Research of Paper Recommendation System Based on Citation Network Model

Sun Jing and Sun Yu(✉)

School of Information Science, Yunnan Normal University, Kunming 650500, Yunnan Province, China
sunyu_km@hotmail.com

Abstract. In view of the increasing number of existing papers, this paper is a study of paper recommendation system. The data set used in this paper is the DBLP citation network in AMiner. First of all, we build a three layers citation network graph model. In this model, we integrate the citation relationship, paper's feature information, co-authorship relationship and research field information into this model. Secondly, we proposed the algorithm PAFRWR. This algorithm combines three layers citation network graph mode with RWR. And, the search vector is constructed by word2vec model. Finally, in this experiment, using Recall@N and NDCG@N as evaluation metric. Then the restart probability of PAFRWR is determined by experiments. And the most effective search vector is determined by comparison. The Recall@N and NDCG@N of PAFRWR are higher than PageRank, LDA and Link-PLSA-LDA through the experiment. So the recommendation model and algorithm in this paper are more accurate and effective.

Keywords: Three layers citation network · RWR · Paper recommendation system

1 Introduction

With the increase of scientific researchers, the number of published papers is also increasing. Reading papers is one of the most important and time-consuming parts of scientific research. When researchers want to get the papers about the field for study, the traditional search engine can only search by keywords and phrases, then get the search results. But the results have a wide range and lack of pertinence. Therefore, the paper recommendation system is produced. This paper is the research about the paper recommendation system based on Citation Network.

In the first part, we conduct a certain amount of research on the recommendation system and the recommendation algorithm. Understanding the development status of researches. In the second part, we constructed a three layers citation network model. We integrate the citation relationship, paper's feature information, co-authorship relationship and research field information into this model. And adopts the DBLP-Citation-network data set of AMiner to build the three layers citation network model. In the third part, we put forward the algorithm of a paper recommendation system based on the three layers

X. Chen et al. (Eds.): ML4CS 2020, LNCS 12488, pp. 237–247, 2020.
https://doi.org/10.1007/978-3-030-62463-7_22

citation network model. We combined the three layers citation network model and RWR algorithm to form the paper recommendation algorithm (PAFRWR). Finally the experimental results show that, the evaluation index value of PAFRWR is better than the other three methods.

1.1 Research Status of Paper Recommendation System

In 1997, Resnick and Varian gave the definition of recommender system for the first time [1]. Then, Bollacker, Lawrence, Giles build the first paper recommendation system in 1998 [2]. This system uses web search engine and heuristic search to find the papers, and through the reference relationships between papers to find the recommendation of related papers.

In 2012, Wang et al. Proposed a method of paper recommendation based on the historical behavior of users [3]. Choochaiwattana [4] studies the application of tags in paper recommendation and proposes a paper recommendation mechanism based tag. Li Ran et al. [5] solved the cold start problem in the paper recommendation system. According to the preferences for research of users, a collaborative topic regression model based on frequent topic set preferences was proposed.

1.2 Research Status of Citation Network

At first, citation network was only used to Library and Information Science [6]. Through the continuous research of citation network, its application scope is more comprehensive. The citation network of the papers is a kind of network information body, which is formed by the reference relationships between the papers [7].

Strohman et al. [8] put forward a global citation recommendation for the first time, and they get the papers of citation recommendation by searching the whole papers. Tang et al. [9] studies a new problem in his paper, that is, citation recommendation based topic, and explores the topic distribution and citation relationship of the papers. In 2011, Shi Jie et al. [10] proposed to form a relationship collection with multiple attribute information in the citations, and then cluster more related citations according to the clustering algorithm, and finally recommend the results to users. Xiao Shibo et al. [11] use the graph model to analyze the relationship between the papers in the citation network and recommend for users. Chen Zhitao et al. [12] proposed a citation recommendation algorithm based on multi feature factor fusion. Based on the traditional citation recommendation model, integrated to the author related factors, overall influence factors and query related factors.

2 Related Work

2.1 Citation Recommendation

A paper needs a large number of references to support its point of view, the citation recommendation can provide appropriate references for researchers. According to the different citation method, citation recommendation can be divided into local citation recommendation and global citation recommendation. The purpose of local citation

recommendation is to recommend relevant papers in the process of writing papers where the citation needs to be added. However the global citation recommendation refers to the recommendation for the whole paper, and provides a reference list for the target paper.

