Combining Time, Keywords and Authors Information to Construct Paper Correlation Graph

Hanwen Liu, Huaizhen Kou, Xiaoxiao Chi, Lianyong Qi*

School of Information Science and Engineering
Qufu Normal University, Rizhao, China

E-mail: lianyongqi@qfnu.edu.cn

***Abstract*—Nowadays, recommender systems have become one of the main tools and methods for users to search for their interested papers from massive candidates. Typically, through analyzing the typed keywords by a user, a recommender system can easily retrieve the papers that cover the keywords, in an efficient and economic manner. However, one paper often only contains partial keywords that the user is interested in; therefore, the recommender system needs to analyze a pre-built paper citation graph and then return a set of papers that collectively satisfy the user's requested keywords. While the existing paper citation graph does not consider the possible self-citations and potential correlations among the papers that are not connected in the paper citation graph but with close publication time. Considering the above drawbacks, in this paper, we propose a link prediction approach that combines time, keywords and authors information for constructing a new relation graph. Finally, a case study is employed to explain our approach step by step and demonstrate the feasibility of our proposal.**

Keywords-link prediction; paper citation graph; paper correlation graph; time; keywords; author information

I. INTRODUCTION

Currently, when searching for interested papers via existing paper search websites, e.g., Google Scholar and Baidu Academic, users can type their preferred keywords and then the websites will recommend appropriate papers that cover the typed keywords to the users [1]. Generally, a paper often contains only partial keywords that a user is interested in; therefore, to meet the user's paper search requirement, a paper recommender system often needs to return the user a set of papers that collectively cover all the requested keywords. However, the keywords of a paper can only represent the paper topics or themes; therefore, considering keywords only in paper search process may generate a set of papers that belong to different research domains and are actually not correlated, which fails to satisfy the original user requirements on deep and continuous research on a certain domain or topic.

Fortunately, paper citation graphs that depict the citation relationships among different papers have provided a promising way to model the paper correlations from both width and depth perspectives. However, current paper citation graphs still face a big challenge, i.e., they do not consider the possible self-citations from authors and potential correlations among the papers not connected in the paper citation graphs but with close publication time.

DOI reference number: 10.18293/SEKE2019-161

Considering this challenge, we propose a novel link prediction approach to improve the traditional paper citation graphs, as link prediction has already been proven the best solution for various link optimization problems in graphs [2][3]. More specifically, link prediction attempts to estimate the likelihood of the existence of a link between two nodes based on the existing properties information of nodes and network structures.

Overall, our contributions in this paper are three-fold:

- We propose a novel link prediction approach to construct new relation graphs among papers (i.e., paper correlation graphs). Our proposal considers a wide range of factors that influence the correlations among different papers, such as paper publication time, paper keywords and paper authors. In addition, our link prediction approach takes the network structure of paper citation graphs into considerations, which makes the predicted results more reasonable and convincing.
- We improve the existing paper citation graphs by reducing the negative influence of intentional self-citations from partial authors.
- At last, we evaluate the feasibility of our proposal through a case study.

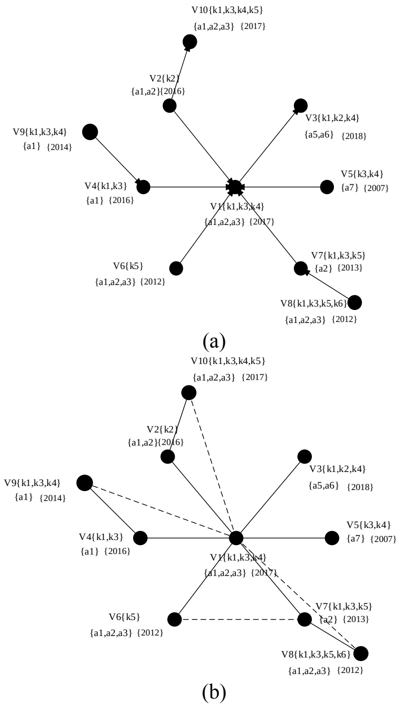
The rest of paper is organized as follows. Related work is presented in Section Ⅱ. In Section Ⅲ, we introduce the research motivation. In Section Ⅳ, the details of our proposed link prediction approach is described. A case study is investigated in Section Ⅴ to demonstrate the effectiveness of our link prediction approach. Finally, in Section Ⅵ, we summarize this paper.

