{\neg i})j} + \mathcal{W}.\beta} \cdot \frac{C^{\mathcal{D}T}_{(d_i)j} + \alpha}{\sum_{t=1}^{K} C^{\mathcal{D}T}_{(d_i)t} + K.\alpha}. \tag{2}$$

In general, the latent topic inference processes in LDA model are controlled by two hyper-parameters (α) and (β). Specifically, the word distributions over generated (j^{th}) latent topic, denoted as $\phi^{(t_j)}$, are controlled by the (β) hyper-parameter and the latent topic distributions over the documents are controlled by the (α) parameter, respectively (as illustrated in Eq. (2)). Following this equation, for each scientific paper in the collection $(\mathcal{D})$, the distributional probability of a latent topic (z), which is assigned for a specific scientific paper (d_i), mainly relies on the set of occurred words in each paper. Such that, we can obtain an approximate maximum-likelihood (AML) for the LDA model's learnable parameters/distributions, including (ϕ) and (θ). There are several proposed sampling approaches, which have been utilized to obtain the AML of the given topic modeling techniques, like Gibbs sampling,[23] Expectation-Propagation,[24] etc.

The application of labeled LDA (LLDA) model[20] for topic taxonomy-based scientific paper representation learning is considered. Also designed upon the paradigm of latent topic distribution inference and extraction of LDA model, the LLDA is

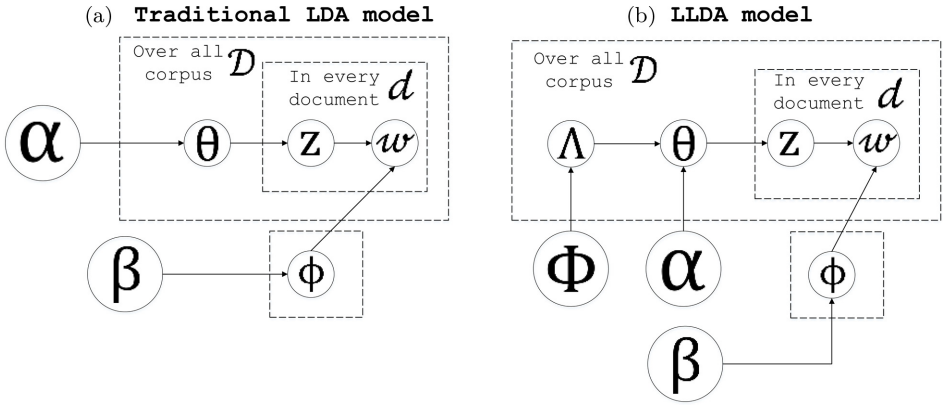


Fig. 2. The comparisons between the generative processes of the traditional LDA and LLDA models.

considered as a semi-supervised topic modeling approach in which it allowed to predefine the set of known keywords/topics into the latent topic inference processes in order to let the model generate more meaningful labeled topic distributions across the scientific paper set. In other words, within the LLDA model, the latent topic generation process is controlled not only by the set of occurred keywords inside but also the predefined labeled topics, which are associated with each paper. It means that the set of generated latent topic, denoted as (T), extracted from the topic model is limited to the set of predefined labeled topics, denoted as $\Lambda^{(d)} = (l_1, l_2, \ldots, l_K)$. By incorporating with the set of predefined labeled topics of each scientific paper, different from the traditional LDA model in which latent topics are ambiguous and represented as a set of keywords, the generated topic distributions can be controlled within a set of topic taxonomy as well as more specified for a knowledge domain scope, like as computer science, engineering, etc. Figure 2 illustrates the main differences in the topic generative process between the traditional LDA and the labeled LDA models.

