Liu et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:233
https://doi.org/10.1186/s13638-019-1561-7

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Link prediction in paper citation network to construct paper correlation graph

Hanwen Liu, Huaizhen Kou, Chao Yan and Lianyong Qi*

Abstract

Nowadays, recommender system has become one of the main tools to search for users' interested papers. Since one paper often contains only a part of keywords that a user is interested in, recommender system returns a set of papers that satisfy the user's need of keywords. Besides, to satisfy the users' requirements of further research on a certain domain, the recommended papers must be correlated. However, each paper of an existing paper citation network hardly has cited relationships with others, so the correlated links among papers are very sparse. In addition, while a mass of research approaches have been put forward in terms of link prediction to address the network sparsity problems, these approaches have no relationship with the effect of self-citations and the potential correlations among papers (i.e., these correlated relationships are not included in the paper citation network as their published time is close). Therefore, we propose a link prediction approach that combines time, keywords, and authors' information and optimizes the existing paper citation network. Finally, a number of experiments are performed on the real-world Hep-Th datasets. The experimental results demonstrate the feasibility of our proposal and achieve good performance.

Keywords: Link prediction, Paper citation network, Paper correlated graph, Time, Keywords, Authors' information

1 Introduction

Currently, users can type their preferred keywords into paper-searching websites (e.g., Google Scholar and Baidu Academic) to search for their interested papers, and then, these websites will recommend appropriate papers to them [1]. Generally, a paper just contains partial keywords that a user is interested in, so a paper recommender system must return a set of papers that collectively cover all requested keywords.

As shown in Fig. 1, here is a brief introduction to a process of user's creation. Figure 1 shows that the user normally achieve his goal by putting the following keywords into research tasks: (1) *link prediction* addresses the sparsity of cited network, (2) *weighting criteria* is applied to the link prediction, (3) *data mining* is concerning about information on the mining paper from a network and is applied to the weighted method, and (4) *citation network* is for studying the cited relationships among papers. Therefore, the user obtains four corresponding manuscript keywords, i.e., link prediction, weighting criteria, data mining, and citation network, and the user needs to do these keywords searching and research tasks before his writing.

As shown in Fig. 1, the user obtains a set of keywords including link prediction, weighting criteria, data mining, and citation network. Then, paper-searching websites usually recommend some papers to their users based on those above keywords. As we all know, the keywords of a paper can only represent papers' topics or themes; therefore, considering keywords only appear in paper-searching process may find a set of papers that belong to different research domains or are actually not correlated, which fails to satisfy the user original requirements on deep and continuous research.

Fortunately, paper citation network that depicts the cited relationships among different papers has provided a promising way to model the correlations among the papers in terms of width and depth perspectives. However, the current paper citation

* Correspondence: lianyongqi@gmail.com
School of Information Science and Engineering, Qufu Normal University, Rizhao, China

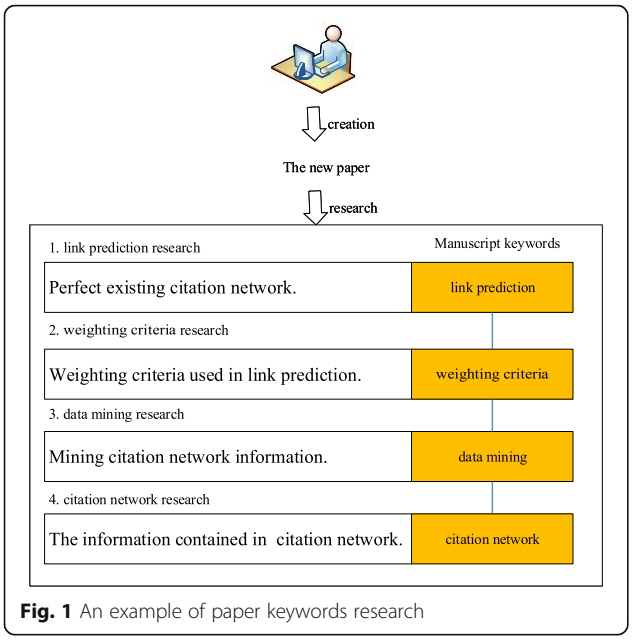

Fig. 1 An example of paper keywords research

network still faces a big challenge, that is, each paper of the existing paper citation network has slight cited relationships with other papers, so that correlated relationships among papers are also very sparse.

Considering this challenge, we will propose a novel link prediction approach to optimize the existing paper citation network. Furthermore, many previous researches proved that link prediction is the best solution to various network optimization problems [2, 3]. More specifically, link prediction attempts to estimate the likelihood of the existence of a link between two nodes because nodes attribute to information and network structures. In addition, when using our proposal to build new paper relationships (i.e., correlated relationships), we also consider the effect of self-citations from authors and potential correlations among papers (i.e., these correlated relationships are not included in the paper citation network as their published time is close).

Overall, our contributions in this paper are concluded into three aspects below:

- We propose a novel link prediction approach to construct new relation graphs. Our proposal considers a wide range of factors that influence the correlations among papers, such as paper published time, paper keywords, and paper authors. Furthermore, our link prediction approach takes the network structure of paper citation network into considerations, which makes the predicted results more reasonable and convincing.
- We optimize the existing paper citation network by reducing the negative influence of intentional self-citations from partial authors.
- At last, extensive experiments are performed on a real-world paper dataset to demonstrate the actual capability of our method of dealing with the network sparsity problem.

The rest of paper is organized as follows. Related work is presented in Section 2. In Section 3, we introduce the research motivation. In Section 4, the detail of our proposed link prediction approach is described. Next, Section 5 discusses the experimental datasets (i.e., Hep-Th) and experimental evaluated metrics and mainly analyzes the experimental results. Finally, in Section 6, we have summarized our proposal as well as future research topics.

2 Related work

Link prediction is a significant research content and approach of optimizing various network. To the best of our knowledge, an essential fact of the link prediction is that node attributes to those known information and network structure features, so link prediction methods can easily find the missing links. Besides, these methods can build new links (i.e., correlated links) between two nodes without connection. Thus, the link prediction can effectively address a core problem of our proposal, i.e., solve the sparsity in the existing paper citation network.

Currently, link prediction has made massive strides and plays an important role in many research areas. For example, new friends through link prediction can be found in social network [4] and protein-protein interactions can also be found [5]. Link prediction approaches can be classified into three categories: similarity-based methods, maximum likelihood approaches, and probabilistic methods [6]. As far as we know, the similarity-based methods can be used to the large-scale networks, which is because it can calculate the similarity scores between two nodes [7]. Although maximum likelihood approaches can obtain specific parameters and probabilistic methods can predict missing links by using the trained model, maximum likelihood approaches and probabilistic methods cannot dispose of the broad-scale networks [8]. Therefore, we mainly consider the similarity-based approach in our research.