II. RELATED WORK

Currently, link prediction has made massive strides in many research areas and played an important role in more and more fields. According to [4], link prediction approaches can be classified into three categories: similarity-based methods, maximum likelihood approaches and probabilistic methods. However, the similarity-based methods can be used to the large-scale networks, which is because it can calculate the similarity score between two nodes [5]; 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 often fail to deal with the large-scale networks [6]. Therefore, in our research we mainly consider the similarity-based approach. In addition, the work in [7] investigated the use of link strength for the link prediction problem, and they proposed the weighting criterion was based absolutely on topological data: the frequency of existing interactions (i.e. the number of edges) between nodes in the social networks. But they don't take full advantage of node information in the weighting criterion.

In view of existing link prediction approaches, a novel the link prediction approach to construct the paper correlation graph, that is, the similarity-based weighting method.

III. Research Motivation



(a)

(b)

Figure 1. (a) Paper citation graph and (b) Paper correlation graph.

An intuitive example is presented in Fig.1 to motivate our paper. Assume that there is a paper citation graph G_C and a paper correlation graph G_p, Fig.1(a) and Fig.1(b) are a part of G_C and G_p, respectively. Fig.1 contains 10 nodes, i.e., $v_1, \ldots, v_{10}$, each of which represents a paper and contains some node attributes (i.e., paper publication time, paper keywords and paper authors). In Fig.1(a), the self-citation relationship between node v_1 and node v_2 in the paper citation graph is generated merely due to the common authors of v_1 and v_2, which is not reasonable and fair for accurate paper recommendation. Therefore, in this paper, we need to reduce the effect of the intentional self-citation phenomenon through a weighted approach. Besides, in Fig. 1(a), node $v_1$and node v_{10} are published in same period and they also have common keywords and common authors, but there is no link (edge) between them. Thus, in Fig.1 (b), we need to establish the new link between node $v_1$and node v_{10} by using the link prediction approach. In view of the aforementioned analyses, a link prediction approach is necessary to improve current paper citation graphs, which will be introduced in detail in Section Ⅳ.

IV. Link Prediction Approach

According to the analysis of the research motivation, we propose a link prediction approach by using the attributes information and network structure of nodes. To the best of our knowledge, the fundamental process of the unsupervised link prediction model follows the task sequence, which was first proposed by Kleinberg [8]. Concretely, our process of link prediction approach can be seen from Fig. 2. This process mainly consists of the following five activities:

Activity 1: Pre-processing of the graph. In our research, the paper citation graph (G_C) is regarded as an undirected paper citation graph (G), which is because it is easier to construct the paper correlation graph.

Activity 2: Graph partition. In this activity, the G is divided in to two parts. One is training sub-graph (G_{train}) and another one is test sub-graph (G_{test}). In the G_{train}, we need to get the Maximum Score from existing pairs of nodes. And in the G_{test}, we need to get the weighted values of the two unconnected nodes.

