Studies in Computational Intelligence 1227

Ford Lumban Gaol
Tokuro Matsuo
Takayuki Ito *Editors*

Advances in Smart Knowledge Computing

Towards Post Artificial Intelligence Era

Springer

Studies in Computational Intelligence

Volume 1227

The series "Studies in Computational Intelligence" (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI AG (Switzerland), zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Ford Lumban Gaol · Tokuro Matsuo · Takayuki Ito
Editors

Advances in Smart Knowledge Computing

Towards Post Artificial Intelligence Era

Springer

Editors
Ford Lumban Gaol
Department of Informatics Engineering
and Information System
BINUS University
Jakarta, Indonesia

Tokuro Matsuo
Advanced Institute of Industrial Technology
Shinagawa, Japan

Takayuki Ito
Department of Social Informatics
Kyoto University
Kyoto, Japan

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-032-01132-9 ISBN 978-3-032-01133-6 (eBook)
https://doi.org/10.1007/978-3-032-01133-6

Contents

Graph-Based Filtering Using Community Selection for Citation Recommendation

Agung Hadhiatma

Abstract Graph-based filtering (GF) methods are suitable for addressing paper/citation recommendation problems. However, these methods still produce biased rankings due to multi-topic, similar, and related topics. Moreover, a scholarly dataset is expanding into big data, causing interesting issues in citation recommendation, especially information overload. Hence, we introduce a topic community selection model to take text and graph analysis into account. The model consists of four stages: transforming datasets into citation networks, detecting communities, identifying topics in communities, and selecting topic communities. The proposed method can significantly limit the search space but still produces good recall accuracy. The experiment shows that the community selection relating to specific topical queries produces five selected sub-graphs (communities) constituting only 18% of the total dataset volume while having 97% recommended paper candidates of the ground truth test. In future research, paper recommendations using GF can be carried out sufficiently on the sub-graph rather than the whole graph.

Keywords Digital library · Citation recommendation · Community selection

1 Introduction

A digital library of scientific articles is vast, rich, unstructured, and complex information. It includes books, dissertations, slides, patents, and scientific articles in proceedings and journals. Examples of scientific article databases/datasets are Web of Science (WoS), Scopus, Google Scholar, IEEE Xplore, and ACM Digital Library. Searching, analyzing, and mining information on scientific article datasets has become a challenging research field because of the need to obtain relevant and fast information that provides bases and insight for research. Several studies on scientific article datasets comprise information retrieval, academic evaluation [3], summarization

A. Hadhiatma (✉)
Department of Informatics, Faculty of Science and Technology, Universitas Sanata Dharma, Yogyakarta, Indonesia
e-mail: agunghad@usd.ac.id

F. Lumban Gaol et al. (eds.), *Advances in Smart Knowledge Computing*, Studies in Computational Intelligence 1227, https://doi.org/10.1007/978-3-032-01133-6_14

[4], topic prediction [1], academic recommendation [5], and others. Some academic recommendations are a paper, reference, author, and venue recommendation.

Information retrieval of scientific articles usually uses keyword queries. The retrieved scientific articles are then still reviewed, selected, and selected. Searching for relevant scientific articles suitable to researchers takes a long time, is an uneasy process (Cai et al., 2018a), and can cause information overload. To overcome this, research with a new approach has emerged, namely a recommendation method approach.

Scientific article recommendation methods are not only using keywords in queries but also using a collection of manuscripts [3], user profiles [11], context [15], and others. Manuscript queries are a draft article or other information in a scientific article consisting of a title, author, abstract, text content, venue, year of publication, and references. Input for recommendations can also be in a graph represented by a paper, co-citation, bibliometric, and heterogeneous network. Research on scientific article recommendations is conducted on several approaches, namely Collaboration-Based Filtering (CF), Content-Based Filtering (CBF), Hybrid-Based Filtering (HF), Feature-Based Filtering (FF), and Graph-Based Filtering (GF).

