Deep Correlation Mining Based on Hierarchical Hybrid Networks for Heterogeneous Big Data Recommendations

Xiaokang Zhou, *Member, IEEE*, Wei Liang, *Member, IEEE*, Kevin I-Kai Wang, *Member, IEEE*, and Laurence T. Yang, *Fellow, IEEE*

***Abstract*— The advancement of several significant technologies, such as artificial intelligence, cyber intelligence, and machine learning, has made big data penetrate not only into the industry and academic field but also our daily life along with a variety of cyber-enabled applications. In this article, we focus on a deep correlation mining method in heterogeneous big data environments. A hierarchical hybrid network (HHN) model is constructed to describe multitype relationships among different entities, and a series of measures are defined to quantify the internal correlations within one specific layer or external correlations between different layers. An intelligent router based on deep reinforcement learning framework is designed to generate optimal actions to route across the HHN. An improved random walk with the restart-based algorithm is then developed with the intelligent router, based on the hierarchical influence across network associated with multiple correlations. An intelligent recommendation mechanism is finally designed and applied to support users' collaboration works in scholarly big data environments. Experiments based on DBLP and ResearchGate data show the practicability and usefulness of our model and method.**

***Index Terms*— Correlation mining, cyber intelligence, heterogeneous big data, hierarchical hybrid network (HHN), reinforcement learning, social influence.**

I. Introduction

THE development of emerging technologies has enabled big data to quickly penetrate into both industry area and academic field [1], [2]. Cyber-social computing in the era of big data is also rapidly emerging and progressing [3]. Social networks can help people quickly publish and obtain the desired information. Due to the highly open and accessible nature of social networks with rapidly developed cyber-social technologies, we are able to mine data and discover association patterns of between users [4]. In particular, with the widespread use of Web 2.0, users of social networks are no longer a single user group. An entire community can join the social network, expanding the scope of information sharing and cyber circles [5]. These intricate but integrated social networks provide us with a more detailed analysis of the heterogeneous big data.

There is one typical type of heterogeneous big data, which is generated by various academic subjects and related to the academic field. Due to the particularity of the staff and work areas, this kind of data shared on various academic websites and stored in different databases are known as the scholarly big data. In particular, these data contain a variety of text contents, academic networks, and digital libraries in the form of, but not limited to, journal articles, conference articles, experimental records, patents, and books. Due to the high popularity of social networking service, along with the rapid development of Web 2.0 technologies, more and more scholars are engaging into their own academic social networks, where they can share their academic activities, research progress, and communicate with other researchers who are interested in the similar fields. Scholarly big data are, thus, continuously increasing at a high growth rate and become a hot area of big data research, which attracts researchers from both scientific institutions and industries, to enrich potential values through scholarly big data mining techniques [6], [7].

In recent years, researchers focused more on studies of multitype relationships among scholarly big data, which resulted in a variety of achievements including the integration of duplication academic information, prediction of research hotspots, and recommendation of related articles [8]–[10]. However, when facing analysis based on heterogeneous data sets, these studies still have several limitations: 1) many mechanisms are developed based on measuring the degree of associations between different entities (such as researchers and articles) for recommendations, but most of them only focus on one single relationship (such as the coauthor relationship). Although

Manuscript received October 30, 2019; revised February 22, 2020; accepted March 27, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0117500, in part by the Natural Science Foundation of Hunan Province of China under Grant 2019JJ40150, in part by the Hunan Provincial Education Department Foundation for Excellent Youth Scholars under Grant 17B146, and in part by the Key Project of Hunan Provincial Education Department under Grant 17A113. *(Corresponding author: Wei Liang.)*

Xiaokang Zhou is with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan (e-mail: zhou@biwako.shiga-u.ac.jp).

Wei Liang is with the with the Key Laboratory of Hunan Province for New Retail Virtual Reality Technology, Hunan University of Technology and Business, Changsha 410205, China (e-mail: weiliang@csu.edu.cn).

Kevin I-Kai Wang is with the Department of Electrical, Computer, and Software Engineering, The University of Auckland, Auckland 1010, New Zealand (e-mail: kevin.wang@auckland.ac.nz).

Laurence T. Yang is with the Department of Computer Science, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada (e-mail: ltyang@ieee.org).

Digital Object Identifier 10.1109/TCSS.2020.2987846

some recommendation systems consider multiple relationships, they measure these relationships separately, rather than combining them in an associative way. 2) When dealing with different types of entities in a constructed network model, most studies calculate the relevance among the same type of entities in a nondiscriminatory way, and thus, do not take multilevel factors into account. 3) Analysis of complex relationships across heterogeneous networks usually requires extracting multimodality features from different data sets. How to effectively incorporate different data sources together and further identify the hidden relations among them have not been well studied.

In this article, we designed and built a hierarchical hybrid network (HHN) model to describe multiple associations among different entities, in order to provide the intelligent recommendation in the context of heterogeneous big data integration from multiple data sources. Compared with other related models, it is improved mainly in the following two aspects: 1) a hierarchical architecture to fully consider the various relationships from multisource data sets for deep correlation mining. 2) A multilevel measurement to quantify relations among different entities (e.g., users and items) considering the internal and external social influence. Based on these, we developed a deep reinforcement learning-based intelligent router with an improved random walk with restart (RWR)-based algorithm, which is applied to provide intelligent recommendations for users' collaboration work support in scholarly big data environments. Specifically, the main contributions addressed in this article can be summarized in the following.

