Subspace Embedding Based New Paper Recommendation

Yi Xie [1], Wen Li [1], Yuqing Sun [1*], Elisa Bertino [2], Bin Gong [1]
[1]*Shandong University, China*, [2]*Purdue University, the United States*
heilongjiangxieyi@163.com, 3375757785@qq.com, sun_yuqing@sdu.edu.cn, bertino@cs.purdue.edu, gb@sdu.edu.cn

2022 IEEE 38th International Conference on Data Engineering (ICDE) | 978-1-6654-0883-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDE53745.2022.00178

***Abstract*—As huge numbers of academic papers are published every year, it is critical to be able to recommend high quality papers. The typical evaluation method for papers is to use citation information, which however is not applicable to new papers. To address such a shortcoming, in this paper, we consider a novel perspective on the association between the content difference of a paper, with respect to other papers, and its innovation. Since innovation has often domain-specific characteristics and forms, we introduce the concept of subspace to describe the commonly recognized aspects of paper contents, namely background, methods and results. A set of expert rules are formalized to annotate the differences between papers, based on which a twin-network is proposed for learning the embeddings of papers in different subspaces. A series of empirical studies show that there are clear correlations between a paper influence and its difference with others in those subspaces. The results also show the characteristics of innovation in different scientific disciplines. To take into account information about academic networks for paper recommendation, we propose a graph convolutional neural method to combine the paper content with other related elements, where user interests and academic influences are modeled asymmetric. Experimental results on real datasets show that our method is more effective than other baseline methods for new paper recommendation. We also discuss the characteristics of scientific disciplines and authors to show the effectiveness of modeling the asymmetric user interests and influences. Finally, we verify the reusability of our method on a patent dataset. The results show that it is also applicable to academic data with low-resource features.**

***Index Terms*—subspace, new paper, recommendation**

I. Introduction

Large numbers of academic papers are published every year, which makes it difficult for researchers to find high-quality works. Innovation and potential academic influence are important factors commonly used for evaluating paper quality [1], [2]. However, since paper contents involve much domain-specific knowledge, it is difficult to quantify them. Citation information is the most important indicator on academic influence, but it is not applicable for new papers without citations. For new papers, some approaches consider the potential citations and give a unified quantitative evaluation [3]–[5]. However, since there are specific innovation characteristics in different academic fields, the evaluation standards on paper innovation are also different [6], [7]. For example, novel findings are more attractive in social sciences, while in physics a new theory always receives a larger number of citations. Therefore an evaluation based on a unified score is not suitable for assessing papers innovations in different academic fields, especially for new papers.

To address the challenge of evaluating the quality of a new paper, we adopt a novel perspective on the association between the content difference of a paper with other papers and its potential influence. Considering the specific characteristics of innovation in different academic fields, we introduce the concept of subspace to describe the commonly recognized aspects that are related to these characteristics, namely the background, methods and results in paper contents. Since there is no clear labels on paper differences that are measured by researchers for quantifying the innovation, we adopt the basic elements of a paper as measure, including the references of the paper, the academic categories it belongs to, the used keywords, etc., which are formalized as a set of expert rules. We propose a twin-network with a contrast loss to embed the paper contents into subspaces and evaluate the difference between two papers as a probability proportional to the distance in each subspace. Compared with supervised models predicting the number of citations of a paper [8], [9], our measure can eliminate the citations variance due to specific characteristics of academic fields or disciplines. A series of empirical analyses show that there are clear correlations between the paper quality and its difference with others in the subspace. The results also show the characteristics of innovation in different disciplines [10], [11]. Based on our subspace embedding method for paper difference analysis, many tasks can be realized, such as academic innovation law analysis, paper quality prediction, expert evaluation, and etc.

The relationships between elements of academic networks reflect the research interests and provide relevant information about academic influence propagation. Therefore such elements are relevant for evaluating papers quality. For new paper recommendation, we propose a method based on a graph convolutional neural network to combine the paper contents with the information from the academic network, where the user interests and the academic influences are modeled asymmetric. The publications of a researcher are used to model the researcher interests, and new papers are recommended based on the correlations between research interests and candidate papers. The experimental results on real datasets show the effectiveness of this method in personalized paper recommendation. We also discuss the characteristics of specific academic fields and authors to show the effectiveness of modeling the asymmetric user interests and influences, and

*Yuqing Sun is the corresponding author.

2375-026X/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDE53745.2022.00178

verify the reusability of the model.

The rest of this paper is organized as follows. Sec. II discusses related works. Sec. III presents the model details on subspace embedding. Sec. IV presents our paper recommendation model and the evaluation results. Sec. V outlines some conclusions.

II. Related Work

A. New Paper Evaluation

The typical methods for evaluating paper quality are generally based on citation information [8], [12], which are not applicable to new papers without any citation. Some methods consider multiple characteristics for inferring the influence of academic papers, such as citation information [3], [13], co-author relationships [14], and writing quality [1], [4]. These methods are appropriate for evaluating the quality of a new paper and achieve comparable results with the number of citations on papers, especially for those written by well-known researchers. But they do not directly evaluate the paper contents. Such an evaluation is similar to evaluating a today event by historical events and may result in some blind spots. Besides, since they adopt a unified evaluation score, they are unable to reflect the various forms of academic innovations.

Measures have been proposed to assess the difference in papers contents, such as checking text [15], [16] or exploring the semantic changes in the historical versions so as to reconstruct the editing process [17], [18]. These approaches focus on different versions of the same text, which are totally different with respect to the analysis of the innovation in academic papers. Moreover, they do not consider the impact of other features on paper quality, such as the authority of author or the influence of the publication venue.

B. Paper Recommendation

Most paper recommendation techniques model user interests or paper contents with the help of deep networks, topic model and other representation learning methods using information about papers and authors on academic service platform [2], [10]. Sugiyama et al. [8] propose a neighborhood-based collaborative filtering algorithm for academic paper recommendations. He et al. [12] propose a method to learn the associations between user interests and papers using both the citation information and the paper content. Wang et al. [9], [19] propose a method adopting graph convolution neural networs for paper recommendation, which takes into account the information in the academic network. These methods have been acknowledged effective in personalized paper recommendation although they use in a unified score without considering the specific characteristics of the different academic fields.

Some methods predict the potential citations of a paper by learning user interests and paper contents [20], [21], or based on the author authority [22] by information from the academic network. However, most of these methods use the citation relationships to infer user interests on papers without considering the asymmetric influence of academic knowledge between them.

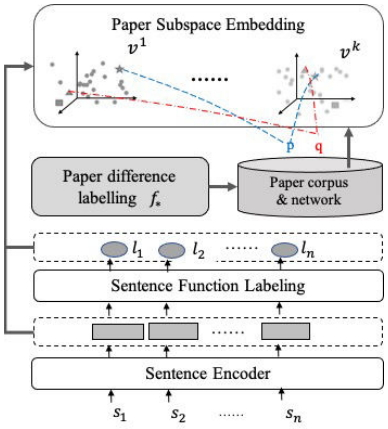


Fig. 1: Paper subspace embedding model.

III. Subspace Based Paper Difference Analysis

Inspired by the fact that the influential papers must have some differences compared with the papers previously published, in this section, we study the correlations between the difference of paper content and the academic influence. To reflect the specific characteristics of innovation in different academic fields, we introduce the concept of subspace to describe the core aspects of paper contributions that are recognized by academia. Our method has three components, shown in Fig. 1. The white boxes at the bottom are the pretrained module, which includes a pretrained sentence encoder for paper content and a pretrained subspace labelling model on sentence function. The gray boxes in the middle represent the annotated data against a set of expert rules. The top part is a twin-network based subspace embeddings model, which is designed to map a paper content to a few subspaces according to the annotated data.

A. Annotation on Paper Difference

We first define some expert rules that reflect the basic consensus of academia on measuring the differences between papers, and design several metrics to quantify them.

1) Score on academic classification: The academic classification is a hierarchically organized classification system (**HCS** denotes this category tree) for research fields created by experts, such as the ACM computing discipline classification system in the computer science (ACM Computing Classification System, referred to as ACM CCS [23]), widely used for journal and conference publications. We adopt a commonly used edit distance on hierarchical structure [24] to evaluate the difference score between papers with respect to academic classification. For paper p and the corresponding tags from **HCS**, r_p represents a set of notes on the path from the root to the node of the paper tag in the category tree **HCS**. The difference score between papers p and q is defined as:

$$f_c(p,q) = \sum_{i \in (r_p \bigcup r_q) - (r_p \bigcap r_q)} \frac{w_{l_i}}{2^{l_i}} \quad (1)$$

where l_i represents the level of the current node i in **HCS**. w_{l_i} is the weight parameter, which satisfies the property that the closer l_i is to the root node in the classification tree, the larger w_{l_i} is.