Generally, the similarity-based approach can also be classified into two categories: the network structure-based similarity methods and the node attribute-based similarity methods. The node attribute-based similarity methods mainly focus on the node attribute to information of finding the similar nodes, so these

methods are a significant way to form node pairs. Furthermore, these methods also solve the cold-start problem for link prediction research, e.g., Wang et al. [9] used the node attribute information (e.g., user profile) to address the cold-start problem on Twitter and Facebook. In addition, the network structure-based similarity method allocates similarity scores to the node pairs according to the structure features of networks. Currently, the network structure-based similarity method mainly contains four categories, i.e., local approaches, global approaches, quasi-local approaches, and community-based approaches [10]. Here, we mainly pay attention to the local similarity-based approaches, because it calculates the similarity scores of two nodes without connection based on the nodes' neighboring structural features; furthermore, some common index of the local approaches can be used in the large-scale networks, e.g., Common Neighbors index (CN), Jaccard Coefficient (JC), Adamic–Adar index (AA), and Resource Allocation index (RA).

Many of link prediction researches only concentrate on unweighted networks, but actually, many real-world networks can be weighted. For example, edge weighting value can represent the strength of connection in brain networks and the number of flights in airline networks [11], respectively. For the social network, the work [12] uses local weighted similarity functions to calculate the weighting value of two nodes without connection. Besides, this work also proves that the weak ties have an effect on link prediction. In addition, [13] shows that the ties of spouses or romantic partners play an important role in the social network, so these ties can be regarded as one of the significant edge-weighted ways in the link prediction. Recently, the work in [14] is carrying out a study into the effects of the strength of link in the social network and proposes weighting criterion for link prediction model according to users' data information and the number of interactions among users. However, in their weighting criterion, their work does not take full advantage of the node and its attribute information.

In view of the above research content, we know that the link prediction is one of the significant approaches to solve network sparsity, as it specializes in predicting the missing/correlated links among two nodes without connection. Thus, we propose a novel link prediction approach to construct the paper correlated graph, that is, the similarity-based weighting method.

3 Research motivation

In our paper, we focus on the following key issue: how to solve the sparsity of the existing paper citation network? As for this problem, link prediction approach is the best solution. Furthermore, in the process of building a correlated relationship on the paper citation network, we consider the effect of self-citations from authors and potential correlations among the papers, which are not included in citation network but with close published time.

An intuitive example is presented in Fig. 2 to motivate our approach. Assume that Fig. 2a and b are two parts: paper citation network G_C and paper correlated graph G_p, respectively. Each of the network contains the same 10 nodes, i.e., v_1, …, v_{10}; each node represents a paper and contains some attribute information (i.e., paper time, paper keywords, and paper authors). As shown in Fig. 2a, nodes v_1 and v_2 have cited relationship that is mainly because they have common authors (i.e., authors a_1 and a_2). This

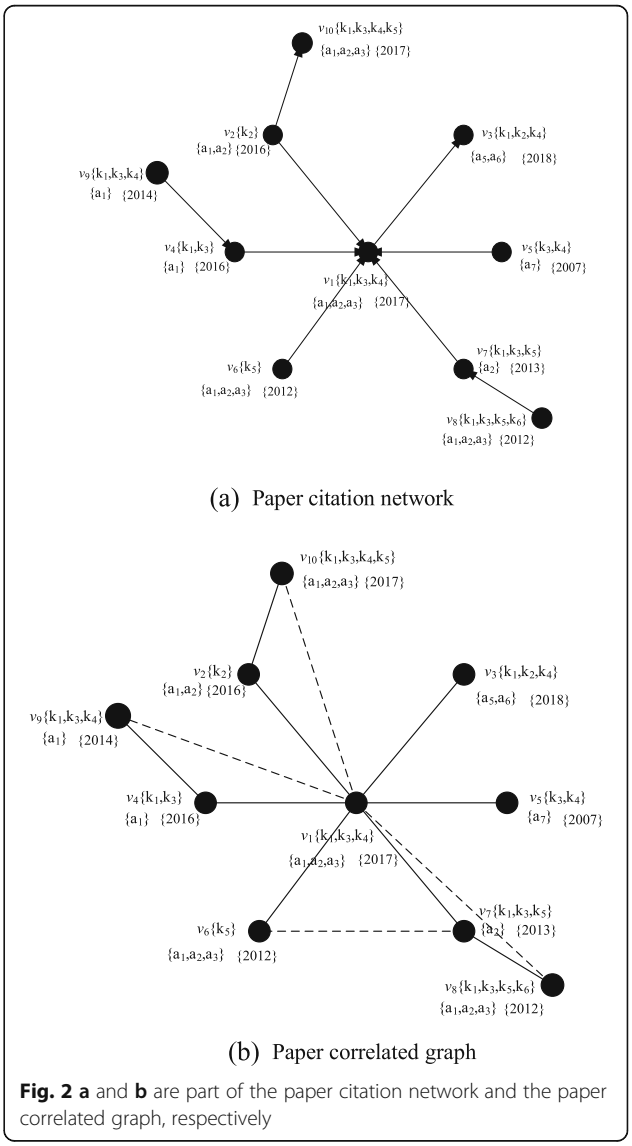

Fig. 2 **a** and **b** are part of the paper citation network and the paper correlated graph, respectively

phenomenon (i.e., v_1 cites v_2) is called self-citation. Therefore, in this paper, we reduce the effect of the intentional self-citations through a weighting model. In addition, nodes v_1 and v_{10} have the same attribute information (i.e., keywords k_1, k_3, and k_4 and authors a_1, a_2, and a_3); however, they do not have direct relationship that is mainly because they have the same published year. Hence, we will establish the new link (i.e., correlated relationship) between two similar nodes by using link prediction approach, e.g., nodes v_1 and v_{10} can build the correlated relationship in Fig. 2b. In view of the analysis mentioned above, we know that the link prediction approach is necessary to optimize current paper citation network, which will be introduced in detail in Section 4.

4 Link prediction method

According to the above analysis, we propose a novel link prediction model to optimize existing paper citation network. As shown in Fig. 3, our link prediction process follows a task sequence [15], and this task sequence mainly consists of the following five activities:

Activity 1: Pre-processing of the network. In order to construct a paper correlated graph, the paper citation network is regarded as an undirected paper citation network (G).

Activity 2: To divide paper citation network. G is partitioned into two parts, i.e., G_{train} and G_{test}. In the G_{train}, we need to get the average score from existing pairs of nodes. Furthermore, in the G_{test}, we need to get the weighting value of the two nodes without connection.

Activity 3: Network to be weighted. In the G_{train}, the weighting value of the two connected nodes are calculated by using the weighting criteria, and the weighting value of two nodes without connection are calculated in the G_{test}.

Activity 4: Score calculation and ranking. (1) Firstly, we use two weighted similarity function formulas WCN and WJC [16] to calculate the weighting value of two nodes without connection in the G_{train}. Next, we can obtain a ranking list in order to descend weighting value. At last, the average score is saved in $w_{\text{average}}(v_{\text{itrain}}, v_{\text{jtrain}})$.