Main keyword-based topic frequency filtering. By utilizing the LLDA topic model, we can achieve topic taxonomy-based representations of all scientific papers within the given bibliographic network; however, with a large number of main keywords, which are defined in the papers' metadata, we might be unable to apply them all in the topic inference process, which might cause much pressures on our system due to the high computational efforts required, in case the size of vocabulary set $(\mathcal{W})$ is normally quite or very large in real-world bibliographic networks. Therefore, in this paper, we define a custom labeled topic filtering mechanism to control the process of selecting labeled topics for each scientific paper, so we can efficiently minimize the size of labeled topics, which are needed to be evaluated during the inference process. Moreover, the common topic filtering mechanism also supports to produce better latent topic representations for all papers in which labeled topics are significant and popular ones within a specific research domain. In order to

do this, we apply the common topic frequency filtering mechanism with each defined main keyword-based topic (t) in each paper has been assigned a frequency score across the paper set ($\mathcal{D}$), denoted as topfreq $(t, \mathcal{D})$. The topic frequency score is identified as the following (as shown in Eq. (3)):

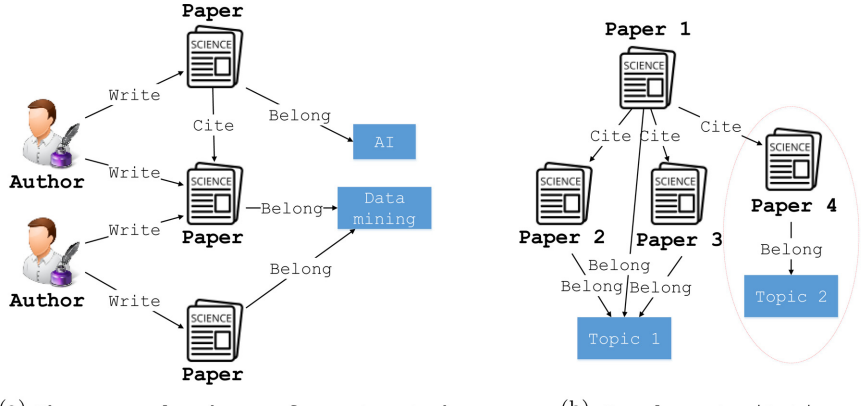
$$\text{topfreq}(t, \mathcal{D}) = \frac{|\{t \in d : d \in \mathcal{D}\}|}{|\mathcal{D}|}. \tag{3}$$

By identifying the frequency score of all the topics, we can choose the top common topics from the paper set ($\mathcal{D}$). In fact, the top-highest frequent topics are considered as important/hot research fields (they appeared in most of the papers), which are popular across the research communities. After selecting the top common labeled topics, we apply the LLDA model to achieve the probabilistic distributions of these topics upon the given paper set. These extracted latent topic distributions are then utilized to construct the paper–topic bibliographic network, which is later described in the next section.

3.2. *Paper–topic bibliographic network construction*

From a given bibliographic network such as AMiner/DBLP, besides paper–author, paper–paper citation relationships, we also add the extra paper–topic relationships. The topic entities are achieved from the set of common labeled topics in the paper set ($\mathcal{D}$), which have been described in the previous section. To establish the relationships between papers and common topics, in each paper, we choose the labeled topic with the highest probabilistic distribution portion as the main topic for that paper. The general schema of our paper–topic bibliographic graph is illustrated in Fig. 3(a).

Irrelevant citation identification problem. Upon the constructed paper–topic bibliographic graph, we utilize it to deal with the irrelevant citation classification problem in which less/non-relevant cited papers are categorized by identifying



(a) The general schema of constructed paper-topic bibliographic graph in this paper

(b) Irrelevant citation classification with constructed paper-topic bibliographic graph

Fig. 3. Illustrations of paper–topic bibliographic network construction and irrelevant citation identification.

their belonged topics. In our research assumption, meaningful cited paper set and the target evaluated papers should belong to the same research topics. Therefore, if the cited references belong to different topics compared with the target evaluated ones, they should be considered as less/irrelevant. This process can be generally illustrated as shown in Fig. 3(b). With the example shown in Fig. 3(b), "paper 4" is considered as less/irrelevant in which it belongs to different topic "topic 2" compared with other cited resources ("paper 2" and "paper 3") and target evaluated paper, the "paper 1", which all belong to the given "topic 1".

3.3. *Identifying the number of latent topic distribution*

For most of the probabilistic topic distribution models, such as the traditional LDA model, choosing the appropriate number of latent topics (K) for different types and sizes of textual datasets is considered as challenging due to the influences of this parameter upon the overall time-consuming and accuracy performance of the given model. Specifically, for a large number of latent topics, it will cause high pressures on the computational efforts for the latent topic inference process. With an insufficient latent topic quantity, the quality of learnt topic distributions from the given text corpus might be reduced accordingly. Within our approach, the labeled topic modeling is applied in which the chosen (K) parameter is identified as the number of labeled keyword-based topic. These set of labeled keyword-based topics are identified as the top-frequent topics (as illustrated in Sec. 3.1—Eq. (3)), which have been declared in the original papers. By doing this, we can sufficiently extract set of keywords/topics, which have been commonly focused by researchers in their published papers.

