# A Survey of the Research on Big-Scholarly-Data Based Scientific Collaborator Recommendation

**Hongwu Qin[a], Penghong Li[b*] and Xiuqin Ma[c]**

Department of Computer Science and Engineering, Northwest Normal University, Gansu, Lanzhou

[a] qinhongwu@nwnu.edu.cn, [b] lphnwnu@163.com, [c] maxiuqin@nwnu.edu.cn

**Keywords:** Big scholarly data, Recommendation, collaborator recommendation, Survey, Scientific collaboration

**Abstract:** At present, collaboration is a very universal phenomenon in the field of scientific research. Finding the right collaborators is essential to the research work for the majority of scholars. However, with the growth of academic big data, it's becoming increasingly difficult to find the appropriate collaborators. In recent years, scientific collaborator recommendation based on big scholarly data has been extensively studied. This paper makes a full survey on the algorithms for scientific collaborator recommendation, and the data sources commonly used in the experiments in detail. Some detailed discussions on research challenges and future directions are also provided at the end of this survey.

## 1. Introduction

Collaboration is a very universal phenomenon in the field of scientific research and academic circle, a paper is often completed by multiple authors. and so is a project. Some studies have confirmed that there is a strong relationship between collaboration and productivity: the fruitful researchers often have more collaboration[1][2],In addition, cooperation can also make up for the differences caused by the imbalance of research infrastructure in different regions. Therefore, Therefore, for researchers, it will be helpful to collaborate with researchers who match their research direction and all aspects[3].

Traditional collaboration is usually limited to a small area, and the collaborators know each other. With the rapid development of information technology, many researchers are more willing to find new or long-distance collaborators who didn't know each other before[4], in order to seek more innovation, breakthrough and expand the scope of their academic exchanges. However, with the rapid development of science and technology, scientific research results continue to emerge, academic entities such as papers, authors, research institutions, citations, and the relationship among them have grown rapidly, with the formation of big-scholarly-data, there is more difficult and time-consuming to find the right collaborators, especially to find cross-domain collaborators, due to the differences between different fields, most researchers are not familiar with other fields, So it is a great challenge to correctly determine the research direction and collaborators in other fields. Therefore, it is necessary to study how to automatically recommend appropriate collaborators for researchers based on big-scholarly-data.

In the past ten years, the recommendation of research collaborators based on big-scholarly-data has been widely studied. with using Microsoft academic, DBLP, Arnetminer and other scholarly data sources, researchers have carried out content-based recommendation, collaborative filtering based recommendation and social network-based recommendation, among which there are not only the recommendation of collaborators in a single domain[5][6], but also in the cross-domain[7][8].From the existing document, there is no document review on the recommendation of scientific research collaborators based on big-scholarly-data. In order to make researchers fully and clearly understand the current situation of the research point, this paper elaborates from the two aspects of the recommendation algorithm, and the data used in the recommendation, analyzes the problems and difficulties existing in the research point, and puts forward some possible development directions and trends in the future.

## 2. Recommendation algorithm

### 2.1. Content-based recommendation algorithm

The content-based recommendation algorithm has been applied earlier in the research of scientific collaborators' recommendation problems. In the early work, Gollapalli, Mitra and Giles[9] extracted information from scholars' papers, citations and academic home pages for topic modeling, and recommended according to the similarity of scholars' topic. Lopes et al[10]built a vector space model with the help of scholars' research fields, calculated the correlation among scholars, and finally recommended it by using the extent of correlation. The core of content-based recommendation algorithm is usually topic modeling technology. Next, we describe the two topic modeling methods——LDA document generation model and Word2vec.

### 2.1.1. Latent dirichlet allocation

The LDA topic model is based on the co-occurrence of words and documents, and gains the connotation topic of documents by calculating the probability distribution of words and documents. Liu et al[11] analyzed the papers published by scholars LDA to construct a topic model of scholars' research interests, and measured similarity between authors according to the similarity of topics. Similarly, Blei and steyvers [12][13] also use LDA to build the research interest topic model of scholars and make recommendations based on it. Masataka[14]used the LDA model to construct the feature vector of scholars' research interest, and according to the similarity of the feature vector, implemented interdisciplinary collaborator recommendation. Yang et al [15] proposed a weighted topic model based on LDA, and integrated this model into greedy algorithm for collaborator recommendation. Gopalan et al[16] proposed a collaborative topic Poisson factorization (CTPF) model based on LDA to extract scholars' topics and research interests. Kong X et al[17] proposed a BCR recommendation model to recommend scholars with high academic level and high impact. The BCR model uses LDA to obtain the distribution of scholars 'topics of interest each year. The highlight is that a time function is introduced to give prominence to the dynamic changes of scholars' research interests.