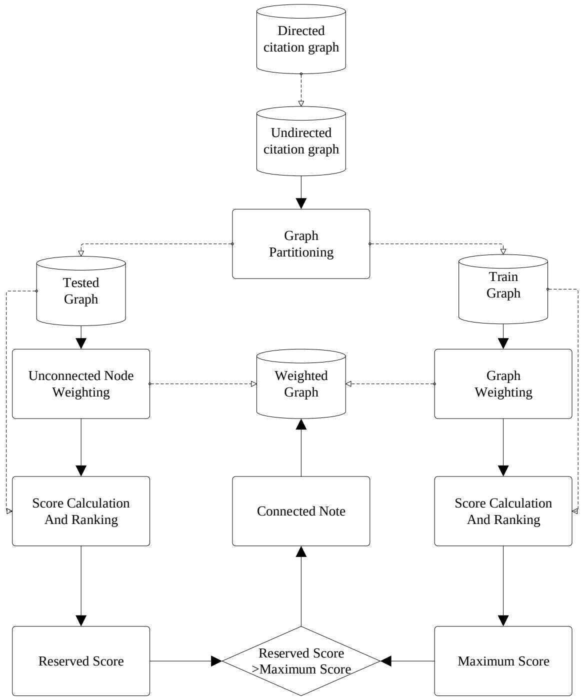

Directed citation graph
Undirected citation graph
Graph Partitioning
Tested Graph
Train Graph
Unconnected Node Weighting
Weighted Graph
Graph Weighting
Score Calculation And Ranking
Connected Note
Score Calculation And Ranking
Reserved Score
Reserved Score >Maximum Score
Maximum Score

Figure 2. Process for weighting-based link prediction.

Activity 3: Graph to be weighted. The weights of the two connected nodes are calculated by using the weighting criteria in the G_{train} and the weights of two unconnected nodes are calculated in the G_{test}.

Activity 4: Score calculation and ranking. (1) Firstly, we use a similarity function formula WCN to calculate a weight value of two unconnected nodes in the G_{train}. Then we produce a ranking list in descending order of score. At last, the Maximum Score is saved in $w_{max}(v_{itrain}, v_{jtrain})$.

The Weighted Common Neighbor - $WCN(v_{itrain}, v_{jtrain})$ and the Maximum Score - $w_{max}(v_{itrain}, v_{jtrain})$:

$$\sum_{v_{ztrain} \in \Gamma(\mathrm{v}_{itrain}) \cap \Gamma(\mathrm{v}_{jtrain})} \frac{w(v_{itrain}, v_{ztrain}) + \mathrm{w}(v_{jtrain}, v_{ztrain})}{2} \tag{1}$$

$$w_{train}^{WCN}\left(v_{itrain}, v_{jtrain}\right) = \frac{WCN\left(v_{itrain}, v_{jtrain}\right)}{\left|\Gamma(\mathrm{v}_{itrain}) \cap \Gamma(\mathrm{v}_{jtrain})\right|} \tag{2}$$

$$w_{\max}\left(v_{itrain}, v_{jtrain}\right) = \arg\max_{i,j=1,N} w_{\mathrm{train}}\left(v_{itrain}, v_{jtrain}\right) \tag{3}$$

Where $\left|\Gamma(\mathrm{v}_{itrain}) \cap \Gamma(\mathrm{v}_{jtrain})\right|$ represents the number of common neighboring nodes of node v_{itrain} and node v_{jtrain}.

(2) In the G_{test}, we will perform score calculation of two unconnected nodes and produce a descending ranking list.

Activity 5: Connecting nodes. LP (link prediction) is defined as in equation (4):

$$LP = \left\{ w_{test}\left(v_{itest}, v_{jtest}\right) > w_{\max}\left(v_{itrain}, v_{jtrain}\right) \right\} \tag{4}$$

A. Proposed Weighting Criteria

Consider that each paper of the G contain paper attributes information (time, keywords and authors). In addition, the link prediction approach offers the similarity functions WCN that can be used for the weight calculation. Therefore, here we consider three sets of those functions: Time, Keyword and Author, and we propose the general weighting model as described in Eq. (5), where $time \in Time$, $keyword \in Keyword$, $author \in Author$ and $x_{time}, x_{keyword}, x_{author} \in \{0,1\}$.

$$w^*\left(v_i, v_j\right) = time^{x_{time}} \times keyword^{x_{keyword}} \times author^{x_{author}} \tag{5}$$

