IEICE TRANS. INF. & SYST., VOL.E104–D, NO.8 AUGUST 2021

PAPER Special Section on Computational Intelligence and Big Data for Scientific and Technological Resources and Services

A Hybrid Approach for Paper Recommendation

Ying KANG[†a)], ***Student Member***, **Aiqin HOU**[†b)], **Zimin ZHAO**[†], ***and*** **Daguang GAN**[††], ***Nonmembers***

SUMMARY Paper recommendation has become an increasingly important yet challenging task due to the rapidly expanding volume and scope of publications in the broad research community. Due to the lack of user profiles in public digital libraries, most existing methods for paper recommendation are through paper similarity measurements based on citations or contents, and still suffer from various performance issues. In this paper, we construct a graphical form of citation relations to identify relevant papers and design a hybrid recommendation model that combines both citation- and content-based approaches to measure paper similarities. Considering that citations at different locations in one article are likely of different significance, we define a concept of citation similarity with varying weights according to the sections of citations. We evaluate the performance of our recommendation method using Spearman correlation on real publication data from public digital libraries such as CiteSeer and Wanfang. Extensive experimental results show that the proposed hybrid method exhibits better performance than state-of-the-art techniques, and achieves 40% higher recommendation accuracy in average in comparison with citation-based approaches.

key words: *paper recommendation, citation graph, hybrid model*

1. Introduction

Literature review is a critical, indispensable task in research conduct in many scientific fields where researchers need to search for relevant publications to understand the state of the arts and spark inspirations for further research. This task has become increasingly challenging due to the rapidly expanding volume and scope of publications in the large literature accumulated over many years. The lack of user profiles in commonly used digital libraries has exacerbated the problem.

Paper recommendation is mainly based on the information typically contained in an academic paper, including authors, abstract, keywords, citations, and so on. Citation-based recommendation, which is among the most widely used, exploits the citation information, for example, by employing collaborative filtering (CF) to establish a rating matrix with direct citation relations [1], [2]. However, CF-based methods may suffer from the problems of cold start and sparse data especially in the presence of big data [3]. Also, some studies utilize one's published work to constitute the latent interests of a researcher [4]. Citation analysis is another important application of citation relations, which are categorized into direct citation, bibliographic coupling, and co-citation [5]. Particularly, the last two relations indicate a high degree of logical correlation and have demonstrated promising performance in paper recommendation [6]. Citation network, which depicts citation relations, is also considered in some efforts [7].

In this paper, we construct a graphical form of citation relations that incorporates various types of the aforementioned citation relations. This citation graph is similar to citation network but with weighted edges and limited step lengths. Furthermore, considering that a single type of information presents only a partial aspect of a paper, we design a hybrid model to measure the similarity between papers by combining citation- and content-based approaches. Traditional citation-based methods typically use a binary label to indicate whether or not there is a citation relation between two papers. However, citations at different locations in one article are likely of different significance. According to the study in [8], the citations in the section of Methodology are generally more important than those in the section of Related Work. Therefore, we take the position of citations into consideration and define a concept of citation similarity with varying weights according to the sections of citations in the hybrid model. For performance evaluation, we test the proposed recommendation method on the dataset crawled from public digital libraries such as CiteSeer [9] and Wanfang [10], and compute Spearman correlation coefficient against the benchmark ranking generated by JensenShannon Divergence (JSD) [11]. We adopt Spearman correlation due to its capability of minimizing the impact of subjective assessment on numerical ratings. Extensive experimental results show that the proposed hybrid method exhibits superior performance and achieves better performance than state-of-the-art techniques.

The contributions of our work are summarized as follows:

- We construct a graphical form of citation relations between papers, which incorporates the relations of direct citation and indirect citation, to identify a collection of candidate relevant papers.
- We define a new concept of citation similarity with varying weights according to the location of citations, and design a hybrid model that combines citation- and content-based approaches to measure paper similarity.

Manuscript received November 11, 2020.
Manuscript revised January 13, 2021.
Manuscript publicized April 26, 2021.

[†]The authors are with the School of Information Science and Technology, Northwest University, Xi'an, Shaanxi 710127, China.
[††]The author is with Wanfang Data Co., Beijing, 100038, China.
a) E-mail: yingkang@stumail.nwu.edu.cn
b) E-mail: houaiqin@nwu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2020BDP0008

- The proposed hybrid method achieves higher recommendation accuracy in average than state-of-the-art methods on real datasets from public digital libraries in terms of Spearman correlation against the benchmark ranking generated by JSD.

The rest of this paper is organized as follows: Sect. 2 conducts a survey of related work for paper recommendation. Section 3 details the design of our hybrid method. Section 4 carries out comparative experiments for performance evaluation. Section 5 concludes our work and sketches a plan for future research.

2. Related Work

We conduct a brief survey of various techniques in four categories for paper recommendation in the literature.

2.1 Citation-Based Approaches

Citation is an important and valuable piece of information in an academic paper, and has been exploited by many researchers for paper recommendation. Some studies applied collaborative filtering to paper recommendation, where both users and items in the rating matrix are research papers, and rating scores are of a binary type (either cited or not) [1], [12]. Tanner *et al.* took one step further to count the number of times, for which the paper cites the same reference, as the weight of citation [13]. Hu *et al.* capture authors' citation relationship using author's dual role citation relationship to obtain papers that relevant to authors [14].

Citation analysis is also a widely used technique to search for relevant papers. Bibliographic coupling is a citation analysis approach to identify papers that cite the same reference as a given paper [6], [15], and co-citation analysis takes an opposite approach to identify papers cited by the same paper [16]. Gipp and Beel *et al.* proposed the CPA approach, which is based on co-citation analysis to assist researchers in finding related work more precisely [17]. In recent years, more work has been done based on citation relations. For example, Khan *et al.* exploited the citation part based on bibliographic coupling and co-citation analysis [18], [19].