Recent research on scientific article recommendations has started to utilize Graph-Based Filtering (GF) models [2]. The GF approach for academic recommendations works on a graph, by converting the dataset to an academic citation network [14], which comprises nodes and links. Nodes may represent scientific articles, authors, or venues, while links signify references that connect one scientific article to another. Scientific article nodes can possess attributes such as titles, keywords, abstracts, and document content. Citation relationships within the academic citation network provide valuable and potent information for ranking scientific article recommendations.

The GF method for recommending scientific articles has a weakness in that it relies solely on citation links, neglecting the actual content of the articles. Consequently, the methods still cause biased and irrelevant recommendations [13]. In addition, an academic citation network has characteristics such as volume, value, velocity, veracity, and variety. Recommending scientific article references in that characteristic, particularly in a growing vast citation network, is challenging because of resulting information overload and topic bias.

To address these challenges, we proposed a topic community selection model for paper recommendations. The model consists of four stages: transforming datasets into citation networks, detecting communities, identifying topics in communities, and selecting topic communities. Utilizing topic communities for article recommendations is anticipated to mitigate semantic issues. Searching for scientific article candidates via topic community selection also aims to narrow the search space, allowing the retrieval of the top k recommended papers to focus solely on the selected community, a sub-graph, rather than the entire graph. This approach will be beneficial when the model operates on large data volumes (Big Data).

2 Literature

A digital library of scientific articles is vast, rich, unstructured, and complex. Digital Library stores and collects published articles, author profiles, references, images, tables, and more. There are various analyses in digital libraries used for conducting research and establishing applications, such as scientific impact evaluation [3], expert findings [17], trend prediction [1], and academic recommendation [5]. These analyses are a statistical, social network, and content analysis described in Fig. 1. Research on academic recommendations includes paper/citation recommendations, author recommendations, and venue recommendations. Research on scientific article recommendations consists of several approaches: Collaboration-Based Filtering (CF), Content-Based Filtering (CBF), Hybrid-Based Filtering (HF), Feature-Based Filtering (FF), and Graph-Based Filtering (GF).

The GF method for recommending scientific articles relies exclusively on citation links, overlooking the content of the articles. Consequently, the recommendation outcomes can be less than satisfactory [13]. Prior research suggests some personalized PageRank-based rankings within the Academic Citation Network, but these still result in inaccurate problems due to topic factors [6]. Topic bias arises because the ranking considers only the global network of textual information without focusing on a particular topic. A topic can be defined as a group of semantically related words, as these words frequently appear together in context. Various methods for topic identification include Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

Several researchers have developed the PageRank method to discuss topic bias. Jardine and Teufel [7] sought to enhance the functionality of PageRank with the Topical PageRank Model (TPM), where ranking work based on specific topic boundaries and particular publication time. Zhang [16] introduced a PageRank model called Collective Topical PageRank (CTPM) to overcome the limitations of TPM, as the CTPM method accommodates correlations between topics and the quality

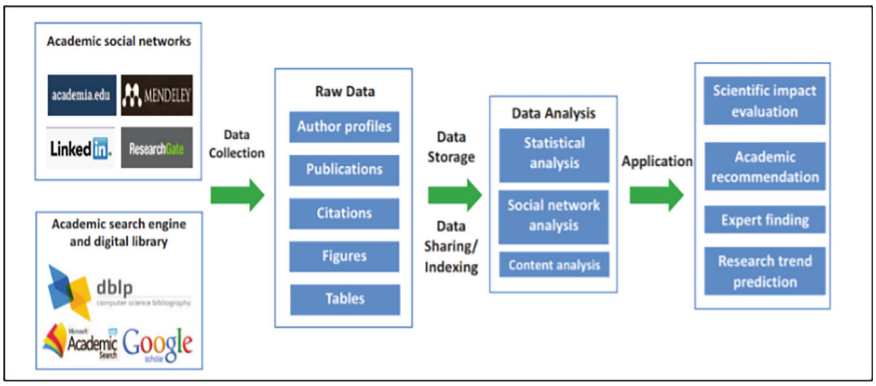

Fig. 1 Data analysis on academic citation network. *Source* [14]

of venues (journals or conferences) relevant to specific topics. Moreover, several other researchers proposed a query-dependent PageRank model to identify significant papers in a graph by factoring in text and topic similarity. Roul and Sahoo [10] developed a query-optimized PageRank approach by integrating TF-IDF and Personalized PageRank methods to yield more stable web rankings.