4. Experiments and Discussions

To demonstrate the effectiveness of our approach within the irrelevant citation classification task, we conducted the extensive experiments in benchmark AMiner/DBLP dataset. Experimental results demonstrated the effectiveness of our proposed ideas within this problem.

4.1. *Dataset description and pre-processing steps*

Within the context of integrated content and relationship-based bibliographic data analysis for irrelevant citation identification problem, we constructed a paper–topic bibliographic graph with different types of relationships between main entities: author, paper and topic. To achieve the information of authors/researchers, papers and their corresponding relations, we utilized the existing data resources from the AMiner data repository. The AMiner data repository also provides the abstract contents of papers as well as set of defined main keywords of each paper, which are utilized as the labeled topic taxonomy in our approach for extracting latent topic distributions. From the set of main keyword-based topics of all papers in the AMiner

Table 1. General statistics on the AMiner/DBLP dataset, which is used for experiments in this paper.

Parameter	Value
Number of authors	1,139,875
Number of papers	756,281
Number of labeled topic in taxonomy	100
Number of citation relationships	1,224,921
Number of paper–topic relationships	756,763
Number of author–paper relationships	2,476,144

dataset,[5] we apply the topic frequency calculation (as described in Sec. 3.1) and select top 100 topics with highest frequency scores as the main labeled topic taxonomy, which are later used for applying in the topic inference process through the LLDA model and paper–topic bibliographic graph constructed (as described in Sec. 3.2). Table 1 illustrates the general statistics of AMiner dataset as well as how we applied this dataset for constructing the paper–topic bibliographic graph for our experiments in this paper. Figures 4 and 5 show the general statistics of the constructed topic-based scientific graph, which is used in our experiments.

Environmental setups. For the setups of our experimental environments, we implemented our TopCite model mainly in Python programming language. The model is set up and run on a single computer with the following configurations: CPU: Intel(R) Core(TM) i5-7500 CPU @ 3.40 GHz, 8 GB Memory, OS: Windows 10.

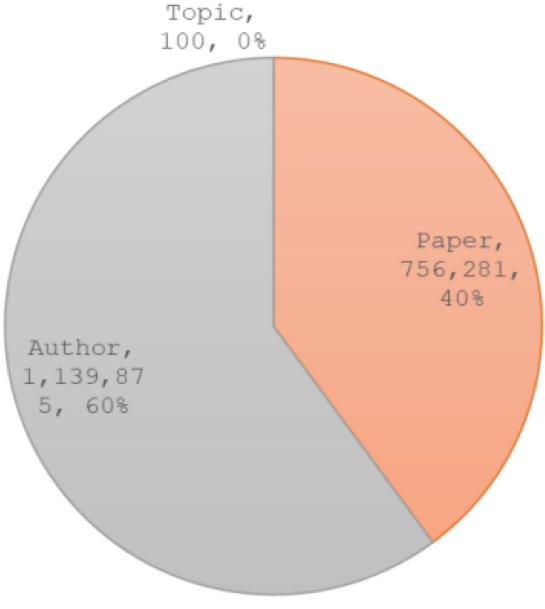


Fig. 4. Statistics on portions/number of main entities in our constructed paper–topic bibliographic graph.

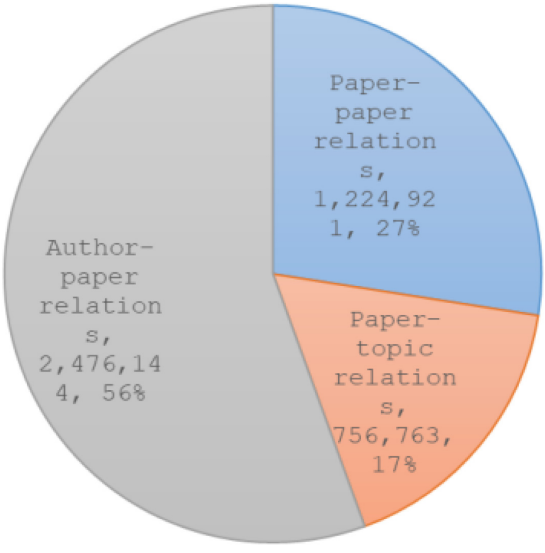


Fig. 5. Statistics on relationship data between entities in paper–topic bibliographic graph.