The proposed general weighting model allows the generation of the different weighting criteria by Eq. (5). In addition, it is significant to emphasize that the product between the weighting criteria in link prediction approach formulation ensures that the selected node attributes must be considered simultaneously. Thus, we propose two different weighting criteria as below:

Keywords and Authors Weighting. In our research, if the number of common keywords and co-authors of two papers increases, the weighted values between the two nodes will be greater. But when there is no common keyword in two papers, the weighted values between the two nodes will decrease as the number of co-authors increases. Such strategies have been adopted to reduce the effect of the self-referencing. Thus, the weighting criteria for a pair of nodes v_i and v_j are defined as in equations (6)-(9):

$$\gamma = \begin{cases} 1 & Contains\ common\ keywords \\ 0 & Otherwise \end{cases} \tag{6}$$

$$cosine(\mathrm{K}_{v_i}^a, \mathrm{K}_{v_j}^a) = \frac{\left|\mathrm{K}_{v_i}^a \cap \mathrm{K}_{v_j}^a\right|}{\sqrt{\left|\mathrm{K}_{v_i}^a\right|} \times \sqrt{\left|\mathrm{K}_{v_j}^a\right|}} \tag{7}$$

$$cosine(A_{v_i}^a, A_{v_j}^a) = \frac{\left|A_{v_i}^a \cap A_{v_j}^a\right|}{\sqrt{\left|A_{v_i}^a\right|} \times \sqrt{\left|A_{v_j}^a\right|}} \tag{8}$$

$$w^{KA}\left(v_i, v_j\right) = C \times \left(r \times \beta^{\left(1-\cos ine(\mathrm{K}_{v_i}^a, \mathrm{K}_{v_j}^a)\right)} \times \alpha^{\left(1-\cos ine(\mathrm{A}_{v_i}^a, \mathrm{A}_{v_j}^a)\right)} + (1-r) \times \beta \times \alpha^{\cos ine(\mathrm{A}_{v_i}^a, \mathrm{A}_{v_j}^a)}\right) \tag{9}$$

Where α/β $(0 < \alpha, \beta < 1)$ is arbitrary damping parameters used to calibrate the importance of authors and keywords in the weighting criteria. $A_{v_i}^a / A_{v_j}^a$ ($K_{v_i}^a / K_{v_j}^a$) is a set of authors (keywords) of the node v_i/v_j. $A_{v_i}^a \cap A_{v_j}^a / K_{v_i}^a \cap K_{v_j}^a$ represents the node v_i and node v_j have same authors/keywords. A constant C is defined for convenience of calculation.

Time, Keywords and Authors Weighting. According to the Keywords and Authors Weighting, if the published time of two papers are relatively close, the weighted values between the two nodes will be greater. Thus, the weighting criteria for a pair of nodes v_i and v_j are defined as in equations (10)-(11):

$$k(\mathrm{t}_p) = \frac{t_c - \max\left(t_{p_{v_i}}, t_{p_{v_j}}\right)}{t_c - \min\left(t_{p_{v_i}}, t_{p_{v_j}}\right)} \tag{10}$$

$$w^{TKA}\left(v_i, v_j\right) = C \times k(\mathrm{t}_p) \times \left(r \times \beta^{\left(1-\cos ine(\mathrm{K}_{v_i}^a, \mathrm{K}_{v_j}^a)\right)} \times \alpha^{\left(1-\cos ine(\mathrm{A}_{v_i}^a, \mathrm{A}_{v_j}^a)\right)} + (1-r) \times \beta \times \alpha^{\cos ine(\mathrm{A}_{v_i}^a, \mathrm{A}_{v_j}^a)}\right) \tag{11}$$

Where $t_{p_{v_i}}/t_{p_{v_j}}$ indicates the time of publication of the paper p_{v_i}/p_{v_j}, t_c is the current time.