2) Score on references: The paper references are the indirect indicator about the paper content. The difference score for papers p and q is defined as:

$$f_r(p,q) = \frac{R(p) \bigcup R(q)}{R(p) \bigcap R(q)} \tag{2}$$

where $R(p)$ and $R(q)$ represent the set of references for p and q, respectively, which is the reciprocal of Jackard coefficient.

3) Score on keywords: Keywords are selected by authors to briefly describe the paper content. We adopt the semantics of keywords to measure the difference, defined as the expectation of the Euclidean distance of keyword vectors, i.e.

$$f_w(p,q) = \mathbf{E}_{x \in W(p), y \in W(q)} D(e(x), e(y)) \tag{3}$$

where $W(p)$ and $W(q)$ are the keyword sets of p and q, $e(x)$ denotes the pretrained word embedding [25] of x, and $D(e(x), e(y))$ is the distance function.

4) Score on abstract: The abstract contains the core elements of a paper contribution such as the problem, method, and results, which are often narrated in a sequential form. The subspace based paper differences are learned mostly depending on this part. The measure uses the semantic embedding of the abstract text generated by the pretrained text encoder. We adopt BERT-base [26] as the content encoder since it is widely acknowledged by academia. For the abstract text $t_1 t_2 \cdots t_n$ of p, where t_i is a sentence, the output of BERT is the vector sequence on sentences $H = h_1, h_2, \ldots, h_n$. The pretrained model [27] is used for sentence-level function labeling of the sentences, namely $\mathbf{l} = l_1, l_2, \ldots, l_n$,$l_i \in 1, \ldots, K$, where K is the number of subspaces.

Based on the text vector H and the subspace label $\mathbf{l}$, the subspace fusion embeddings are obtained. The sentence vectors with the same label are mapped to the same subspace, denoted as $C_p = c_p^1, c_p^2, ..., c_p^K$, where $c_p^k \in R^d$ is the embedding vector in the subspace k, and d is the vector dimension. We adopt the expectation of the sentence vector in the same subspace $c_p^k = \mathbf{E}_{i \in [1..n]}(h_i \circ I(l_i = k))$, where $I(l_i = k)$ is the indicator function. The final score on subspace difference is defined as $f_t(p,q) = D(c_p^k, c_q^k)$, where D is the distance function.

B. The Subspace Based Paper Embedding

By the above expert rules, the difference between two papers is computed, where $f_c(p,q)$, $f_r(p,q))$, $f_w(p,q)$ are the indicators of whole paper difference that are applicable to all subspaces. For convenience in the discussion, we use a unified form $f_*^k(p,q)$ to mark such difference, and $D^k(p,q)$ is specific for subspace k.

However, since the above quantification only provides an indication of the content closeness rather than an exact measure, we introduce the possibility that paper differences are proportional to the score differences. That is,

$$\begin{aligned} P(D^k(p,q) > D^k(p,q')) &\propto P(f_*^k(p,q) - f_*^k(p,q')) \\ &= \exp(f_*^k(p,q) - f_*^k(p,q')) \end{aligned} \tag{4}$$

where P represents the probability distribution function.

Then we adopt a multi-layer perceptron neural network with the global attention and pooling for subspace embeddings, denoted by $\hat{\mathbf{c}}_\mathbf{k}$. The embedding after combining the subspace embeddings and context information is denoted as $\tilde{\mathbf{c}}_\mathbf{k}$. Details are given below:

$$\mathbf{x_i^k} = h_i \circ I(l_i = k) \tag{5}$$

$$\mathbf{x^k} = [\mathbf{x_1^k}; \cdots; \mathbf{x_i^k}; \cdots; \mathbf{x_n^k}] \tag{6}$$

$$\mathbf{h_1} = \tanh(W^1 \mathbf{x^k} + b_1) \tag{7}$$

$$\mathbf{h_l} = \tanh(W^l \mathbf{h_{l-1}} + b_l) \tag{8}$$

$$\hat{\mathbf{c}}_\mathbf{k} = \mathbf{m^k} \tanh(M \mathbf{h_l} + b) \tag{9}$$

$$\tilde{\mathbf{c}}_\mathbf{k} = \sum_{j \in [1..K], j \neq k} \mathbf{a_j} * \hat{\mathbf{c}}_\mathbf{j} \tag{10}$$

$$\mathbf{a_i} = \frac{\exp(\mathbf{c_k^T c_i})}{\sum_{j=1, j \neq k}^{K} \exp(\mathbf{c_k^T c_j})} \tag{11}$$

$$\mathbf{c_k} = [\hat{\mathbf{c}}_\mathbf{k}; \tilde{\mathbf{c}}_\mathbf{k}] \tag{12}$$

where $\mathbf{x_i^k}$ is the subspace embedding of the i-th sentence in subspace k, $\mathbf{x^k}$ denotes the subspace embeddings of n sentences in subspace k. W^i and b_i are the weight and bias parameters of multi-layer perceptron neural network, respectively. $\mathbf{m^k}$ is the weight matrix of subspace k, M and b are the weight and bias parameters based on the global attention mechanism. $\mathbf{a_i}$ is the weight of subspace i with respect to k. $\mathbf{c_k}$ is the embedding vector in subspace k by merging $\hat{\mathbf{c}}_\mathbf{k}$ and $\tilde{\mathbf{c}}_\mathbf{k}$.

Finally, we adopt a twin neural network with the contrast loss for fine-tuning. It accepts $\mathbf{c_p^k}$ and $\mathbf{c_q^k}$ as input. For papers p, q, and q', if there is $D^k(p,q) > D^k(p,q')$,

$$\ell^k(\theta) = \sum_{(p,q,q')} D^k(p,q) - D^k(p,q') \tag{13}$$

We use the indicator $D^k(p,q) = -c_p^k \cdot c_q^k$ to measure the paper difference. There are other choices for $D^k(p,q)$, such as Euclidean distance or inner product, which are out of the scope for discussion here. We adopt the hinge loss form and add a regularization term:

$$\ell^k(\theta) = \max\{0, D^k(p,q) - D^k(p,q') + \epsilon\} + \lambda * ||\theta|| \tag{14}$$

Compared with other score based quantification methods of paper quality, our comparative learning method can better eliminate the impact of numerical deviations, such as citations, caused by different criteria concerning paper innovation and characteristics of the specific disciplines. Our method can eliminate the scoring bias of different expert rules while integrating expert knowledge. In addition, it supports an increasing number of expert rules for labeling papers, which is more robust in practice.

C. Datasets and Metrics

We adopt three datasets to evaluate our model. The first is the ACM dataset, which contains 2 million computer science papers in the ACM Digital Library [28], and is often used for paper recommendation. The abstract of papers contains 6.34 sentences on average. The second is the PubMedRCT dataset [29], which contains 200,000 biomedical papers and is often used in searches for research on sequence classification. The average number of sentences in each abstract is 11.5. The third dataset is Scopus [30], which contains multidisciplinary papers and is the largest world wide academic paper database, covering 400,000 papers from 27 disciplines, such as pharmacy, social sciences, computer science and etc.. On average, the abstract contains 5.92 sentences. The metadata in the above three datasets includes title, abstract, citation, and field label.

The PubMedRCT dataset contains sentence-level function tags on paper abstracts, such as background, method, and conclusion. Each sentence in the abstract is marked by a category. Since the abstract of ACM and Scopus do not have function tags, we tag 100 abstracts for each dataset to train the label classifier using the 10-fold cross-validation method. The number of subspaces K is set to 3, namely the background, methods and results. In practical applications, the number of the subspaces can be adjusted according to the characteristics of the academic field.

We adopt ACM CCS as the academic classification system. We select three disciplines in Scopus and 200 papers for each. The papers published in 2013 were regarded as new papers, and the number of their citations up to 2017 was used as the measure on the influence of paper quality. We select the papers published before 2013 in the same fields as the historical comparison papers.

We adopt the Gaussian mixture clustering method to perform spatially independent clustering. This method can fit data distributions of any shape and is more robust than other clustering methods. The number of clusters is set according to the Bayesian information criterion [31]. The specific method is to select the closely related papers using the subspace embeddings. The higher the Local Outlier Factor (LOF for short) value [32] of paper, the more difference the paper has with other papers.

D. Training Process and Experimental Design

The training process includes two parts, namely the subspace sequence labeling and the subspace embedding. We adopt pretrained BERT-base as the text encoder with the conditional random field for labeling sentence subspace [27]. The length of the sentence is set to 30 words and the dimension of the sentence vector output is 768.

Then we train the subspace embedding model. For three papers p, q, q', we compare the difference scores between each pair of them. We select the larger pair as a positive sample , while the smaller pair as a negative sample. Without loss of generality, p is regarded as the reference. The fusion function values $f^k(p,q)$ and $f^k(p,q')$ in each subspace are calculated as $f^k(p,q) = \sum_i a_i f_i(p,q)$, where a_i is the weight parameter learned along with training. For two papers p and q, if the difference in subspace k is greater than p, q', (p,q) is a positive sample pair, and (p,q') is a negative sample pair, $f^k(p,q) > f^k(p,q')$. We calculate the loss function according to equation 14, update the BERT network weights, subspace semantic fusion network parameters, and rule fusion function weights to obtain the paper subspace embedding vector of the fusion of multiple rules.