Table 2. The general configurations for the LLDA model, which is used for latent topic extraction in our approach.

Model's configuration	Value
Number of latent topic (K)	100
Hyper-parameter (α)	0.01
Hyper-parameter (β)	0.002
Number of training iterations	10

To store all data as well as associated configurations of the constructed paper–topic bibliographic graph, we mainly used the Neo4J Community Edition 4.3.12. For the implementation of LLDA model, which is utilized for the latent topic distribution extraction process (as described in Sec. 3.1), we mainly utilized the Python-versioned implementations of Ramage *et al.*[20] at this GitHub repository.[6] Table 2 shows general configurations of the LLDA model, which is implemented in our approach for all experiments in this paper.

Evaluation metric usage. In this paper, we mainly focus on the problem of identifying the irrelevant citations from papers, which majorly depended on the belonged topic relationships between target evaluated papers and their citing references. In the form of a primitive classification problem, for evaluating the experimental results in our experiments, we mainly utilized the F-measure (F-1) accuracy metric. Within the F-measure approach, the F-1-based metric is the most frequent one, which is used to compute the harmonic average precision and recall

values. Specifically, within our approach, the F-1 score is computed for each paper as the following (as shown in Eq. (4)):

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F\text{-}1 = 2.\frac{P.R}{P + R}. \tag{4}$$

In this equation, the TP (true-positive) presents for the number of correct references in a paper, which are classified to their ground-truth (relevant/irrelevant citation labels). Similar to that, the FP (false-positive) and FN (false-negative) are the number of expected references, which are categorized to specific citation labels but not correct and the number of references, which are not classified by their actually ground-truth citation labels, respectively.

4.2. *Experimental results and discussions*

In this section, we conducted extensive experiments to evaluate how our proposed TopCite technique can support to classify irrelevant citations from papers, which can further assist to leverage the accuracy performance of integrated content-based and citation relation-based retrieval and recommendation problems. In order to assess the accuracy performance of our model, we randomly constructed a set of papers from given paper–topic bibliographic graph with their associated reference list. From the selected paper set, we manually evaluated and labeled references as relevant or irrelevant classes. These labeled references of the selected paper set are considered as the ground-truth labels for evaluating against the predicted classes of references, which are generated by our proposed TopCite model. The prediction outputs are then assessed against the ground-truth labels to identify the accuracy scores in forms of F-1 evaluation metric. We apply the top@k (where k is the number of papers that will be evaluated each time) assessment strategy to test the effectiveness of our proposed model for dealing with the irrelevant citation identification task. Table 3 illustrates the general processes of categorizing irrelevant citations from a target paper and accuracy performance evaluation against the ground-truth labels.

As shown from the experimental outputs in Fig. 6, our proposed technique achieves the average accuracy performance about 73.7% in terms of F-1 accuracy metric for all top@k evaluation strategies. Specifically, with the top@10 evaluation strategy, our proposed TopCite technique achieved the highest accuracy performance, with F-1-based score approximately 79.02%. For the other top@k strategies, we achieved light fluctuations in the accuracy performance of our proposed techniques about 76.37%, 68.71% and 70.69% in terms of F1 evaluation metric for the values of $k = 20$, 50, and 100, respectively. With the above 70% average accuracy performance for all top@k assessment strategy, it can be assumed that our technique is quite promising for handling irrelevant citation categorization task in which both the content/topic-based and citation relationship aspects are taken in consideration.

Table 3. The illustrations of how we apply the TopCite technique to identify irrelevant citations as well as evaluating the accuracy performance against the ground-truth labels.

Target paper: Mouradian, C., Naboulsi, D., Yangui, S., Glitho, R. H., Morrow, M. J., & Polakos, P. A. (2017). A comprehensive survey on fog computing: State-of-the-art and research challenges. IEEE communications surveys & tutorials, 20(1), 416-464.