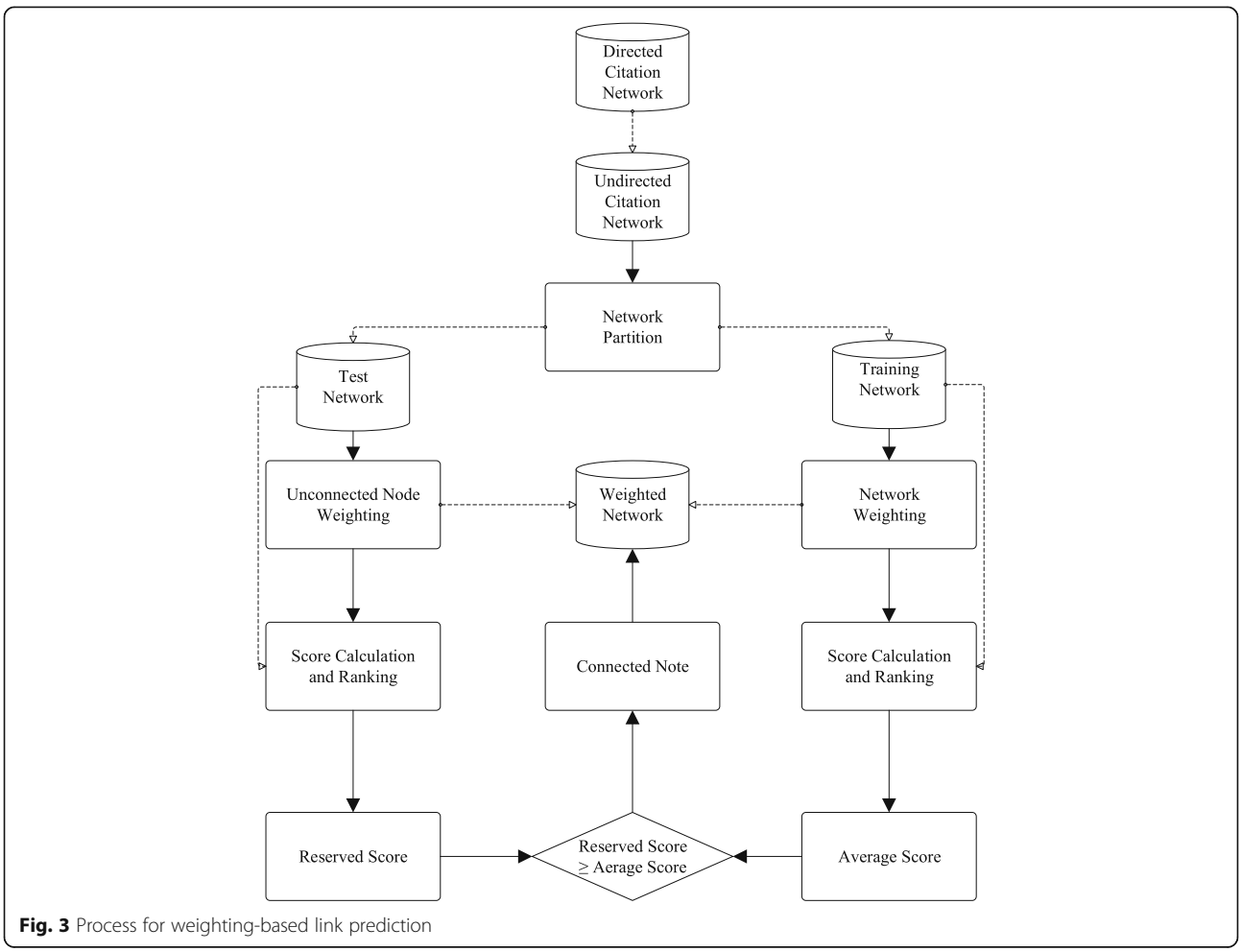

Fig. 3 Process for weighting-based link prediction

Two weighted similarity functions are as follow:

(A) Weighted Common Neighbor - WCN($v_{i\text{train}}$, $v_{j\text{train}}$). Which is:

$$\sum_{v_{z\text{train}}\in\Gamma(v_{i\text{train}})\cap\Gamma\left(v_{j\text{train}}\right)}\frac{w(v_{i\text{train}},v_{z\text{train}})+w\left(v_{j\text{train}},v_{z\text{train}}\right)}{2} \tag{1}$$

$$w_{\text{train}}^{\text{WCN}}\left(v_{i\text{train}},v_{j\text{train}}\right)=\frac{\text{WCN}\left(v_{i\text{train}},v_{j\text{train}}\right)}{\left|\Gamma(v_{i\text{train}})\cap\Gamma\left(v_{j\text{train}}\right)\right|} \tag{2}$$

where Eq. (2) calculates the actual weighting value between nodes $v_{i\text{train}}$ and $v_{j\text{train}}$, $|\Gamma(v_{i\text{train}}) \cap \Gamma(v_{j\text{train}})|$ represents the number of common nodes between nodes $v_{i\text{train}}$ and $v_{j\text{train}}$.

(B) Weighted Jaccard Coefficient - WJC($v_{i\text{train}}$, $v_{j\text{train}}$). Which is:

$$\frac{\sum_{v_{z\text{train}}\in\Gamma(v_{i\text{train}})\cap\Gamma\left(v_{j\text{train}}\right)}\frac{w(v_{i\text{train}},v_{z\text{train}})+w\left(v_{j\text{train}},v_{z\text{train}}\right)}{2}}{\sum_{v_i'\in\Gamma(v_{i\text{train}})}w\left(v_{i\text{train}},v_i'\right)+\sum_{v_j'\in\Gamma\left(v_{j\text{train}}\right)}w\left(v_{j\text{train}},v_j'\right)} \tag{3}$$

$$w_{\text{train}}^{\text{WJC}}\left(v_{i\text{train}},v_{j\text{train}}\right)=\frac{\text{WJC}\left(v_{i\text{train}},v_{j\text{train}}\right)}{\left|\Gamma(v_{i\text{train}})\cap\Gamma\left(v_{j\text{train}}\right)\right|} \tag{4}$$

where Eq. (4) calculates the actual weighting value between nodes $v_{i\text{train}}$ and $v_{j\text{train}}$.

$$w_{\max}\left(v_{i\text{train}},v_{j\text{train}}\right)=\arg\max_{i,j=\overline{1,N}}w_{\text{train}}\left(v_{i\text{train}},v_{j\text{train}}\right) \tag{5}$$

$$w_{\min}\left(v_{i\text{train}},v_{j\text{train}}\right)=\arg\min_{i,j=\overline{1,N}}w_{\text{train}}\left(v_{i\text{train}},v_{j\text{train}}\right) \tag{6}$$

Equations (5) and (6) are used to find the maximum value and the minimum value in the G_{train}.

Since the available paper citation datasets are very sparse, the greatest challenge is how to find correlated relationships among papers. Besides, if we select high threshold value, the correlated links among papers will not be predicted and built in the existing paper citation network, i.e., the chances to find correlated papers decrease for papers. To guarantee the accuracy of predicting and building these correlated links among papers, we select the threshold value same as the average score, i.e., $w_{\text{average}}(v_{i\text{train}}, v_{j\text{train}})$. The average score is as follows:

$$w_{\text{average}}\left(v_{i\text{train}},v_{j\text{train}}\right)=\frac{w_{\max}\left(v_{i\text{train}},v_{j\text{train}}\right)+w_{\min}\left(v_{i\text{train}},v_{j\text{train}}\right)}{2} \tag{7}$$

(2) In the G_{test}, we will perform score calculation of two nodes without connection and produce a descending ranking list.