1) An HHN model is constructed to represent multitype relationships among different entities in heterogeneous big data environments, in which the internal correlations within one specific layer or external correlations between different layers are quantified with a series of measures.
2) An intelligent router based on deep reinforcement learning framework is proposed to generate optimal actions to route across the HHN model, which efficiently improves the RWR process when handling the heterogeneous network.
3) An improved RWR-based algorithm is developed to provide intelligent recommendations, which is applied in the scholarly big data environments to support users' collaboration works.

The rest of this article is organized as follows. Section II presents an overview of related works. We present the modeling of HHN and define a series of measures for the quantification of multitype correlations in the constructed model in Section III. Algorithms are developed to provide intelligent recommendations in Section IV. In Section V, we discuss experiment and evaluation results based on the DBLP and ResearchGate data. In Section VI, we summarize this research and give a promising perspective on future research.

II. Related Work

Studies on social network and correlation analysis, issues of cyber intelligence mining with the RWR method, and researches on the intelligent recommendation in the cyber-social system are addressed, respectively, in this section.

A. Social Network Modeling and Correlation Analysis

Network modeling and correlation or relationship analysis among users have become the fundamental and even indispensable part for application and system development in cyber-social computing environments. In particular, Xu *et al.* [11] built a two-stage discriminant framework to model the group-oriented decision-making process. Based on the analysis of users' preferences and their latent social connections, the dynamic mutual influence was considered to improve the prediction of potential event participation. Cepni *et al.* [12] considered a new paradigm in social computing as a social internet of vehicles. They constructed a social sensing model to describe user observations and social consensus on Facebook through the comment thread network. The reliability was analyzed based on users' varying behaviors, relationships, and features on Facebook. Yu *et al.* [13] combined the three factors: mobility influence, content similarity, and social relationship together to identify the on-site user in mobile social networks. Through the transformation and location projection between users and social events, individual location privacy could be protected efficiently. Wang *et al.* [14] focused on crowd behavior analysis and proposed a multiview based framework for group detection. They analyzed individuals' properties within a structural context and extract cluster features based on the integration of their motions and context similarities. Lin *et al.* [15] studied the correlation of users' stress states and social interactions and presented a hybrid model to analyze users' textual, visual, and social attributes. A convolutional neural network was integrated with a factor graph model to detect stress using a combination of tweet contents, social structure, and interaction information. Zhao *et al.* [16] fused the social network and mobile network together in a social-physical graph model and designed a three-phase approach for social-aware data dissemination within mobile-to-mobile communications. Mechanisms were developed for message selection and cooperation pairing based on users' altruistic and selfish behaviors.

B. Cyber Intelligence Mining With RWR Method

As one of the famous graph-based methods, the RWR method plays an important role in graph mining in social and biological systems, due to its powerful ability in detecting missing associations and inferring structural properties. It has been widely applied in link prediction, Web ranking, text mining, and personalized recommendation, which leads to a significant way for representation learning in terms of cyber intelligence hidden in the large-scale structured network model. Specifically, Liao *et al.* [17] built a two-phase model to recommend a group event in event-based social networks. A global trust network was constructed according to multiple factors, including topological network structures, users' online social behaviors, and event participation records. The RWR method was applied to predict the user's preferences along with the unexperienced events. Xia *et al.* [6] presented

an RWR-based recommendation method, in which coauthor order, latest collaboration time, and times of collaboration were considered as three core academic factors to define the link importance, in order to improve the academic collaboration in scholarly big data environments. Liu *et al.* [18] proposed a method to recommend successive events based on graph entropy. To cope with the cold start problem in event-based social networks, they built a primary graph according to entities and their relations from an event-based social network, and a feedback graph according to user feedback from event reservations. The RWR-based algorithm was then adopted to calculate user-event similarity scores from these two graphs with their corresponding graph entropies, and finally, generated the recommendation list. Zhang *et al.* [19] developed a three-stage hierarchical community detection algorithm based on random walks, in which they used the partial transition matrix in terms of the error function to determine the number of random walk steps. Recently, the RWR model was also employed in heterogeneous network environments to deal with hybrid graphs. For example, Yu *et al.* [20] developed a birandom walk algorithm within a hybrid graph, which was composed of two types of nodes and considered their hierarchical structure and interactions in the modern biology system. Fang and Lei [21] improved the RWR model with the k-nearest neighbor algorithm, to handle a heterogeneous network, which consisted of two similarity subnetworks and one association subnetwork. The influence of the neighbor in the constructed graph was taken into account to find some potential associations.