Reference	Ground-truth label	Predicted label by TopCite model
Sarkar, S., & Misra, S. (2016). Theoretical modelling of fog computing: a green computing paradigm to support IoT applications. *Iet Networks, 5*(2), 23-29.	0	0
Maier, Martin, *et al.* "The tactile internet: vision, recent progress, and open challenges." *IEEE Communications Magazine* 54.5 (2016): 138-145.	0	0
Yangui, Sami, and Samir Tata. "An OCCI compliant model for PaaS resources description and provisioning." *The Computer Journal* 59.3 (2016): 308-324.	0	0
Bonomi, Flavio, *et al.* "Fog computing and its role in the internet of things." *Proceedings of the first edition of the MCC workshop on Mobile cloud computing.* 2012.	1	1
Yangui, S., & Tata, S. (2015). The SPD approach to deploy service-based applications in the cloud. *Concurrency and Computation: Practice and Experience*, 27(15), 3943-3960.	1	1
...	...	...

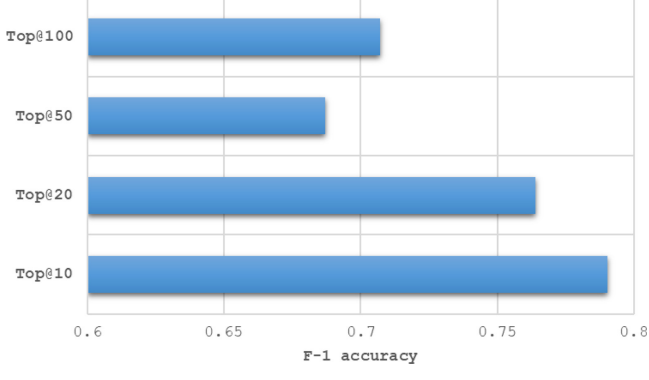


Fig. 6. Experimental outputs for the top@10, top@20, top@50 and top@100 based irrelevant citation classification task in terms of F-1 accuracy metric.

4.3. *Optimizations for paper–topic bibliographic graph storage and querying*

In addition, for the purpose of optimizing for the data storage and querying processes, we implemented and stored our paper–topic bibliographic graph in the Neo4J

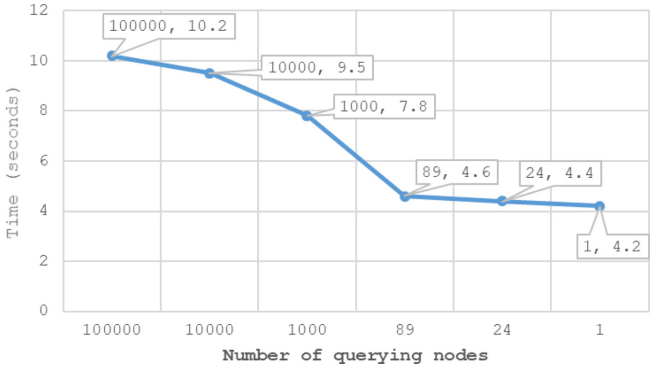


Fig. 7. Evaluating on the time-consuming performance of querying data of entities/nodes from the constructed paper–topic bibliographic graph, which are stored in Neo4J.

graph database. By using graph-structural organization and indexing mechanisms, we can achieve better time-consuming performance for the data querying process for further processing/analysis tasks, such as paper similarity measurement, irrelevant citation classification, co-authorship evaluation, etc. In order to assess the time-consuming performance of applying graph dataset in our approach, we evaluated it with two different types of data querying, including node data only based and node-relationship based queries and reported the operational times with different size of data volumes. Figures 7 and 8 show the time-consuming performances of different data querying types in our paper–topic bibliographic graph. As shown from the experimental outputs, both node-only and integrated node-relationship based queries are quite scalable with reasonable consuming times even with the rapid incremental increases of extracted data volumes. Thus, it proves the effectiveness of applying graph database in handling the data storage and querying of our constructed paper–topic bibliographic network for better time-consuming performances in data storing, indexing, querying and processing/mining tasks.