TABLE I: Correlation between paper difference and citations in Scopus.

Model type	Model	Computer Science	Medicine	Sociology
Paper quality prediction	CLT	0.27	0.21	0.39
	CSJ	0.20	0.16	0.08
	HP	0.33	0.39	0.31
Our method	SEM-B	0.56	0.49	0.62
	SEM-M	**0.87**	0.31	**0.68**
	SEM-R	0.72	**0.70**	0.51

E. Correlation Between Paper Difference and Citations

In order to fairly compare different evaluation methods, the prediction results of each method are ranked, and the ranking results of the methods are compared. The ranking results of the actual citations of the papers are used as a reference to evaluate the performance of the method. We compare the following methods:

- **CLT** [4] is a paper quality evaluation method based on the text readability, language quality, fluency, semantic complexity and other characteristics of papers.
- **CSJ** [1] is a paper writing quality evaluation method using expert evaluation indicators on linguistic knowledge.
- **HP** [3] is a paper influence evaluation method based on h-index, which measures the core degree of papers in academic network. Since there is no citation data for new papers, we use the citation relationship within one year after publication to rank the paper's influence.
- **SEM** is our Subspace Embedding Method (Subspace Embedding Method). SEM-B, SEM-M and SEM-R are used to represent the background, method, and result subspace embedding. The papers are sorted in the descending order of their content differences using our method.

The evaluation is as follows. The actual paper citations are the ground truth. The testing papers are ranked based on the above models. To calculate the consistency of the calculated paper rankings and the actual citation ranking, we adopt the Spearman correlation coefficient [33], which is designed as a measure of a monotonic association between two ranks.

With respect to the degree of correlation between the predicted results and the paper quality, for our method a higher degree of correlation means that the outliers of subspace embeddings better reveal the paper difference. For other methods, the more accurate the quantification of the quality of the paper, the higher the correlation between their quality scores and the actual citation counts.

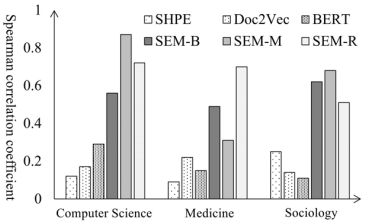


Fig. 2: Correlation between paper outlier and citations of different embedding methods on Scopus

The comparison results are shown in Tab. I. Compared with the other methods, the difference rankings calculated by our method show a higher correlation with the actual paper citations, that is, SEM based paper potential influences are more accurate. Another meaningful conclusion is that, the paper difference reflects the discipline specific innovation in certain subspaces, such paper always arouse more interests. For example, papers with innovative model design in computer science tend to obtain high citations, as shown in bold in Tab. I. By contrast, pharmacy pays more attention to groundbreaking results, and social science tends to novel research methods.

F. Analyzing the Necessity of Subspace

For the ablation analysis, we consider two parts, i.e. the introduction of subspace and the influence of the twin network. The comparison methods are the embedding ones that perform prominently in downstream tasks such as paper recommendation, and model the paper representation vector without distinguishing the subspaces.

- **SHPE** [34] combinnes the word embedding vector generated by Word2Vec technology with the TF-IDF vector of the paper linearly to generate the final paper representation vector.
- **Doc2Vec** [20] is a document embedding method, it adopts paper abstracts as the entire documents to model their embeddings.
- **BERT** [26] learns the sentence vector in the abstract based on the BERT method, and calculates the average value as the paper representation.
- **SEM** Methods in this paper.

Using the Scopus data collection of computer science, pharmacy, and social science papers, we analyze the relationship between the differences in the subspace of papers with different citations in various disciplines and the actual citations of the papers. We take 200 papers of various fields published in 2013 for difference analysis, thus a total of 600 papers, and the publications in various fields before 2013 as a comparation collection. We count the number of citations of these papers up to 2017 as a basis for paper quality evaluation. We use the Spearman's correlation coefficient to calculate the consistency of difference ranking and citation ranking.

The comparison results are shown in Fig. 2. The difference of the paper embedding vector calculation using the characteristics of the expert difference rules, proposed in this paper, is positively correlated with paper quality, and the correlation degree is higher than other frontier embedding models. Our method is more suitable for analyzing paper differences and mining innovation criteria. Using the pre-trained language model to generate text embeddings, the calculated differences and citations are very small, and it is difficult to analyze academic innovations. Therefore, the fusion of expert knowledge and twin network learning based on pre-trained text embeddings is critical for paper difference analysis. In addition, the method in this paper can better reflect the innovation aspects of a paper than methods based on modeling in a single semantic space. For example, high-quality papers in the field of computer science generally show higher differences in methods, pharmacy papers tend to have higher differences in results, and social sciences pay more attention to background and methods.

We further analyze the innovative characteristics of different research directions in the same field. We select 80 papers in the field of Information Systems under the ACM CCS, perform Gaussian mixture clustering on the paper subspace embeddings, and then use the T-SNE method to perform dimensionality reduction on them. The right three figures in Fig. 3 show the clustering results in each subspace. Nodes with different colors indicate papers clustered in different clusters.

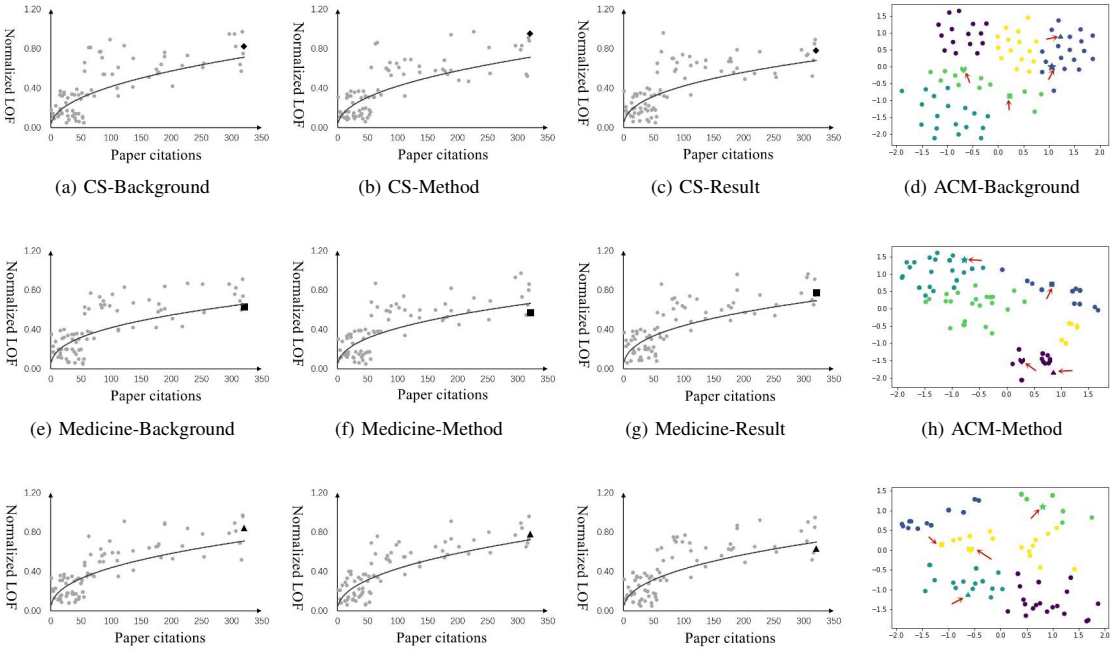
The results show that papers belonging to the same cluster species in one subspace may be in different categories in other subspaces. In the figure, the four papers marked with triangles, squares, stars, and hearts indicate different subspaces. Subspace embedding has learned different knowledge, which also shows the necessity of the concept of subspace proposed in this paper, that is, the analysis of differences in academic papers requires semantic understanding of issues, theories, and technologies at different levels.

G. Analyzing the Characteristics of Different Disciplines

1) Characteristics on disciplines: We now analyze the subspace differences and distribution rules of highly cited papers and the characteristics of knowledge innovation in different disciplines, and visualize the distribution of subspace embeddings in different disciplines. We consider the fields of computer, pharmacy and social sciences, and selecte 80 papers with different citations from each field. We used the normalized LOF value as an index to analyze the correlation between the differences and the citations of the papers.

The result are in the left 9 figures of Fig. 3. The horizontal axis is the citation number of the paper, and the vertical axis is the normalized LOF value. Each node in the graph represents the difference of a paper in a certain subspace. On the whole, the differences and citations of papers in the three subspaces show a positive correlation. Papers with higher differences have a higher probability of obtaining high citations. High-quality papers are generally innovative in all subspaces.

