Hindawi
Advances in Multimedia
Volume 2018, Article ID 2058670, 16 pages
https://doi.org/10.1155/2018/2058670

Research Article

Temporal Activity Path Based Character Correction in Heterogeneous Social Networks via Multimedia Sources

Jun Long,[1] Lei Zhu,[1] Zhan Yang,[1] Chengyuan Zhang,[1] and Xinpan Yuan[2]

[1]*School of Information Science and Engineering, Central South University, Changsha 410083, China*
[2]*School of Computer, Hunan University of Technology, Zhuzhou 412007, China*

Correspondence should be addressed to Chengyuan Zhang; cyzhang@csu.edu.cn

Received 23 November 2017; Accepted 26 February 2018; Published 24 May 2018

Academic Editor: Meng Fang

Vast amount of multimedia data contains massive and multifarious social information which is used to construct large-scale social networks. In a complex social network, a character should be ideally denoted by one and only one vertex. However, it is pervasive that a character is denoted by two or more vertices with different names; thus it is usually considered as multiple, different characters. This problem causes incorrectness of results in network analysis and mining. The factual challenge is that character uniqueness is hard to correctly confirm due to lots of complicated factors, for example, name changing and anonymization, leading to character duplication. Early, limited research has shown that previous methods depended overly upon supplementary attribute information from databases. In this paper, we propose a novel method to merge the character vertices which refer to the same entity but are denoted with different names. With this method, we firstly build the relationship network among characters based on records of social activities participating, which are extracted from multimedia sources. Then we define temporal activity paths (TAPs) for each character over time. After that, we measure similarity of the TAPs for any two characters. If the similarity is high enough, the two vertices should be considered as the same character. Based on TAPs, we can determine whether to merge the two character vertices. Our experiments showed that this solution can accurately confirm character uniqueness in large-scale social network.

1. Introduction

In the past decade, the mobile Internet and social multimedia applications have become an indispensable part of social life, and huge multimedia data are being produced and consumed [1]. For instance, Facebook reports 350 million photos uploaded daily as of November 2013; 100 hours of video are uploaded to YouTube every minute, resulting in more than 2 billion videos totally by the end of 2013 [2]. Social Media Networks allow people to communicate, share, comment, and observe different types of multimedia content [3]. As social activities are becoming more frequent, social networks have been larger and much more complex. Generally, we extract information and construct social transaction databases from vast amount of multimedia data, such as text, images [4, 5], videos, and audios [6], to construct large-scale social networks which are modelled by graphs [7] with node-edge representation [8]. Multimedia data, generally, can be described in multiviews [9, 10] such as color view and textual view [11, 12]. In social networks, the relations between character vertices are tangled by the time difference of transaction, incompletion of personal information record, anonymous phenomena, and difference of information pattern and structure. It is distinctly difficult to maintain one-to-one mapping between characters in relation networks and people in real life. Besides, characters are marked up as difference vertices by former and present name. These vertices have the same personal information, structure, and attributes of relation. For social networks, these vertices and relationships are redundant, which will severally perturb the results of social network analysis. Therefore, character vertices ambiguity has become a key problem in social network analysis.

In relational databases, we can use multidimensional personal information to confirm uniqueness of characters, such as name, gender, and date of birth. In big data environment, however, multimedia data is mainly from unstructured data storage. Its scale is vast and types are multifarious [13, 14], such as text, images, videos, and audios. Besides, the data is

not complete and consistent generally, which caused the uniqueness of characters to be difficult to determine. Most of the large-scale multimedia data, nevertheless, are stored externally and included people's participation in social activities, such as vocational and educational experience and participating in service club. Social relationships are generated by using multimedia data of these activities, and then the social networks can be built up. It is a key problem to confirm the character uniqueness in social networks analysis and application. We propose to measure the similarity of characters by network characteristics to conduct character correction by vertices merging with computing structure error of networks. However, social networks have temporal attributes generally, and relations extracted from them also have obvious temporal characteristics. We regard temporal attributes as a key factor of relations, which is used in computing the similarity of vertices. Accordingly, it boosts the accuracy of uniqueness conforming. We put forward new notions of character activity path and transaction activity network with heterogeneous features and then use temporal activity path similarity evaluation to improve the reliability of character correction.