C. Intelligent Recommendation in Cyber-Social System

Currently, the design of an intelligent recommendation model has become more and more important in various hotly discussed social application topics, such as social sensing, user identification, community discovery, and so on. Multiple social factors and properties have been taken into account to improve the recommendation effectiveness, especially when facing billions of heterogeneous user-item information with the complex context in cyber-social systems. To overcome the cold-start recommendation in the edge computing environment, Zhou *et al.* [22] proposed an intelligent recommendation method based on the inverse memory-based collaborative filtering. According to the inference rules in social balance theory, they selected the opposite users and indirect friends for the target user, based on which the quality-optimal services could be recommended as the candidate result. Zhong *et al.* [23] developed a hyperlink-induced topic search algorithm and designed the weighted association rules to calculate the apps' authority and the users' hub score, which could reduce the number of malicious user ratings with increasing intelligent devices in a mobile computing environment. Lin *et al.* [24] exploited a weighted interest degree recommendation algorithm with an association rule mining algorithm, aiming at obtaining more accurate recommendation content for the improvement of intelligent navigator in-vehicle networks. Mao *et al.* [25] presented a recommendation mechanism based on the modeling of multirelational social networks. A propagation model using random walk algorithm was built to extract the original data from a single social network, and a multigraph ranking model was constructed to find out the nearest neighbors of particular users based on the calculation of closeness between users. Meng *et al.* [26] considered the spatial-temporal features in-service recommendations and adopted tensor factorization to discover latent factors from the constructed time-sensitive and region-sensitive quality of service metrics. Two clustering algorithms were employed to handle the data sparsity issue, and a fast-distributed asynchronous mechanism was applied to balance the accuracy of prediction and speed of convergence in big data environments. Sun *et al.* [27] designed a framework for the group recommendation based on a compositive consideration of social relationships and social behaviors. A group preference model was built according to social profiles from external experts, and a joint tolerance and altruism model was built to describe user personalities. Algorithms were then developed to provide online recommendations for group users under different social contexts.

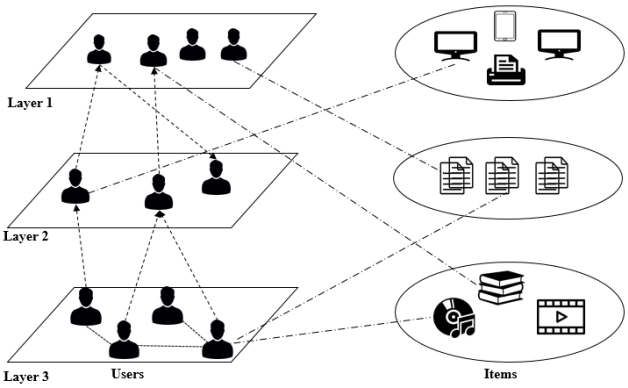


Fig. 1. Conceptual architecture of HHN.

III. Modeling of Hierarchical Hybrid Network

A. Model Description

Three basic relations among users and items are considered to build the HHN model. Definitions can be expressed as follows:

$$G_{\mathrm{HHN}}(N, E, C) \tag{1}$$

where $N = \mathrm{UN} \cup \mathrm{IN}$ indicates a combination of users and items in the heterogeneous network model, which is basically used to form a user network UN and an item network IN. Each network in this hierarchical model is composed of a nonempty set of nodes, respectively. A conceptual image is shown in Fig. 1.

As shown in Fig. 1, the whole architecture of HHN is composed of a multilayer user network, which describes the hierarchical user correlations, and a group of item networks, which describes a variety of items categorized according to their relevance or similarities. Users are connected within each layer, or between different layers according to their various relationships. For example, the directed links between different layers can be used to describe the potential influence among

users connected within their social networks. The undirected links among different users within each layer can be used to describe the similarity-based correlations among them. The link between a user and an item indicates that this user is interested in the item according to the constructed network model. The details of UN and IN are addressed as follows.

$\mathrm{UN} = \{u_1, u_2, \ldots, u_m\}$ represents a nonempty set of nodes (i.e., the users) in the HHN model. In particular, u_i represents a specific user, and $u_i = (\mathrm{UID}_i, \mathrm{Int}_i, \mathrm{LYR}_i)$, in which UID_i is the user ID, Int_i is a vector with a set of keywords that indicates the interests of u_i, and LYR_i is the specific layer that u_i belongs to in UN.

$\mathrm{IN} = \{\mathrm{itm}_1, \mathrm{itm}_2, \ldots, \mathrm{itm}_n\}$ represents a nonempty set of nodes (i.e., the items) in the HHN model. In particular, itm_i indicates a specific item, and $\mathrm{itm}_i = (\mathrm{IID}_i, F_i, \mathrm{Cui}_i, \mathrm{Ctg}_i)$, in which IID_i is the item ID, F_i is a vector with a set of keywords to represent and describe the semantic features for itm_i, Cui_i is a vector with a set of users who are interested in itm_i, and Ctg_i indicates the specific category that itm_i belongs to in IN.

$E = (E_{\mathrm{UN}}, E_{\mathrm{IN}}, E_{\mathrm{UI}})$ indicates a combination of edges which connect different nodes in UN and IN. In particular, it includes three important relations as: 1) E_{UN} indicates the direct or indirect relationships between u_i and u_j; 2) E_{IN} indicates one kind of semantic-aware relationships between itm_i and itm_j; and 3) E_{UI} indicates one kind of interest-based relationships between u_i and itm_j.

$C = \{C_{ij} | \text{if } \exists e_{ij} \in E\}$ indicates a series of measures, which are used to represent the multiple correlations between different users and items within the HHN model.

B. Influence-Based Hierarchical User Relations

Users may connect and benefit from each other through direct and indirect interactions in social networks. They will conduct more collaborative behaviors when having similar targets. In the HHN model, different kinds of interactions within a layer or between layers can reflect different social influences among users. Following our previous study [28], the social influence hidden in users' interactive behaviors based on their similarities is involved in analyzing their multilayer relationships in a hierarchical way. In particular, the three relations mentioned above can be measured as follows.