Also, from the slope of the regression line, we can see that different disciplines tend to focus on different innovation aspects. Taking computer science as an example, as shown in Fig. 3a, 3b, 3c, the difference in subspace methods and results



(a) CS-Background (b) CS-Method (c) CS-Result (d) ACM-Background

(e) Medicine-Background (f) Medicine-Method (g) Medicine-Result (h) ACM-Method

(i) Sociology-Background (j) Sociology-Method (k) Sociology-Result (l) ACM-Result

Fig. 3: The figures in the left three columns show the paper subspace outliers of different citations on Scopus. The figures in the right column show the paper clustering results in different subspace on ACM.

TABLE II: Paper subspace outlier in different research topics of Computer Science on ACM.

ACM CCS	Information Systems		Theory of Computation		General Literature		Hardware	
High\low citation	Low citation	High citation	Low citation	High citation	Low citation	High citation	Low citation	High citation
Background	2.07	3.12	2.65	2.73	1.66	2.97	2.53	2.87
Method	3.85	4.91	3.56	4.01	3.24	4.15	2.74	3.05
Result	1.98	2.15	1.06	2.58	2.45	2.68	1.9	2.71

is more relevant to the amount of citation than the background subspace. This shows that in the field of computer science, innovative methods and results are more likely to receive attention and recognition. Similarly, the regression line trends show that pharmaceutical research pays more attention to innovative research results, while pioneering research methods in social sciences receive more attention.

Then, we analyzed the representative papers in detail. We selected highly cited papers in different fields and marked them with solid big nodes. For example, the paper [35] in computational science, labeled as a solid diamond, its difference in the three subspaces is higher than the difference regression value of other papers with similar citations, as shown in Fig. 3a, 3b, and 3c; thus this paper is innovative in all subspaces. In pharmacy, we analyzed the paper [36], labeled as a solid square, which shows relatively low differences in the method subspace. The reason is that its method is based on conventional statistical analysis, and its value is more reflected by the research conclusions. Analyzing the highly cited paper [37] in the social science, labeled as a solid triangle, we found that this paper is based on social phenomena to trace the root cause, compared to the conclusion generally accepted by the public. It's background and methods are more innovative.

2) Characteristics on various topics of the computing disciplines: In order to verify that the embedding method proposed in this paper can result in innovative discoveries in the fine-grained research topics, we analyzed the semantic differences in subspaces of high-cited and low-cited papers in the same discipline. We select papers published in 2015 in each ACM CCS field from ACM dataset as samples, namely, in each field, we select 200 high-cited papers with more than 300 citations since published, and 200 low-cited papers with less than 5 citations. Papers published before 2015 are used as a comparative collection. Based on the representation vectors of the above papers in each subspace, we use Gaussian mixture clustering method to cluster the papers, and calculate the local outlier factor values (LOF value, %) of high-cited and low-cited papers. The results are shown in Tab. II. It can be seen

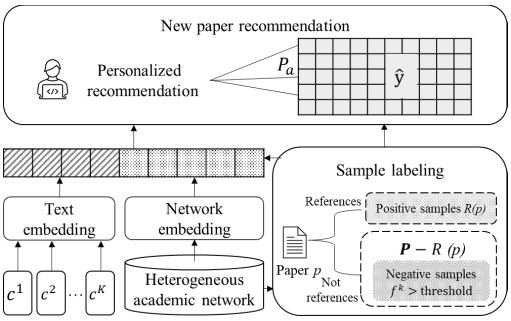


Fig. 4: New paper recommendation model.

that the differences of highly cited papers in each subspace are generally higher than those of low-cited papers, which is consistent with the general perception that "highly cited papers are more likely to be highly innovative work".

IV. New Paper Recommendation

In this section, We propose a graph convolutional networks based neural method to combine the paper contents with the information in the academic network, as well as a strategy for new paper recommendation (NPRec). There are three parts of our model, shown in Fig. 4. The left-bottom part is the graph convolutional neural method to combine the paper content with other related elements in the academic network. The second is the right-bottom part, which is the sample strategy based on paper relevance. The part on top is the recommendation strategy on new papers based on the user interests and the potential influence of papers.

After experimental verification and discussion, the pre-trained paper subspace embedding contains the different features between papers that is highly correlated with paper citations. Therefore, we use the paper subspace embeddings $C_p = \mathbf{c_p^1}, \mathbf{c_p^2}, ..., \mathbf{c_p^K}$ as the text embedding, and uses the attention mechanism to fuse all subspace embeddings to obtain the text representation vector of p by $\mathbf{c_p} = \sum_{k\in[1..K]} \lambda_k \mathbf{c_p^k}$, where $\mathbf{c_p^k} \in R^d$ is the embedding of p on the subspace k, d is the vector dimension.

A. Integrating Information in Academic Network

To recommend new papers, we adopt academic network to learn user interests and paper influence, since paper potential academic impact can be inferred based on the paper's elements in academic network.

The heterogeneous academic network contains a series of entities and the relationships between entities. We use $\mathbf{G} = (E, R, T_E, T_R)$ to denote the academic network, where E and R denote entities and the relationships, respectively. $\forall r_{i,j} = (e_i, e_j) \in R, e_i, e_j \in E$ indicates that there is a relation between entities e_i and e_j. T_E represents the type of entity E, where there are 7 types, including paper, user/author, author unit, venue, specialty classification, keyword, and publication year. T_R is the type of R, marking the relationship between different types of entities, specifically, including "cite", "published in", "written", "published year is", "unit is", "keywords include", "specialty classification is".

It is a usual way to use academic network to learn user interests for paper recommendation [38]. Different with them, we model some element relations as asymmetric, since the different characteristics between user interests and academic Influence. In our model, T_R is divided into one-way and two-way relations according to the dissemination form of academic knowledge and influence. Among them, the citation relationship between paper entities is a one-way association. For two papers $p, q \in E$, $r_{p,q} = (p, q) \in R$ means p cites q, this relationship implies that q is interesting to p, and q's academic influence on p. The academic influence reflected by the other six associations are symmetrical. For example, let e_1 and e_2 be the paper and author, respectively; the authoritative characteristics of e_2 affect the quality of e_1, and the quality of e_1 affects the authority of e_2; similarly, when e_2 is a venue, its research field and authoritative characteristics are reflected in the published papers, and the characteristics of the venue will also be implicit in the embedding of the publication. Therefore, in addition to the paper citation relationship, the other 6 types of relationships are two-way associations.

For the entities e_1, e_2 in $\mathbf{G}$, the function $g(e_1, r_{1,2}, e_2)$ reflects the possibility relation $r_{1,2}$ between them. When both e_1 and e_2 are papers, the score is asymmetric, which implies the academic influence of e_2 on e_1. Considering the association of other types of entities, the function g is symmetric, that is, $g(e_1, r_{1,2}, e_2) = g(e_2, r_{2,1}, e_1)$. To facilitate discussion, use $\pi_{e_2}^{e_1}$ to represent the scoring function $g(e_1, r_{1,2}, e_2)$.

The neighborhood of entity e is denoted by $N(e)$. The embedding of $N(e)$ is a linear combination of all its neighbors, calculated by:

$$v_{N(e)} = \sum_{e' \in N(e)} \tilde{\pi}_{e'}^{e} \cdot v_{e'} \tag{15}$$

Among them, $\tilde{\pi}_{e'}^{e}$ is the normalized weight, which is used to adjust the degree of influence of all it neighbors, calculated by:

$$\tilde{\pi}_{e'}^{e} = \frac{\exp(\pi_{e'}^{e})}{\sum_{e' \in N(e)} \exp(\pi_{e'}^{e})} \tag{16}$$

The final representation vector of the entity e is the fusion of v_e and $v_{N(e)}$, denoted as:

$$v_e^1 = \sigma(W^1 \cdot (v_e + v_{N(e)}) + b^1) \tag{17}$$

$$v_e^H = \sigma(W^H \cdot (v_e^{H-1} + v_{N(e)}^{H-1}) + b^H) \tag{18}$$

where H represents the number of iterations.