B. Paper Correlation Graph

Definition1. Paper correlation graph: Paper correlation graph is represented by $G_p = \{V_p, E_p\}$, where V_p and E_p denotes its sets of nodes and edges respectively. In addition, the paper correlation graph is undirected. Meanwhile, for each paper, the paper correlation graph G_p has a corresponding node v, and for each of nodes pair $\left(v_i, v_j\right)$, the paper correlation graph contains the edge e$\left(v_i, v_j\right)$ between v_i and v_j.

V. A CASE STUDY

In this section, a case study is discussed to demonstrate the process of link prediction approach. Due to the limitation of the length of the paper, the case study only considers the first weighting criteria (i.e., the Keywords and Authors Weighting) for the link prediction task. Thus, the process of constructing the paper correlation graph is demonstrated as follows:

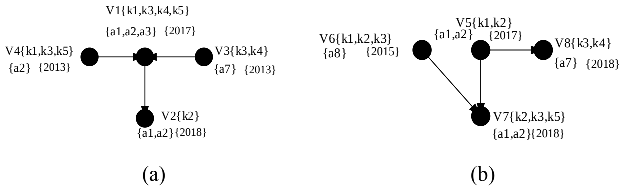


Figure 3. Paper citation graph.

Step1: Pre-processing of the graph. In our case, we regard the paper citation graph of Fig.3 as an undirected citation graph.

Step2: Graph partition. In our case, the G is divided into two parts. One is training sub-graph (G_{train})and another one is test sub-graph (G_{test}). Therefore, in the Fig. 3, the (a) is the G_{train} and the (b) is the G_{test}.

Step3: Graph to be weighted. We use the Keywords and Authors Weighting to calculate the weight of two connected nodes in the (a) and two unconnected nodes in the (b). Meanwhile, we range the values of α $(0 < \alpha < 1)$ from 0.5 to 0.7 with step 0.2, and β=0.5 (see Table 1). Here, we set up the different parameters value that will obtain the different weight value.

TABLE I. PARAMETERS SET.

Similarity function	Parameters set	
	α	β
$w^{KA}(v_i, v_j)$	0.5	0.5
$w^{KA}(v_i, v_j)$	0.7	0.5

(1) we use: β=0.5, α=0.5 and C=1:

Weighted calculation in training sub-graph:

$$(v_1,v_2): r=0; cosine(\mathrm{K}_{v_1}^a, \mathrm{K}_{v_2}^a)=0; cosine(\mathrm{A}_{v_1}^a, \mathrm{A}_{v_2}^a)\approx 0.81; w^{KA}(v_1,v_2)\approx 0.29$$
$$(v_1,v_3): r=1; cosine(\mathrm{K}_{v_1}^a, \mathrm{K}_{v_3}^a)\approx 0.71; cosine(\mathrm{A}_{v_1}^a, \mathrm{A}_{v_3}^a)=0; w^{KA}(v_1,v_3)\approx 0.41$$
$$(v_1,v_4): r=1; cosine(\mathrm{K}_{v_1}^a, \mathrm{K}_{v_4}^a)\approx 0.87; cosine(\mathrm{A}_{v_1}^a, \mathrm{A}_{v_4}^a)\approx 0.58; w^{KA}(v_1,v_4)\approx 0.68$$

Weighted calculation in test sub-graph:

$$(v_5,v_6): r=1; cosine(\mathrm{K}_{v_5}^a, \mathrm{K}_{v_6}^a)\approx 0.82; cosine(\mathrm{A}_{v_5}^a, \mathrm{A}_{v_6}^a)=0; w^{KA}(v_5,v_6)\approx 0.44$$
$$(v_6,v_8): r=1; cosine(\mathrm{K}_{v_6}^a, \mathrm{K}_{v_8}^a)\approx 0.41; cosine(\mathrm{A}_{v_6}^a, \mathrm{A}_{v_8}^a)=0; w^{KA}(v_6,v_8)\approx 0.33$$
$$(v_7,v_8): r=1; cosine(\mathrm{K}_{v_7}^a, \mathrm{K}_{v_8}^a)\approx 0.41; cosine(\mathrm{A}_{v_6}^a, \mathrm{A}_{v_8}^a)=1; w^{KA}(v_7,v_8)\approx 0.66$$