$E_{\mathrm{UN}} \subset U \times U$ indicates one kind of hierarchical influence-based relation between two connected users. In particular, $E_{\mathrm{UN}} = \{\langle u_i, \mathrm{CUU}_{ij}, u_j\rangle \mid u_i, u_j \in \mathrm{UN}\}$ can be determined based on social activities between u_i and u_j. The weight CUU_{ij}, which is quantified based on the direct or indirect social influence between u_i and u_j, can be expressed, respectively, as follows:

$$\mathrm{CLR}_{ij} = e^{-|\mathrm{LYR}_i - \mathrm{LYR}_j|} \tag{2}$$

where CLR_{ij} represents the direct influence-based correlation, and $e^{-|\mathrm{LYR}_i - \mathrm{LYR}_j|}$ measures the correlation of two reachable u_i and u_j from different layers.

$$\mathrm{CFR}_{ij} = \frac{\mathrm{sm}_i \cdot \mathrm{sm}_j}{\|\mathrm{sm}_i\| \cdot \|\mathrm{sm}_j\|} \tag{3}$$

where CFR_{ij} represents the indirect influence-based correlation between u_i and u_j within the same layer. It can be measured according to the similarity of u_i and u_j's interesting items. Assuming there are in total K items interested by u_i and u_j, the indicator τ_{ik} is introduced to indicate whether u_i/u_j is interested in $\mathrm{itm}_k(\tau_{ik} = 1)$ or not $(\tau_{ik} = 0)$. $\mathrm{sm}_i = \langle \mathrm{Int}_i, \tau_{i1} \cdot F_1, \tau_{i2} \cdot F_2, \ldots, \tau_{ik} \cdot F_k\rangle$ is the semantic measurement of u_i to the K items.

C. Item Relations Associated With Semantic Attribute and Social Influence

Another important factor is to measure the relationships among items within IN. Given a specific itm_i, the semantic attribute which can be represented from its F_i, and the social influence, which can be inferred from its connections with users are utilized to analyze their relationships.

$E_{\mathrm{IN}} \subset I \times I$ indicates one kind of inner relation between two items. In particular, $E_{\mathrm{IN}} = \{\langle \mathrm{itm}_i, \mathrm{CII}_{ij}, \mathrm{itm}_j\rangle \mid \mathrm{itm}_i, \mathrm{itm}_j \in \mathrm{IN}\}$ can be determined based on the similarities in terms of semantic attributes and social influence between itm_i and itm_j. The detailed vector can be expressed as follows:

$$\mathrm{LFM}_i = \langle F_i, \mathrm{Cui}_i, \mathrm{Ctg}_i\rangle \tag{4}$$

Therefore, the correlation CII_{ij} between two items can be measured base on their similarities as follows:

$$\mathrm{CII}_{ij} = \frac{\mathrm{LFM}_i \cdot \mathrm{LFM}_j}{\|\mathrm{LFM}_i\| \cdot \|\mathrm{LFM}_j\|} \tag{5}$$

Furthermore, the user-item connection is analyzed to represent relations between UN and IN.

$E_{\mathrm{UI}} \subset U \times I$ indicates one kind of interest-based relation between users and items. In particular, $E_{\mathrm{UI}} = \{\langle u_i, \mathrm{CUI}_{ij}, \mathrm{itm}_j\rangle \mid u_i \in \mathrm{UN} \wedge \mathrm{itm}_j \in \mathrm{IN}\}$, where CUI_{ij} can be measured to indicate whether u_i is interested in itm_j $(\mathrm{CUI}_{ij} = 1)$ or not$(\mathrm{CUI}_{ij} = 0)$.

IV. Intelligent Recommendation in Heterogeneous Data Environments

A. Random Walk With Restart Framework

Given a randomly selected user or item as an initial vertex in the HHN model, the goal of intelligent recommendation is to detect the optimal users in UN or items in IN. The RWR model is employed to extract the structure-aware features and figures out the optimal node based on their relevance in a constructed HHN model.

The basic framework of the RWR model for recommendations from the HHN can be expressed as follows:

$$\mathrm{HR}^{(t+1)} = \lambda M * \mathrm{HR}^{(t)} + (1 - \lambda)q \tag{6}$$

where λ, ranging from 0 to 1, is a damping coefficient. HR^t indicates a ranking score vector at the iteration step t. q is the initial vector when starting the RWR model, which is constructed according to the initial state of the network. Specifically, q is initialized as $[0, 0, \ldots, 1, \ldots, 0, 0]$ and we set $\mathrm{HR}^0 = q$, in which "1" indicates the target vertex v_t at the beginning. M is a transfer matrix, which indicates the probability of each vertex to transfer to one another.