For each paper $p \in E$, the network embedding contains two parts, namely the interest characteristics and the influence characteristics. The neighborhoods of implicit interest features include entities connected by two-way association relationships and paper entities cited by p, which are recorded as $\overleftrightarrow{N(p)}$; neighborhoods with implicit influence features $\overrightarrow{N(p)}$ includes entities connected by p bidirectional relationship and paper entities that reference p. The interest expression is

obtained by fusion of p's own network embedding and its neighborhood $\overleftarrow{N(p)}$. The calculation method is:

$$\overrightarrow{v_p^1} = \sigma(W^1 \cdot (\overrightarrow{v_p} + \overleftarrow{N(p)}) + b^1) \tag{19}$$

$$\overrightarrow{v_p^H} = \sigma(W^H \cdot (\overrightarrow{v_p^{H-1}} + \overleftarrow{N(p)^{H-1}}) + b^H) \tag{20}$$

Similarly, the influence of p is expressed by the integration of p's own network embedding and its neighborhood $\overrightarrow{N(p)}$:

$$\overleftarrow{v_p^H} = \sigma(W^H \cdot (\overleftarrow{v_p^{H-1}} + \overrightarrow{N(p)^{H-1}}) + b^H) \tag{21}$$

The embedding of new papers considers the textual semantics of the papers, the interest representation based on heterogeneous academic networks, and the asymmetric academic influence. For the paper p in $\mathbf{G}$, the vector link method is used to fusion to generate its interest representation vector $\overrightarrow{v_p} = [\mathbf{c_p}; \overrightarrow{v_p^H}]$. Considering the asymmetric academic influence of $q \in E$, we compute the influence representation vector of q is $\overleftarrow{v_q} = [\mathbf{c_q}; \overleftarrow{v_p^H}]$.

B. New Paper Recommendation Method

The probability of the asymmetric correlation between papers p and q is related to their research contents, the research interest of p, and the academic influence of q. It is calculated by function f:

$$\hat{y}(p,q) = f(\overrightarrow{v_p}, \overleftarrow{v_q}) \propto \overrightarrow{v_p} \cdot \overleftarrow{v_q} \tag{22}$$

The objective function is to make the predicted correlation degree between papers $\hat{y}$ close to the true marked correlation y. The objective is in the form of a cross-entropy loss function, with an added regularization term to limit the parameters, thus preventing over-fitting, λ is the regularization parameter, which is used to control the influence of regularization term:

$$\ell(\theta) = -\sum_{E \bigcup \bar{E}} y \cdot \log(\hat{y}) + \lambda * ||\theta|| \tag{23}$$

The paper embeddings imply the multi-level semantic characteristics from paper contents, the potential influence features from structured data, and the differentiated knowledge from expert marks, so that recommendation methods can satisfy users' relevance and potential requirements on paper influence.

We use $\hat{y}$ to recommend new papers according to users' requirements. Specifically, for user $a \in \mathbf{G}$, P_a denotes his/her published or cited papers. The calculation method of the user's interest in all candidate papers I_a is the vector expectation corresponding to the paper P_a in $\hat{y}$, denoted as $I_a = |P_a|^{-1} \sum_{p \in P_a} \hat{y}_p$, $\hat{y}_p$ reflects the multi-level correlation between p and candidate papers, and the potential influence of candidate papers relative to p.

C. Dataset and Sample Strategy

We use ACM and Scopus datasets to evaluate our paper recommendation model. In order to verify whether our method is applicable to other academic data with low-resource features, we adopt the US patent dataset (PT) [39] to verify the accuracy of the recommendation model. Each patent contains ownership (author), references, history of updates and maintenance, etc. The patent data set does not contain information such as venues and keywords, only patents and author entities are considered when constructing an academic network. After data cleaning, the statistical results of the three datasets are shown in Tab. III.

TABLE III: Statistics on experimental datasets.

	Paper/patent	Author	Publication year	
ACM	3,056,388	1,752,401	2000-2019	
Scopus	1,304,907	482,602	2008-2017	
PT	182,260	73,974	2017	
	Keyword	Venue	Class	Affiliation
ACM	354,693	11,397	11	15,376
Scopus	127,630	7,653	27	-
PT	-	-	-	-

The citation relationship between papers reflects the scientific researcher's recognition of the relevance and quality of the cited papers. The paper recommendation approaches usually train the models based on the positive and negative samples labeled by citation relationships [40]–[42]. However, there are also cases where the research content of two or more papers is related, even though the papers that do not have a citation relationship. For example, paper p cites q, q cites q', then p and q' are likely to be highly correlated. As another example, p and q refer to q' at the same time, and the probability of a correlation between p and q is greater. Authors often fail to cite these highly relevant works due to space limitations or just because of missing such papers. Some approaches use the above-mentioned indirect references to mark the relevance between papers, but this type of method is generally computationally expensive [43]–[45]. It is difficult to classify fuzzy samples that are judged to be highly relevant under the guidance of expert knowledge without citation relationship as positive samples or negative samples, and requires a lot of labor and calculation costs. Therefore, we propose the following sample strategy for defuzzing samples to improve the model training effect:

- Given paper corpus $P \in E$, for any paper $p \in P$, use $R(p) \subseteq P$ to represent the reference set of p. $\forall q \in R(p)$, mark (p, q) as a positive correlation sample, denoted as $y(p, q) = 1$;
- Filter negative samples from $P - R(p)$ based on the subspace difference fusion function value $f^k(p, q)$ in section III-D. If the difference between p and q in all subspaces is greater than the threshold, it means that there is a large gap between their research content, and (p, q) is marked as a negative sample, recorded as $y(p, q) = 0$.

D. Comparison Methods and Metrics

Comparison methods include recommendation methods based on matrix factorization and on random walks, and entity embedding methods based on contrastive learning or graph convolution:

- **SVD** [46] is a paper recommendation method based on collaborative filtering, in which the scoring matrix is constructed using the data of the author's cited papers.
- **WNMF** [47] is based on matrix factorization, which uses the user's citation relationship to construct the matrix; the number of features is set to 10.
- **NBCF** [8] uses the neighborhood-based collaborative filtering algorithm.
- **MLP** [12] is based on a multi-layer perceptron, which learns the non-linear interaction function between user interest and paper embedding from the citation relationships.
- **JTIE** [2] is a paper recommendation method based on the joint embedding of paper text and influence.
- **KGCN** [19] is a academic graph convolution based paper recommendation model, which predict the user's potential interest on the indirectly connected paper nodes.
- **KGCN-LS** [9] introduces a label propagation model on the basis of KGCN.
- **RippleNet** [21] is a paper recommendation strategy based on random walk. Use the interested papers as seed nodes, and spread out to other papers on the academic network.
- **NPRec** is our proposed recommendation method.

All parameters of the baseline methods are set to optimal values based on experience.

We prepare k candidate papers for each user. Each candidate set contains at least one paper that is actually cited by the user. The candidates are sorted according to the correlation between user interests and paper vectors. We choose $nDCG@k$ as the measurement [48]. $nDCG@k = \frac{DCG@k}{IDCG}$ is often used to measure the effectiveness of online search engine algorithms. Considering the relevance between the user and each actually cited paper is the same, $DCG@k$ is calculated as $DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$. If the i-th paper is indeed cited by researcher, we set $rel_i = 5$ based on experience, otherwise $rel_i = 0$. The position discount $\log_2(i+1)$ is used to distinguish user preference on different rankings. $IDCG = \sum_{i=1}^{|Ref|} \frac{5}{\log_2(i+1)}$ is the ideal discounted cumulative income. $|Ref|$ refers to the number of papers actually cited by the researcher in the candidate papers. The larger the value of $nDCG@k$, the higher the ranking of the papers that users actually cite among the candidates, and the better the recommendation effect.

E. Personalized New Paper Recommendation Comparison

The experiment in this section verifies the accuracy of NPRec and the comparison methods in personalized recommendation tasks. We randomly select 300 and 100 researchers from the ACM and Scopus datasets to verify the models. The data set is divided into two parts according to the publication year: papers published before year Y are used to train the model, and the papers published after year Y are used for testing. For each user, we prepare k candidate papers. We set Y to 2014.

The evaluation results on the two datasets are shown in Tab. IV. It can be seen that our method effectively improves the performance of new paper recommendations. KGCN and other methods based on graph convolution take into account the various structural features of academic papers, and the recommendation effect is better than contrastive learning, matrix decomposition and other methods; but because the multi-level correlation between user interests and the text of the paper is ignored, the recommendation effect is slightly worse than our NPRec method. The performance of each model on the Scopus data set is relatively poor. The reason is that the amount of data is small and the data set lacks some features, such as the author unit, which makes the model unable to accurately model user interests or the characteristics of the paper. In addition, as the number of candidate papers increases, the ranking of the actually cited papers decreases, and the value of $nDCG@k$ of each model decreases.

TABLE IV: New paper recommendation comparison.

nDCG@k	ACM			Scopus		
	k=20	k=30	k=50	k=20	k=30	k=50
SVD	0.6814	0.6582	0.6033	0.6510	0.6097	0.5766
WNMF	0.8265	0.7892	0.7316	0.7889	0.7725	0.7052
NBCF	0.8331	0.7994	0.7322	0.7932	0.7856	0.7272
MLP	0.8391	0.8011	0.7649	0.8263	0.8201	0.7305
JTIE	0.8693	0.8512	0.8053	0.8309	0.8257	0.7399
KGCN	0.8731	0.8592	0.8437	0.8507	0.8365	0.7592
KGCN-LS	0.9093	0.9010	0.8904	0.8660	0.8548	0.8063
RippleNet	0.9217	0.9088	0.8970	0.9040	0.8673	0.8465
NPRec	**0.9736**	**0.9688**	**0.9645**	**0.9576**	**0.9349**	**0.9021**

Compared with the traditional new paper recommendation strategies, our sample labeling strategy incorporating expert rules avoids the problem of model under-fitting and supports the expansion of expert rules. Besides, our embedding method is more robust in dealing with the balance of correlation and potential influence.