Activity 5: Connected nodes. LP (link prediction) is defined as in Eq. (8):

$$\text{LP}=\left\{w_{\text{test}}\left(v_{i\text{test}},v_{j\text{test}}\right)\geq w_{\text{average}}\left(v_{i\text{train}},v_{j\text{train}}\right)\right\} \tag{8}$$

4.1 Proposed weighting criteria

Consider the case that undirected paper citation network (G) contains node attribute information (paper time, paper keywords, and paper authors). Furthermore, the existing link prediction approach provides some similarity functions for our proposal. Thus, our proposed weighting model will be described in Eq. (9), $time \in Time$, $keyword \in Keyword$, $author \in Author$ and x_{time}, $x_{keyword}$, $x_{author} \in \{0, 1\}$.

$$w^*\left(v_i,v_j\right)=\text{time}^{x_{\text{time}}}\times\text{keyword}^{x_{\text{keyword}}}\times\text{author}^{x_{\text{author}}} \tag{9}$$

Here, we propose weighting model which can generate two disparate weighting criteria based on the Eq. (9). Please note that the product between paper's data in weighting criteria emphasizes the fact that the selected data information must be concurrently calculated. Thus, the two disparate weighting criteria are as follows:

4.1.1 Keywords and authors' weighting criteria

In our research, if the number of common keywords and co-authors of two papers increases, the weighting value between two nodes will be greater. However, when two papers do not have common keywords, the value between the two papers will decrease as the number of co-authors increases. Such strategies have been adopted to reduce the effect of self-citations. Therefore, the weighting criteria for a pair of nodes v_i and v_j are defined as in Eqs. (10)–(13):

$$r=\begin{cases}1 & \text{Contains common keywords}\\ 0 & \text{Otherwise}\end{cases} \tag{10}$$

$$\text{cosine}\left(K_{v_i}, K_{v_j}\right) = \frac{\left|K_{v_i} \cap K_{v_j}\right|}{\sqrt{\left|K_{v_i}\right|} \times \sqrt{\left|K_{v_j}\right|}} \tag{11}$$

$$\text{cosine}\left(A^a_{v_i}, A^a_{v_j}\right) = \frac{\left|A^a_{v_i} \cap A^a_{v_j}\right|}{\sqrt{\left|A^a_{v_i}\right|} \times \sqrt{\left|A^a_{v_j}\right|}} \tag{12}$$

$$\begin{aligned} w^{\text{KA}}\left(v_i, v_j\right) = C \times \Big(r \times \beta^{\left(1-\text{cosine}\left(K_{v_i}, K_{v_j}\right)\right)} \\ \times \alpha^{\left(1-\text{cosine}\left(A^a_{v_i}, A^a_{v_j}\right)\right)} + \beta \\ \times \alpha^{\text{cosine}\left(A^a_{v_i}, A^a_{v_j}\right)} \times (1-r)\Big) \end{aligned} \tag{13}$$

where α and β $(0 < \alpha, \beta < 1)$ are arbitrary damping parameters and they are used to calibrate the importance of paper authors and paper keywords in the weighting criteria. $A^a_{v_i}$ $(A^a_{v_j})$ and K_{v_i} (K_{v_j}) are a set of authors and keywords, respectively. $A^a_{v_i} \cap A^a_{v_j}$ and $K_{v_i} \cap K_{v_j}$ represent the co-authors and common keywords, respectively. cosine$(A^a_{v_i}, A^a_{v_j})$ and cosine(K_{v_i}, K_{v_j}) denote an approach that computes the similarity between the two nodes v_i and v_j, respectively. A constant C is defined for convenience of calculation.

4.1.2 Time, keywords, and authors' weighting criteria

According to the above analysis, we know the effect of paper keywords and paper authors on the weighting criteria. Here, we then try to find the role of paper time in the weighting model, i.e., if the published time of two papers ia relatively close, the weighting value between the two nodes will be greater. Therefore, the weighting criteria of a pair of nodes v_i and v_j is defined with Eqs. (14)–(15):

$$k(t) = \begin{cases} 0.5 & \text{if } t_{v_i} = t_{v_j} \\ \dfrac{1}{1 + e^{-\left|t_{v_i} - t_{v_j}\right|}} & \text{if } t_{v_i} \neq t_{v_j} \end{cases} \tag{14}$$

$$\begin{aligned} w^{\text{TKA}}\left(v_i, v_j\right) = C \times \lambda^{k(\text{t})} \times \Big(r \\ \times \beta^{\left(1-\text{cosine}\left(K_{v_i}, K_{vj}\right)\right)} \\ \times \alpha^{\left(1-\text{cosine}\left(A^a_{v_i}, A^a_{v_j}\right)\right)} + \beta \\ \times \alpha^{\text{cosine}\left(A^a_{v_i}, A^a_{v_j}\right)} \times (1-r)\Big) \end{aligned} \tag{15}$$

where parameter λ $(0 < \lambda < 1)$ adjusts the effect of paper time on the weighting criteria. Furthermore, t_{v_i} and t_{v_j} indicate the published time of nodes v_i and v_j.

Note that, if there are no common keywords and co-authors among two papers, the weighting value of them will be set to the fixed value, namely $w_{n-co} =$ 0.1. In addition, as for synonymy, word inflections and polysemy are tackled with automatic query expansion techniques [17]. However, it is out of the scope of this paper research.

4.2 Paper correlated graph

Here, we define the undirected relation network as a paper correlated graph.

Definition 1. Paper correlated graph: Paper correlated graph is represented by $G_p = \{V_p, E_p\}$, where V_p and E_p denote its set of nodes and edges, respectively. Furthermore, for each of node pairs (v_i, v_j), the paper correlated graph has a corresponding edge e(v_i, v_j).

5 Experiments

In this section, large-scale experiments are designed and tested on a real-world paper citation dataset which demonstrates the usefulness and effectiveness of our link prediction approach.

5.1 Experimental environments

5.1.1 Experimental tools

The proposed link prediction approach is implemented in PyCharm and executed under the environment of Intel(R) Core(R) CPU @3.0 GHz, 16 GB RAM, and Windows 10 @ 1809, 64bit operating system.

5.1.2 Dataset

A paper citation network was extracted from the available Hep-Th dataset [18]. In the paper citation network, a node represents a paper and an edge indicates that two specific nodes have cited relationship. Furthermore, each node stores the paper's published time, keywords, and authors' information. Besides, we will use the information of each paper that the title and abstract are used to construct a set of keywords; here, we mainly use RAKE (Rapid Automatic Keyword Extraction algorithm) method to construct a set of keywords, which is because this method analyzes the frequency of words appearing and their co-occurrence with other words is used to identify keywords or phrases in the body of a text.