Moreover, a complex Big Scholarly Dataset is growing into very vast data, namely big data. This dataset raises gaps and fascinating challenges in citation recommendation research. Hence, the research requires new approaches for analyzing data sources in big data, such as information mining, feature extraction [8], and topical analysis [16]. The approach should simultaneously consider some aspects of text and graph structure analysis.

Unlike the previous studies, we introduce a paper recommendation method that considers text and graph analysis by applying the topic community selection approach. The selection of topic communities aims to (1) obtain scientific articles that align with a similar topic and related topics to queries and (2) minimize search space.

3 Identification and Selection of Topic Communities

We propose a community selection method that consists of four stages: transforming datasets into citation networks, detecting communities, identifying topics in communities, and selecting topic communities (Fig. 2). The chosen topic communities are recommended paper candidates, which can be utilized for subsequent research, for instance, for ranking the candidate papers.

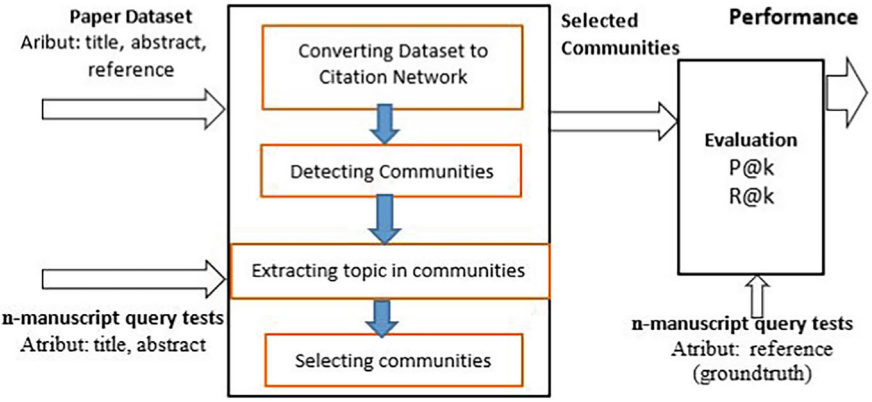

Fig. 2 The model of community selection

3.1 Detecting Communities in an Academic Citation Network

Scientific article community detection (paper community detection) is a clustering process in the academic citation network into several communities (subgraphs). Community identification in the Academic Citation Network is to look for partitions in a graph where the link density in the community is more than the link density between communities by maximizing the modularity function. Modularity optimization is implemented using the Louvain algorithm and the modularity formula is defined as Eq. 1.

$$Q(C) = \frac{1}{2m} \sum_{i \in V} \sum_{j \in V} (a_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j) \tag{1}$$

where A_{ij} is the edge weight from vertex i to vertex j, m is the number of edges on the graph, k_i is the number of degrees from vertex i, k_j is the number of degrees from vertex j. $\delta(C_i, C_j)$ is the Kronecker function which will have a value of 1 if vertex i and vertex j are in the same community, otherwise it will have a value of 0.

3.2 Identifying Topics in Communities

The next step is to identify topics in the scientific article community using the LDA topic model. This identification results in a distribution of topics in communities and papers. The generative Topic Modeling LDA method (reference) identifies a set of topic proportion features for each scientific article defined as θ_{P} and identifies a set of word term proportions on each topic defined as φ_T, where T is the set of Topics. The input to the LDA algorithm is the text feature F_P. The text feature F_P is obtained from a pre-processing step on the set of texts (title + abstract) in set P. In this study, estimates of the parameters of the LDA model θ_{P} and φ_T as latent feature vectors were obtained using Gibbs Sampling.