2.2 Content-Based Approaches

Academic papers are mainly comprised of texts and many research efforts have been made based on contents, for example, by extracting keywords and abstracts from papers to measure the relevance of papers. As such, content-based approaches are rooted in information retrieval, and TFIDF, which is one of the most popular methods in information retrieval, is also widely used for content-based recommendation [20].

However, due to the difficulty for accurate PDF extraction and the limitation of text processing methods, content-based approaches have met with limited success. In addition to TFIDF, among commonly used methods for text similarity measurement are Cosine similarity, Jaccard distance, and a few others. Choi *et al.* use deep neural networks to train the feature vectors built by title and abstract for patent citation recommendation [21]. Another study proposes a novel embedding-based neural network model for citation recommendation that captures the relatedness and importance of words in the context [22]. The study of Ali *et al.* reviews the application of deep learning in the domain of citation recommendation [23]. With the advance in text processing technologies in machine learning, content-based approaches are expected to yield better performance.

2.3 Graph-Based Approaches

Graph-based approaches build graphs to represent relations among citations, keywords, topics, and information of authors or users for paper recommendation. Citation network is one widely used graph-based approach established by citation relations to references. Tanner *et al.* computed the relevance of citations based on AAN citation network [13], [24]. Pan *et al.* constructed a heterogeneous graph using the connection between citations and key-terms [25], combining citation- and content-based information. Unlike existing works which transfer the heterogeneous graphs into simple subgraphs or homogeneous graph, Ma *et al.* directly learn the embeddings of all types of nodes from the original heterogeneous graph [26]. Huang *et al.* proposed a two-layer graph of book layer and customer layer with links of purchase between two layers for book recommendation [27]. Amami *et al.* built researcher profiles based on topics of interest, and constructed researcher graphs to recommend papers based on a researcher-paper rating matrix [28]. However, the profiling-based method has a performance limitation due to the scarcity of user information and the diversity of work conducted by the same researcher.

2.4 Hybrid Approaches

Many existing approaches including CF-based, content-based, and social networks have been widely used for paper recommendation. However, with the increasing number of academic papers, the rating matrix based on CF has various issues such as data sparsity and cold start; content-based approaches critically rely on the technique of text extraction, and lack diversity in recommendation due to the sole use of text similarity; social networks require a wide variety of user information, which is not readily available in most digital libraries. The obvious advantage of hybrid approach is that it can use the combination of different recommendation techniques and more information from many sources [29]. Therefore, hybrid approaches have emerged as a trend for performance improvement. Several studies investigated hybrid models combining citation- and content-based approaches for paper recommendation [30]–[32], which exhibit better performance than traditional ap-

proaches based on a single type of information. A combination of traditional approaches with machine learning or deep learning also yields good performance [33], [34]. [35] uses the Microsoft Academic Graph (MAG), titles, and abstracts of research papers to build a recommendation list for all documents, which combines co-citation and content based approaches. Wang *rt al.* recommend citations in heterogeneous academic information networks, and creates the paper rating matrix based on attributed citation network representation learning to address the challenge of insufficient paper rating matrix [36].

Such hybrid models are built upon an ensemble of multiple approaches to make up for the shortcomings of a single approach, and consider a comprehensive set of paper features to provide more accurate recommendation. Our work also adopts a hybrid model that constructs a weighted graphical form of section-aware citation relations and employs a combination of citation- and content-based approaches for paper recommendation.

3. A New Hybrid Approach Based on Citation, Content, and Graphical Representation

This section details our hybrid model based on citation, content, and graphical representation, whose framework is illustrated in Fig. 1. We first identify candidate papers through a graph of weighted citation network, then measure the relevance between two papers as the weight of their edge considering the position of citations and keyphrases, and finally generate a list of ranked candidate papers based on the weighted graph.

3.1 Citation Analysis

Citation relation includes direct citation and indirect citation. Many recent studies are focused on indirect citation relations such as bibliographic coupling and co-citation, without considering relevant papers among direct citations. In most cases, some cited references are of high relevance to the citing paper, but others may be not. Although most digital libraries provide a complete list of references, it is not straightforward for readers to sift out relevant references. Therefore, we build a citation graph similar to a citation network to identify candidate papers.

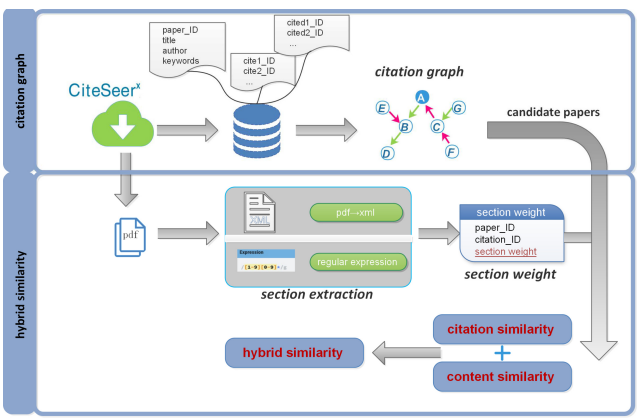


Fig. 1 Framework of a new hybrid model for paper recommendation.

3.1.1 Data Collection through Crawling

We collect data from CiteSeer, which is a public digital library that provides free access to a large database of publications and various paper-related information, through a three-step process to crawl: i) Information of the target paper. We crawl various data items ranging from the title, authors, keyphrases, to links of citations (including papers cited by and citing the target paper). ii) Information of direct citations. We crawl the information of citations resulted from Step i) through the links of citations. iii) Information of indirect citations. We repeat the crawling for data items with links acquired by the previous step. A citation graph is then constructed using the target paper and all citations in the collected dataset. The rest of the information is used at a later stage for computing similarity.