Our experimental process of link prediction also follows the task sequence that is depicted in Fig. 3. First, the existing paper citation network is partitioned into two parts according to their published time. So, the existing paper citation network from the Hep-Th dataset is partitioned into $G_{\text{train}} = [1997, 1999]$ and $G_{\text{test}} = [2000, 2002]$. The G_{train} contains 7304 nodes (papers) and 56,376 edges; likewise, the G_{test} contains 8721 nodes (papers) and 70,045 edges. Next, we need to configure parameters in our experiment (i.e., α, β, and λ). In this experimental process, we need to finetune each parameter of two different

weighting criteria to find the accurate and credible parameter values. Thus, for parameter values α, β, and λ, we first range the values of α from 0.3 to 0.9 with step 0.2, we range the values of β from 0.3 to 0.9 with step 0.1; and finally, we set the values of λ to 0.3 and 0.9 to reflect the factors of published time. In addition, in order to calculate a value of two nodes without connection in the G_{train}, we select two disparate weighted similarity functions, i.e., WCN and WJC, as well as these two disparate weighted similarity functions are usually used in different link prediction research. Next, these two weighted similarity functions will be combined with two different weighting criteria (i.e., Keywords & Authors (KA) and Time & Keywords & Authors (TKA)). Thus, we will obtain four disparate functions (i.e., WCN^{KA}, WCN^{TKA}, WJC^{KA}, WCN^{TKA}) and apply these functions into our experiments.

5.1.3 Evaluation criteria

(1) AUC (area under the receiver operating characteristic curve [19]). The area under the ROC (receiver operating characteristic) curve can demonstrate link prediction methods accuracy, so the AUC can be regarded as measure index. In our research, we first assign a weighting value for each correlated links and non-existent links in the paper citation network. Next, we will randomly select correlated links and non-existent links and compare their values. Finally, we can obtain AUC value by acting the Eq. (16).

$$\text{AUC} = \frac{n' + 0.5n''}{n} \tag{16}$$

where n demonstrates independent comparisons, the randomly selected correlated links are given higher weighting value n' and the same weighting value n'' times.

(2) Edge number. For link prediction approach, we argue that the newer generated edges it has, the approach will be better. Therefore, we test the number of the new generated edges of the output graph (i.e., the paper correlated graph).

In this paper, we focus on the link prediction approach based on the paper time, paper keywords, and paper authors' information of the paper citation network. As far as we know, there are few existing approaches that solve the same problem. Therefore, we compare our proposal approach with the following two approaches:

1) Random [20]: On the test network, if two nodes without connection have more than half of the number of same keywords, these two nodes without connection will generate new edges.
2) Maximum [20]: If the weighting value of two nodes without connection in the test network is greater than the maximum value found in the training network, these two nodes without connection will generate new edges.

5.2 Experimental results

5.2.1 Profile 1: the AUC value of five approaches

In this profile, we compare five different approaches by using the AUC. Figures 4, 5, 6, and 7 show that the experimental results are mostly different in terms of different weighted similarity functions (i.e., WCN and WJC) and different parameters α, β, and λ. That is because the different weighted similarity functions and the different parameters values play a vital influence on different approaches.

As shown in Figs. 4, 5, 6, and 7, for our proposal (i.e., Our-KA and Our-TKA) and the maximum approach (i.e., Maximum-KA and Maximum-TKA), with the same weighted similarity functions, the value of AUC generally increases as the parameters value increases. That is because the node attribute information plays an increasingly important effect on link prediction during the process of parameter value increasing. However, for the random approach, the value of AUC, it is not affected by the parameters value and the weighted similarity functions. In addition, Figs. 4, 5, 6, and 7 show that, with the same weighted similarity functions and the parameters value, the AUC values obtained by our proposal are

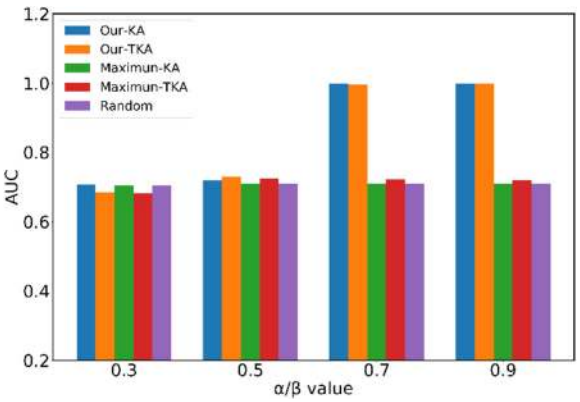

Fig. 4 The different parameters are employed in the weighted similarity function WCN, and $\lambda = 0.3$ to obtain AUC values

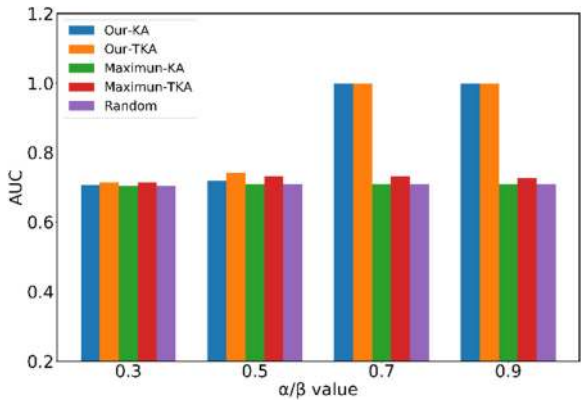

Fig. 5 The different parameters are employed in the weighted similarity function WCN, and $\lambda = 0.9$ to obtain AUC values

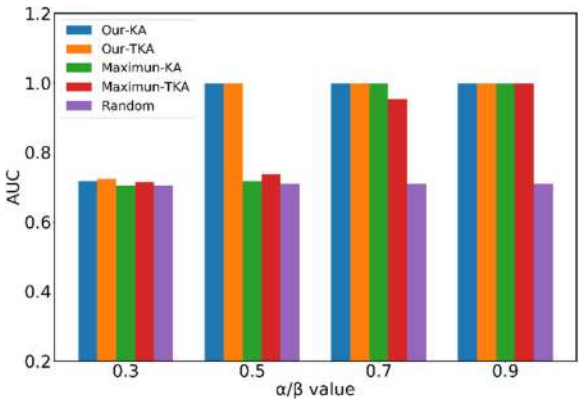

Fig. 7 The different parameters are employed in the weighted similarity function WJC and $\lambda = 0.9$ to obtain AUC values

generally greater than those obtained by the maximum and random approaches. Further, it shows that our proposal is superior to other approaches. As far as we know, the larger value of AUC means that our proposal can better improve the existing paper citation network, i.e., our proposal played a significant role in lightening the sparsity of the existing paper citation network.