Input: The HHN network $G_{HHN}(N, E, C)$;
A given target node v_t
Output: A set of routing decisions Act

Initialize function R and target function $\hat{R}$ by weight θ;
Initialize parameters: *NumOfEpisode, T,* $Act = \emptyset$;
Calculate the correlation weights CUI_{ij}, CUU_{ij}, CII_{ij}on each edge in *HHN*;
Traverse each two vertices v_i and v_j in $UN \cup IN$:
  for eps = 1 to *NumOfEpisode* **do**
    Initialize $S = \{s_1\}$, and $\Phi_1 = \Phi(s_1)$;
    for t = 1 to T **do**
      Execute a random routing decision act_{ij}^t and then observe reward $Rwd^t = R(s^t, act_{ij}^t)$;
      Set $s^{t+1} = s^t$, act_{ij}^t and process $\Phi^{t+1} = \Phi(s^{t+1})$;
      Store transition $(\Phi^t, act_{ij}^t, Rwd^t, \Phi^{t+1})$ in D;
      Conduct a gradient descent optimizer on $R(\Phi^t, act_{ij}^t, \theta)$;
      Sample minibatch of transition $(\Phi^t, act_{ij}^t, r^t, \Phi^{t+1})$ from D;
      Calculate optimal action $\widehat{act_{ij}^t}$ by Φ^t;
    end for
  end for
  $Act = Act \cup \widehat{act_{ij}^t}$;
return Act;

Fig. 2. Algorithm for IR based on deep reinforcement learning.

Input: HHN network $G_{HHN}(N, E, C)$;
Action set $\{Act\}$ from the IR Algorithm;
A randomly selected target node v_t
Output: top-n recommendation node list according to v_t

Initialize $HR^{(0)} = q$ with target node v_t ;
Initialize parameters *NumOfIteration*, δ, *diff* for RWR framework;
for each node a_i in network *IN* **do**
  Construct LFM_i for a_i by Eq. (4);
end for
for each W_{ij} in M
  Calculate CUI_{ij}, CUU_{ij}, CII_{ij};
  Calculate W_{ij} with the following rules:
  case: edges in E_{UN}:
    $W_{ij} = CUU_{ij} * act_{ij}$;
  case: edges in E_{UI}:
    $W_{ij} = CUI_{ij} * act_{ij}$;
  case: edges in E_{IN}:
    $W_{ij} = CII_{ij} * act_{ij}$;
end for
for *iter* = 1 to *NumOfIteration*
  Calculate $HR^{(iter+1)}$ and $HR^{(iter)}$ by Eq. (6);
  if $HR^{(iter+1)} - HR^{(iter)} < \delta$:
    break loop;
end for
Rearrange the nodes in *UN* and *IN* respectively based on their ranking scores *HR*;
return top-n recommendation node list for the input v_t;

Fig. 3. Algorithm for intelligent recommendation.

Significantly, an intelligent router based on deep reinforcement learning scheme is introduced to improve the transfer matrix, which can optimize the RWR process and make it more suitable to solve the problem in this article.

B. Intelligent Router Based on Deep Reinforcement Learning

Two key issues that need to be considered when designing the routing scheme in hybrid networks during the RWR process are: 1) how to set a reasonable routing rule within the complex heterogeneous network during the RWR navigation and 2) how to deal with the lack of labels when learning the routing rule automatically. Therefore, an intelligent router via the reinforcement learning method is developed to find an optimal strategy to maximize the reward during the routing process. With a giving reward, routing rules are learned by the agent to improve the performance of navigation classifier with HHN.

In this article, a tuple $(S, \text{Act}, \text{Rwd})$ is defined to evaluate the mapping from the latent state to the decision action, in which S is a set of latent states, Act is a set of actions, and Rwd is a set of reward determined by the chosen action and the current state. $R(s, \text{act})$ is an action-value function to compute the reward for an action act $\in$ Act at state $s \in S$. The purpose is to find an optimal decision $\hat{\text{act}}$ at each step by estimating the state s. The intelligent routing (IR) algorithm is shown in Fig. 2.

C. Mechanism for Intelligent Recommendation

To improve the navigation within the HHN and choose the optimal solution for recommendations, both the correlation between two connected vertices (i.e., user or item) and routing action are taken into account to construct the transfer matrix M during the RWR process. Specifically, given a constructed HHN model, the routing weight W_{ij} for two vertex v_i and v_j can be defined as follows:

$$W_{ij} = C_{ij} * \text{act}_{ij} \tag{7}$$

where C_{ij} denotes a specific correlation between v_i and v_j. act_{ij} indicates the corresponding routing action from v_i to v_j. For example, if there is an edge between item_i and item_j, W_{ij} in M can be calculated by CII_{ij} in (5) and act_{ij} via the IR algorithm.

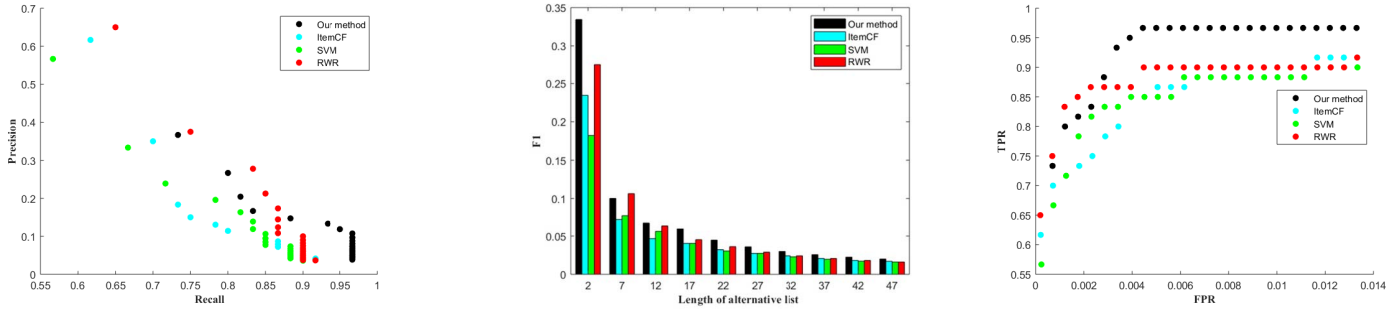
Accordingly, the intelligent recommendation is shown in Fig. 3, which is able to find out more optimal nodes based on the improved transfer matrix M via deep reinforcement learning scheme.