Considering the influence of the number of the user representative papers, our method learns user research interests from different numbers of published papers and makes personalized recommendations. To verify our model more comprehensively, we adopt the metrics based on binary correlation [49] to evaluate the recommendation results besides $nDCG@20$, such as MRR and MAP. The results are shown in Tab. V. We can see that the increase in the number of representative papers helps the model to better learn user interests, and thus the recommendation performance is higher. Compared with other methods, our approach achieves the best recommendation results for experts with different publication volumes. Through the results on MRR and MAP, we found that the actually cited papers are usually in the top 3 of the recommendation lists using our method, which is better than baselines by at least 2 rankings.

For the amount of data labeling of the de-blurred samples, we use different positive and negative sample ratios to train the each model; the results are shown in Tab. VI. When the ratio of the positive sample to the negative sample is 1:10, the performance of each model is optimal. Compared with other methods, our method has the highest recommendation accuracy under different sample ratios.

TABLE V: Comparison on different publication numbers.

Dataset	ACM				Scopus	
Metric	nDCG@20		MRR	MAP	nDCG@20	
#rp	3	5	5	5	3	5
WNMF	0.760	0.790	0.15	0.33	0.715	0.761
NBCF	0.769	0.821	0.21	0.40	0.721	0.782
MLP	0.853	0.871	0.24	0.44	0.747	0.805
JTIE	0.861	0.874	0.35	0.53	0.752	0.828
KGCN	0.881	0.892	0.36	0.65	0.857	0.862
KGCN-LS	0.916	0.922	0.46	0.67	0.854	0.863
RippleNet	0.921	0.928	0.58	0.71	0.894	0.915
NPRec	**0.969**	**0.975**	**0.71**	**0.82**	**0.932**	**0.959**

TABLE VI: Comparison on different ratios between positive and negative samples.

nDCG@20	ACM			Scopus		
	1:1	1:10	1:50	1:1	1:10	1:50
WNMF	0.761	0.793	0.773	0.732	0.778	0.716
NBCF	0.775	0.806	0.798	0.742	0.780	0.735
MLP	0.821	0.864	0.815	0.775	0.809	0.796
JTIE	0.869	0.905	0.892	0.801	0.874	0.853
KGCN	0.852	0.879	0.857	0.831	0.860	0.790
KGCN-LS	0.878	0.902	0.878	0.849	0.866	0.808
RippleNet	0.883	0.931	0.897	0.852	0.895	0.846
NPRec	**0.946**	**0.974**	**0963**	**0.907**	**0.958**	**0.924**

F. Ablation Experiments

We now analyze the importance of each module in NPRec and the influence of different parameter settings for each model. We analyze the importance of de-fuzzing sample strategy, subspace embedding and network embedding, as well as the influence of the neighbor number settings K and the maximum depth of graph convolution H. The model variants are as follows:

- **NPRec+SC** uses the deblurring sample strategy to label the data, and the presentation vector of the paper adopts the subspace embedding method. NPRec+SC is not affected by the parameters K and H.
- **NPRec+SN** uses the deblurring sample strategy to label data, and the presentation vector of the paper adopts graph convolution based on heterogeneous academic network.
- **NPRec+CN** labels positive and negative samples according to the citation relationship between the papers, and the representation vector of the paper adopts the combined embedding.
- **NPRec** is our model, which includes three modules: de-blurring samples, subspace embedding and network embedding.

The experimental results are shown in Tab. VII and VIII. We can see that the recommendation method using the three modules jointly is the best. The paper embedding method that combined subspace and heterogeneous academic network features can accurately learn the multiple data features related to the paper, and the data labeling strategy of de-fuzzing samples effectively improves model training. For the number of neighbor nodes K in the combined embedding part, the best results are when the values are 8 or 16, because within this range, it can cover multiple feature nodes that are most relevant to the paper, and ignore suspected related or irrelevant features. The result is best when the maximum depth of the graph convolution H is equal to 2. The model can accurately learn multiple features of papers and avoid over-fitting.

TABLE VII: Comparison on model variants with different K.

$nDCG$@20	K =2	K =4	K =8	K =16	K =32
NPRec+SC	0.898	-	-	-	-
NPRec+SN	0.900	0.886	0.892	0.884	0.904
NPRec+CN	0.918	0.919	0.919	0.943	0.908
NPRec	**0.952**	**0.958**	**0.968**	**0.974**	**0.947**

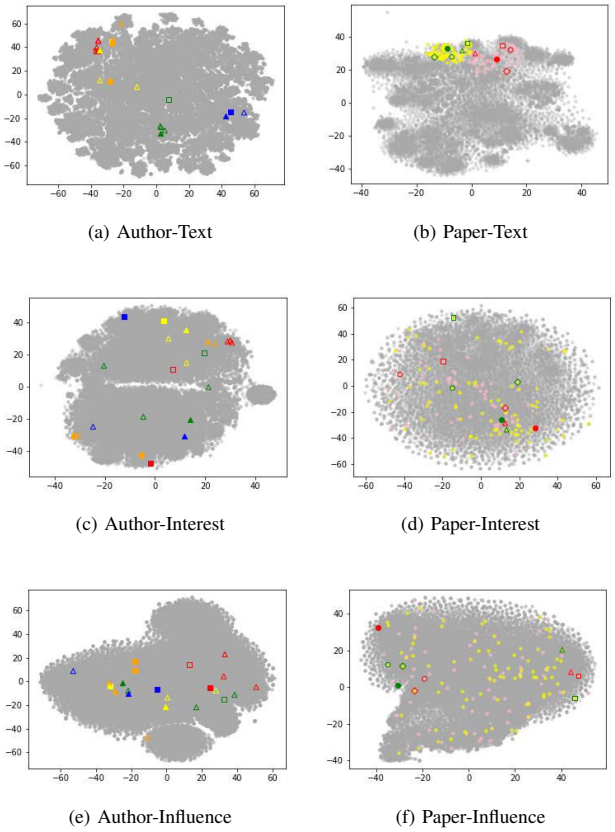
TABLE VIII: Comparison on model variants with different H.

$nDCG$@20	H =1	H =2	H =3	H =4
NPRec+SC	0.898	-	-	-
NPRec+SN	0.882	0.896	0.871	0.897
NPRec+CN	0.934	0.949	0.897	0.881
NPRec	**0.961**	**0.968**	**0.946**	**0.951**

G. Discussion on Different Author Types

In the personalized recommendation scenario, the embedding of the author contains three semantic features: the research content based on the text of the published papers, the research interest reflected by citing other works, and the academic influence of one's own work when it is cited. The research topics of authors, who have published multiple papers, are relatively focused. In this paper, we analyze the researchers who have multiple publications. In order to verify the necessity of our embedding method for fusing text and heterogeneous academic networks, we analyze the combined embedding semantics of different types of authors. For each author, we calculate the expectation of the combined embedding of his/her historical published papers as the user embedding in the personalized recommendation scenario. We use T-SNE [50] to reduce the dimensionality of the author's embedding in the ACM dataset to a two-dimensional space, as shown in the left three figures of Fig. 5.

The nodes of the same color in Fig. 5a, which denote the co-authors, show a certain degree of aggregation, because they tend to study similar content. Authors marked in red are more clustered than authors marked by other colors. This is because the number of papers co-authored by these authors is larger than the number of papers co-authored by other authors, and the research content is more relevant. The embedded research content of some co-authors is relatively scattered. This is because they often work in different scientific research institutions/units/projects and have different scientific research experiences except for co-authored papers. For example, the co-authors of the paper [51] (in yellow) are from different units and institutions, such as the International Computing Research Center and the Department of Electronic Engineering and Computer Science of the University of California, Berkeley,



(a) Author-Text (b) Paper-Text

(c) Author-Interest (d) Paper-Interest

(e) Author-Influence (f) Paper-Influence

Fig. 5: The left and right parts show the author and paper combined embeddings based on NPRec, respectively.

the Department of Computer Science at the University of California, Los Angeles, and the University of Southern California. Their main scientific research areas are also different.

The difference in research content between authors is also reflected in the citation habits of co-authors when they write papers, as shown in Fig. 5c, where there is a greater difference in interest embeddings between co-authors marked with the same color. Some authors with a large number of published papers and a high number of citations have close interest embeddings. Consider the papers [52] (in orange) and [53], [54] (in red) as examples. The highly cited authors of the paper are marked as orange and red solid squares, respectively, located at the bottom of Fig. 5c. The research directions of these co-authors are centered on wireless network protocols, while highly productive and highly cited authors are often authoritative experts in their fields, with excellent and consistent citation patterns. We can see that the author's interest embedding not only reflects the research direction, but also implies the citation habit.