The remainder of this paper is organized as follows. We introduce related work in Section 2 and then describe an academic network building process in Section 3. In Section 4, we present character vertices merging principle and structure error algorithm based on network structure. In Section 5, we introduce transaction activity networks, activity paths algorithm, and vertices screening and then analyze the experiment results. The conclusion and future work are included in Section 6.

2. Related Work

2.1. Social Analysis via Multimedia. The advent of social networks and cloud computing has made social multimedia sharing in social networks become easier and more efficient [15]. With the rapid increasing of volume of multimedia data, social networks analysis and mining via multimedia data attract attention of a number of researchers recently. Zhuhadar et al. proposed combination of social learning network analysis and social learning content analysis in studying the impact of the social multimedia systems cyberlearners [16]. They presented evidence obtained from the analysis that Social Multimedia System impacts the communication between faculty and students. To deal with the challenges of event detection rom massive social media data in social networks, Zhao et al. [17] proposed a novel real-time event detection method named microblog clique to explore the high correlations among different microblogs, which was supported by social multimedia data. Sang and Xu [2] proposed to analyze into variety of big social multimedia from the perspective of various sources. Laforest et al. [18] present a new kind of social networks named spontaneous and ephemeral social networks (SESNs) which allow people to collaborate spontaneously in the production of multimedia documents. In order to find overlapping communities from multimedia social networks, Huang et al. [19] proposed an efficient algorithm named LEPSO for overlapping communities discovery, which is based on line graph theory, ensemble learning, and particle swarm optimization.

2.2. Name Ambiguity and De-Anonymity. Recently, name ambiguity and de-anonymity have been widely studied. There are several methods to identify characters in social networks, which can be divided into three categories, for internal relational database, Internet webpage, and topology structure of social networks.

Name ambiguity and de-anonymity are with the same essential features. In the past, the identity of the characters is determined by the accurate attribute information in internal databases of enterprises. In 2008 Narayanan and Shmatikov proposed the method to process high dimensional data [20], such as personal attribute, recommendation, and transaction information. Users can identify characters in the anonymous database with limited personal information. This method has strong robust even though background information is inaccurate or disturbed. However, internal database is localized and static; it cannot describe feature of characters thoroughly. Therefore, these methods are not suitable for name ambiguity problem in big data with complexity, dynamicity, and cross platform.

Name ambiguity is more prominent in the Internet. In 2008 Tang et al. proposed a standard probability framework to recognize the independence of observed objects [21]. But when we search name on the Internet, numerous webpages containing one same name can be returned, and it is not certain whether these pages belong to the same people. Bekkerman and McCallum proposed two statistical methods to solve this problem in 2005 [22]. One is based on link structure of webpages and the other is on multiway distributional clustering method, which is unsupervised frameworks and only needs a few of prior knowledge, and the experiments show that their solution outperform traditional clustering. However, the above methods are deeply subject to the uncertainty of web information.

At present, name disambiguation becomes even more prominent in social networks modeling and analysis. In 2008, Liu and Terzi analyzed character-centered social networks and then pointed out that features of relation structure can expose characters' identity [23]. For identity hiding in social relations network, they defined graph-anonymization and proposed the algorithm based on k-degree anonymous graph and node degree sequence. Narayanan and Shmatikov defined privacy of social networks in 2009 [24] and designed a novel reidentification algorithm which can implement de-anonymity and identifying node by using the topological structure of network. For dynamic evolution of social networks, Ding et al. proposed the "threading" technique and used the connection between released data to implement de-anonymizing [25]. And they proposed to combine structure information and attributes of nodes to reidentify anonymous nodes. Korayem and Crandall worked on de-anonymizing method specially in cross platform social networks [26], which can recognize that different accounts belong to one user by extracting time sequence data, text features, geographic location, and social relation characteristics. In 2012,

Srivatsa and Hicks introduced de-anonymizing users' mobile trace information based on graph structure of social networks [27]. As contact graph between characters consists of vast quantities of mobile trace, they proposed structure similarity of interuser correlation, which was used to map contact graph and social network.