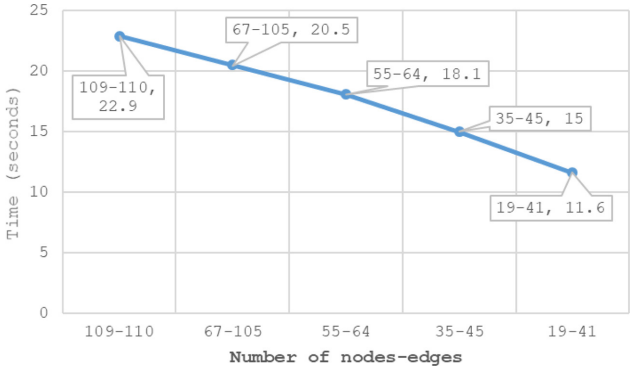


Fig. 8. Evaluating on the time-consuming performance of complex node-relationship extraction from the constructed paper–topic bibliographic graph.

5. Conclusions and Future Works

To sum up, in this paper, we propose a novel approach of citation relationship graph-based labeled topic modelling for extracting topic-varied distributions from the scientific papers. The learnt topic distributions from papers can be used for handling multiple tasks related to paper retrieval as well as recommendation tasks. Within this paper, we combine the extracted labeled topic distributional features with the citation relationships to identify irrelevant citations from scientific papers, which can directly assist to improve the accuracy performance for other paper similarity measurement related tasks. The integration between content/topic-based and citation relationship analysis has provided a promising direction for further researches in the scientific/bibliographic network mining domain. To demonstrate the effectiveness of our approach, we conducted extensive experiments in real-world AMiner dataset. The experimental outputs show the necessary as well as the potentiality of our proposed ideas in this paper in the context of filtering less/irrelevant citation problem.

Limitation and further research analysis on number of latent topic selection. As filtering top-100 frequent keyword-based topics from the given dataset, in our approach, the number of latent topics (K) are configured as 100. Even the top-common frequent topics within the given text corpus can be focused in this case, it still suffered several challenges, which are related to the ignorance of less-popular topics that existed in the given text corpus. Therefore, in our future works, we intend to conduct further analysis on how to identify the appropriated top-frequency threshold for topics, which are covered in the given dataset. Moreover, similar topics, which are covered within common keyword sets, such as LDA, VAE, etc., are also taken into consideration during the processes of choosing common topics for the model inference and representation learning processes. To overcome this challenge, we intend to extend our works on applying advanced neural topic generative architectures,[24,25] which can be a substitute for the current labeled topic modelling approach. These deep learning-based topic generative models can support to simultaneously extract the distributions of latent topics over documents as well as learning the correlations between topics to properly infer similar distributions. By doing this, it can help to overcome the limitation of identifying the single top-distributed topic for each document, which is currently used in our approach.

In addition, in our future works, we also intend to integrate our proposed TopCite model with graph neural network (GNN)-based architectures to achieve better dynamic network-structured representations[26] of the constructed paper–topic bibliographic networks, which are later utilized to integrate with content/topic-based representations of papers to facilitate multiple downstream tasks in bibliographic network analysis domain, such as scientific paper similarity measurement, irrelevant citation filtering, paper/citation recommendation, etc.

Acknowledgment

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2019.323.

