# Paper Recommendation Based on Authorpaper Interest and Graph Structure

Hao L
School of Software
Shandong University
Jinan, China
lestis@mail.sdu.edu.cn

Shijun Liu
School of Software
Shandong University
Jinan, China
lsj@sdu.edu.cn

Li Pan
School of Software
Shandong University
Jinan, China
panli@sdu.edu.cn

Abstract—The recommendation system can recommend information to users efficaciously, which helps many users to obtain information in different fields. The paper recommendation is a research topic to provide authors with personalized papers of interest. However, most existing approaches equally treat title and abstract as the input to learn the representation of a paper, ignoring the author's interest and structure information of the academic network. In the paper recommendation system, authors and papers and the interaction of their information have a crucial impact on the efficiency and accuracy of the recommendations. However, most recommendation systems are usually designed based only on users. Therefore, we propose a method based on the author's periodic interest and academic graph network structure to obtain as much effective information as possible to recommend papers. Extensive offline experiments on large-scale real data show that our method outperforms the representative

Index Terms—Recommendation system, Academic network, author-paper interest, Graph structure

#### I. INTRODUCTION

In the context of the era of big data, the recommendation system can gather historical behavior data of users to obtain personalized recommendations for users, and recommend items or information [32]. The ever-increasing number of research papers have been published over the last decades, resulting in a problem known as 'information overload'. Tens of thousands of papers are published every day, and the number of papers published worldwide has a very high growth trend [16].

Researchers have to spend more time searching for articles they are interested in. Therefore, the paper recommendation is more important than before. Many paper recommendation systems are very limited. Besides Web of Science, Microsoft Academic, Google Scholar, and a few others, the many search engines are limited to particular research domains [33] Also, some proposed user recommender systems employ collaborative filtering for generating user recommendations [21].

In order to mitigate this problem, many academic search engines have begun to add recommendation systems [17] [18]. Collaborative filtering (CF) has been widely adopted in recommendation systems, but CF often generates poor performance since the user-item interaction matrix is very sparse in many

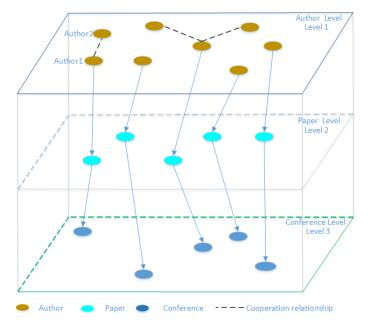


Fig. 1: The hierarchy of academic network. The academic network includes author nodes, paper nodes, and conference nodes. This figure is a typical bipartite graph of authors and conferences through the linking of papers.

fields. So, the auxiliary information is introduced to enhance recommendation performance. Nonetheless, many maverick groups and researchers are still enthusiastically developing tools to help research papers with suggestions for knowledge discovery and citation assistance [19] [20].

Recently, many approaches have been advanced to develop diverse auxiliary information. For paper recommendations, two types of auxiliary information are widely adopted for better recommendations, including structural and textual information. The former type refers to paper citation relationships, i.e., papers that a paper cites or those cite it [1].

The structure of paper citations may indicate the influence of one paper on another, but it ignores the real content and semantics of the paper. [12] points out another issue that some newly published papers may not be cited and some researchers prefer to cite their own less relevant papers. The sequential

order of author-paper interrelations can disclose time-evolving and chronological author behaviors in numerous recommender systems. The paper that the author will interconnect with may rely on the papers visited in the past. But, the notable addition of authors or papers makes sequential recommender systems still confront the challenge that the hardness of modeling the period author interests [35].

The recent development of GNN can be categorized into spectral and non-spectral approaches. Spectral approaches employ the tools in signal processing and transform the convolutional operation in the graph domain to much simpler operations of the Laplacian spectrum, and various approaches have been proposed to localize the convolution in either the graph or spectral domain [30]. Non-spectral methods expound convolutions diametrically on the graph within spatially close nodes. Graph attention network provides a promising framework by combining graph neural networks with attention mechanism in handling graphs with arbitrary structures [31]. Groundbreaking manifestation has been obtained on tasks of node embedding and regimentation [36]. Most of the gnn methods mentioned above only computed mainly based on the content of the nodes. However, in social networks, a community or pathway is oftentimes composed of nodes that are densely interconnected with each other but several hops away. Therefore, the structure of the graph is also very important [37].