The process of identifying topics in the community is as follows. $\mathrm{P_c}$ is the set of communities, $C = \{ C|1, 2, \ldots, k\}$, k is the total number of communities. The c community with m papers is defined as P_{cm} $\{p_{c1}, p_{c2}, \ldots, p_{cm}\}$. A set θ_{P} is aggregated to communities C represented as θ_{cm}. All paper texts in each P_i combined form an aggregated paper text defined as $AP_i = \bigcup_{j=1}^{m} p_{ij}$. Paper text is taken from the title and abstract variables. The set of all communities containing an aggregated set of paper texts can be written $TP_C = \{ AP_i|i = 1, 2, \ldots, k\}$. The next process is inferencing topics on TP_C using the generated LDA topic model. The result of topic inference is a set of community-topic distributions defined as θ_C. Figure 3 illustrates community features as a result of identifying topic communities.

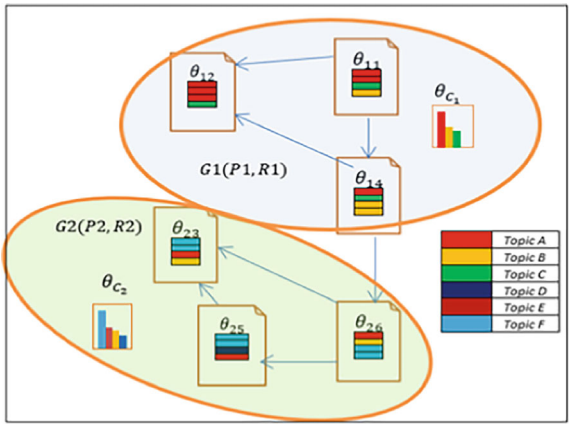

Fig. 3 Topic communities

3.3 Selecting Topic Communities

The multi-topic community selection process compares the features of multi-topic paper queries with multi-topic communities. The result consists of some relevant multi-topic communities. The selected communities combined to form a merged graph are assumed to contain the most pertinent recommended papers in a citation network. The sequence of selecting communities is as follows:

(a) **Sorting topics in community-topic distributions**

1. The sum of the texts in a set of query texts Q represented as $\bigcup_{i=1}^{n} q_i$.
2. Inferencing topics on $\bigcup_{i=1}^{n} q_i$ by the LDA topic model for forming query topic distributions $\overrightarrow{tq}$.
3. Ranking topic query distributions $\overrightarrow{tq}$ to find the dominant topic query, saved as a vector $\vec{e}$.
4. Arranging topics in each community-topic distribution θ_C based on the dominant topic query $\overrightarrow{e}$ to form a set of n dominant topics in each community $c_i \in C$, which the results are stored in set $Y_C = \{\vec{y}_{c_1}, \vec{y}_{c_2}, \ldots, \vec{y}_{c_k}\}$ where k is the total community.

(b) **Filtering multi-topic communities**

5. Looping steps 6 and 7 for each $\vec{y}_{c_i \in C}(\vec{y}_{c_i} \in Y_C)$, i from 1 to k total community).
6. Finding the most dominant topic (rank 1) community in each $\vec{y}_{c_i \in C}$, saved as $dt_i \in DT$
7. Selecting communities in $c_i \in$ C if the dominant topic in the communities ($dt_i \in DT$) is the same as the dominant topic query of the 1st ranks $\vec{e}[0]$ or the 2nd ranks $\vec{e}[1]$ or the 3rd ranks $\vec{e}[2]$ of a topic query. The chosen communities are stored in the set Y_S.