5.2.2 Profile 2: the number of new edges built by five approaches

In this profile, we compare five different approaches by using the number produced by new edges. Tables 1, 2, 3, and 4 present the predicted results by combining different parameters which are originated from the weighted similarity functions WCN and WJC. As shown in Tables 1, 2, 3, and 4, the random approach can obtain the same results, i.e., the numbers of building new edges are 4, which further indicates that the different weighted similarity functions and the different parameters value seldom have impact on this approach. However, for other approaches, Tables 1, 2, 3, and 4 show that we will mostly gain different experimental results under different weighted similarity functions and parameters values, which is because the different weighted similarity functions and the different parameter values have an important impact on building the new edges.

As shown in Tables 1, 2, 3, and 4, for our proposal (i.e., Our-KA and Our-TKA) and the maximum approach (i.e., Maximum-KA and Maximum-TKA), under the same weighted similarity functions, the number of new edges would increase as the parameter value increases. That is because the node attribute information plays an increasingly important role in link prediction as the parameter value increases. Furthermore, the new edges built by our approach are larger than the maximum and random approaches, which show that our proposal is superior to

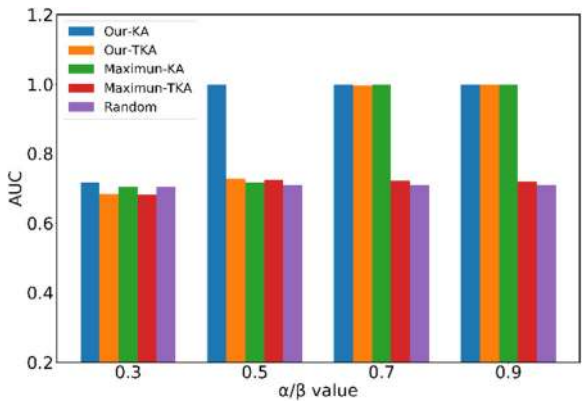

Fig. 6 The different parameters are employed in the weighted similarity function WJC, and $\lambda = 0.3$ to obtain AUC values

Table 1 The different parameters are employed in the weighted similarity function WCN, and $\lambda = 0.3$ to obtain the edge number values

Approaches	α/β/edge number			
	0.3	0.5	0.7	0.9
Random	4	4	4	4
Maximum-KA	12	12	12	10
Maximum-TKA	16	14	14	14
Our-KA	647	4640	26,342,452	26,344,476
Our-TKA	380	2606	8,802,878	21,618,518

Table 2 The different parameters are employed in the weighted similarity function WCN, and $\lambda = 0.9$ to obtain the edge number values

Approaches	α/β/edge number			
	0.3	0.5	0.7	0.9
Random	4	4	4	4
Maximum-KA	12	12	12	10
Maximum-TKA	18	18	18	16
Our-KA	647	4640	26,342,452	26,344,476
Our-TKA	638	4588	12,339,824	26,344,484

Table 4 The different parameters are employed in the weighted similarity function WJC, and $\lambda = 0.9$ to obtain the edge number values

Approaches	α/β/edge number			
	0.3	0.5	0.7	0.9
Random	4	4	4	4
Maximum-KA	128	4122	26,342,452	26,344,476
Maximum-TKA	6	8	14	740
Our-KA	5384	26,342,452	26,344,478	26,344,476
Our-TKA	4534	12,340,892	26,344,484	26,344,484

other approaches. In addition, for our proposal, with the same parameter value and the same weighting criteria (i.e., KA and TKA), we find that the number of the new edges built by the weighted similarity function WJC is generally greater than that built by WCN. It shows that the WJC achieves better results in link prediction than the WCN. According to the above analysis, our proposal can effectively solve sparsity of the existing paper citation network.

5.2.3 Profile 3: investigate the performance in different parameter value setting and the weighted similarity functions

In this profile, we compare the different parameters value setting and the weighted similarity functions by using the AUC. Figures 8 and 9 present the predicted results by parameters α and β combined within the same weighting criteria. Here, we only consider the effect of paper keywords and paper authors' information within the weighting criteria. As shown in Fig. 8, under the same weighted similarity function, three curves (i.e., $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$) show that the value of AUC increases with the parameter value (β) increasing. Besides, the curve of $\alpha = 0.7$ can converge. Likewise, in Fig. 9, $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$, these three curves also show that the value of AUC would increase as the parameter value (β) increases, and these curves all converge. Here, Figs. 8 and 9 indirectly show the reason why the Our-KA approach in Figs. 4, 5, 6, and 7 suddenly increases sharply with the increasing parameter value. In addition, according to the comparison between Figs. 8 and 9, under the same parameters, the AUC value obtained by WJC^{KA} is generally greater than that obtained by WCN^{KA}, which further shows that the WJC in our proposal achieves better results than the WCN.

Figures 10, 11, 12, and 13 present the prediction results by different parameter value combinations on the weighting criteria. Here, we mainly consider the combined effect of paper time, paper keywords, and paper authors' information within the weighting criteria. According to the analysis of Figs. 8 and 9, we know the effect on the paper keywords and paper authors; therefore, we further analyze the role of paper published-time information within the weighting criteria. According to the comparison between Figs. 10 and 11, or Figs. 12 and 13, the value of AUC increases as the parameter value λ increases. Here, Figs. 10, 11, 12, and 13 also indirectly show the reason why the Our-TKA approach in Figs. 4, 5, 6, and 7 suddenly

Table 3 The different parameters are employed in the weighted similarity function WJC, and $\lambda = 0.3$ to obtain the edge number values

Approaches	α/β/edge number			
	0.3	0.5	0.7	0.9
Random	4	4	4	4
Maximum-KA	128	4122	26,342,452	26,344,476
Maximum-TKA	6	8	14	740
Our-KA	5384	26,342,452	26,344,478	26,344,476
Our-TKA	182	2074	8,818,608	26,344,484

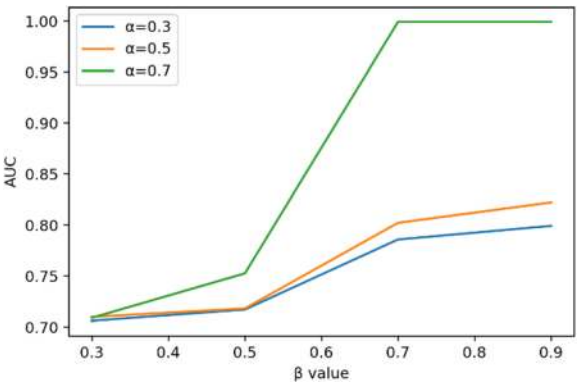

Fig. 8 The different parameters are employed in the function WCN^{KA} to obtain AUC values. $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$

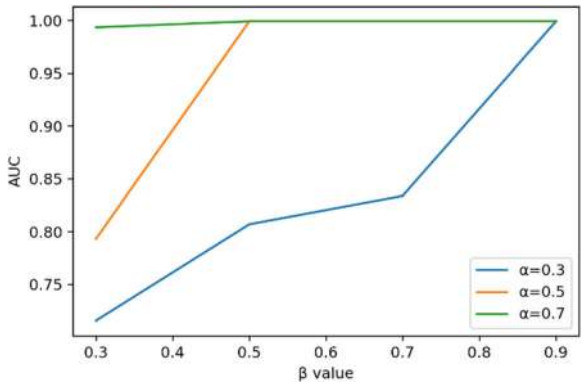

Fig. 9 The different parameters are employed in the function WJC^{KA} to obtain AUC values. $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$