3.1.2 Citation Graph

Figure 2 illustrates the structure of an example graph of citations, where each node represents a paper and the link between two nodes represents a citation relation. An edge incident from a node is associated with $+1$ denoting the relation of citing, while an edge incident to a node is associated with -1 denoting the relation of cited-by. Taking the target node A in Fig. 2 as example, the edge from A has $+1$, while the edge to A has -1. Starting from A, we can reach node B through the edge of $+1$.

3.1.3 Candidate Papers

We take two steps to acquire a collection of candidate papers from citation graph: i) The nodes of direct citations are reached through the traversal of one edge. In Fig. 2, from the target paper A, the traversal of one edge of $+1$ and -1 reaches two direct neighbor nodes B and C, respectively. ii) The nodes of indirect citations are reached by the traversal of two consecutive edges of $(+1,+1)$, $(+1,-1)$, $(-1,-1)$ or $(-1,+1)$, where $(+1,-1)$ finds the node of bibliographic coupling, $(-1,+1)$ finds the node of co-citation, and $(+1,+1)$ and $(-1,-1)$ may help find more potentially relevant papers. In Fig. 2, from node A, traversing $(+1,+1)$,

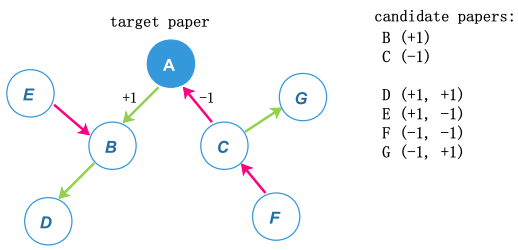


Fig. 2 An example citation graph.

$(+1,-1)$, $(-1,-1)$ and $(-1,+1)$ reaches indirect neighbors D, E, F and G, respectively. Combining both direct and indirect citations, we obtain a set of candidate papers consisting of B, C, D, E, F and G for the target paper A.

To obtain more candidate papers, we can collect papers by traversing more edges. As described above, since the traversal of one edge is $+1$ or -1, the traversal of two edges has $2^2 = 4$ different ways, which are the combination of two edges traversals: $(\pm1, \pm1)$. Similarly, there are $2^3 = 8$ different traversals to collect more indirect citations through three edges: $(\pm1, \pm1, \pm1)$, each edge has two choices to traverse. Hence, the traversal of k edges has 2^k different ways: $(\underbrace{\pm1, \pm1, \pm1, \cdots}_{\text{k}})$.

We denote the collection of n candidate recommended papers as $RP(t) = \{(r_1, sim_1, p_1), (r_2, sim_2, p_2), \ldots, (r_n, sim_n, p_n)\}$, where t is the target paper, for which we recommend papers, r_i is a relevant paper among the candidates, sim_i is the similarity measurement between t and r_i, and p_i is the path from t to r_i, whose length is calculated as the sum of component edges. For example, in Fig. 2, the path from A to D is $(+1, +1)$, and the sum of edges is $+2$. Therefore, the collection of candidate recommended papers for A is denoted as $RP(A) = \{(B, 0, +1), (C, 0, -1), (D, 0, +2), (E, 0, 0), (F, 0, -2), (G, 0, 0)\}$, where the weight of each edge is initialized to be 0 for now.

Based on the collection of candidate papers obtained through edge traversal in the citation graph, we propose a hybrid model that combines citation- and content-based approaches to compute the similarity between each candidate paper and the target paper. The hybrid similarity is measured as the sum of section-aware citation similarity and content similarity.

3.2 Section-Aware Citation Similarity

We consider assigning a section weight to reflect the relevance of a citation at a specific location (section) in the paper, which is also set as the weight of the citation edge. We compute the citation similarity of each candidate paper through the weighted edge in the citation graph, as detailed below.

3.2.1 Section Extraction

We extract the sections in the PDF documents downloaded from public libraries such as CiteSeer through the use of PDFx that converts PDF to XML format [37]. XML documents have the section element with the tag "section" and the citation element. Sections where citations are located can be fetched through these elements, as illustrated in Fig. 3.

However, according to our study, only about 60% of PDFs can be converted into the standard XML format. Therefore, we also use another approach of regular matching to extract sections and citations directly from PDFs that cannot be converted correctly. Despite the limitation of PDF processing techniques and the inaccuracy of regular matching, we believe that this approach is viable because the extracted results of section names such as introduction and related work are consistent in most papers. The technical section typically has different names, which are handled separately later.

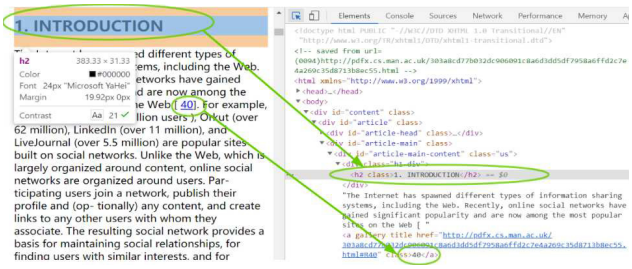
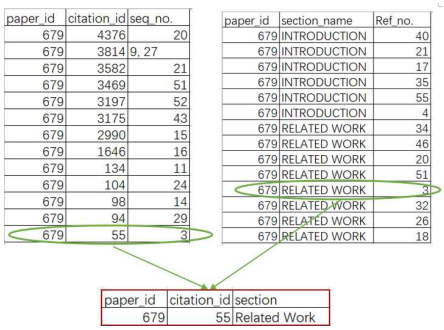


Fig. 3 Convert PDF to XML format to extract sections of citations.



paper_id	citation_id	seq_no.
679	4376	20
679	3814	9, 27
679	3582	21
679	3469	51
679	3197	52
679	3175	43
679	2990	15
679	1646	16
679	134	11
679	104	24
679	98	14
679	94	29
679	55	3

paper_id	section_name	Ref_no.
679	INTRODUCTION	40
679	INTRODUCTION	21
679	INTRODUCTION	17
679	INTRODUCTION	35
679	INTRODUCTION	55
679	INTRODUCTION	4
679	RELATED WORK	34
679	RELATED WORK	46
679	RELATED WORK	20
679	RELATED WORK	51
679	RELATED WORK	3
679	RELATED WORK	32
679	RELATED WORK	26
679	RELATED WORK	18

paper_id	citation_id	section
679	55	Related Work

Fig. 4 Determine the section of citation.