Finally, the ranked top-n nodes can be generated. Given a specific user or item as the target node, the improved RWR framework incorporating with the intelligent router can find the more valuable nodes, which can be provided as alternative recommendations to improve the intelligent navigation in heterogeneous big data environments.

V. Experiment and Analysis

A. Data Set

To evaluate the effectiveness of the proposed HHN model in heterogeneous big data environments, we collected a real-world data set from the scholarly social network. A crawling



(a) PR-Curve (b) F1 (c) ROC

Fig. 4. Performance evaluation for intelligent recommendation. (a) PR-curve. (b) F1. (c) ROC.

program developed by Python 3.6 is deployed to collect data from DBLP and ResearchGate, respectively. DBLP is an online open-data service for bibliographic information. As of the beginning of 2019, DBLP indexed over 4 million publications by more than 2 million authors, including more than 40 000 journal volumes and 39 000 conference and workshop proceedings. ResearchGate is another online scientific, social community service which contains over 140 million publications and connects with over 15 million users. In addition, ResearchGate captures a variety of relationships for both publications and users in cyberspace, including the following relationship, citations, co-authorships, and so on. Therefore, two entities, namely, researchers and articles, were employed to construct the proposed HHN model based on the collected DBLP and ResearchGate data.

We crawled more than 370 000 articles and 2 million citation records generated by these articles. In addition, 13 100 researchers, along with their profiles, are collected, along with nearly 800 thousand following relationships among them. We separated the whole data set into two parts, in which 70% of the data is the training data, and the other 30% is the test data.

B. Experiment Configuration and Evaluation Metrics

To demonstrate the effectiveness of our HHN model in real-world applications, an experiment scenario is designed as given a randomly selected target object (researcher/article) in the test data set; we evaluate our method on recommending it with a top-n valuable collaborative object list.

Four widely used metrics in recommendation evaluation, namely, precision, recall, F1, and receiver operating characteristic (ROC), are employed in this article to evaluate the performance of our proposed method. We set the length of lists of recommendation alternatives from 2 to 50. The basic RWR recommendation [6], ItemCF-based recommendation [29], and SVM recommendation [30] are selected to compare recommendations in heterogeneous big data environments, with the proposed method.

Different optimizers were tested for deep neural network, and the gradient descent performs the best when considering the speed of reaching min loss. The ReLu activation function and SoftMax classifier are applied in the deep neural network. To avoid overfitting for the model, dropout function with a rate of 0.5 is applied. To find the most valuable collaboration objects in HHN, a reward (i.e., $+1$) is set for every right collaboration hit. The batch size of the learning process is 30. A server of Intel Dual-Core Xeon E5-2630V3 8C/16T 2.40-GHz 20-MB L3, 28-GB/DDR4/2133 RAM, CentOS, TensorFlow r2.0 is utilized to conduct the experiments.

TABLE I
EFFICIENCY OF IR

Episodes	Rewards	
	Our method	Q-Learninng
50	10	10
100	140	130
200	231	188
300	**1000**	450
400	1000	**1000**
500	1000	1000
600	1000	1000
700	1000	1000

C. Performance Evaluation

To evaluate the efficiency of IR, we performed a test on how fast the proposed method can achieve the highest reward. The Q-learning method was tested as a baseline to quantitatively compare the learning efficiency. Table I shows how rewards increase with learning episodes for our proposed learning scheme and Q-learning scheme.

As shown in Table I, we observed that both our proposed learning scheme and Q-learning method could reach the same rewards of 1000 finally. This result indicates that the semi-supervised learning model can improve the efficiency of the IR within the heterogeneous network. In addition, our deep reinforcement learning scheme with the HHN model reaches the maximum rewards at 300 episodes, which performs more efficiently than Q-learning in this application scenario.

Furthermore, to evaluate the effectiveness of the proposed method, especially when handling intelligent recommendations with heterogeneous data, we tested both the user-based and item-based intelligent recommendations in the experiment with the scholarly big data. Given a randomly selected target object (researcher or article) with a maximum of 50 alternatives, the proposed method was used to generate the optimal academic collaborations alternatives, and the results were compared with the aforementioned four baseline methods.

Experiments were conducted according to the previously mentioned recommendation scenario, and the performances of

the four methods evaluated according to the metrics introduced earlier are shown in Fig. 4(a)–(c), respectively.

First, the downdrift PR-curves of the four methods shown in Fig. 4(a) indicate how the precision and recall metrics change and their correlations in this scenario. Based on these, the proposed method outperforms the others in terms of the recommendation results. It can be explained as the reinforcement learning-based intelligent router can automatically decide and help the RWR algorithm to transmit the more relevant nodes in different layers across the HHN model.