LDA topic model has been very mature in content-based collaborator recommendation applications[18]. This method has been proved to be effective.

### 2.1.2. Words2vec model

Unlike LDA, the Word2Vec model is mainly based on the context information of words, that is, semantics and grammar to calculate scholars' research topics. Word2vec can generate more accurate feature descriptions for researchers[19]. Kong et al[19] proposed a collaborator recommendation model called CCREC. This model first uses Word2vec method to extract research topics from scholars' paper titles, then partitions scholars based on the similarity of research topics, and finally combines random walks to achieve collaborator recommendation. The innovation of CCREC recommendation model is to recommend new collaborators with high similarity, to recommend the most potential collaborators.

In general, the accuracy of content-based recommendation algorithms depends heavily on the quality of acquired scholarly topics, and the extracted topics must be valuable and well-structured. When scholars' research interests change, it will affect the performance of recommendations. Therefore, in the research process of collaborators' recommendation problems, content-based recommendation is often used in combination with other methods to improve the accuracy of recommendation.

### 2.2. Collaborative filtering-based recommendation

Collaborative filtering-based method is popular in the field of recommender system.The core of the similar users have similar interests with the target uaer. Heck, Peters & Stock et al[20] proposed co-citation and literature coupling to detect author similarity，to achieve collaborator recommend. Kim et al[21] used collaborative labeling to obtain scholars' research preferences and make collaborator recommendations based on similarities in preferences. Balabanovic [22] exploited a system called fab, which combines content-based filtering and collaborative filtering. Jamali et al [23] proposed a random walk model based on trust and collaborative filtering to implement collaborator recommendation. Kautz et al[24]developed a system called referralweb, which combines collaborative filtering and social networks.

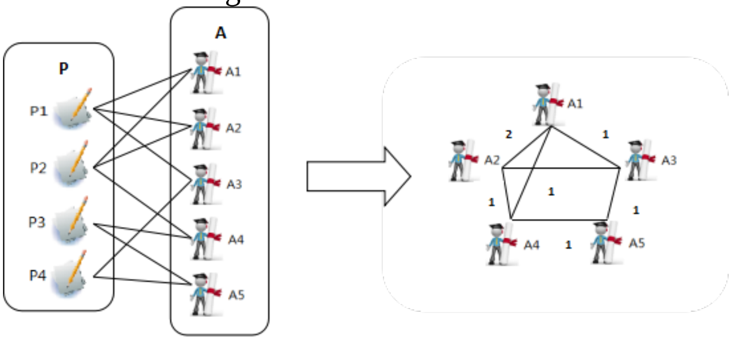
Collaborative filtering-based recommendation algorithms do not need to use natural language processing technology to analyze content, and are easier to achieve. In addition, for new scholars, there is a lack of sufficient information to find similar scholars, so there is a cold start problem [25].

## 2.3. Social network-based recommendation

At present, the research on the recommendation of academic collaborative is more focused on academic social networks. Among, the network application of co-authors is very universal. This section will introduce the network and several algorithms based on it for recommendation, including random walks and Network Embedding.

### 2.3.1. Co-author Network

The co-author network is a weighted graph extracted from the relationship between the paper and the author. The construction process is shown in Figure 1. In Figure 1 (a), P represents the collection of papers, A represents the collection of authors, and the line indicates which authors completed a certain paper. Each vertex in Fig. 1 (b) represents an author. If two authors have cooperated in a certain paper, they connect an edge, and the weight of the edge represents the number of papers that have cooperated. For example, A1 and A2 have cooperated in P1 and P2, so the weight of edge A1A2 is 2. Edge weight is an important basis in the process of collaborative recommendation. Chen [3] and others developed a system called collabseer to help the author find potential collaborators, which is based on the network structure of Co-author and the research interests of author. Lee[6]and others used the author's research expertise and their co-author network to construct a hybrid recommendation algorithm. Experimental results show that the hybrid algorithm is superior to other similar algorithms.