Since a large number of mobile trajectories can be used to build the contact graph of characters, the structural similarity is employed to find out the corresponding nodes in the contact graph and social network. The de-anonymity with mobile trace is implemented by mapping character nodes between the two networks. The methods mentioned above aim to solve the problem of anonymity in traditional networks in which nodes and relations have only one category. However, the social network created on big data is comprehensive, such as category diversity of relationship and nodes and temporality. In addition, these methods need to add attributes to supplement topology information of social networks.

From enterprise databases to webpages databases and then to social networks databases, this is a developing process from local data to global data. Previous methods rely on attribute details of local data; namely, it needs much more auxiliary information to identify characters, but the efficiency is low. However, big data is multisourced [28], time-variable, global, and macroscopic. The social networks are built on global data, and it is impossible to look back upon distributed sources. In this type of networks, intrinsic relationship structure of vertices is a key factor to measure uniqueness of characters. Since different characters have different social relationships, we can identify characters by network structure features. For the heterogeneity and temporality of big data networks, we propose uniqueness correction method and the notion of activity path similarity based on heterogeneous temporal networks [29, 30], to promote the efficiency and accuracy of character identification.

3. Social Network Modeling

A large-scale social network is based on diversified multimedia data which is multimodality [31]; for instance, an image can be described by color modality or shape modality. It contains information of multifarious and complex social activities. We can build this kind of network by extracting relations from transaction activity information which is extracted from multimedia datasets. As academic network is a typical case of social network, we use it as an example to describe the process of social networks mining.

In general, academic relations mainly include teacher-student relationship, classmate relationship, project partnership, and coauthor relationship. These relations are contained in education experiences, research and work experiences, cooperation and coauthor experiences, and academic activities and conferences experiences. The information of academic activities is contained in project proposals, project progress and concluding reports, degree certificates, award certificates, photographs or videos concerning conference, and other scientific information documents. Therefore, we extract academic activities information from the multimedia data and then construct academic transaction activity network. It is the base of mining and analysis of academic network between scholars.

Figure 1 shows a general view of framework of academic relation network construction. First, we use academic activity transaction extract method from multimedia sources which contain texts, images, audios, and videos to collect individual resume information of scholar and team members information. Then we construct an academic activity transaction database. This database contains personal information, study and work experience information, and project and publication information. After that, we build academic activity relation network containing heterogeneous vertices and relations. On this basis, we create academic transaction activity networks. In this kind of networks, there are several types of transaction activities, such as study experiences ("graduated from Tsinghua University," "studying in Cambridge University," "was conferred doctor's degree," etc.), work experiences ("worked at Microsoft MSRA," "teaches in Central South University," etc.), publication information ("published (μ + λ) Evolutionary Strategy for 3D Modeling and Segmentation with Super quadrics," etc.), and research experiences ("took over The Association Rules Mining of Time Series and Knowledge Discovery for Recognition of Expert Academic Activities Track project," etc.).

These transaction networks are 2-mode networks which consist of two types of vertices: character vertices and entity vertices and their activity relations. These vertices represent scholars or researchers and academic entities, respectively. We can mine alumni relationship, workmate relationship, project cooperation, and coauthor relationship from them. The character vertices and academic relation constitute academic relationship network which is a kind of homogeneous 1-mode network. We proposed vertices merging method based on structure error of network to implement uniqueness correction in this 1-mode network.

4. Evaluating Uniqueness of Character Vertices Based on Structure Error

Redundant information of vertices and relation is generally carried out by nonunique character vertices. Thus, correct structure merging is a key process to remove redundant information from social networks. Theoretically, structure of networks will not be changed after redundant vertices and relations merging. We evaluated uniqueness of character vertices by merging test and then screened out redundant vertices candidates.

4.1. Evaluating Uniqueness of Character Vertices. In a social network, we consider the character vertices which have the same neighbor as suspicious redundant vertices. Some of them containing redundant information are nonunique, and the others with a high similarity may not be redundant. Thus, we call suspicious redundant vertices as redundant vertices candidates.