(c) **Selecting the j-best topic communities**

8. Measuring the similarity of n-dominant topic communities $\vec{y}_{c_i \in S}(\vec{y}_{c_i} \in Y_S)$ and n-dominant topic queries $\vec{e}$ with Jenson Shannon Divergence (JSD) sim $(\vec{e}, \vec{y}_{c_i \in S})$, for each $\vec{y}_{c_i \in S}$, with the aim of finding the j-best topic communities using Eq. 2 and Eq. 3. The search results are stored in a set X. The optimum j value is sought from experiments.
9. Merging a number of the best j topic communities in the set X to one community namely *merged graph* $G_X(H, \mathrm{R}_H)$, H = a set of recommended article candidates.

$$JSD\left(\vec{\mathrm{e}}, \vec{y}_{c_i \in S}\right) = \frac{1}{2} KLD(\vec{y}||\vec{u}) + \frac{1}{2} KLD(\vec{e}||\vec{u}) \tag{2}$$

$$\vec{u} = \frac{1}{2}(\vec{y} + \vec{e}),\ KLD(\vec{y}||\vec{u}) = \sum_{i=1}^{n} y_i \log \frac{y_i}{u_i} \text{ and } KLD(\vec{e}\ ||\ \vec{u}\ ) = \sum_{i=1}^{n} e_i \log \frac{e_i}{u_i} \tag{3}$$

4 Experiment and Evaluation

4.1 Dataset

The dataset obtained was first developed based on an article from [12]. The dataset taken is the (ACM)-Citation-Network V4 version (DBLP, https://www.aminer.cn/citation) containing collected papers in the field of computer science (1,511,035 scientific articles and 2,084,019 citations). The information structure of scientific articles consists of title, author, year of publication, venues, ID number, references and abstract, and data format structure. This research utilized only a subset of the dataset comprising papers from five topics: Information Retrieval, Machine Learning, Network and Communication, Computer Vision and Computer Security in some venues (proceedings and journals of 46,870 papers).

4.2 Results of Community Features

The model of identifying topic communities produces several features, such as a set of communities C, a set of community-topic distributions θ_{C}, a set of paper-topic distributions θ_{cm}, and a set of topic-word distributions φ_T. A set of communities C is shown in Table 1, showing a community, total paper, and papers in a community. A community and a paper are named by the ID number.

Each community has a topic distribution. The community-topic distributions (Table 2) represent a variety of research topics along with their proportional weights in each community.

Table 1 A sample of communities C

Community ID	The total number of papers in a community	Paper ID as a member of a community
'1'	821	'9187', '984349', '175167', '982402', '388932', …
'4'	4672	'12476', '600449', '613 083', '613113', …
'7'	1504	'11418', '818137', '499240', '286290', '645835', …
'9'	4580	'17965', '1025089', '1032775', '1031167', …

Table 2 A sample of community-topic distributions C

Community ID	Community-topic distributions					
'1'	Topic 4 0.661032	Topic 19 0.094678	Topic 11 0.091	Topic 12 0.0608	Topic 20 0.036	Topic 1 0.0187
'4'	Topic 19 0.47063	Topic 4 0.2636	Topic 29 0.1161	Topic 11 0.0359	Topic 12 0.0315	Topic 1 0.0121
'7'	Topic 20 0.75439	Topic 19 0.08559	Topic 1 0.01659	Topic 16 0.01647	Topic 12 0.01487	Topic 2 0.01092
'10'	Topic 2 0.54006	Topic 12 0.25077	Topic 19 0.04654	Topic 4 0.0424	Topic 11 0.03892	Topic 1 0.0266

Each community has a set of paper-topic distributions θ_{cm} shown in Table 3.

Each topic consists of some keywords in the same knowledge domain. Each keyword has a weight value to form topic-word distributions φ_T, as shown in Table 4. We interpret that some extracted keywords for Topic 14 are "information retrieval" topic.