Furthermore, the F1-score curves shown in Fig. 4(b) demonstrate the general performance of the four methods. It shows that the performances of all the methods decrease with the increasing length of recommendation lists. It is noted our method using the reinforcement learning-based intelligent router can achieve about 5% improvement than the other methods and reach the peak result of 0.33 when two recommendation alternatives are provided.

Moreover, the four ROC curves rise from left to right and locate at the position of the upper left corner beyond the diagonal line in Fig. 4(c), which indicates all the methods can obtain a reasonable result for the experiment scenario. It is observed that the proposed method gains 10% higher performance than the other three methods. The result explains the IR scheme introduced in our method can efficiently improve the recommendation effectiveness in the hierarchical model. The above-mentioned results indicate that the semisupervised scheme is a practical way to deal with the heterogeneous big data using our proposed HHN model.

VI. Conclusion

In this article, we proposed an HHN model for deep correlation mining, which can be applied for intelligent recommendations in the heterogeneous big data environments.

We designed a hierarchical architecture to capture multitype relationships among different users and items across multiple layers. The internal social influence within one specific layer or external social influence between different layers was taken into account to quantify their correlations with a series of measures. An intelligent router was then developed based on the deep reinforcement learning to obtain optimal decisions of whether jump into or jump out across the hierarchical network during the RWR process. An RWR-based algorithm was improved to provide an intelligent recommendation. As an application scenario, our proposed model was applied in the scholarly network analysis, to deal with real-world heterogeneous big data. Experiments based on DBLP and ResearchGate data demonstrated the practicability and effectiveness of the proposed HHN model in providing users with collaboration work support.

For future studies, the policy of IR within a heterogeneous network can be further investigated. More evaluations on different kinds of data set need to be conducted to examine and improve the algorithm for different kinds of practical situations using the proposed HHN model.

References

[1] X. Wang, L. T. Yang, H. Liu, and M. J. Deen, "A big data-as-a-service framework: State-of-the-art and perspectives," *IEEE Trans. Big Data*, vol. 4, no. 3, pp. 325–340, Sep. 2018.

[2] J. Qi, P. Yang, L. Newcombe, X. Peng, Y. Yang, and Z. Zhao, "An overview of data fusion techniques for Internet of Things enabled physical activity recognition and measure," *Inf. Fusion*, vol. 55, pp. 269–280, Mar. 2020.

[3] W. Yang, X. Liu, L. Zhang, and L. T. Yang, "Big data real-time processing based on storm," in *Proc. 12th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Melbourne, VIC, Australia, Jul. 2013, pp. 1784–1787.

[4] P. Yang *et al.*, "DUAPM: An effective dynamic micro-blogging user activity prediction model towards cyber-physical-social systems," *IEEE Trans. Ind. Informat.*, early access, Dec. 16, 2019, doi: 10.1109/TII.2019.2959791.

[5] P. Yang, J. Liu, J. Qi, Y. Yang, X. Wang, and Z. Lv, "Comparison and modelling of country-level microblog user and activity in cyber-physical-social systems using Weibo and Twitter data," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 6, pp. 1–24, Dec. 2019.

[6] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 364–375, Sep. 2014.

[7] X. Zhou, W. Liang, K. I.-K. Wang, R. Huang, and Q. Jin, "Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data," *IEEE Trans. Emerg. Topics Comput.*, early access, Jul. 26, 2018, doi: 10.1109/TETC.2018.2860051.

[8] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "PPHOPCM: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing," *IEEE Trans. Big Data*, early access, May 5, 2017, doi: 10.1109/TBDATA.2017.2701816.

[9] Q. Zhang, L. T. Yang, Z. Yan, Z. Chen, and P. Li, "An efficient deep learning model to predict cloud workload for industry informatics," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3170–3178, Jul. 2018.

[10] X. Zhou, W. Liang, S. Huang, and M. Fu, "Social recommendation with large-scale group decision-making for cyber-enabled online service," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 1073–1082, Oct. 2019.

[11] T. Xu, H. Zhu, H. Zhong, G. Liu, H. Xiong, and E. Chen, "Exploiting the dynamic mutual influence for predicting social event participation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 6, pp. 1122–1135, Jun. 2019.

[12] K. Cepni, M. Ozger, and O. B. Akan, "Event estimation accuracy of social sensing with facebook for social Internet of Vehicles," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2449–2456, Aug. 2018.

[13] Z. Yu, F. Yi, Q. Lv, and B. Guo, "Identifying on-site users for social events: Mobility, content, and social relationship," *IEEE Trans. Mobile Comput.*, vol. 17, no. 9, pp. 2055–2068, Sep. 2018.

[14] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 46–58, Jan. 2020.

[15] H. Lin *et al.*, "Detecting stress based on social interactions in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1820–1833, Sep. 2017.

[16] Y. Zhao, W. Song, and Z. Han, "Social-aware data dissemination via device-to-device communications: Fusing social and mobile networks with incentive constraints," *IEEE Trans. Services Comput.*, vol. 12, no. 3, pp. 489–502, May 2019.

[17] G. Liao, X. Huang, M. Mao, C. Wan, X. Liu, and D. Liu, "Group event recommendation in event-based social networks considering unexperienced events," *IEEE Access*, vol. 7, pp. 96650–96671, 2019.