4.1.1. Uniqueness of Vertices. Let $G = \langle V, R, \Phi \rangle$ be a 1-mode network in which the vertices represent characters. Two

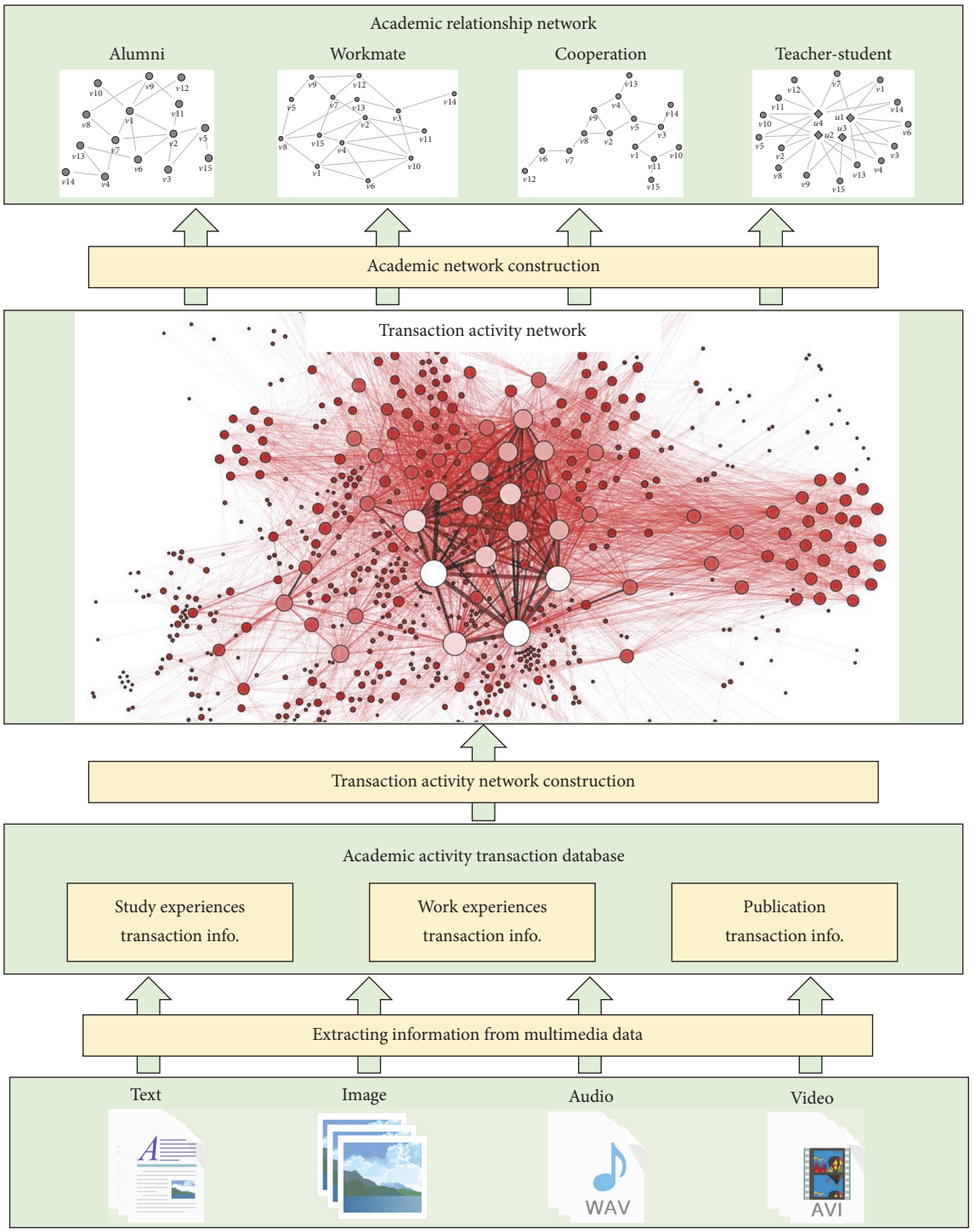

Figure 1: Framework of academic relationship network construction.

nonempty finite sets V and R are character vertices set and relations set. We denote the mapping from relations set to vertices set as $\Phi : R \to \{\langle v_x, v_y \rangle \mid v_x \in V \cap v_y \in V\}$. The set which has I vertices to be tested is denoted as V_I and $\widehat{V}_J$ is set of J neighbor vertices, and $V_I \cap \widehat{V}_J = \varnothing$, $\{v_1, v_2, \ldots, v_I\} \subseteq V_I$, $\{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_J\} \subseteq \widehat{V}_J$. The relation set which contains K relations between character vertices and neighbor vertices is denoted as $R = \{r_1, r_2, \ldots, r_K\}$ and $|R| = K$. The mapping from character vertices set to relations set is denoted as $\Psi : V_I \to R$, and the mapping from character vertices to its neighbor vertices is denoted as $\Lambda : V_I \to \widehat{V}_J$.