References

1. S. Ma, C. Zhang and X. Liu, A review of citation recommendation: From textual content to enriched context, *Scientometrics* **122**(3) (2020) 1445–1472.
2. Z. Ali, P. Kefalas, K. Muhammad, B. Ali and M. Imran, Deep learning in citation recommendation models survey, *Expert Syst. Appl.* **162** (2020) 113790.
3. J. D. Guerrero-Sosa, V. H. Menéndez-Domínguez, M. E. Castellanos-Bolaños and L. F. Curi-Quintal, Analysis of internal and external academic collaboration in an institution through graph theory, *Vietnam J. Comput. Sci.* **7**(4) (2020) 391–415.
4. Z. Ali, I. Ullah, A. Khan, A. Ullah Jan and K. Muhammad, An overview and evaluation of citation recommendation models, *Scientometrics* **126**(5) (2021) 4083–4119.
5. M. D. Ekstrand, P. Kannan, J. A. Stemper, J. T. Butler, J. A. Konstan and J. T. Riedl, Automatically building research reading lists, in *Proc. Fourth ACM Conf. Recommender Systems* (ACM, 2010), pp. 159–166.
6. B. Kazemi and A. Abhari, A comparative study on content-based paper-to-paper recommendation approaches in scientific literature, in *Proc. 20th Commun. Netw. Symp.* (ACM, 2017), pp. 1–10.
7. C. Bhagavatula, S. Feldman, R. Power and W. Ammar, Content-based citation recommendation, in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (ACL, 2018), pp. 238–251.
8. Y. K. Ng, Research paper recommendation based on content similarity, peer reviews, authority, and popularity, *2020 IEEE 32nd Int. Conf. Tools with Artificial Intelligence (ICTAI)*, 9–11 November 2020, Baltimore, MD.
9. H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele and F. Xia, Context-based collaborative filtering for citation recommendation, *IEEE Access* **3** (2015) 1695–1703.
10. M. V. Murali, T. G. Vishnu and N. Victor, A collaborative filtering based recommender system for suggesting new trends in any domain of research, *2019 5th Int. Conf. Advanced Computing & Communication Systems (ICACCS)*, 15–16 March 2019, Coimbatore, India.
11. N. Sakib, R. B. Ahmad and K. Haruna, A collaborative approach toward scientific paper recommendation using citation context, *IEEE Access* **8** (2020) 51246–51255.
12. Y. Liang, Q. Li and T. Qian, Finding relevant papers based on citation relations, in *Int. Conf. Web-age Information Management*, 2011, pp. 403–414.
13. W. Tanner, E. Akbas and M. Hasan, Paper recommendation based on citation relation, *2019 IEEE Int. Conf. Big Data (big data)*, 9–12 December 2019, Los Angeles, CA, USA.
14. L. Pan, X. Dai, S. Huang and J. Chen, Academic paper recommendation based on heterogeneous graph, *Chinese Computational Linguistics and Natural Language Processing based on Naturally Annotated Big Data*, 13–14 November 2015, Guangzhou, China, pp. 381–392.
15. M. Amami, R. Faiz, F. Stella and G. Pasi, A graph based approach to scientific paper recommendation, in *Proc. Int. Conf. Web Intelligence* (ACM, 2017), pp. 777–782.
16. Z. Ali, G. Qi, P. Kefalas, W. A. Abro and B. Ali, A graph-based taxonomy of citation recommendation models, *Artif. Intell. Rev.* **53**(7) (2020) 5217–5260.

17. R. Manrique and O. Marino, Knowledge Graph-based Weighting Strategies for a Scholarly Paper Recommendation Scenario, *KaRS@ RecSys*, 7 October 2018, Vancouver, Canada, pp. 1–4.
18. H. Tang, B. Liu and J. Qian, Content-based and knowledge graph-based paper recommendation: Exploring user preferences with the knowledge graphs for scientific paper recommendation, *Concur. Comput. Prac. Exp.* **33**(13) (2021) e6227.
19. T. K. T. Ho, Q. V. Bui and M. Bui, Information diffusion on complex networks: A novel approach based on topic modeling and pretopology theory, *Vietnam J. Comput. Sci.* **6**(3) (2019) 285–309.
20. D. Ramage, D. Hall, R. Nallapati and C. D. Manning, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, in *Proc. 2009 Conf. Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2009), pp. 248–256.
21. P. Fournier-Viger, J. C. W. Lin, B. Vo, T. T. Chi, J. J. Zhang and H. B. Le, A survey of itemset mining, *Interdiscip. Rev. Data Min. Knowl. Discov.* **7**(4) (2017) e1207.
22. D. M. Blei, A. Y. Ng and M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.* **3** (2003) 993–1022.
23. M. Steyvers and T. Griffiths, Probabilistic topic models, in *Handbook of Latent Semantic Analysis* (Psychology Press, 2007), pp. 424–440.
24. P. Gupta, Y. Chaudhary, T. Runkler and H. Schuetze, Neural topic modeling with continual lifelong learning, in *Int. Conf. Machine Learning* (ACM, 2020), pp. 3907–3917.
25. A. B. Dieng, F. J. Ruiz and D. M. Blei, Topic modeling in embedding spaces, *Trans. Assoc. Comput. Linguist.* **8** (2020) 439–453.
26. P. Pham, L. T. Nguyen, N. T. Nguyen, W. Pedrycz, U. Yun and B. Vo, ComGCN: Community-driven graph convolutional network for link prediction in dynamic networks, *IEEE Trans. Syst. Man Cybernet. Syst.* (2021).