(2) we use: β=0.5, α=0. 7 and C=1:

Weighted calculation in training sub-graph:

$$(v_1,v_2): r=0; cosine(\mathrm{K}_{v_1}^a, \mathrm{K}_{v_2}^a)=0; cosine(\mathrm{A}_{v_1}^a, \mathrm{A}_{v_2}^a)\approx 0.81; w^{KA}(v_1,v_2)\approx 0.37$$
$$(v_1,v_3): r=1; cosine(\mathrm{K}_{v_1}^a, \mathrm{K}_{v_3}^a)\approx 0.71; cosine(\mathrm{A}_{v_1}^a, \mathrm{A}_{v_3}^a)=0; w^{KA}(v_1,v_3)\approx 0.57$$
$$(v_1,v_4): r=1; cosine(\mathrm{K}_{v_1}^a, \mathrm{K}_{v_4}^a)\approx 0.87; cosine(\mathrm{A}_{v_1}^a, \mathrm{A}_{v_4}^a)\approx 0.58; w^{KA}(v_1,v_4)\approx 0.79$$

Weighted calculation in test sub-graph:

$$(v_5,v_6): r=1; cosine(\mathrm{K}_{v_5}^a, \mathrm{K}_{v_6}^a)\approx 0.82; cosine(\mathrm{A}_{v_5}^a, \mathrm{A}_{v_6}^a)=0; w^{KA}(v_5,v_6)\approx 0.62$$
$$(v_6,v_8): r=1; cosine(\mathrm{K}_{v_6}^a, \mathrm{K}_{v_8}^a)\approx 0.41; cosine(\mathrm{A}_{v_6}^a, \mathrm{A}_{v_8}^a)=0; w^{KA}(v_6,v_8)\approx 0.47$$
$$(v_7,v_8): r=1; cosine(\mathrm{K}_{v_7}^a, \mathrm{K}_{v_8}^a)\approx 0.41; cosine(\mathrm{A}_{v_6}^a, \mathrm{A}_{v_8}^a)=1; w^{KA}(v_7,v_8)\approx 0.66$$

Step4: Score calculation and ranking.

(1) For KA weighting criteria, when $\beta = 0.5$ and α=0.5, we can get the Maximum Score, i.e., $w_{max}(v_{itrain}, v_{jtrain}) = w_{train}(v_2, v_4)$ ≈0.55. When $\beta = 0.5$ and α=0.7, we can get the Maximum Score, i.e., $w_{max}(v_{itrain}, v_{jtrain}) = w_{train}(v_2, v_4)$ ≈0.68.

(2) In the G_{test}, we can get such a ranking list that $w^{KA}(v_7, v_8) > w^{KA}(v_5, v_6) > w^{KA}(v_7, v_8)$.

Step5: Connecting nodes. Seen from the Step 4, When $\beta = 0.5$ and α =0.5, we can get such a ranking result that $w^{KA}(v_7, v_8) > w_{max}(v_{itrain} \quad , \quad v_{jtrain}) > w^{KA}(v_5, v_6) > w^{KA}(v_7, v_8)$. Therefore, we can draw a conclusion that v_7 with v_8 constructs a new link. And we can construct the paper correlation graph by connecting a pair of nodes v_7 with v_8.

VI. CONCLUSIONS

In this paper, we mainly put forward a novel link prediction approach to construct the paper correlation graph. In addition, we investigated whether the combination of time, keywords and authors information in the weight computation could reduce the effect of the self-citation. Finally, the feasibility of this the link prediction approach is validated by a case study. In the future work, we will design and deploy a set of real-world experiments to further prove the feasibility of our proposal. Besides, as recommendation process often involves the data privacy issues [9-18], we will further refine our work by considering the privacy-preservation.

ACKNOWLEDGMENT

This research is partially supported by the National Science Foundation of China (No. 61872219).