Table 1 Section mapping relationship.

Generic sections	*Mapped sections*
Introduction	Introduction \ Overview
Related Work	Related Work \ Background
Methodology	(other sections)
Result	Result \ Experiment \ Evaluation
Conclusion	Conclusion \ Future Work \ Discussion

For illustration, Fig. 4 plots the sequence numbers of references in the citing paper, which are crawled directly from CiteSeer, and the corresponding extraction results including all reference numbers and sections where the references are cited. The section of citation can be obtained by combining these two lists of reference numbers.

3.2.2 Section Mapping

We map the extracted sections into a generic list of sections based on the section names. According to the study in [38], most academic papers contain certain sections shown in Table 1 as "generic sections". Hence, we map the extracted sections to five generic sections using the mapping relationship in Table 1.

Table 2 Section weight values.

generic sections	*weight*
Introduction	2
Related Work	1
Methodology	3
Result	3

3.2.3 Citation Weight

It has been well recognized that different sections in an academic paper are of different significance [39]. The sections of Methodology and Results are usually of primary importance, so the citations in these sections are more relevant than in other sections. The section of Related Work typically contains a significant number of citations, but most of them are just some supporting work to the citing paper. Therefore, this section is usually considered less important. The following inequality qualitatively describes the importance of different sections in an academic paper:

$$w_{Methodology}/w_{Result} > w_{Introduction} > w_{RelatedWork} \tag{1}$$

where w_i denotes the weight of section. We determine weight values to sections in Table 2 based on Eq. (1) through experiments. The experimental result is measured by Spearman coefficient, the computation of which is detailed in Sect. 4. To get started, the weight of the 'Related Work' section is initialized to be 1, and the sets of weights are assigned increasingly from 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, ..., 2.0, 2.5, 3.0, ..., to 5.0. Therefore, the first weight set of the 'Related Work', 'Introduction', 'Methodology' and 'Result' sections is 1, 1.2, 1.4 and 1.4 (increment of 0.2), and the second weight set is 1, 1.4, 1.8 and 1.8 (increment of 0.4). The results of section-based similarity are qualitatively similar, and are around 0.73 as the weights increase from 0.8 to 1.0. However, with hybrid similarity, which combines content similarity, the weight set of 1, 2, 3 and 3 (increment of 1.0) performs better with coefficient of 0.77. We also initialize the 'Related Work' section with larger values of 2, 3, 4, ..., 10, and the weights of other sections are assigned by increasing 1.0. The results of section-based similarity are close, while the results of hybrid similarity are around 0.74. We observe that larger section weights may be less sensitive to the content similarity. Generally, more than twice as much content similarity is taken to obtain the same hybrid similarity of the weight set of 1, 2, 3 and 3. Our results of weight determination are the same as the weight values in [11]. The value represents the weight of citation in one section, and is also the weight of each edge in the constructed citation graph. An example weighted citation graph is provided in Fig. 5 for illustration.

3.2.4 Citation Similarity

In the weighted citation graph, there are five different values of path p_i in the collection of candidate recommended papers RP, i.e., $+1$, -1, 0, $+2$, and -2. We compute the citation similarity for candidate papers according to the length of path, as follows:

$$sim_i(citation) = \begin{cases} w(edge), & p = +1, -1 \\ w(edge_1) \times a + w(edge_2) \times (1-a), & p = 0 \\ w(edge_1) \times b + w(edge_2) \times (1-b), & p = +2, -2 \end{cases} \tag{2}$$

where $w(edge)$ denotes the weight of edge, and the citation similarity is classified into three conditions according to path length p_i. The first condition is for direct citations, where the citation similarity is measured by the weight of edge directly when $p_i = +1$ or -1. The other two conditions are for indirect citations, where the citation similarity is computed by the weights of both the first edge and the second edge, namely, $edge_1$ and $edge_2$. Past studies have shown the importance of bibliographic coupling and co-citation structures, whose path lengths are both 0, but no comparison of difference has been made between these two structures of indirect citations with path lengths of 0 and ±2. Therefore, we discuss the proportions of two edges for these two structures separately.

We conduct an empirical study to test the proportions of two edges, as shown in Fig. 6, under the conditions of path lengths of ±2 and 0. We determine the proportions a and b of path lengths of 0 and ±2 by comparing the ranking of citation similarities with the ranking of JensenShannon Divergence (JSD) for candidate papers in terms of Spearman correlation coefficient. There are five possible values of a and b, respectively, in Fig. 6. We observe a result of 0.73 when the proportion a is 0.3 and the proportion b is also 0.3, while the results of other values of proportions are between

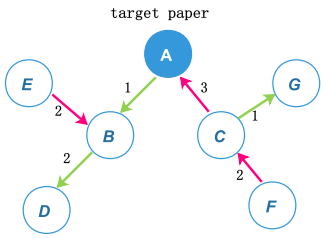


Fig. 5 An example weighted citation graph.

path = 0

edge1 (a)	edge2 (1-a)
0.1	0.9
0.2	0.8
0.3	0.7
0.4	0.6
0.5	0.5

path = ±2

edge1 (b)	edge2 (1-b)
0.1	0.9
0.2	0.8
0.3	0.7
0.4	0.6
0.5	0.5

Fig. 6 Proportion values.

0.2 and 0.5. Therefore, we assign $a = b = 0.3$, which means that we set the proportion of 0.3 to the first edge and 0.7 to the second edge in both cases of path lengths of 0, +2 and −2. It is also verified that the indirect citations whose path lengths are ±2 are of equal importance as those of path lengths 0.