2.2 Random Walk with Restart

Random Walk with Restart (RWR) is a model proposed by Grady [13] in 2006. It mainly measures the similarity between network nodes through the relationship between the topology structure. The main ideas of RWR are as follows: (1) Starting from any vertex or any set of vertices in the graph, walk randomly along the edge of the graph to the next vertex. (2) In the random walk process, any vertex randomly selects the next adjacent node with a certain probability to move or selects the starting point to return to for random walk again. (3) After repeating the random walk process for a finite times and iterating for many times, the probability value of vertices in each graph tends to be stable, and the iteration ends. (4) Finally, the probability value of each vertex can be regarded as the similarity between the current vertex and the selected starting vertex.

The formula for RWR is as follows:

$$\overrightarrow{r_i} = c\,\tilde{W}\,\overrightarrow{r_i} + (1-c)\overrightarrow{e_i} \tag{1}$$

Here, $\overrightarrow{r_i} = \left[r_{i,j}\right]$ is the relevance score, c is the restart probability, $W = \left[w_{i,j}\right]$ is the weighted graph adjacency matrix, $\tilde{W}$ is matrix of W by standardizing, $\overrightarrow{e_i}$ is the identity matrix.

3 Analysis and Construction of Three-Layer Citation Network Graph Model

3.1 Analysis of Three Layers Citation Network Model

This part will analyze and build a three layers citation network model.

1. Citation network of papers
   A paper contains multiple references, and each reference, as an independent paper, also has its own references, which constitutes a citation network. The Fig. 1 is the citation network structure of the papers. As shown in the figure, paper P_1 quoted P_3, P_4 and P_5. If a paper P_i quoted a paper P_j, there are directed edges to connect them. So $P_iP_j = 1$ when we building the paper citation network.
   In this paper, we get the word vector of feature information of the paper by word2vec. Then we calculate the mean value of the word vector, and the mean reflects the characteristic information of a paper. The formula for calculation is as follow: N is the total vocabulary of title and abstract of paper i, w_j is the word vector of a word j in paper i.

$$R_i = \frac{1}{N}\sum_{j=1}^{n} w_j \tag{2}$$

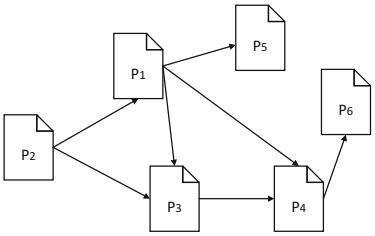

Fig. 1. Citation network of the papers

2. Co-authorship network
   For an academic paper, it can have many authors, so there is a co-authorship relationship in the paper. The Fig. 2 is the co-authorship network diagram. As shown in the figure, author A_1 and authors A_2, A_4, A_5 write the same paper. If there is a co-authorship between author A_i and author Aj, there is an undirected edge between them. So $A_iA_j = 1$ when we building the co-authorship network.

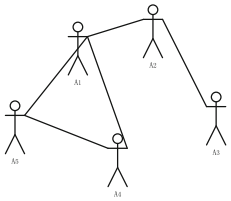

Fig. 2. Co-authorship network diagram

3. Relationship between papers and research fields
   Research fields can help researchers directly locate the topic and research direction of the paper when they study the papers. A paper can correspond to one or more research fields. The Fig. 3 is the diagram of relationship between papers and research fields. As shown in the figure, if the paper P_i corresponds to a research field F_j, there is an edge to connect them. So $P_iF_j = 1$ when building relationships between papers and research fields.
4. Three layers citation network model
   In this paper, we selects the feature information, the author and the research field of the paper, and builds a three layers citation network model with the relationship among them. The Fig. 4 is the three citation network model diagram.

3.2 Construction of Three Layers Citation Network Model

Before building the model, explain the symbols used. As follows in this Table 1, it is the definition of related symbols.