According to the author influence embedding in Fig. 5e, we observe that the highly productive and highly cited authors, marked as solid squares, are highly clustered. This is because most researchers have the habit of citing authoritative expert papers when writing papers, so authors who have accumulated a high number of citations have similar academic influence, especially authoritative experts with similar research directions, such as the highly cited authors of paper [51] and [55] (in green), that are marked with yellow and green solid nodes, respectively. Their research directions are centered on data flow management, and they have high visibility and influence in their research fields.

According to the above three-part results of author embedding, we found that only the text embedding of the papers published by the user is not enough to reflect the user's interest and academic influence in the paper recommendation scenario. The academic influence spread between users through the paper citation relationship is asymmetric. When users cite other work or their own work is cited, the academic influence received or disseminated is quite different. Therefore, in the academic paper recommendation task, it is necessary to consider the asymmetric academic influence and embed a variety of features from the paper.

H. Discussion on Highly-cited Papers

Similarly, to show the necessity of three paper feature embeddings separately, we reduce the dimensionality of the paper embeddings in the ACM dataset, as shown in the right three figures of Fig. 5. We mark the two highly cited papers as solid nodes, mark their related papers as hollow nodes, and mark the other papers as gray nodes.

First, we analyze the necessity of content embeddings. We mark the two highly cited papers [56], [57] as solid red and green nodes respectively. In order to mark the works that are closest to the research content of the highly cited papers, we screened out the 50 papers closest to the text embedding of each highly cited paper from the collection of papers under the CCS to which the highly cited papers belong, and marked them as pink and yellow nodes. As shown in Fig. 5b, these papers with similar research content are distributed around the highly cited papers. From these 100 papers, 8 papers were randomly selected for further analysis of the paper characteristics, and marked as hollow nodes with the same color as the related highly cited papers. The interest embeddings corresponding to these papers are shown in Fig. 5d. We can see that the aggregation of the original text embedding work has changed in the interest embeddings. For example, the interest embeddings of paper [58], [59] are still similar to that of the highly cited paper [56] (red hollow diamond and triangle), while the research interests shown by paper [60], [61] are not similar to paper [56] (red hollow square and circle). This is because some researchers focus on innovative technologies in the same research context, and some papers tend to cite the work of a fixed scientific research team or journal, and so on. Even for papers with similar research content, they pay attention to different citation characteristics when citing other works.

Then we analyze the importance of paper interest embeddings, as shown in Fig. 5f. Take the highly cited paper [57](green solid dot) with 993 citations as an example. Two papers with similar research content [62], [63](green hollow diamond and circle) have similar influence ranges. The citations of them are 219 and 183, respectively; the influence ranges of the paper [64], [65](green hollow diamond and circle) are different from that of the highly cited paper, which are located in the upper right corner of the figure, and having 97 and 187 citations, respectively. We found that the influence embeddings which are close to the highly cited paper tend to be clustered with the publication venue of the highly cited paper, which means their audiences are similar, and the scopes of academic influence are similar. For example, the research content of paper [57] is object-oriented programming systems and languages, while similar papers were published in the Programming Language Design and Implementation Conference PLDI and the Supercomputing Conference ICS, both of which are authoritative publications in the area of computer systems. Highly cited papers disseminate similar academic knowledge when they are cited, so they are more clustered. The papers with influence embedded far away from the highly cited papers were published in the data management conference SIGMOD and the very large database conference VLDB. Although they were also from top international conferences, the research content tends to focus on database management, which is different from the interests of the researchers who are affected by paper [57]. Since the academic knowledge propagated is different when the papers are referred, there is a long distance between the paper influence embeddings.

We further analyze the influence characteristic of the paper. Take the highly cited paper [56] with 10,723 citations as an example (red solid node). Two papers with similar research content [59], [60] (red hollow cricle and diamond) have similar academic influences. The influence range of the paper [58], [61] (red hollow square and triangle) is different from paper [56]. In fact, paper [59], [60] have 360 and 36 citations, respectively, lower than the paper [58], [61], which are 760 and 282. This is because the research contents of paper [59], [60] are closer to paper [56], which is about distributed computing, and the audiences are similar. We can see that the influence characteristics of new papers embedded in our method not only reflect the potential authority, but also imply the scope of the paper audiences and the academic knowledge spread when it is referred. The citation relationship between papers is asymmetric, and the academic influence received when papers are referred is quite different with the influence propagated when one cite other papers. Therefore, it is necessary to consider the asymmetric academic influence and embed a variety of features in paper embeddings for recommendation.

I. Discussion on Model Reusability

We now focus on verifying whether our recommendation method is applicable to other types of academic resources with few types of features. We use the US patent data set PT to verify the personalized recommendation on patents.

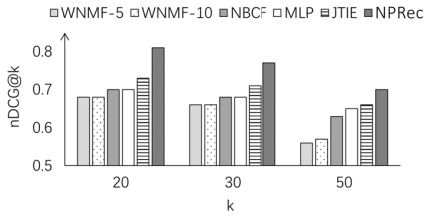


Fig. 6: Performance comparison on personalized patent recommendation on PT.

The author preferences are learned from the patents published from January to October in 2017. We use the patent citations from November to December in 2017 to verify the accuracy of the recommendation. The number of authors selected as experimental samples is 50, and $nDCG@20$ is used as an evaluation indicator. The experimental results are shown in Fig. 6. We can see that NPRec can still learn potential influence features and achieve a higher accuracy on recommendation, which confirm the applicability of our model in dealing with academic resources with fewer features.

V. Conclusion

The quality evaluation of new papers is an important factor in the academic recommendation task. In this paper, we introduce the concept of subspace to describe the commonly recognized aspects of paper contents, namely background, methods and results. A set of expert rules are formalized to annotate the differences between papers, based on which a twin-network is proposed for learning the embeddings of papers in different subspaces. A series of empirical studies show that there are clear correlations between a paper influence and its difference with others in the subspaces. The results also show the characteristics of innovation in different scientific fields. To recommend new paper, we propose graph convolutional neural method to combine the paper content with other related elements from academic networks, where the user interests and the academic influences are modeled asymmetric. Experimental results on real datasets show that our method is more effective compared with other baseline methods for paper recommendation. We also discuss the characteristics of scientific disciplines and authors to show the effectiveness of modeling the asymmetric user interests and influences. Finally, we verify the reusability of our method for academic data with low-resource features.

Acknowledgment

This work was supported by the Major Project of NSF Shandong Province (ZR2018ZB0420), the Key Research and Development Project of Shandong Province (2019JZZY010107) and the Ouma Research Project. The scientific calculations in this paper have been done on the HPC Cloud Platform of Shandong University.

REFERENCES

[1] A. Louis and A. Nenkova, "A corpus of science journalism for analyzing writing quality," *Dialogue & Discourse*, vol. 4, no. 2, pp. 87–117, 2013.

[2] Y. Xie, S. Wang, W. Pan, H. Tang, and Y. Sun, "Embedding based personalized new paper recommendation," in *CCF Conference on Computer Supported Cooperative Work and Social Computing*. Springer, 2020, pp. 558–570.

[3] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, "The h-index of a network node and its relation to degree and coreness," *Nature communications*, vol. 7, no. 1, pp. 1–7, 2016.

[4] P. Glasziou, J. Vandenbroucke, and I. Chalmers, "Assessing the quality of research," *Bmj*, vol. 328, no. 7430, pp. 39–41, 2004.

[5] P.-H. Lin, J.-R. Chen, and C.-H. Yang, "Academic research resources and academic quality: a cross-country analysis," *Scientometrics*, vol. 101, no. 1, pp. 109–123, 2014.

[6] L. Xia, J. Xu, Y. Lan, J. Guo, and X. Cheng, "Modeling document novelty with neural tensor network for search result diversification," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 395–404.

[7] ——, "Learning maximal marginal relevance model via directly optimizing diversity evaluation measures," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 113–122.

[8] K. Sugiyama and M.-Y. Kan, "A comprehensive evaluation of scholarly paper recommendation using potential citation papers," *International Journal on Digital Libraries*, vol. 16, no. 2, pp. 91–109, 2015.

[9] H. Wang, F. Zhang, M. Zhang, J. Leskovec, M. Zhao, W. Li, and Z. Wang, "Knowledge-aware graph neural networks with label smoothness regularization for recommender systems," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 968–977.

[10] Y. Zhu, Q. Lin, H. Lu, K. Shi, P. Qiu, and Z. Niu, "Recommending scientific paper via heterogeneous knowledge embedding based attentive recurrent neural networks," *Knowledge-Based Systems*, vol. 215, p. 106744, 2021.

[11] J. G. Foster, A. Rzhetsky, and J. A. Evans, "Tradition and innovation in scientists' research strategies," *American Sociological Review*, vol. 80, no. 5, pp. 875–908, 2015.