(a) (b)

Figure 1. Co-author network construction process

### 2.3.2. Random walk

The idea of random walk is to calculate the structural similarity between nodes in the co-author network according to transition probability. Jamali [23] proposed a random walk model for recommendation, which combined trust-based and collaborative filtering methods. They used the random walk model to define and measure the confidence of the recommendation. Fouss [26] et al put forward a new viewpoint to describe the similarity between elements in database or undirected weighted graph, the method principle is based on a random walk Markov chain model. Random Walk with Restart is used to measure the correlation between two nodes in the graph. Araki [27] applied topic model and restart random walk method to implement cross domain collaborator recommendation. Konstas[28]created a collaborator recommendation system. Their experiments compared the random walk with restart and collaborative filtering methods. The experimental results show that the random walk with restart model is superior to the collaborative filtering method. Li J [29] used the random walk with restart model to make academic collaborator recommendations. Xia [30] extended Li's work, applying some new academic attributes to restart random walk model to recommend the most valuable collaborators. Kong [17] [19] et al improved random walk model by adjusting transition matrix, using scholars' dynamic research interests, academic influence and publications content to achieve collaborators' recommendation.

### 2.3.3. Network Embedding

Network Embedding learning is a feature learning method based on deep learning that has been applied to the recommendation task [31]. Ganguly and Pudi [32] and others proposed Paper2vec model, it combines doc2vec and deepwalk to obtain the text information of the paper and citation network structure. Xu[33] proposed a DeepWalk model named Seren2vec to implement collaborator recommendation, the model learns the vector representation of each node in the co-author network and improves the conversion probability of random walks in DeepWalk. Kong [34] proposed a collaborator recommendation system called TNERec, based on the research interests and network structure of scholars, the system learns the vector representation of scholars and generates top-N recommendation according to the vector of scholars. In summary, content-based recommendation algorithms focus on extracting topics from text, collaborative filtering-based recommendation algorithms focus on mining research preferences of authors, social network-based recommendation algorithms focus on the relationship between authors. From the experimental results, the combination of multiple algorithms is often better than the recommendation of a single algorithm.

## 3. Scholarly data

At present, scholarly data sources are mainly divided into two categories, one is academic search services provided by mainstream search engines, such as Google Scholar, Baidu Scholar, Microsoft Scholar, etc .; The other is digital libraries or citation databases provided by publishers or research institutions, such as DBLP , Arnetminer , PubMed, CiteULike, CiteSeer, ACM Digital Library Wait. Existing research collaborators recommend research issues, There are three types of academic data sources: Microsoft Academic, DBLP and Arnetminer, which are described below. We summarize the existing research work according to the data set used, As shown in Table 1, DBLP and aminer are the two most used datasets.

Table 1 summarizes the results of existing work according to the data set used

| reference | author | year | social networks | topic model | data set |
|---|---|---|---|---|---|
| [10] | Lopes | 2010 | √ | | DBLP |
| [37] | Brandao.M | 2013 | √ | | DBLP |
| [4] | Xia F | 2014 | √ | | DBLP |
| [29] | Jing li | 2014 | √ | | DBLP |
| [17] | Kong X | 2016 | √ | √ | DBLP |
| [33] | Xu z | 2019 | √ | | DBLP |
| [7] | Tang J | 2012 | √ | | Aminer |
| [10] | Lee DH | 2011 | √ | | Aminer |
| [31] | Dong Y | 2012 | √ | | Aminer |
| [30] | Zheng Liu | 2018 | | √ | Aminer |
| [8] | Guo Y | 2013 | √ | | Aminer |
| [34] | Xiangjie | 2018 | √ | √ | APS |
| [27] | Araki M | 2017 | √ | | Kaken |
| [9] | Gollapalli | 2012 | | √ | Microsoft |

## 4. Existing problems and development trend

Although research on research collaborators' recommendations has yielded some results, However, there are still some problems in the following aspects, which need to be concerned by the majority of scientific researchers.

(1) The collaborators 'recommendation process did not consider the matching of scholars' academic level. The research on the recommendation of collaborators should not be limited to recommendations, but should also ensure that the collaboration is feasible.

(2) The existing algorithm does not consider the active degree of scholars. Existing collaborator recommendation algorithms do not take into account the active degree of scholars, leading to some scholars who have a high academic level in the past but are not active at the moment, which reduces the accuracy of recommendation.