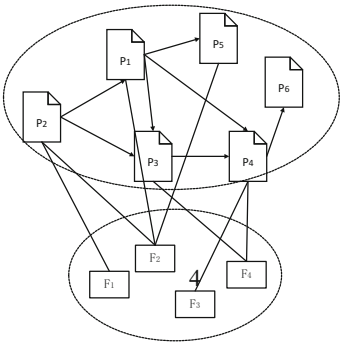

Fig. 3. Research field diagram

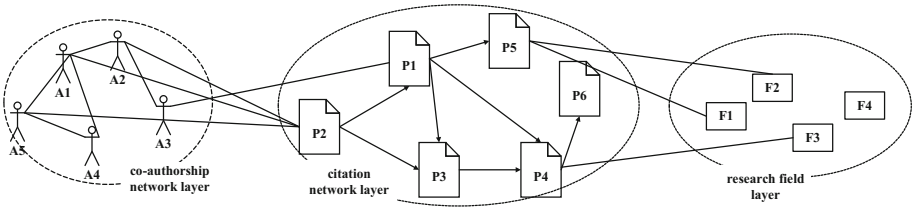

Fig. 4. Three layers citation network model

Table 1. Definition of related symbols

Symbol	Symbol definition
A	Author sets
P	Paper sets
F	Field sets
M_{AA}	Co-authorship matrix
M_{PP}	Citation matrix
M_{AP}	Matrix between papers and authors
M_{PF}	Matrix between papers and fields

When building this model, the matrix is used to express the relationship among the three layers networks. The following matrix is the matrix of three layers citation network graph model.

$$\mathrm{M} = \begin{array}{c} \\ \mathrm{P} \\ \mathrm{A} \\ \mathrm{F} \end{array} \begin{array}{c} \begin{array}{ccc} P & A & F \end{array} \\ \left| \begin{array}{ccc} M_{PP} & M_{PA} & M_{PF} \\ M_{AP} & M_{AA} & M_{AF} \\ M_{FP} & M_{FA} & M_{FF} \end{array} \right| \end{array} \tag{3}$$

4 Recommendation Algorithm Based on Citation Network Model

4.1 Algorithm Design

In this paper, we combines the RWR algorithm with the proposed three layers citation network graph model, it formed the recommendation algorithm based on Citation Network Model, and we call it PAFRWR. Compared with the random walk algorithm, RWR can make the node walk around the initial node, not aimless walk in the random walk model. So RWR algorithm is more accurate and efficient for the determination of similar nodes.

In order to implement the paper recommendation algorithm based on the three layers citation network model, search and recommendation are combined. In this paper, S is used to represent the set of search vectors. Training the search information by word2vec and structure search vector. We normalize the transition probability matrix for assign a reasonable weight to each node.

The recommended algorithm in this paper is as follows:

$$C^{(t+1)} = (1-\beta)MC^{(t)} + \beta s \tag{4}$$

Here C is stationary distribution probability of each node. It represent the relevance between user search and nodes in the three layers citation network model. β is the restart probability. M is a transition probability matrix. s is the initial search vector.

The pseudo code of the algorithm is shown below:

Algorithm: PAFRWR algorithm pseudocode.

Input: The matrix of three layer citation network model matrix ***M***, Search vector set S, restart probability β.
Output: The N recommendation papers
1. Normalization the transition probability matrix
2. Normalization the search vector
3. Initial $C^0 = s$;
4. times=0;
5. d=0;
6. **while** (1):
7. $C^{(t+1)} = (1-\beta)MC^{(t)} + \beta s$;
8. $d = C^{(t+1)} - C^{(t)}$;
9. **if** $d <$ Minimum convergence threshold：
10. **break**
11. **if** times > Maximum number of iterations：
12. **break**
13. times = times+1;
14. **break**
15. **return** Top n Papers with maximum C value;

4.2 Experiment and Analysis

1. Data
   In this paper, The data set used in this paper is DBLP-Citation-network downloaded from AMiner. Then, data of 63469 non repetitive papers from 2013 to 2019 are selected, and it include 152586 authors.
2. Evaluation index
   We use Recall and NDCG as evaluation indexes.
   Recall is an important index to evaluate recommendation results. In the field of information retrieval. The following is the calculation formula of recall. Here, N is the total number of recommended results. Q is the total number of search. $R(p)$ is the set of recommended results produced. $T(p)$ is the set of references of the current tested papers. $\mathrm{R}(p) \cap T(p)$ is the set of papers recommended correctly.

$$\mathrm{Recall@N} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{R(p) \cap T(p)}{T(p)} \tag{5}$$

   In the paper recommendation system, it is considered that the recommendation results are in the current references of the tested papers, and the higher the recommendation results are in the list, indicating that the performance of the recommendation algorithm is better. NDCG is the value normalized by IDCG.