[12] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[13] I. Kanellos, T. Vergoulis, D. Sacharidis, T. Dalamagas, and Y. Vassiliou, "Ranking papers by their short-term scientific impact," in *2021 IEEE 37th International Conference on Data Engineering*. IEEE, 2021, pp. 1997–2002.

[14] M. Song, G. E. Heo, and S. Y. Kim, "Analyzing topic evolution in bioinformatics: investigation of dynamics of the field with conference data in dblp," *Scientometrics*, vol. 101, no. 1, pp. 397–428, 2014.

[15] G. Barabucci, "diffi: diff improved; a preview," in *Proceedings of the ACM Symposium on Document Engineering 2018*, 2018, pp. 1–4.

[16] P. Ciancarini, A. D. Iorio, C. Marchetti, M. Schirinzi, and F. Vitali, "Bridging the gap between tracking and detecting changes in xml," *Software: Practice and Experience*, vol. 46, no. 2, pp. 227–250, 2016.

[17] C. Zhu, Y. Li, J. Rubin, and M. Chechik, "A dataset for dynamic discovery of semantic changes in version controlled software histories," in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*. IEEE, 2017, pp. 523–526.

[18] J. Zhang, Q. Su, B. Tang, C. Wang, and Y. Li, "Dpsnet: Multitask learning using geometry reasoning for scene depth and semantics," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[19] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo, "Knowledge graph convolutional networks for recommender systems," in *Proceedings of The Web Conference*, 2019, pp. 3307–3313.

[20] X. Ma and R. Wang, "Personalized scientific paper recommendation based on heterogeneous graph representation," *IEEE Access*, vol. 7, pp. 79 887–79 894, 2019.

[21] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, and M. Guo, "Ripplenet: Propagating user preferences on the knowledge graph for recommender systems," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 417–426.

[22] N. Sakib, R. B. Ahmad, and K. Haruna, "A collaborative approach toward scientific paper recommendation using citation context," *IEEE Access*, vol. 8, pp. 51 246–51 255, 2020.

[23] N. Coulter, J. French, E. Glinert, T. Horton, N. Mead, R. Rada, A. Ralston, C. Rodkin, B. Rous, A. Tucker *et al.*, "Computing classification system 1998: current status and future maintenance report of the ccs update committee," *Computing Reviews*, vol. 39, no. 1, pp. 1–62, 1998.

[24] P. Bille, "A survey on tree edit distance and related problems," *Theoretical computer science*, vol. 337, no. 1-3, pp. 217–239, 2005.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*, 2013, pp. 1–12.

[26] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[27] J. LAFFERTY, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference. on Machine Learning*, 2001, pp. 282–289.

[28] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 990–998.

[29] F. Dernoncourt and J. Y. Lee, "Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2017, pp. 308–313.

[30] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses," *The FASEB journal*, vol. 22, no. 2, pp. 338–342, 2008.

[31] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: clustering, classification and density estimation using gaussian finite mixture models," *The R journal*, vol. 8, no. 1, pp. 289–317, 2016.

[32] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*. ACM, 2000, pp. 93–104.

[33] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, p. 1763–1768, 2018.

[34] A. Kanakia, Z. Shen, D. Eide, and K. Wang, "A scalable hybrid research paper recommender system for microsoft academic," in *The world wide web conference*, 2019, pp. 2893–2899.

[35] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.

[36] M. T. Ruel, H. Alderman, Maternal, C. N. S. Group *et al.*, "Nutrition-sensitive interventions and programmes: how can they help to accelerate progress in improving maternal and child nutrition?" *The lancet*, vol. 382, no. 9891, pp. 536–551, 2013.

[37] H. LaFollette and M. L. Woodruff, "The righteous mind: Why good people are divided by politics and religion," *Philosophical Psychology*, vol. 28, no. 3, pp. 452–465, 2015.

[38] H. Shao, D. Sun, J. Wu, Z. Zhang, A. Zhang, S. Yao, S. Liu, T. Wang, C. Zhang, and T. Abdelzaher, "paper2repo: Github repository recommendation for academic papers," in *Proceedings of The Web Conference 2020*, 2020, pp. 629–639.

[39] J. Kim and S. Lee, "Patent databases for innovation studies: A comparative analysis of uspto, epo, jpo and kipo," *Technological Forecasting and Social Change*, vol. 92, pp. 332–345, 2015.

[40] Y. Zhang and Q. Ma, "Dual attention model for citation recommendation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3179–3189.

[41] T. Saier and M. Färber, "Semantic modelling of citation contexts for context-aware citation recommendation," *Advances in Information Retrieval*, vol. 12035, pp. 220–233, 2020.

[42] R. Nogueira, Z. Jiang, K. Cho, and J. Lin, "Navigation-based candidate expansion and pretrained language models for citation recommendation," *Scientometrics*, vol. 125, no. 3, pp. 3001–3016, 2020.

[43] H. Liu, H. Kou, C. Yan, and L. Qi, "Keywords-driven and popularity-aware paper recommendation based on undirected paper citation graph," *Complexity*, vol. 2020, pp. 1–15, 2020.

[44] S. Ma, C. Zhang, and X. Liu, "A review of citation recommendation: from textual content to enriched context," *Scientometrics*, vol. 122, no. 3, pp. 1445–1472, 2020.
[45] L. Ma, D. Song, L. Liao, and J. Wang, "A hybrid discriminative mixture model for cumulative citation recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 617–630, 2019.
[46] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 426–434.
[47] S. Zhang, W. Wang, J. Ford, and F. Makedon, "Learning from incomplete ratings using non-negative matrix factorization," in *Proceedings of the 2006 SIAM international conference on data mining*. SIAM, 2006, pp. 549–553.
[48] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu, "A theoretical analysis of ndcg ranking measures," in *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, vol. 8. Citeseer, 2013, p. 6.
[49] B. McFee and G. R. Lanckriet, "Metric learning to rank," in *ICML*, 2010.
[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
[51] S. Ratnasamy, B. Karp, L. Yin, F. Yu, D. Estrin, R. Govindan, and S. Shenker, "Ght: a geographic hash table for data-centric storage," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, 2002, pp. 78–87.
[52] P. Levis, E. Brewer, D. Culler, D. Gay, S. Madden, N. Patel, J. Polastre, S. Shenker, R. Szewczyk, and A. Woo, "The emergence of a networking primitive in wireless sensor networks," *Communications of the ACM*, vol. 51, no. 7, pp. 99–106, 2008.
[53] J. Liu, Y. Yuan, D. M. Nicol, R. S. Gray, C. C. Newport, D. Kotz, and L. F. Perrone, "Simulation validation using direct execution of wireless ad-hoc routing protocols," in *Proceedings of the eighteenth workshop on Parallel and distributed simulation*, 2004, pp. 7–16.
[54] ——, "Empirical validation of wireless models in simulations of ad hoc routing protocols," *Simulation*, vol. 81, no. 4, pp. 307–323, 2005.
[55] T. Leighton, F. Makedon, S. Plotkin, C. Stein, E. Tardos, and S. Tragoudas, "Fast approximation algorithms for multicommodity flow problems," *Journal of Computer and System Sciences*, vol. 50, no. 2, pp. 228–243, 1995.
[56] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content-addressable network," in *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, 2001, pp. 161–172.
[57] T. Harris and K. Fraser, "Language support for lightweight transactions," *ACM Sigplan Notices*, vol. 49, no. 4S, pp. 64–78, 2014.
[58] K. Petersen, M. J. Spreitzer, D. B. Terry, M. M. Theimer, and A. J. Demers, "Flexible update propagation for weakly consistent replication," in *Proceedings of the sixteenth ACM symposium on Operating systems principles*, 1997, pp. 288–301.
[59] L. Lazos and R. Poovendran, "Serloc: Robust localization for wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 1, no. 1, pp. 73–100, 2005.
[60] J. R. Lange and P. A. Dinda, "Transparent network services via a virtual traffic layer for virtual machines," in *Proceedings of the 16th international symposium on High performance distributed computing*, 2007, pp. 23–32.
[61] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson, "Inferring link weights using end-to-end measurements," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, 2002, pp. 231–236.
[62] C. CaBcaval and D. A. Padua, "Estimating cache misses and locality using stack distances," in *Proceedings of the 17th annual international conference on Supercomputing*, 2003, pp. 150–159.
[63] T. M. Chilimbi, "Efficient representations and abstractions for quantifying and exploiting data reference locality," *ACM SIGPLAN Notices*, vol. 36, no. 5, pp. 191–202, 2001.
[64] M. K. Aguilera, W. Golab, and M. A. Shah, "A practical scalable distributed b-tree," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 598–609, 2008.
[65] S. J. White and D. J. DeWitt, "Quickstore: A high performance mapped object store," in *Proceedings of the 1994 ACM SIGMOD international conference on Management of data*, 1994, pp. 395–406.