Table 3 A sample of paper-topic distributions θ_{cm}

Paper ID	Community ID	Paper-topic distributions					
'473813'	'16'	Topic 12 0.64683	Topic 11 0.12451	Topic 7 0.10107	Topic 22 0.03162	Topic 15 0.03009	Topic 3 0.01798
'20308'	'16'	Topic 12 0.9239	Topic 1 0.04642	–	–	–	–
'472388'	'16'	Topic 12 0.60753	Topic 11 0.19773	Topic 7 0.10645	Topic 10 0.01710	Topic 18 0.06085	-
'107746'	'14'	Topic 11 0.91825	Topic 15 0.06796	–	–	–	–

Table 4 A sample of topic-word distributions φ_T

Topic 14						
query (0.097248)	document (0.083266)	web (0.068311)	search (0.067179)	retrieval (0.066948)	datum (0.065034)	information (0.058248)
user (0.053995)	model (0.051803)	text (0.047448)	method (0.046875)	database (0.043042)	classification (0.041961)	system (0.040772)
algorithm (0.040434)	language (0.037415)	semantic (0.036918)	learning (0.035647)	approach (0.035618)	feature (0.034746)	

4.3 Evaluations of Topic Community Selection

The proposed model addresses selecting the communities with their features (C, θ_C, θ_{cm}, and φ_T), that are relevant to user query topics. The community selection results are recommended paper candidates that can be utilized for subsequent studies, particularly for a recommendation problem using a graph-based filtering approach. Consequently, analyzing, processing, and ranking recommended paper candidates do not consider the entire large graph but will focus solely on selected relevant communities. Processing data on selected communities rather than on the whole graph aims to reduce computational complexity and enhance the performance of paper recommendations in future research.

The stages of conducting community selection are how to (1) determine the optimal number of the *j*-communities (neither too many nor too few) and (2) rank the *j*-selected communities that are most relevant to the query. Finding the j-communities is done by filtering topic communities with a recall value exceeding 90%. When the topic communities share the same recall values, we choose the highest precision value. The recall and precision calculation are at Eq. (4) and subsequently (5). The variable n(j) represents the number of relevant references that match the ground truth test in j communities, and the variable g denotes the number of article references used as a ground truth test.

$$\text{Recall} = \text{n(j)}/\text{g} \tag{4}$$

$$\text{Precision} = \text{n(j)}/\text{p(j)} \tag{5}$$

Meanwhile, ranking the filtered j communities relevant to a query based on multi-topic similarity employs the Eqs. (2) and (3).

We conducted experiments on the DBLB dataset and query tests on information retrieval. The IR topics taken from the preceding venues are ACL, ECIR, SIGIR, COLING, and NAACL. The following outlines the process for determining the number and ranking of selected communities tested in the IR field. A set of 50 test queries in this domain has 341 article references serving as the ground truth test. Table 5 presents the experimental result for determining the number of selected communities using Recall and Precision metrics related to the IR topic query. The

number of communities that achieve a satisfactory percentage value (a minimum of 90%) is five, with the recall value at 97% (333/341) and the precision value at 0.026. With j determined to be five, the next step is identifying the five best community ranks relevant to the test query.

Table 6 shows the experimental result for rankings of the five selected communities measured by topic similarity between the topic community and topic IR query. The similarity level is high if the value is close to 0. The selection result is ranks of the chosen communities that best match the query topic, ordered community ID '14', '49', '48', '15', and '13'.

The community ID '14' includes 4260 articles, of which 252 are relevant recommended references that match the ground truth test. This precision value is (252/341) 73% of the total references considered ground truth within the community ID '14'. Five selected communities combine into a new graph (a merged graph) generally dealing with the topic of Information Retrieval (IR) because the dominant topic in five selected communities is Topic 14 (Table 7). Topic 14 denoted as the Information Retrieval field shown in Table 4. This merged graph comprises 12,663 articles for recommendation candidates related to the IR-topic query (4260 + 1392 + 1516

Table 5 Experimental results of determining the number of selected community

The number of selected communities j	The number of papers in j communities P(j)	The number of relevant citations in j communities N(j)	Recall	Precision
j = 6	13,302	333	0.97	0.025
j = 5	*12,663*	*333*	*0.97*	*0.026*
j = 4	11,113	283	0.83	0.025
j = 3	10,174	283	0.83	0.027
j = 2	9089	257	0.81	0.028
j = 1	4395	252	0.73	0.057

Table 6 Experimental results of ranking the five selected communities

Rank	Community ID	Similarity value
1	*'14'*	*0.004206*
2	*'49'*	*0.053871*
3	*'48'*	*0.077497*
4	*'15'*	*0.202284*
5	*'13'*	*0.235112*
6	'22'	0.290117
7	'78'	0.314114
8	'18'	0.385652
9	'1'	0.408453
10	0	0.451868

+ 5409 + 1983). This merged graph of 12,663 articles narrows the search space compared to 67,850 scientific articles in the dataset. As a result, the search space for further processing of paper recommendations now represents only (12,663/67850) 18% of the total dataset volume. Meanwhile, the overall precision accuracy across the five selected communities stands at 97%.