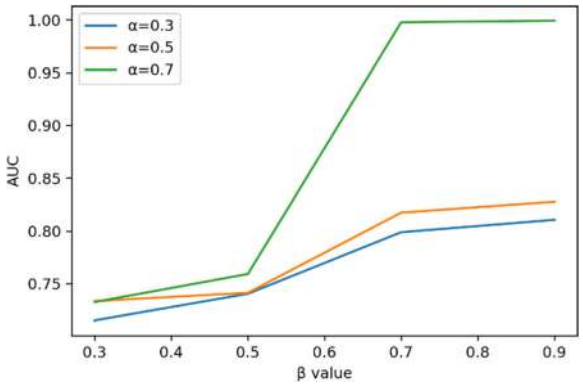

Fig. 11 The different parameters are employed in the function WCN^{TKA}, and $\lambda = 0.9$ to obtain AUC values. $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$

increases sharply with increasing parameters. In addition, according to the comparison between Figs. 10 and 12, or Figs. 11 and 13, under the same parameters setting, the AUC value obtained by WJC^{TKA} is also generally greater than that obtained by WCN^{TKA}, which also further shows that the WJC in our proposal achieves better results than the WCN.

5.2.4 Profile 4: performance comparison in different functions

In this profile, under the same weighted similarity functions, we compare the different weighting criteria by using the values of AUC and the edges' number. According to the experimental results, we get the best results of the function WCN^{KA}, i.e., the value of AUC is 0.9992 and the value of the edge number is 26344478. Similarly, we also acquire the best results of the other three kind of functions (i.e., WCN^{TKA}, WJC^{KA}, WJC^{TKA}). Therefore, Table 5 shows the best results of four functions in terms of AUC and the edge's number. As shown in Table 5, the values of AUC and the edge's number within the TKA weighting criteria is greater than that within the KA weighting criteria, which indicates that the weighting criteria of TKA in our proposal achieves better results than the weighting criteria of KA. Hence, the experimental results suggest that combining paper time, paper keywords, and paper authors' information can enhance link prediction performance significantly and optimize the existing paper citation network effectively.

5.3 Further discussions

However, there are still some shortages in our proposed link prediction approach. Firstly, there are many attributes of node on the existing paper citation network, and it is hard to obtain these attribute information [21,

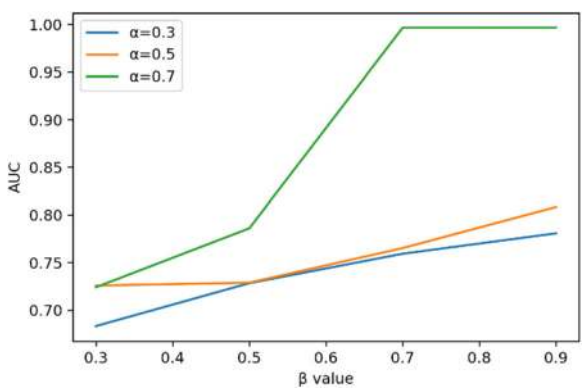

Fig. 10 The different parameters are employed in the function WCN^{TKA}, and $\lambda = 0.3$ to obtain AUC values. $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$

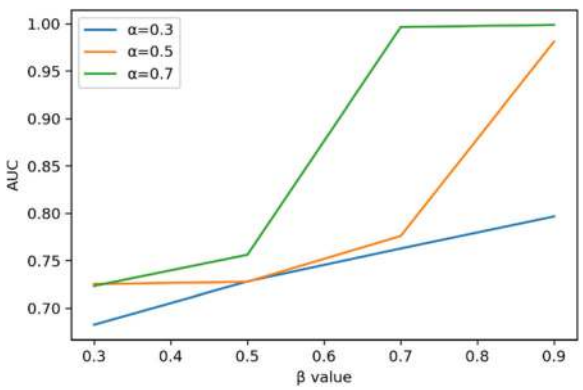

Fig. 12 The different parameters are employed in the function WJC^{TKA}, and $\lambda = 0.3$ to obtain AUC values. $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$

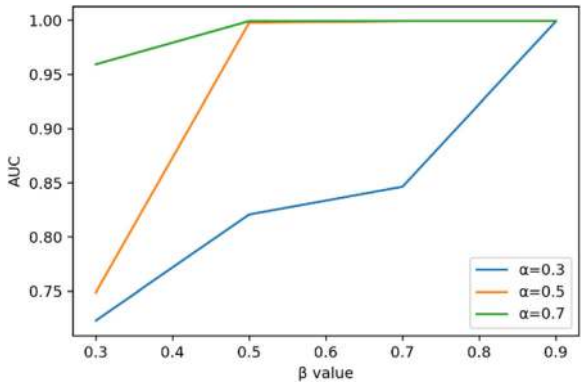

Fig. 13 The different parameters are employed in the function WJC^{TKA}, and $\lambda = 0.9$ to obtain AUC values. $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$

22] to further optimize paper citation network. Secondly, since the process of obtaining data may involve some privacy issues [23–33], our work will further consider the privacy-preservation effects when making link prediction. Finally, more complex multi-dimensional or multi-criterion application scenarios [34–44] should be considered in the future to make our proposal more comprehensive.

6 Conclusions

Predicting whether two correlated papers will build correlated links in an existing paper citation network is a significant analysis task, which is regarded as a link prediction problem. To find and build correlated links in the existing paper citation network, we put forward a novel link prediction approach. The novel link prediction approach not only has advantages of predicting and building correlated links, but also helps with alleviating the current paper citation network sparsity. Furthermore, we also use the combination of paper time, paper keywords, and paper authors' information to reduce the effect of the self-citations. Since the weighting value of nodes pair in the paper citation network is obtained from calculating its attribute information, the experimental results can reflect the actual weighting value of a node pair as accurately as possible. Finally, the feasibility of our proposal is validated by a real-world dataset.

In the future, we will continue to refine our work by considering more complex scenarios, such as privacy-aware or multi-dimensional link prediction problems.

Table 5 The result of performance metrics in the different functions

Similarity function	Performance	
	AUC	Edge number
WCN^{KA}	0.9992	26,344,478
WCN^{TKA}	0.9993	26,344,484
WJC^{KA}	0.9992	26,344,478
WJC^{TKA}	0.9993	26,344,484

Abbreviations

AUC: Area under the curve; KA: Keywords & Authors; RAKE: Rapid Automatic Keyword Extraction algorithm; ROC: Receiver operating characteristic; TKA: Time & Keywords & Authors; WCN: Weighted Common Neighbor; WJC: Weighted Jaccard Coefficient

Acknowledgements

Not applicable.

Authors' contributions

HL finished the algorithm and English writing of the paper. HK and CY finished the experiments. LQ put forward the idea of this paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (No. 2017YFB1400600) and the Natural Science Foundation of China (No. 61872219).

Availability of data and materials

The recruited experiment dataset Hep-Th is available at snap.stanford.edu.