To fully exploit rich, high-order structural details in graph attention networks, we need to contextualize each node within a local receptive field composed of its high-order neighbors.

The major contributions of this paper are summarized as follows:

- To model the period interests of the author, we present a period interest graph neural network to catch papers' period-term contextual information.
- To computing the structural attention and obtain the information of the graph structure, we developed a method of graph structure attention and combined it with the author's interest.
- Extensive offline experiments on large-scale real data show that our method outperforms the representative baselines.

#### II. RELATED WORK

The first type of paper recommendation is based on the citation structure, i.e., the papers it cites and those citing it. The constructed paper graph is further mined to calculate paper similarity and generate paper recommendations. For example, [1] construct paper representation based on TF-IDF technology, which is heavily relied on the term frequency. The similarity between papers based on citation references is used as weights to build user and paper profiles. However, not all relevant works can be fully covered in one paper. To alleviate this issue, [2] further improve their previous model by extending a paper's reference list with the involvement of the top-N relevant papers. Moreover, [3] build a basic paper graph based on the reference citations. A random walk algorithm

is devised to generate recommendation items. To sum up, the underlying assumption of this research line stresses that the citation topology can accurately reflect paper relatedness. However, in many cases, such an assumption cannot hold because: (1) most recently published papers cannot be referred to by previous papers; (2) some valuable references may be missing due to the unawareness of researchers; and (3) some irrelevant or less relevant papers may be adopted in the reference list, for example, some other papers from the same authors.

To handle sequential data more effectively, a memory network (MN) has been proposed to memorize long-term dependencies [22]. MN introduces a memory matrix to store historical information by updating it when new information is available. Recently, MN has been adopted in the recommendation system. [23] propose a RUM model that integrates the insights of CF with MN, constructing user preference by accommodating long-term historical interactions. [24] combine GRU and MN to enhance sequential recommendation, where the knowledge base is leveraged to better update the MN.

Guibing and Bowei propose a two-level attentive neural network called TAAS to capture the semantic correlation between title and abstract for paper ranking and recommendation [25].

To settle the challenges in the intent recommendation, Fan and Shaohua method interactions in intent recommendation system with a Heterogeneous Information Network(HIN) and provide a novel metapath-guided GNN model for intent ranking and suggestion, called MEIRec [15].

## III. METHOD

An author's interest describes the author's present preference and is depended on some lately accessed papers in a period. The paper's author will interact within the near future are likely to be related to the papers she just visited. Therefore, it is very important in the sequential recommendation to efficaciously method the author's interest, as incarnated by recently accessed papers.

First, we must explicitly model the author's interest, our method conduct a sliding window strategy to split the paper sequence into fine-grained sub-sequences. For each author u, we extract every |L| successive papers as input  $L_u$   $l=(I_l,I_{l+1},\ldots,I_{l+|L|-1})$  is the 1-th sub-sequence of author u. GNNs has their ability to structure learning, so it's a good match for the task of aggregating the papers in  $L_u$ ,1 to learn author interests.

Then, we demand to set up a graph to seize the connections between papers because paper sequences are not inherently graphs for GNN training. We extract several subsequent papers (two papers in our experiments) and add edges between them.

We do this for each author and count the number of edges of extracted paper pairs across all authors. After this performance, we row-normalize the adjacency matrix. After the above process, we can extract relevant papers that appear closer to one another in the sequence. As the fig3 shows, we designate the extracted adjacency matrix as A,  $A_i$  and k express the normalized node weight of paper k regarding

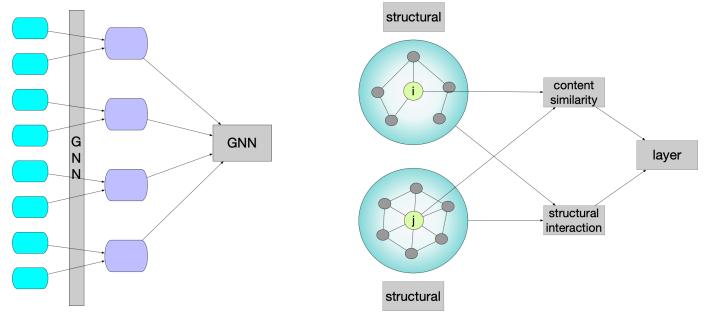


Fig. 2: The frame of our model.