The following experiment is a paper recommendation test using the proposed community-based model and non-community-based methods. Paper recommendation testing is carried out sequentially using the CBF (Content-Based Filtering) approach, GF (Graph-Based Filtering) Personalized PageRank (GF-PPR), and GF Personalized PageRank based on community selection (PPR_TC) references. The CBF method is represented by the Do2Vect, Cosine TF-IDF, and Bert methods, while the GF-PPR approach is represented by the PPR Restart, Query-PPR, and Edge Weight-PPR methods. The PPR_TC method was tested on the Information Retrieval (IR) topic community. The test results appear in Table 8.

The average performance of various recommendation methods, evaluated using Recall@n, is ranked as follows: PPR_TC > GF-PPR > CBF. The CBF method relies solely on text content, while PPR incorporates both text and graph content. In contrast, PPR_TC combines text and graph content with community topic selection. The experimental results indicate that paper recommendations utilizing community topic selection enhance the effectiveness of the Personalized PageRank (PPR) model [9].

Table 7 Five selected topic communities

Ranks	Community ID	Topic (weights)			The number of papers in a community	Relevant references in a community	Recall accuracy
1	*'14'*	*Topic 14*	*Topic 10*	*Topic 22*	*4260*	*252*	*(252/341) = 73%*
		0.577	*0.2819*	*0.0325*			
2	'48'	Topic 14	Topic 10	Topic 7	1392	5	(5/341) = 1.4%
		0.513	0.248	0.095			
3	'49'	Topic 14	Topic 7	Topic 20	1516	26	26/341 = 7.6%
		0.783	0.0422	0.042			
4	'15'	Topic 14	Topic 7	Topic 20	5409	49	(49/341) = 14%
		0.830	0.056	0.0261			
5	'13'	Topic 14	Topic 22	Topic 7	1983	1	(1/341) = 0.2%
		0.7098	0.0851	0.067			
Total					12,663	331	97%

Table 8 Comparison of recommendation methods based on community and non-community

Citation recommendation approach		Recall @25	Recall @50	Recall @75	Recall @100	MAP @100	MRR @50
CBF (content)-based filtering	Doc2Vect	0.051	0.077	0.096	0.15	0.026	0.049
	Cosine TFIDF	0.176	0.251	0.288	0.316	0.078	0.06
	BERT	0.151	0.326	0.406	0.462	0.028	0.031
GF-PPR	PPR restart	0.067	0.101	0.123	0.183	0.012	0.043
(Content + Graph)	Query-PPR	0.277	0.391	0.451	0.525	0.073	0.138
Based filtering	Edge weight PPR	0.241	0.367	0.405	0.494	0.095	0.138
PPR_TC (content + community graph)-based filtering	PPR_TC	0.279	0.379	0.459	0.513	0.1	0.167

5 Conclusion

The graph feature extraction method has produced a set of communities, community-topic distributions, paper-topic distributions, and topic-word distributions used to select relevant communities for paper recommendations. The community selection method can constrain the search space (forming a local graph represented in the chosen communities) but still results in good recall accuracy. The experiment shows that the community selection relating to IR topic queries produces the five selected sub-graphs (communities). These selected communities have only 18% of the total dataset volume while still having 97% of recommended paper candidates from the ground truth test. So, in future research, the recommendation process applying graph-based filtering can be conducted only on the sub-graph rather than the whole graph. The experimental results indicate that paper recommendations utilizing community topic selection enhance the effectiveness of the Personalized PageRank (PPR) model.