$$\begin{aligned} \mathrm{NDCG@N} &= \frac{1}{Q} \sum_{j=1}^{Q} \frac{DCG@N}{IDCG@N} \\ \mathrm{DCG@N} &= \sum_{i=1}^{N} \frac{2^{r_i} - 1}{log_2(i+1)} \\ \mathrm{IDCG@N} &= \sum_{i=1}^{Rel} \frac{2^{rel_i} - 1}{log_2(i+1)} \end{aligned} \tag{6}$$

3. Restart probability experiment
   In this section, the restart probability parameter β is determined by algorithm experiment. The following figure shows the change trend of Recall@N and NDCG@N under different restart probability β. In this experiment, N is selected as 50, 75 and 100. It can be seen that Recall@N and NDCG@N are the highest when $\beta = 0.3$. And the value of Recall@100 and NDCG@100 are the highest under different . So we choose the $\beta = 0.3$ as the restart probability (Fig. 5).
4. Search vector contrast experiment
   In PAFRWR algorithm, there are author search vectors, paper search vectors and field search vectors. Different search vectors have different effects on the recommendation results of this paper. Four search vectors are defined as follows:

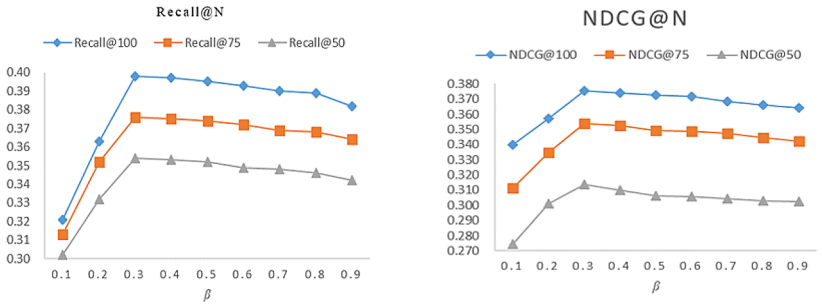

Fig. 5. Recall@N and NDCG@N under different β values

(1) $s_1 = [s_p, s_a, s_f]$ include paper search vectors, author search vectors and field search vectors.
(2) $s_2 = [0, s_a, s_f]$ include author search vectors and field search vectors.
(3) $s_3 = [s_p, s_a, 0]$ include paper search vectors, author search vectors.
(4) $s_4 = [0, s_a, 0]$ include only author search vectors.
The following table shows the results of the search vector comparison experiment (Table 2):
In this experiment, the N is 25, 50, 75, 100. Among them, the Recall@N and NDCG@N of search vector S_1 reach the maximum, which containing all three kinds of information. The results of S_2 and S_3 are very similar, but S_3 has higher Recall@N and NDCG@N than S_2. This shows that the search vector containing the content information of the paper can give the user better recommendations. So the search vector S_1 will be selected as the search vector of the PAFRWR in this paper.
(5) PAFRWR algorithm comparison
We compare PAFRWR with PageRank, LDA and Link-PLSA-LDA. The following figure is a comparison of four algorithms. The Recall@N and NDCG@N of PAFRWR are higher than PageRank, LDA and Link-PLSA-LDA.
PageRank only considers the citation relationship between papers, but does not consider the author, content subject, research field and other specific information of the paper. Therefore, the Recall@N and NDCG@N of PageRank are the lowest among the four algorithms. LDA and Link-PLSA-LDA both build the theme model of the paper. The Link-PLSA-LDA combines the reference relationship between papers based on the topic model. So the result of Link-PLSA-LDA is slightly higher than that of LDA. Overall, the algorithm of this paper (PAFRWR) has better recommendation results (Fig. 6).

Table 2. Recall@N and NDCG@N under different search vectors

	Recall@25	Recall@50	Recall@75	Recall@100	NDCG@25	NDCG@50	NDCG@75	NDCG@100
s_1	**0.297**	**0.354**	**0.376**	**0.398**	**0.2875**	**0.3136**	**0.3538**	**0.3753**
s_2	0.252	0.325	0.349	0.387	0.2527	0.3008	0.3425	0.3587
s_3	0.263	0.338	0.357	0.392	0.2633	0.3027	0.3496	0.3631
s_4	0.208	0.297	0.314	0.335	0.2118	0.2715	0.3038	0.3223