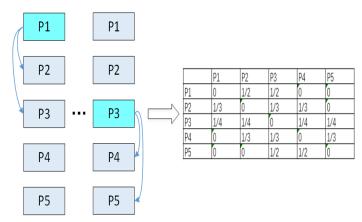


Fig. 3: Paper adjacent matrix construction example.

paper i. And the neighboring papers of paper i is denoted as  $N_i$ .

We next use a GNN to aggregate the neighboring papers in  $L_u$ ,  $L_l$  for learning the author's interest representation. For a paper j in the 1-th short-term window  $L_u$ ,  $L_l$ , its input embedding is represented as  $e_j \in R^d$ . So we can acquire author interest:

$$\mathbf{h}_{i} = \tanh\left(\mathbf{W}^{(1)} \cdot \left[ \sum_{k \in \mathcal{N}_{i}} \mathbf{e}_{k} A_{i,k}; \mathbf{e}_{i} \right] \right), \forall i \in L_{u,l} \quad (1)$$

$$\mathbf{p}_{u,l}^{S} = \tanh\left(\mathbf{W}^{(2)} \cdot \left[\frac{1}{|L|} \sum_{i \in L_{u,l}} \mathbf{h}_{i}; \mathbf{p}_{u}\right]\right)$$
(2)

S denotes that the representation is from the author interest and  $W^{(1)}$  and  $W^{(2)}$  are the parameters. ; represent vertical

concatenation,  $p_{u,l}^S$  denotes which papers are closely relevant to the papers in  $L_{u,l}$  by aggregating neighbors of papers in  $L_{u,l}$ . So, we can acquire the author will access the next paper by summarized the author's term interest.

For each graph entered, each node with its features and structural fingerprints. In content, features of the two nodes will be used to compute their content similarity; In structure, structural fingerprints of the two nodes will be used to evaluate their interaction.

Specifically, given a graph of n nodes G = (V,E), V represent set of the nodes, and E represent set of the edges; then we need to evaluate the content similarity and structural interaction between nodes, for example:

$$e_{ij} = \mathcal{A}_{fea} \left( \mathbf{W} \mathbf{h}_i, \mathbf{W} \mathbf{h}_j \right)$$
 (3)

$$s_{ij} = \mathcal{A}_{str}\left(F_i, F_j\right) \tag{4}$$

$$\mathcal{A}_{fea}\left(\mathbf{W}\mathbf{h}_{i}, \mathbf{W}\mathbf{h}_{j}\right) = \mathbf{a}^{\top}\left(\mathbf{W}\mathbf{h}_{i} \| \mathbf{W}\mathbf{h}_{j}\right) \tag{5}$$

 $A_{fea}$  to acquire similarity between the feature of  $h_i$  and  $h_j$ ,  $A_{str}\left(F_i,F_j\right)$  quantifies the interaction between two fingerprints W denotes the transformation that maps the node features to a latent space. Next,  $w_i$  and  $w_i$  denotes weights of the fingerprints for node i and j, then we can evaluate the structural interactions by Jacard similarity:

$$\mathcal{A}_{str}(F_i, F_j) = \frac{\sum_{p \in (V_i \cup V_j)} \min(w_{ip}, w_{jp})}{\sum_{p \in (V_i \cup V_j)} \max(w_{ip}, w_{jp})}$$
(6)