(3) The calculation of the possibility of interdisciplinary collaboration among scholars. The recommendation of cross-field collaborators needs to calculate the possibility of cooperation between the recommended scholars and the research points of each scholar in the target area. The scholars corresponding to the most probable research points will be recommended as collaborators.

In view of the above problems, we think the future research direction can be summarized as follows:

(1) In the design of the collaborator recommendation algorithm, we must fully consider the matching of scholars' academic level, Improve recommendation accuracy and make collaboration more practical.

(2) In the design of the collaborator recommendation algorithm, the active degree of scholars must be fully considered. In order to improve the accuracy of recommendation, we should add features such as publication time and signature order to measure the current activity of scholars.

(3) Cross-domain collaboration is developing in an increasing trend, but there are few studies recommended by cross-domain collaborator. This will be an important research direction in the future.

## 5. Conclusion

In recent years, the recommendation of scientific research collaborators based on academic big data is in the ascendant. According to the different recommendation algorithms, the existing research collaborator recommendation work is divided into three categories: Content-based recommendations, collaborative filtering-based recommendations, and social network-based recommendations. Content-based recommendation algorithms focus on extracting topics from text, collaborative filtering-based recommendation algorithms focus on mining authors' research preferences, social network-based recommendation algorithms focus on the relationship between authors. Recent technology development tends to combine multiple algorithms to improve the recommendation effect. Relatively speaking, Cross domain collaborators recommend less research, which is more difficult. In addition, in the existing research collaborators recommending problems, more academic data sources are used, including Microsoft Academic, DBLP and Arnetminer.

Academic data is increasing, and academic characteristic data is changing, Academic research is in a dynamic process. There are still many problems and challenges in the research of research collaborators based on academic big data. This article also summarizes and analyses the existing problems, and puts forward possible future research directions, hoping to be beneficial to researchers in this field.

## References

[1] Lee S. and Bozeman B. The impact of research collaboration on scientific productivity, Soc. Stud. Sci., vol. 35, no. 5, pp. 673_702, 2005.

[2] Katz J. S. and Martin B. R. What is research collaboration? Res. Policy, vol. 26, no. 1, pp. 1_18, 1997.

[3] Chen H.-H., Gou L., Zhang X., and Giles C. L. CollabSeer: A search engine for collaboration discovery, in Proc. 11th Annu. Int. ACM/IEEE Joint Conf. Digit. Libraries (JCDL), Ottawa, ON,

Canada, Jun. 2011, pp. 231_240.

[4] Xia F., Chen Z., Wang W., Li J., Yang L.T. Mvcwalker: Random walk based most valuable collaborators recommendation exploiting academic factors. IEEE Transactions on Emerging Topics in Computing, 2(3):364--375, 2014.

[5] Lee DH, Brusilovsky P, Schleyer T (2011) Recommending collaborators using social features and mesh terms. Proceedings of the American Society for Information Science and Technology 48: 1–10.

[6] Wang C. and Blei D. M., Collaborative topic modeling for recommending scientific articles. In KDD’11, pages 448–456, 2011.

[7] Tang J., Wu S., Sun J., and Su H. Cross-domain collaboration recommendation, in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), Beijing, China, 2012, pp. 1285_1293.

[8] Guo Y., and Chen X., Cross-domain Scientific Collaborations Prediction Using Citation, IEEE/ACM ASONAM’13, Niagara, Ontario, CAN, 2013, pp. 765- 770.

[9] Gollapalli SD, Mitra P, Giles CL. 2012. Similar researcher search in academic environments. In: Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries. NewYork:ACM, 167–170 DOI 10.1145/2232817.2232849.

[10] Lopes G. R., Moro M. M., Wives L. K., and Oliveira J. P. M. de Collaboration recommendation on academic social networks,in AdvancesinConceptualModeling_Applications and Challenges. Berlin, Germany: Springer-Verlag, 2010, pp. 190_199.

[11] Ping Liu, Kailun Zheng et al.. A Research on Science Research Recommendation Based on LDA Model[J]. Information Studies: Theory & Practice, 2015,38(9):79-85.

[12] Wang C. and Blei D. M., Collaborative topic modeling for recommending scientific articles. In KDD’11, pages 448–456, 2011.