REFERENCE

[1] L. Pan, X. Dai, S. Huang, J. Chen, Academic Paper Recommendation Based on Heterogeneous Graph. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer. (2015), pp.381–392.

[2] B. Yan, S. Gregory, Finding missing edges in networks based on their community structure, Phys. Rev. E 85 (5) (2012), pp. 056112-056117.

[3] P. Wang, B. Xu, Y. Wu, X. Zhou, Link prediction in social networks: the state-of-theart, Sci. China Inf. Sci. 58 (1), (2015), pp.1–38.

[4] L. Lü, T. Zhou. Link prediction in complex networks: a survey. Phys. A. 390 (6), (2011), pp.1150–1170.

[5] A. Clauset, C. Moore, M.E. Newman. Hierarchical structure and the prediction of missing links in networks. Nature 453, (2008), pp.98-101.

[6] R.H. Li, J.X. Yu, J. Liu, Link prediction: the power of maximal entropy random walk, Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM ’11, (2011), pp.1147-1156.

[7] C. P. Muniz, R. Goldschmidt, and R. Choren, Combining contextual, temporal and topological information for unsupervised link prediction in social networks, Knowledge-Based Syst., vol. 156, (2018), pp.129–137.

[8] P.M. Chuan, L. H. Son, M. Ali, T. D. Khang, N. Dey, Link prediction in co-authorship networks based on hybrid content similarity metric. Applied Intelligence, 48(8), (2018), pp.2470-248.

[9] L. Qi, X., W. Dou, Q. Ni. A Distributed Locality-Sensitive Hashing based Approach for Cloud Service Recommendation from Multi-Source Data. IEEE Journal on Selected Areas in Communications, 35(11): 2616-2624, 2017.

[10] Y. X, H. Liu, C. Yan. A privacy-preserving exception handling approach for dynamic mobile crowdsourcing applications in EURASIP Journal on Wireless Communications and Networking, 2019, pages:113.

[11] Y. Xu, L. Qi, W. Dou, J. Yu. Privacy-preserving and Scalable Service Recommendation based on SimHash in A Distributed Cloud Environment. Complexity, Volume 2017, Article ID 3437854, 9 pages, 2017.

[12] L. Qi, R. Wang, S. Li, Q. He, X. Xu, C. Hu. Time-aware Distributed Service Recommendation with Privacy-preservation. Information Sciences, 480: 354-364, 2019.

[13] L. Qi, Y. Chen, Y. Yuan, S. Fu, X. Zhang, X. Xu. A QoS-Aware Virtual Machine Scheduling Method for Energy Conservation in Cloud-based Cyber-Physical Systems. World Wide Web Journal, 2019.

[14] W. Gong, L. Qi, Y. Xu. Privacy-aware Multidimensional Mobile Service Quality Prediction and Recommendation in Distributed Fog Environment. Wireless Communications and Mobile Computing, vol. 2018, Article ID 3075849, 8 pages, 2018.

[15] L. Qi, W. Dou, W. Wang, G. Li, H. Yu, S. Wan. Dynamic Mobile Crowdsourcing Selection for Electricity Load Forecasting. IEEE ACCESS, 6: 46926-46937, 2018.

[16] C. Yan, X. Cui, L. Qi, X. Xu, X. Zhang. Privacy-aware Data Publishing and Integration for Collaborative Service Recommendation. IEEE ACCESS, 6: 43021-43028, 2018.

[17] L. Qi, X. Zhang, W. Dou, C. Hu, C. Yang, J. Chen. A Two-stage Locality-Sensitive Hashing Based Approach for Privacy-Preserving Mobile Service Recommendation in Cross-Platform Edge Environment. Future Generation Computer Systems，88: 636-643, 2018.

[18] L. Qi, S. Meng, X. Zhang, R. Wang, X. Xu, Z. Zhou, W. Dou. An Exception Handling Approach for Privacy-preserving Service Recommendation Failure in A Cloud Environment. Sensors, 18(7): 1-11, 2018.