When $a = b = 0.3$, taking the weight in Fig. 5 as example where B and C are direct citations, we have $sim_B(citation) = w(AB) = 1$, and $sim_C(citation) = w(AC) = 3$. Also, since the other nodes D, E, F and G are indirect citations, we have $sim_D(citation) = w(AB) \times 0.3 + w(BD) \times 0.7 = 1.7$, $sim_E(citation) = 1.7$, $sim_F(citation) = 2.3$, and $sim_G(citation) = 1.6$. With the above assignment of a and b, Eq. (2) can be rewritten as

$$sim_i(citation) = \begin{cases} w(edge), & p = +1, -1 \\ w(edge_1) \times 0.3 + w(edge_2) \times 0.7, & p = 0, +2, -2 \end{cases} \quad (3)$$

3.3 Content Similarity

There could be many citations in one section, so citations may have the same citation similarity. Therefore, we also compute content similarity, which is combined with citation similarity for more accurate recommendation. The content similarity between the keyphrases of the citing paper and the cited paper is measured by cosine similarity as follows:

$$sim_i(content) = \frac{A \cdot B}{||A|| \cdot ||B||}. \quad (4)$$

where A is the keyword frequency vector of relevant paper r_i, and B is the keyword frequency vector of the whole collection of relevant papers for the target paper. Each item of vector A or B is the word frequency calculated by TF, the term frequency of word i in the keywords of paper r_j is $tf_{i,j} = \frac{n_{i,j}}{\sum_k^m n_{k,j}}$, $n_{i,j}$ is the number of word i in the keywords of paper r_j, and $\sum_k^m n_{k,j}$ is the sum of numbers of m words in the keywords of paper r_j. $||A||$ and $||B||$ are the norms of vectors A and B, respectively. For the diversity of recommended papers, we use the keyphrases of all candidate papers as vector B instead of the keyphrases only of the target paper. We use the average of the content similarities of the candidate papers as the content similarity of those papers that lack keyphrases in the dataset.

doi_paper	doi_citation	sim(citation)	sim(content)	similarity	path
10.3969j.is	10.13413j.c	2	0.6107472	2.610747	1
10.3969j.is	10.3969j.iss	1	0.2685467	1.268547	1
10.3969j.is	10.3979j.iss	1	0.2770089	1.277009	1
10.3969j.is	10.11772j.i	3	0.3870371	3.387037	1
10.3969j.is	10.11896j.i	2	0.6451852	2.645185	1
10.3969j.is	10.3969j.iss	2	0.3705156	2.370516	1
10.3969j.is	10.3969j.iss	1	0.5025135	1.502514	-1
10.3969j.is	10.3969j.iss	2	0.5133162	2.513316	-1
10.3969j.is	10.13413j.c	2	0.4503229	2.450323	-1
10.3969j.is	10.3969j.iss	1.7	0.7117547	2.411755	2
10.3969j.is	10.3778j.iss	1.7	0.5970131	2.297013	0
10.3969j.is	10.11907rj	1.7	0.4659119	2.165912	0
10.3969j.is	10.13328j.c	2	0.3766222	2.376622	2
10.3969j.is	120-124	2	0.660574	2.660574	2
10.3969j.is	U46 TV9	2.7	0.6801361	3.380136	0
10.3969j.is	10.3969j.iss	1.3	0.2007054	1.500705	0

Fig. 7 Similarity results.

3.4 Hybrid Similarity

We compute hybrid similarity sim_i to rank candidate recommended papers $RP(t)$ through a hybrid approach by combing citation similarity and content similarity, as follows:

$$sim_i = sim_i(citation) + sim_i(content). \quad (5)$$

A partial list of the similarity computing results based on real-world data is provided in Fig. 7, the third column is citation similarity and forth column is content similarity, hybrid similarity in the fifth column is the sum of two similarities. The relevant papers ranked by hybrid similarity is the final list of paper recommendations.

4. Experimental Results

4.1 Performance Evaluation with Respect to JSD

We conduct experiments on a dataset crawled from CiteSeer, which contains 1,100 documents with 18 target papers. For a better illustration of the experimental results, we divide 18 target papers into 10 sets according to the size of candidate papers, which are collected from the citation graph of each target paper. The resulted ten sets of target papers are plotted in Fig. 8, where the x-axis is the number of candidate papers, and y-axis is the number of target papers. Note that the PDF documents of papers are not always available, so some target papers may have only a small number of candidate papers.

There exist various methods and standards for recommendation evaluation, including those proposed by McNee *et al.* [1] and Sugiyam *et al.* in the field of information retrieval [12]. In this work, we employ Spearman correlation to evaluate the accuracy of recommendation [11] against the ranking produced by JensenShannon Divergence (JSD) as the benchmark. JSD computes the distance between the word distribution probability of a target paper and a candidate paper, and generates a ranking of relevant papers using the JSD value of each pair. To accurately measure the relevance between papers, we use the full texts of papers to calculate the word distribution probability. Therefore, this JSD-based procedure is prohibitively time-consuming, especially when processing a large number of PDF documents.

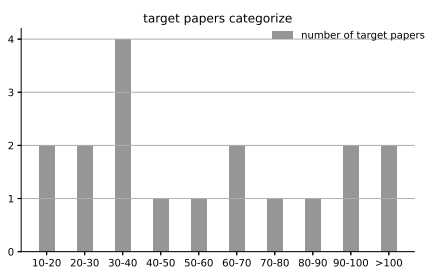


Fig. 8 Ten sets of target papers.

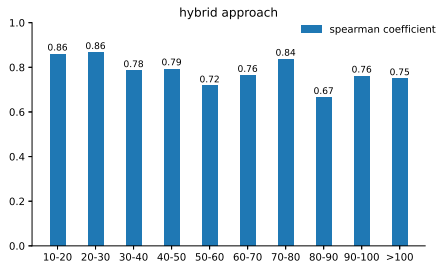


Fig. 9 Spearman coefficients of the proposed hybrid approach w.r.t. JSD.