References

1. K. Asatani, J. Mori, M. Ochi, I. Sakata, *Detecting Trends in Academic Research from a Citation Network Using Network Representation Learning* (2018), pp. 1–13
2. X. Cai, J. Han, W. Li, R. Zhang, S. Pan, L. Yang, A Three-layered mutually reinforced model for personalized citation recommendation. EEE Trans. Neural Netw. Learn. Syst. **29**(12), 6026–6037 (2018). https://doi.org/10.1109/TNNLS.2018.2817245
3. X. Cai, Y. Zheng, L. Yang, T. Dai, L. Guo, Bibliographic network representation based personalized citation recommendation. IEEE Access **7**, 457–467 (2019). https://doi.org/10.1109/ACCESS.2018.2885507
4. A. Cohan, *Scientific Article Summarization Using Citation-Context and Article ' s Discourse Structure. September* (2015), pp. 390–400
5. T. Dai, L. Zhu, Y. Wang, H. Zhang, X. Cai, Y. Zheng, Joint model feature regression and topic learning for global citation recommendation. IEEE Access **7**, 1706–1720 (2019). https://doi.org/10.1109/ACCESS.2018.2884981

6. M. Dunaiski, J. Geldenhuys, W. Visser, On the interplay between normalisation, bias, and performance of paper impact metrics. J. Informet. **13**(1), 270–290 (2019). https://doi.org/10.1016/j.joi.2019.01.003
7. J. Jardine, S. Teufel, Topical pagerank: a model of scientific expertise for bibliographic search, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (2014), pp. 501–510
8. C. Jeong, Jang, S., Shin, H., Park, E., & Choi, S. (2020). A Context-Aware Citation Recommendation Model with BERT and Graph Convolutional Networks. *Scientometrics*, *124*(3), 1907–1922. https://doi.org/10.48550/arXiv.1903.06464
9. D. Mu, L. Guo, X. Cai, F. Hao, Query-focused personalized citation recommendation with mutually reinforced ranking. IEEE Access **6**, 3107–3119 (2018). https://doi.org/10.1109/ACCESS.2017.2787179
10. R.K. Roul, J.K. Sahoo, A novel approach for ranking web documents based on query-optimized personalized pagerank. Int. J. Data Sci. Anal. **11**, 37–55 (2021). https://doi.org/10.1007/s41060-020-00232-2
11. K. Sugiyama, M. Kan, Towards higher relevance and serendipity in scholarly paper recommendation. Proceedings of the ACM International Conference on Digital Libraries (2015). https://doi.org/10.1145/2719943.2719947
12. J. Tang, J. Zhang, ArnetMiner : Extraction and mining of academic social networks, in *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), pp. 990–998. https://doi.org/10.31235/osf.io/me8zd
13. W. Waheed, M. Imran, B. Raza, A.K. Malik, H.A.L.I. Khattak, S. Member, A hybrid approach toward research paper recommendation using centrality measures and author ranking. IEEE Access **7**, 33145–33158 (2019). https://doi.org/10.1109/ACCESS.2019.2900520
14. F. Xia, W. Wang, T.M. Bekele, H. Liu, Big scholarly data: a survey. IEEE Trans. Big Data **3**(1), 18–35 (2017). https://doi.org/10.1109/TBDATA.2016.2641460
15. L. Yang, Y.U. Zheng, X. Cai, D. Mu, L. Guo, T.A.O. Dai, A LSTM based model for personalized context-aware citation recommendation. IEEE Access **6**, 59618–59627 (2018). https://doi.org/10.1109/ACCESS.2018.2872730
16. Y. Zhang, Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers. Scientometrics (2017). https://doi.org/10.1007/s11192-017-2626-1
17. F. Zhao, Y. Zhang, J. Lu, O. Shai, Measuring academic influence using heterogeneous author—citation networks. Scientometrics **118**, 1119–1140 (2019). https://doi.org/10.1007/s11192-019-03010-5