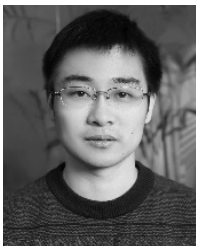
[18] S. Liu, B. Wang, and M. Xu, "SERGE: Successive event recommendation based on graph entropy for event-based social networks," *IEEE Access*, vol. 6, pp. 3020–3030, 2018.

[19] W. Zhang, F. Kong, L. Yang, Y. Chen, and M. Zhang, "Hierarchical community detection based on partial matrix convergence using random walks," *Tsinghua Sci. Technol.*, vol. 23, no. 1, pp. 35–46, Feb. 2018.

[20] G. Yu, G. Fu, J. Wang, and Y. Zhao, "NewGOA: Predicting new GO annotations of proteins by bi-random walks on a hybrid graph," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1390–1402, Jul. 2018.

[21] Z. Fang and X. Lei, "Prediction of miRNA-circRNA associations based on k-NN multi-label with random walk restart on a heterogeneous network," *Big Data Mining Anal.*, vol. 2, no. 4, pp. 261–272, Dec. 2019.

[22] Y. Zhou, Z. Tang, L. Qi, X. Zhang, W. Dou, and S. Wan, "Intelligent service recommendation for cold-start problems in edge computing," *IEEE Access*, vol. 7, pp. 46637–46645, 2019.
[23] X. Zhong, Y. Zhang, D. Yan, Q. Wu, Y. T. Yan, and W. Li, "Recommendations for mobile apps based on the HITS algorithm combined with association rules," *IEEE Access*, vol. 7, pp. 105572–105582, 2019.
[24] F. Lin, Y. Zhou, I. You, J. Lin, X. An, and X. Lu, "Content recommendation algorithm for intelligent navigator in fog computing based IoT environment," *IEEE Access*, vol. 7, pp. 53677–53686, 2019.
[25] M. Mao, J. Lu, G. Zhang, and J. Zhang, "Multirelational social recommendations via multigraph ranking," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4049–4061, Dec. 2017.
[26] S. Meng, H. Wang, Q. Li, Y. Luo, W. Dou, and S. Wan, "Spatial–temporal aware intelligent service recommendation method based on distributed tensor factorization for big data applications," *IEEE Access*, vol. 6, pp. 59462–59474, 2018.
[27] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 609–618, Mar. 2017.
[28] X. Zhou, B. Wu, and Q. Jin, "Analysis of user network and correlation for community discovery based on topic-aware similarity and behavioral influence," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 6, pp. 559–571, Dec. 2018.
[29] C. M. Rodrigues, S. Rathi, and G. Patil, "An efficient system using item & user-based CF techniques to improve recommendation," in *Proc. 2nd Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Dehradun, Uttarakhand, 2016, pp. 569–574.
[30] F. Ali *et al.*, "Merged ontology and SVM-based information extraction and recommendation system for social robots," *IEEE Access*, vol. 5, pp. 12364–12379, 2017.



Xiaokang Zhou (Member, IEEE) received the Ph.D. degree in human sciences from Waseda University, Tokyo, Japan, in 2014.

From 2012 to 2015, he was a Research Associate with the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University. He is currently an Associate Professor with the Faculty of Data Science, Shiga University, Hikone, Japan. He has also been a Visiting Researcher with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo, since 2017. He is engaged in interdisciplinary research works in the fields of computer science and engineering, information systems, and social and human informatics. His recent research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, cyber-physical-social-system, cyber intelligence, and cyber-enabled applications.

Dr. Zhou is a member of the IEEE Computer Society (CS), Association for Computing Machinery (ACM), USA, the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence (JSAI), Japan, and the China Computer Federation (CCF), China.

Wei Liang (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2005 and 2016, respectively.

He is currently with the Key Laboratory of Hunan Province for Mobile Business Intelligence, Hunan University of Technology and Business, Changsha, China. He has published more than 20 articles at various conferences and journals. His research interests include information retrieval, data mining, and artificial intelligence.

Dr. Liang is a member of the IEEE Computer Society (CS) and the China Computer Federation (CCF), China.

Kevin I-Kai Wang (Member, IEEE) received the B.E. degree (Hons.) in computer systems engineering and the Ph.D. degree in electrical and electronics engineering from the Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand, in 2004 and 2009, respectively.

He was a Research Engineer with The University of Auckland, where he was involved in designing commercial home automation systems and traffic sensing systems from 2009 to 2011, where he is currently a Senior Lecturer with the Department of Electrical and Computer Engineering. His current research interests include wireless sensor network-based ambient intelligence, pervasive healthcare systems, human activity recognition, behavior data analytics, and bio-cybernetic systems.

Laurence T. Yang (Fellow, IEEE) received the B.E. degree in computer science and technology and B.Sc. degree in applied physics from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of Victoria, Victoria, BC, Canada.

He is currently a Professor and the W. F. James Research Chair with the Department of Computer Science, St. Francis Xavier University, Antigonish, NS, Canada. His research was supported by the National Sciences and Engineering Research Council, Canada, and the Canada Foundation for Innovation. His research interests include parallel and distributed computing, embedded and ubiquitous/pervasive computing, and big data.