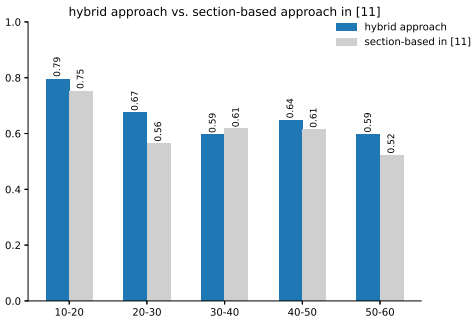


Fig. 10 Hybrid approach vs. section-based approach [11] on bibliographic coupling.

For example, processing 100 PDF documents takes nearly half an hour in average. In contrast, our proposed hybrid approach only processes minimal texts such as abstract and keywords, and hence runs much faster.

We compute the consistency between the ranking of our approach and the ranking of JSD using Spearman correlation coefficient [40]. Figure 9 plots the average Spearman coefficients in 10 sets of papers. The average correlation coefficient of all papers is 0.77.

4.1.1 Hybrid Approach vs. Section-Based Approach

We compare our hybrid approach in Fig. 10 with the section-based approach in [11], which is the state of the art using section position for paper recommendation. Since this section-based approach is focused on bibliographic coupling, we conduct this comparison on the indirect citations of bibliographic coupling. The results show that our approach outperforms the state-of-art technique in average.

4.1.2 Section-Based Citation Similarity vs. Traditional Citation Similarity

Traditional citation-based approaches do not consider the location of citation, while our section-based citation approach differentiates the section position of citation with different weights. We compare the results of our section-based approach and traditional citation-based approach, as shown in Fig. 11, where x-axis represents the size of relevant papers, and y-axis represents the consistency measured by Spearman coefficient. These results clearly show that our section-based approach has superior performance over the traditional approach.

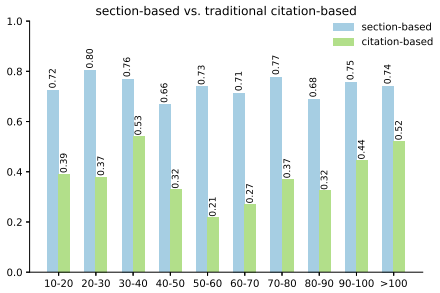


Fig. 11 Section-based approach vs. traditional citation-based approach.

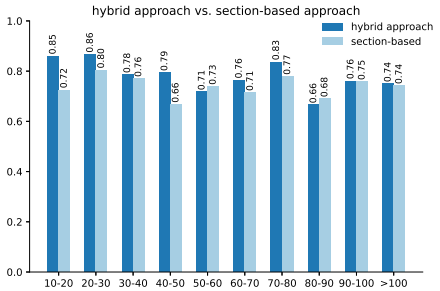


Fig. 12 Hybrid approach vs. section-based approach.

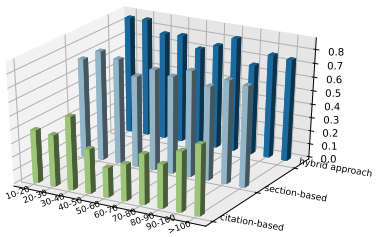


Fig. 13 Comparison of three approaches.

4.1.3 Hybrid Approach vs. Section-Based Approach

Our hybrid model is built upon a combination of the section-based approach and the content-based approach. We compare the recommendation performance of hybrid similarity with section-based citation similarity, as shown in Fig. 12. These results show that the proposed hybrid approach outperforms the section-based approach in most of the cases we studied.

The overall results in Fig. 13 illustrate that the proposed hybrid approach exhibits superior performance over the approaches using citations only. It achieves about 40% higher recommendation accuracy compared with the traditional citation-based approach, as shown in Fig. 14.

4.2 Performance Evaluation with Respect to Manual Ranking

We also conduct experiments on Wanfang digital library, which has the largest collection of Chinese papers [10]. We

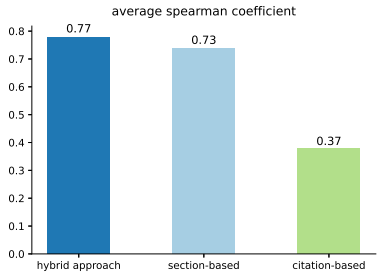


Fig. 14 Average Spearman coefficient of three approaches.

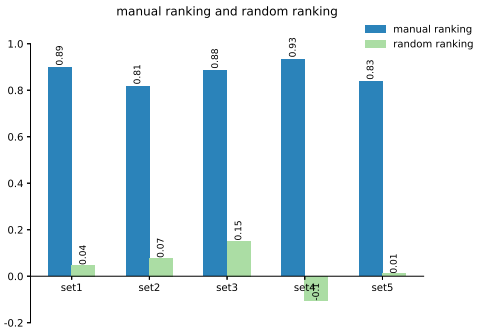


Fig. 15 Spearman coefficient of manual ranking and random ranking on Wanfang dataset.

collect about 500 papers from Wanfang, including 5 sets of relevant papers for 5 target papers. Since most of the papers from Wanfang are in Chinese, and cannot be directly processed by JSD to measure text similarity, we evaluate the performance of our approach with respect to manual ranking. In our study, the control group contains 12 participants who are all postgraduates, each of whom is assigned with 2 sets of candidate papers. The average manual ranking for candidate papers produced from the participants is used as the benchmark for performance comparison. We conduct experiments to verify the effectiveness and consistency of manual ranking with 12 participants. As shown in Fig. 15, we calculate the coefficient between each manual ranking and average manual ranking on each set of papers. The average coefficient of each set is above 0.8, which verifies the consistency in manual ranking. For comparison, we generate 5 random rankings. The low correlation coefficients indicate that average manual ranking is statistically different from random ranking, which verifies the effectiveness of our manual ranking.