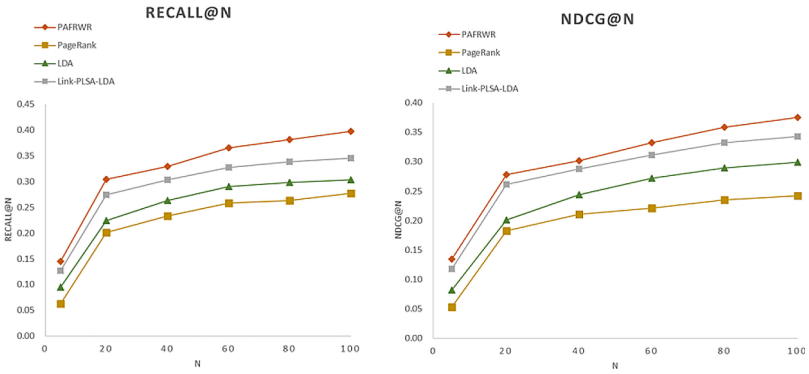

Fig. 6. Comparison of Recall@N and NDCG@N of different algorithms

5 Conclusions

In this paper, we constructs a three layers citation network graph model based on the DBLP citation network data set of AMiner. Which combines the content information, author information and research field information of the papers. Secondly, we proposed the algorithm PAFRWR. This algorithm combine three layers citation network graph mode with RWR. The restart probability $\beta = 0.3$ is determined by experiments, and the most effective search vector is determined. Finally, the Recall@N and NDCG@N of PAFRWR are higher than PageRank, LDA and Link-PLSA-LDA through the experiment.

For the paper recommendation method, it can also be improved from the following aspects. First, we can refine the structure and content of the network model, for example, you can add the same journal relationship of the papers, the same organization relationship of authors, and the relationship between research fields. Secondly, it can combine the user's historical behavior in the paper recommendation. According to the shortcomings, we can build more effective paper recommendation model and algorithm in the future work.

Acknowledgments. This work was supported by the project is the Yunnan Provincial Smart Education Key Laboratory Project,Key Laboratory of Education Informalization for Nationalities of Ministry of Education and the Yunnan University Innovation Research Team Project.

References

1. Resnick, P., Varian, H.R.: Recommender systems. Commun. ACM **40**(3), 56–58 (1997)
2. Bollacker, K.D., Lawrence, S., Giles, C.L.: CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications. In: Proceedings of the 2nd International Conference on Autonomous Agents, pp. 116–123 (1998)
3. Wang, Y., Liu, J., Dong, X., Liu, T., Huang, Y.: Personalized paper recommendation based on user historical behavior. In: Zhou, M., Zhou, G., Zhao, D., Liu, Q., Zou, L. (eds.) NLPCC 2012. CCIS, vol. 333, pp. 1–12. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34456-5_1

4. Choochaiwattana, W.: Usage of tagging for research paper recommendation. In: International Conference on Advanced Computer Theory and Engineering (2010)
5. Ran, L., Hong, L.: Academic paper recommendation algorithm based on frequent topic set preference. Appl. Res. Comput. (9) (2019)
6. Haifeng, W., Yiming, S.: On status QUO OF citation network research and the overview on its development. Comput. Appl. Softw. **29**(2), 164–168 (2012)
7. Yaru, D.: Structural modeling of citation network systems. Libr. Inform. Serv. **4**, 58–61 (1996)
8. Strohman, T., Croft, W.B., Jensen, D.: Recommending citations for academic papers. In: International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2007)
9. Tang, J., Zhang, J.: A discriminative approach to topic-based citation recommendation. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS (LNAI), vol. 5476, pp. 572–579. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01307-2_55
10. Jie, S., Derong, S., Tiezheng, N., et al.: A citation recommendation method based on multiple factors. J. Comput. Res. Dev. (s2) (2011)
11. Shibo, X., Sheng., F.: Research on intelligent recommendation algorithm of research papers based on citation graph model. Comput. Knowl. Technol. **15**(03), 196–198 (2019)
12. Zhitao, C., Shuqin, L., Bin, L., et al.: Citation recommendation algorithm based on multi-feature factor fusion. Comput. Eng. Des. **39**(7), 103–111 (2018)
13. Grady, L.: Random walks for image segmentation. Pattern Anal. Mach. Intell. **28**(11), 1768–1783 (2006)