[13] Araki M , Katsurai M , Ohmukai I , et al. Interdisciplinary Collaborator Recommendation Based on Research Content Similarity[J]. IEICE Transactions on Information and Systems, 2017, E100.D(4):785-792.

[14] Konstas I., Stathopoulos V., & Jose J. M., On social networks and collaborative recommendation, In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, July, 2009, pp.195-202.

[15] C. Yang, J. Ma, X. Liu, J. Sun, T. Silva, and Z. Hua, “A weighted topic model enhanced approach for complementary collaborator recommendation,” in Pacifific Asia Conference on Information Systems, PACIS 2014, 01 2014.

[16] P. Gopalan, L. Charlin, and D. M. Blei, “Content-based recommendations with poisson factorization,” in International Conference on Neural Information Processing Systems, 2014, pp. 3176–3184.

[17] Kong X, Jiang H, Wang W, Bekele TM, Xu Z, Wang M. 2017. Exploring dynamic research interest and academic influence for scientific collaborator recommendation.

[18] Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3(Jan):993–1022.

[19] Kong X, Jiang H, Yang Z, et al. Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation[J]. Plos One, 2016, 11(2):e0148492.

[20] Heck T, Peters I, Stock WG. 2011. Testing collaborative filtering against co-citation analysis and bibliographic coupling for academic author recommendation. In: Proceedings of the 3rd ACM RecSys’ 11 workshop on recommender systems and the social web. New York: ACM, 16–23.

[21] Kim HN, Ji AT, Ha I, Jo GS (2010) Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. Electron Commer Res Appl 9: 73–83. doi: 10.1016/j.elerap.2009.08. 004

[22] Cogo E, Donko D. Clustering approach to collaborative filtering using social networks[C]// IEEE International Conference on Electronics Information & Emergency Communication. 2013.

[23] Jamali M. and Ester M., TrustWalker: A random walk model for combining trust-based and item-based recommendation, in Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), Paris, France, 2009, pp. 397-406.

[24] Kautz H., Selman B., and Shah M., Referral web: Combining socialnetworks and collaborative filtering. Communications of the ACM, 40(3):63–65, 1997.

[25] M. Deshpande and G. Karypis. 2004. Item-based top-n recommendation algorithms. ACM Transactions on Information Systems.

[26] Fouss F., Pirotte A., Renders J.-M., and Saerens M., Random-walk computation of similarities between nodes of a graph with application tocollaborative recommendation, IEEE Trans. Knowl. Data Eng., vol. 19, no. 3, pp. 355_369, Mar. 2007.

[27] Araki M , Katsurai M , Ohmukai I , et al. Interdisciplinary Collaborator Recommendation Based on Research Content Similarity[J]. IEICE Transactions on Information and Systems, 2017, E100.D(4):785-792.

[28] Konstas I., Stathopoulos V., and Jose J. M., On social networks and collaborative recommendation, in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2009, pp. 195-202.

[29] Li J., Xia F., Wang W., Chen Z., Asabere N. Y., and Jiang H. ACRec: A co-authorship based random walk model for academic collaboration recommendation, in Proc. 23rd Int. Conf. World Wide Web Companion (WWW Companion), Seoul, Korea, 2014, Art. ID 4.

[30] Zheng Liu, Xing Xie, and Lei Chen. 2018. Context-aware Academic Collaborator Recommendation. In KDD ’18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19– 23, 2018, London, United Kingdom. ACM, New York, NY, USA, 10 pages.

[31] Tian H, Zhuo HH. 2017. Paper2vec: citation-context based document distributed representation for scholar recommendation. ArXiv preprint. arXiv:1703.06587.

[32]S. Ganguly and V. Pudi, “Paper2vec: Combining graph and text information for scientifific paper representation,” in Advances in Information Retrieval, J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, and J. Tait, Eds. Cham: Springer International Publishing, 2017, pp. 383–395.

[33] Xu Z, Yuan Y, Wei H, Wan L. 2019. A serendipity-biased Deepwalk for collaborators recommendation.

[34] PeerJComput. Sci. 5:e178 http://doi.org/10.7717/peerj-cs.178.Xiangjie kong.Mengyi Mao.et al,TNERec:Topic-Aware Network Embedding for Scientific Collaborator Recommendation.In IEEE,2018.

[35] Tang J., Lou T., and Kleinberg J., Inferring social ties across heterogeneous networks, in WSDM’12, pp:743-752, 2012.