The consistency between our proposed approach and manual ranking is still measured by Spearman correlation coefficient. The results in Fig. 16 show that the average correlation coefficient of all papers is 0.73. The comparison of our hybrid approach with the section-based approach in [11] is shown as Fig. 17.

We further compare our hybrid approach with the section-based and traditional citation-based approaches. Since Wanfang dataset provides comprehensive keywords, we also compare with the content-based approach. The results in Fig. 18 show that the proposed hybrid method outperforms the other three approaches in most of the cases we studied.

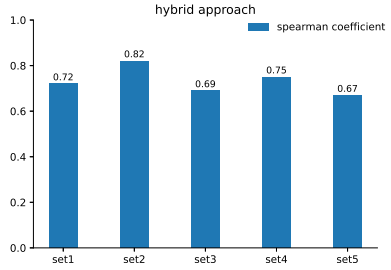


Fig. 16 Spearman coefficient of the hybrid approach on Wanfang dataset.

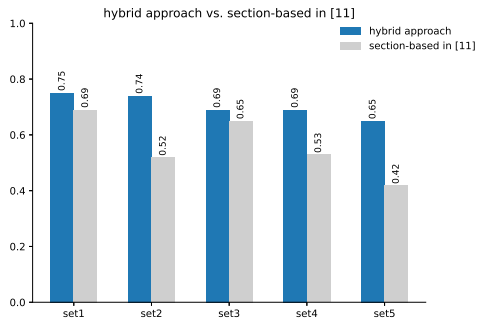


Fig. 17 Hybrid approach vs. section-based approach [11] on bibliographic coupling of Wanfang dataset.

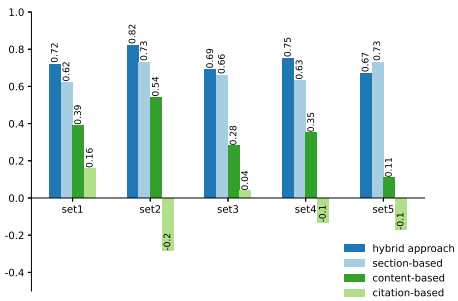


Fig. 18 Comparison of four approaches on Wanfang dataset.

4.3 Scalability Evaluation

To illustrate the scalability of our approach, we conduct another experiment on Wanfang dataset. The citation graph in our approach can be extended to a deeper level to obtain more relevant citations when there are very few citations. However, if there exist adequate citations, the citation graph may become very complex when using a path length of 3 or 4, instead of 2. More importantly, there is no need to generate the citation graph with a path length more than 2 if there are adequate citations. We compute the recall of top 50 papers for 5 sets of Wanfang data with a path length of 1, 2 and 3, respectively. As shown in Fig. 19, the recall@50 with a path length of 2 and 3 in these 5 sets are almost the same, which indicates that a path length of 2 is sufficient.

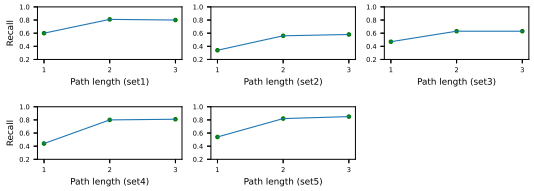


Fig. 19 Recall@50 of different path on 5 sets of Wanfang dataset.

5. Conclusion

In this paper, we constructed a citation graph based on citation relations to identify candidate relevant papers, and proposed a hybrid model consisting of section-based and content-based approaches to compute the similarity of relevant papers. Our hybrid approach recommends relevant papers using direct citations and indirect citations, and measures similarity based on section position and keyword information. This approach is able to find papers that are highly relevant to the target paper. Meanwhile, the extraction of sections through XML documents and the similarity computing of keywords are much faster than other approaches that typically require intensive text computing. Our approach was evaluated with real-life datasets, and the results show its performance superiority over existing methods.

The section-based approach requires the section position of citations, but it is challenging to extract accurate section information. We will explore new approaches to improve the accuracy of section extraction. Also, content similarity in our work is measured by a traditional method. It is of our interest to employ emerging techniques in machine learning and deep learning to measure content similarity more effectively.

Acknowledgments

This research is sponsored by National Key Research and Development Plan of China under Grant No. 2017YFB1400301. Thanks for the collaboration of Dr. Liqiong Chang.

References

[1] S.M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S.K. Lam, A.M. Rashid, J.A. Konstan, and J. Riedl, "On the recommending of citations for research papers," Proceedings of the 2002 ACM conference on Computer supported cooperative work, pp.116–125, 2002.

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," Knowledge-based systems, vol.46, pp.109–132, 2013.

[3] X. Dong, L. Yu, Z. Wu, Y. Sun, L. Yuan, and F. Zhang, "A hybrid collaborative filtering model with deep structure for recommender systems," Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[4] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," Proceedings of the 10th annual joint conference on Digital libraries, pp.29–38, 2010.

[5] K.W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," Journal of the American Society for information Science and Technology, vol.61, no.12, pp.2389–2404, 2010.

[6] M.M. Kessler, "Bibliographic coupling between scientific papers," American documentation, vol.14, no.1, pp.10–25, 1963.

[7] X.Y. Liu and B.-C. Chien, "Applying citation network analysis on recommendation of research paper collection," Proceedings of the 4th multidisciplinary international social networks conference, pp.1–6, 2017.

[8] H. Voos and K.S. Dagaev, "Are all citations equal? or, did we op. cit. your idem?," Journal of Academic Librarianship, vol.1, no.6, pp.19–21, 1976.

[9] C.L. Giles, K.D. Bollacker, and S. Lawrence, "Citeseer: An automatic citation indexing system," Proceedings of the third ACM conference on Digital libraries, pp.89–98, 1998.

[10] J. Xia, J. Wright, and C.E. Adams, "Five large chinese biomedical bibliographic databases: accessibility and coverage," Health Information & Libraries Journal, vol.25, no.1, pp.55–61, 2008.