Normalize feature similarities  $\boldsymbol{3}$  and the structural interactions as  $\boldsymbol{4}$ 

$$\bar{e}_{ij} \leftarrow \frac{\exp\left(\operatorname{LeakyRelu}\left(e_{ij}\right)\right)}{\sum \exp\left(\operatorname{LeakyRelu}\left(e_{ik}\right)\right)}, \bar{s}_{ij} \qquad \frac{\exp\left(s_{ij}\right)}{\sum \exp\left(s_{ik}\right)} \quad (7)$$

combine them to compute the final attention:

$$a_{ij} = \frac{\alpha(\bar{e}_{ij})\,\bar{e}_{ij} + \beta(\bar{s}_{ij})\,\bar{s}_{ij}}{\alpha_{(\bar{e}_{ij}) + \beta(\bar{s}_{ij})}} \tag{8}$$

 $\alpha$  and  $\beta$  denote transfer functions, we can adjust feature similarity and structure interaction scores before combining them. Then, we perform message passing to update the features of each node as:

$$\mathbf{h}_{i}^{(t+1)} = \sigma \left( \sum_{j \in \mathcal{N}_{i}} \alpha_{ij} \mathbf{W} \mathbf{h}_{j}^{(t)} \right)$$
(9)

our model simultaneously calculates two scores: the content-based  $e_{ij}$  and structure-based  $s_{ij}$ , and combine them together.

#### IV. EXPERIMENTS

#### A. Datasets

ACL Anthology Network (AAN) involves papers relevant to Natural Language Processing. The papers belong to disparate venues involving EMNLP, COLING, ACL, and EACL. After excluding several papers with absent information, we possess 21,455 papers, 312 venues, 17,342 authors, and 113367 citation relations. For the labeling job, we accepted a like way as applied to the DBLP.

In order to test the capability, we gather real-world data from DBLP and Aminer. Especially, we choose two disparate conference series as the experimental subjects: AI and Datamining. Specifically, we gather the following information: (1) user information, involving the user id name; (2) conference information, involving the conference id and the accepted papers; (3) paper information, involving the paper id, the title, the authors, and the abstract.

Then, in the DBLP, we delete the users who don't possess any tags in these conferences or don't possess publications. Finally, we have a valid dataset.

The dataset citeulike-a is extracted from CiteULike3, and the other dataset PRSDataset comes from CSPubGuru4. For both datasets, we remove the items with missing and defective abstracts as well as their relative interactions. We also filter out the users who have interactions with at most one item. Finally, the citeulike-a dataset is composed of 5548 users, 10987 items (papers), and 134510 user-item pairs. The PRSDataset dataset consists of 2453 users, 21940 items, and 35969 user-item pairs. Basic text processing is adopted to remove stop words from the title and abstract, as well as the segmented (abstract) sentences with less than 20 characters.

Cora is a single graph of 2.7K nodes, TU-IMDB has 1.5K graphs with 13 nodes on average and TU-MUTAG has 188 molecules with 18 nodes. Although small datasets are useful as sanity checks for new ideas, they can become a liability in the long run as new GNN models will be designed to overfit

the small test sets instead of searching for more generalizable architectures [27] [28].

#### B. Baselines

In this segment, we narrate details about baseline paper recommendation methods and revisions of the offered method used for contrast. We mainly compare our model with the following paper recommendation methods.

CDL [13] attempts to combine an auto-encoder neural model (for better item representation based on textual information) and a traditional collaborative filtering method.

ConvMF [14] applies a convolution neural network to learn the representation of items, and then jointly model user preference by integrating with a traditional matrix factorization model.

LDA-TM [9]. The method utilizes LDA modeling to the detail of papers to produce paper recommendations. In order to produce topics, a Gensim wrapper of LDA topic modeling from MALLET is adopted to the titles and abstracts of research papers. The method produces good results on the informed parameters settings.

MEIRec [15] objects and interactions in intent ranking and recommendation system with a HIN and propose a novel metapath-guided GNN method for the intent recommendation. MEIRec utilizes metapath-guided neighbors to exploit rich structural information.

TAAS [25] can capture the semantic correlation between title and abstract for paper or conference ranking and recommendation.