Competing interests

The authors declare that they have no competing interests.

Received: 10 July 2019 Accepted: 11 September 2019
Published online: 16 October 2019

References

1. L. Pan et al., *Academic Paper Recommendation Based on Heterogeneous Graph. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (Springer, Guangzhou, 2015), pp. 381–392
2. X. Zhou et al., Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. IEEE Trans. Emerg. Top. Comput. (2018). https://doi.org/10.1109/TETC.2018.2860051
3. X. Zhou et al., Analysis of user network and correlation for community discovery based on topic-aware similarity and behavioral influence. IEEE Trans. Hum. Mach. Syst. **48**(6), 559–571 (2018). https://doi.org/10.1109/THMS.2017.2725341
4. L.M. Aiello et al., Friendship prediction and homophily in social media. ACM Trans. Web (TWEB) **6**(2), 9 (2012)
5. C. Lee et al., How to assess patent infringement risks: A semantic patent claim analysis using dependency relationships. Tech. Anal. Strat. Manag. **25**, 23–38 (2013)
6. L. Lü et al., Link prediction in complex networks: A survey. Phys. A **390**(6), 1150–1170 (2011)
7. A. Clauset et al., Hierarchical structure and the prediction of missing links in networks. Nat. Publ. Group **453**, 98–101 (2008)
8. R.H. Li et al., in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM'11*. Link prediction: The power of maximal entropy random walk (2011), pp. 1147–1156
9. Z. Wang et al., An approach to cold-start link prediction: Establishing connections between non-topological and topological information. IEEE Trans. Knowl. Data Eng. **28**(11), 2857–2870 (2016)
10. E. Bastami et al., A gravitation-based link prediction approach in social networks. Swarm Evol. Comput. Available online (2018)
11. M.E.J. Newman et al., The structure and function of complex networks. SIAM Rev. **45**(2), 167–256 (2003)
12. L. Lü et al., Link prediction in weighted networks: The role of weak ties. Europhys. Lett. Assoc. **89**(1), 18001 (2010)

13. L. Backstrom et al., in *Proc. 17th ACM Conf. Comput. Supported Cooperative Work Social Comput*. Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on Facebook (2014), pp. 831–841
14. C.P. Muniz et al., Combining contextual, temporal and topological information for unsupervised link prediction in social network. Knowl.-Based Syst. **156**, 129–137 (2018)
15. D. Liben-Nowell et al., The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
16. P.M. Chuan et al., Link prediction in co-authorship networks based on hybrid content similarity metric. Appl. Intell. **48**(8), 2470–2248 (2018)
17. L. Qi et al., Time-Aware IoE Service Recommendation on Sparse Data. Mob. Inf. Syst. **2016**, 4397061, 12 Pages (2016)
18. J. Leskovec et al., in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations (2005)
19. Z. Wen et al., *h*-index-based link prediction methods in citation network. Scientometrics (2018)
20. L. Qi et al., Finding all you need: Web APIs recommendation in web of things through keywords search. IEEE Trans. Comput. Soc. Syst. (2019). https://doi.org/10.1109/TCSS.2019.2906925
21. W. Li et al., On improving the accuracy with auto-encoder on conjunctivitis. Appl. Soft Comput. **81**, 105489 (2019)
22. S. Ding et al., Image caption generation with high-level image features. Pattern Recogn. Lett. **123**, 89–95 (2019)
23. Z. Gao et al., Adaptive fusion and category-level dictionary learning model for multi-view human action recognition. IEEE Internet Things J. (2019)
24. L. Qi et al., Time-aware distributed service recommendation with privacy-preservation. Inf. Sci. **480**, 354–364 (2019)
25. S. Zhang et al., A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services. Futur. Gener. Comput. Syst. **94**, 40–50 (2019)
26. X. Wang et al., Improved multi-order distributed HOSVD with its incremental computing for smart city services. IEEE Trans. Sustain. Comput. (2018). https://doi.org/10.1109/TSUSC.2018.2881439
27. W. Gong et al., Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment. Wirel. Commun. Mob. Comput. **2018**, 3075849, 8 pages (2018)
28. S. Zhang et al., A trajectory privacy-preserving scheme based on a dual-K mechanism for continuous location-based services. Inf. Sci. (2019). https://doi.org/10.1016/j.ins.2019.05.054
29. X. Xu et al., An edge computing-enabled computation offloading method with privacy preservation for internet of connected vehicles. Futur. Gener. Comput. Syst. **96**, 89–100 (2019)
30. L. Qi et al., A distributed locality-sensitive hashing based approach for cloud service recommendation from multi-source data. IEEE J. Sel. Areas Commun. **35**(11), 2616–2624 (2017)
31. Y. Xu et al., Privacy-preserving and scalable service recommendation based on SimHash in a distributed cloud environment. Complexity **2017**, 3437854, 9 pages (2017)
32. X. Wang et al., A cloud-edge computing framework for cyber-physical-social services. IEEE Commun. Mag. **55**(11), 80–85 (2017)
33. L. Qi et al., A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment. Futur. Gener. Comput. Syst. **88**, 636–643 (2018)
34. Y. Zhang et al., Service recommendation based on quotient space granularity analysis and covering algorithm on Spark. Knowl.-Based Syst. **147**, 25–35 (2018)
35. Z. Gao et al., Cognitive-inspired class-statistic matching with triple-constrain for camera free 3D object retrieval. Futur. Gener. Comput. Syst. **94**, 641–653 (2019)
36. S. Wan et al., Multi-dimensional data indexing and range query processing via Voronoi diagram for internet of things. Futur. Gener. Comput. Syst. **91**, 382–391 (2019)
37. X. Wang et al., A tensor-based big data-driven routing recommendation approach for heterogeneous networks. IEEE Netw. Mag. **33**(1), 64–69 (2019)
38. S. Ding et al., A long video caption generation algorithm for big video data retrieval. Futur. Gener. Comput. Syst. **93**, 583–595 (2019)
39. L. Qi et al., A QoS-aware virtual machine scheduling method for energy conservation in Cloud-based Cyber-Physical Systems. World Wide Web J. (2019). https://doi.org/10.1007/s11280-019-00684-y
40. S. Zhang et al., Enhancing privacy through uniform grid and caching in location-based services. Futur. Gener. Comput. Syst. **86**, 881–892 (2018)
41. L. Qi et al., Dynamic mobile crowdsourcing selection for electricity load forecasting. IEEE Access **6**, 46926–46937 (2018)
42. Y. Zhang et al., Covering-based web service quality prediction via neighborhood-aware matrix factorization. IEEE Trans. Serv. Comput. (2019). https://doi.org/10.1109/TSC.2019.2891517
43. S. Zhang et al., A dual privacy preserving scheme in continuous location-based services. IEEE Internet Things J. **5**(5), 4191–4200 (2018)
44. X. Wang et al., NQA: A nested anti-collision algorithm for RFID systems. ACM Trans. Embed. Comput. Syst. **18**(4) (2019). https://doi.org/10.1145/3330139

Publisher's Note