[11] R. Habib and M.T. Afzal, "Sections-based bibliographic coupling for research paper recommendation," Scientometrics, vol.119, no.2, pp.643–656, 2019.

[12] K. Sugiyama and M.-Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation," Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, pp.153–162, 2013.

[13] W. Tanner, E. Akbas, and M. Hasan, "Paper recommendation based on citation relation," 2019 IEEE International Conference on Big Data (Big Data), pp.3053–3059, IEEE, 2019.

[14] D. Hu, H. Ma, Y. Liu, and X. He, "Scientific paper recommendation using author's dual role citation relationship," International Conference on Intelligent Information Processing, vol.581, pp.121–132, Springer, 2020.

[15] O. Küçüktunç, E. Saule, K. Kaya, and Ü.V. Çatalyürek, "Recommendation on academic networks using direction aware citation analysis," arXiv preprint arXiv:1205.1143, 2012.

[16] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," Journal of the American Society for information Science, vol.24, no.4, pp.265–269, 1973.

[17] B. Gipp and J. Beel, "Citation proximity analysis (cpa): A new approach for identifying related work based on co-citation analysis," ISSI'09: 12th International Conference on Scientometrics and Informetrics, pp.571–575, 2009.

[18] A.Y. Khan, A.S. KHATTAK, and M.T. Afzal, "Extending co-citation using sections of research articles," Turkish Journal of Electrical Engineering & Computer Sciences, vol.26, no.6, pp.3345–3355, 2018.

[19] A.M. Khan, A. Shahid, M.T. Afzal, F. Nazar, F.S. Alotaibi, and K.H. Alyoubi, "Swics: Section-wise in-text citation score," IEEE Access, vol.7, pp.137090–137102, 2019.

[20] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE transactions on knowledge and data engineering, vol.17, no.6, pp.734–749, 2005.

[21] J. Choi, S. Jang, J. Kim, J. Lee, J. Yoona, and S. Choi, "Deep learning-based citation recommendation system for patents," arXiv preprint arXiv:2010.10932, 2020.

[22] Y. Zhang and Q. Ma, "Dual attention model for citation recommendation," arXiv preprint arXiv:2010.00182, 2020.

[23] Z. Ali, P. Kefalas, K. Muhammad, B. Ali, and M. Imran, "Deep learning in citation recommendation models survey," Expert Systems with Applications, vol.162, p.113790, 2020.

[24] D.R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The acl anthology network corpus," Language Resources and Evaluation, vol.47, no.4, pp.919–944, 2013.

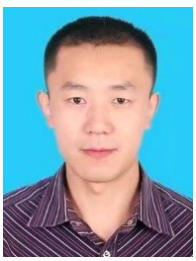
[25] L. Pan, X. Dai, S. Huang, and J. Chen, "Academic paper recom-

mendation based on heterogeneous graph," Chinese computational linguistics and natural language processing based on naturally annotated big data, vol.9427, pp.381–392, Springer, 2015.
[26] X. Ma and R. Wang, "Personalized scientific paper recommendation based on heterogeneous graph representation," IEEE Access, vol.7, pp.79887–79894, 2019.
[27] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, "A graph-based recommender system for digital library," Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, pp.65–73, 2002.
[28] M. Amami, G. Pasi, F. Stella, and R. Faiz, "An lda-based approach to scientific paper recommendation," International conference on applications of natural language to information systems, vol.9612, pp.200–210, Springer, 2016.
[29] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," IEEE Access, vol.7, pp.9324–9339, 2019.
[30] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," Communications of the ACM, vol.40, no.3, pp.66–72, 1997.
[31] C. Basu, H. Hirsh, W. Cohen, et al., "Recommendation as classification: Using social and content-based information in recommendation," Aaai/iaai, pp.714–720, 1998.
[32] I. Soboroff and C. Nicholas, "Combining content and collaboration in text filtering," Proceedings of the IJCAI, pp.86–91, sn, 1999.
[33] G. Nandi and A. Das, "A survey on using data mining techniques for online social network analysis," International Journal of Computer Science Issues (IJCSI), vol.10, no.6, p.162, 2013.
[34] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp.1235–1244, 2015.
[35] A. Kanakia, Z. Shen, D. Eide, and K. Wang, "A scalable hybrid research paper recommender system for microsoft academic," The World Wide Web Conference, pp.2893–2899, 2019.
[36] W. Wang, T. Tang, F. Xia, Z. Gong, Z. Chen, and H. Liu, "Collaborative filtering with network representation learning for citation recommendation," IEEE Transactions on Big Data, 2020.
[37] A. Constantin, S. Pettifer, and A. Voronkov, "Pdfx: fully-automated pdf-to-xml conversion of scientific literature," Proceedings of the 2013 ACM symposium on Document engineering, pp.177–180, 2013.
[38] T. Hengl and M. Gould, "Rules of thumb for writing research articles," Enschede, Sept. 2002.
[39] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," Proceedings of the 2006 conference on empirical methods in natural language processing, pp.103–110, 2006.
[40] C. Spearman, "The proof and measurement of association between two things," The American journal of psychology, vol.100, no.3/4, pp.441–471, 1987.

Ying Kang received the B.S. degree in School of Software Engineering from Northwest University. Her research focuses on recommendation system and distributed system. She is now a graduate student in School of Information Science and Technology from Northwestern University.

Aiqin Hou received the Ph.D. degree in school of information science and technology from Northwest University in 2018. She is currently an Associate Professor with the School of Information Science and Technology, Northwest University of China. Her research interests include big data, high performance network, and resource scheduling.

Zimin Zhao received the B.S. degree in Electronic Information Science from Northwest University in 2020. His research focus on automated machine learning and automated deep learning.



Daguang Gan received the M.S. degree from the China institute of science and technology in 2008. He is an assistant to the general manager of Beijing wanfang software co., LTD., mainly engaged in knowledge organization, information retrieval and information analysis.