### C. Results

From the results, our method is better than other methods, which illustrates the superiority of our model. LDA-TM performs unsuccessfully in all baseline methods because it uses barely the content of papers and does not think over papers' citation propinquity or additional supplementary information sources. CDL is a relatively shallow model, and ConvMF applies a convolution neural network (CNN) to learn the representation of papers, the two methods ignore the author's interest and cannot find effective information. The MEIRec places too much emphasis on structural information compared to other methods. TAAS focuses on textual information, the feature of nodes, and does not involve structure. Compared with other methods, our method is better.

#### V. CONCLUSION

In this paper, we presented a method that relied on the author's periodic interest and academic graph network structure to obtain as much effective information as possible to recommend papers. We model the period interests of the author and computing the structural attention, our method developed a way of graph structure attention and combined it with the author's interest. Extensive offline experiments on real data show that our method outperforms the representative others.

TABLE I: RESULTS ON THE DATASETS

Menthod	AAN   DBLP	citeulike-a	Aminer   Cora
CDL	0.269   0.487	0.369	0.554   0.578
ConvMF	0.301   0.511	0.455	0.475   0.377
LDA-TM	0.372   0.451	0.496	0.387   0.475
MEIRec	0.266   0.604	0.587	0.786   0.675
TAAS	0.432   0.624	0.722	0.774   0.755
Our Method	0.477   0.678	0.743	0.814   0.799

#### VI. ACKNOWLEDGEMENT

The authors would like to acknowledge the support provided by the National Natural Science Foundation of China under Grant 61872222, the Natural Science Foundation of Shandong Province (ZR2020LZH011), and the Young Scholars Program of Shandong University.

#### REFERENCES

- Sugiyama, K., and Kan, M.-Y. 2010. Scholarly paper recommendation via user's recent research interests. In Proceedings of the 10th Annual Joint Conference on Digital Libraries(JCDL), 29–38.
- [2] Sugiyama, K., and Kan, M.-Y. 2013. Exploiting potential citation papers in scholarly paper recommendation. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), 153–162.
- [3] Mohammadi, S.; Kylasa, S.; Kollias, G.; and Grama, A. 2016. Context-specific recommendation system for predicting similar pubmed articles. In IEEE 16th International Conference on Data Mining Workshops (ICDMW), 1007–1014.
- [4] A. Gazdar, L. Hidri, A new similarity measure for collaborative filtering based recommender systems, Knowl.Based Syst. 188 (2020) 443–452.
- [5] C. Wang, D.M. Blei, Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 448–456.
- [6] T. Bansal, D. Belanger, A. McCallum, Ask the GRU: Multi-task learning for deep text recommendations, in: Proceedings of the 10th ACM Conference on Recommender Systems, New York, NY, USA, 2016, pp. 107–114.
- [7] M.J. Pazzani, D. Billsus, Content-based recommendation systems, in: The Adaptive Web, vol. 4321, Springer, 2007, pp. 325–341.
- [8] K. Sugiyama, M.-Y. Kan, Exploiting potential citation papers in scholarly paper recommendation, in: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), 2013, pp. 153–162.
- [9] M. Amami, G. Pasi, F. Stella, R. Faiz, An LDA-based approach to scientific paper recommendation, in: Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language To Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings, Springer International Publishing, 2016, pp. 200–210.
- [10] T. Achakulvisut, D.E. Acuna, T. Ruangrong, K. Kording, Science concierge: A fast content-based recommendation system for scientific publications, PLOS ONE 11 (7) (2016) 1–11.
- [11] C. Bhagavatula, S. Feldman, R. Power, W. Ammar, Content-based citation recommendation, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 238–251.
- [12] Hassan, H. A. M. 2017. Personalized research paper recommendation using deep learning. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP), 327–330.

- [13] Wang, H.; Wang, N.; and Yeung, D.-Y. 2015. Collaborative deep learning for recommender systems. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1235–1244.
- [14] Kim, D.; Park, C.; Oh, J.; Lee, S.; and Yu, H. 2016. Convolutional matrix factorization for document context-aware recommendation (recsys). In Proceedings of the 10th ACM Conference on Recommender Systems, 233–240.
- [15] Fan, Shaohua, Junxiong Zhu, Xiaotian Han, Chuan Shi, Linmei Hu, Biyu Ma, and Yongliang Li. "Metapath-Guided Heterogeneous Graph Neural Network for Intent Recommendation." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, 2478–86. Anchorage AK USA: ACM, 2019.
- [16] Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. 2017. A Century of Science: Globalization of Scientific Collaborations, Citations, and Innovations. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1437–1446.
- [17] Kurt D Bollacker, Steve Lawrence, and C Lee Giles. 1999. A system for automatic personalized tracking of scientific literature on the web. In Proceedings of the fourth ACM conference on Digital libraries. ACM, 105–113.
- [18] Roberto Torres, Sean M. McNee, Mara Abel, Joseph A. Konstan, and John Riedl. 2004. Enhancing digital libraries with TechLens+. In Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries. ACM, 228–236.
- [19] Joeran Beel, Akiko Aizawa, Corinna Breitinger, and Bela Gipp. 2017. Mr. DLib: recommendations-as-a-service (RaaS) for academia. In Digital Libraries (JCDL), 2017ACM/IEEE Joint Conference on. IEEE, 1–2.
- [20] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-Based Citation Recommendation. In NAACL HLT 2018: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 238–251.
- [21] Tsung Teng Chen and Maria R. Lee. 2018. Research Paper Recommender Systems on Big Scholarly Data.. In PKAW. 251–260.
- [22] Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. Computing Research Repository.
- [23] Chen, X.; Xu, H.; Zhang, Y.; Tang, J.; Cao, Y.; Qin, Z.; and Zha, H. 2018. Sequential recommendation with user memory networks. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM), 108–116.
- [24] Huang, J.; Zhao, W. X.; Dou, H.; Wen, J.-R.; and Chang, E. Y. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In The 41st Inter- national ACMSIGIR Conference on Research; Development in Information Retrieval (SIGIR), 505–514.
- [25] Guo, Guibing, Bowei Chen, Xiaoyan Zhang, Zhirong Liu, Zhenhua Dong, and Xiuqiang He. "Leveraging Title-Abstract Attentive Semantics for Paper Recommendation." Proceedings of the AAAI Conference on Artificial Intelligence 34, no. 01 (April 3, 2020): 67–74.
- [26] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, et al.. "The graph neural network model" IEEE Transactions on Neural Networks Vol. 20 Iss. 1 (2009) p. 61 - 80.
- [27] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. Information Retrieval, 3(2):127–163.

- [28] Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. 2020. Benchmark data sets for graph kernels.
- [29] Li, Z.; Ding, X.; and Liu, T. 2018. Constructing narrative event evolutionary graph for script event prediction.
- [30] Michael Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In Advances in Neural Information Processing Systems.
- [31] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In International Conference on Learning Representations, 2017.
- [32] Nascimento C, Laender A H F, da Silva A S, et al. A source independent framework for research paper recommendation[C]//Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. 2011: 297-306.
- [33] Philip S, Shola P, Ovye A. Application of content-based approach in research paper recommendation system for a digital library[J]. International Journal of Advanced Computer Science and Applications, 2014, 5(10)
- [34] Zhao, Weidong, Ran Wu, and Haitao Liu. "Paper recommendation based on the knowledge gap between a researcher's background knowledge and research target." Information processing management 52.5 (2016): 976-988.
- [35] Hassan, H. A. M., Sansonetti, G., Gasparetti, F., Micarelli, A., Beel, J. (2019). Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation?. In RecSys (Late-Breaking Results) (pp. 6-10).
- [36] Ohta, M., Hachiki, T., Takasu, A. (2011, August). Related paper recommendation to support online-browsing of research papers. In Fourth international conference on the applications of digital information and web technologies (ICADIWT 2011) (pp. 130-136). IEEE.
- [37] Siebert, S., Dinesh, S., Feyer, S. (2017). Extending a research-paper recommendation system with scientometric measures. In Proceedings of the 5th Workshop on Bibliometric-enhanced Information Retrieval (BIR) (pp. 112-121).