Property 1. In a 1-mode network G, if vertices $v_1, v_2, \ldots, v_I$ in the vertices set V have uniqueness, then $\Lambda(v_1) \neq \Lambda(v_2) \neq \cdots \neq \Lambda(v_I)$ and the values of structure error between V are not zero.

4.1.2. Redundant Vertices Candidates. In theory, character vertices which are nonunique have selfsame or nearly identical relation structure. Redundant relations and vertices are generated by this situation and they should be merged so as to remove redundant information. We introduce the notion of structure error to describe the difference of network structure between vertices. The vertices with selfsame or highly similar structure are referred to as redundant vertices candidates. They contain redundant relation information.

Definition 2 (redundant vertices candidates). In a network $G = \langle V, R, \Phi \rangle$, character vertices $v_1, v_2, \ldots, v_I$ are redundant vertices candidates if the values of structure error between them are zero, and the redundant vertices candidate set is denoted as $\mathrm{H} = \{v_1, v_2, \ldots, v_I\}$.

Definition 3 (redundant vertices). Let a vertices set be $\widetilde{\mathrm{H}} = \{\widetilde{\mathrm{H}}_1, \widetilde{\mathrm{H}}_2, \ldots, \widetilde{\mathrm{H}}_n\}$, $\widetilde{\mathrm{H}}_1 = \{v_1, v_2, \ldots, v_{\mu_1}\}$, $\widetilde{\mathrm{H}}_2 = \{v_{\mu_1+1}, v_{\mu_1+2}, \ldots, v_{\mu_2}\}, \ldots, \widetilde{\mathrm{H}}_n = \{v_{\mu_{n-1}+1}, v_{\mu_{n-1}+2}, \ldots, v_{\mu_n}\}$. If the vertices in $\widetilde{\mathrm{H}}_1, \widetilde{\mathrm{H}}_2, \ldots, \widetilde{\mathrm{H}}_n$ are nonunique, the set $\widetilde{\mathrm{H}}$ is referred to as redundant vertices set. The number of all vertices in $\widetilde{\mathrm{H}}$ is denoted as μ_n.

If $\widetilde{\mathrm{H}} = \{\widetilde{\mathrm{H}}_1, \widetilde{\mathrm{H}}_2, \ldots, \widetilde{\mathrm{H}}_n\}$ is the redundant vertices set, the vertices merging process is $\forall v_i \in \widetilde{\mathrm{H}}_1, v_{i+1} \in \widetilde{\mathrm{H}}_2, \ldots, v_{i+n} \in \widetilde{\mathrm{H}}_n$, and calculate $\mathrm{N}_1 = \widetilde{\mathrm{H}}_1 - \{v_i\}$, $\mathrm{N}_2 = \widetilde{\mathrm{H}}_2 - \{v_{i+1}\}, \ldots, \mathrm{N}_n = \widetilde{\mathrm{H}}_n - \{v_{i+n}\}$ and $\mathrm{E}_1 = \{r_1, r_2, \ldots, r_K\} - \Psi(v_i)$, $\mathrm{E}_2 = \{r_1, r_2, \ldots, r_K\} - \Psi(v_{i+1}), \ldots, \mathrm{E}_n = \{r_1, r_2, \ldots, r_K\} - \Psi(v_{i+n})$. Then calculate $V' = V - (\mathrm{N}_1 \cup \mathrm{N}_2 \cup \cdots \cup \mathrm{N}_n)$ and $R' = R - (\mathrm{E}_1 \cup \mathrm{E}_2 \cup \cdots \cup \mathrm{E}_n)$ to get the merged network $G' = \langle V', R', \Phi \rangle$. This new network does not contain any redundant information because $\mu_n - n$ vertices have been removed by vertices merging.

In Figure 2, network G_1 contains six character vertices and nine relations, we denote them as $V = \{v_1, v_2, v_3, \widehat{v}_1, \widehat{v}_2, \widehat{v}_3\}$ and $R = \{r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8, r_9\}$; the neighbors of vertices v_1, v_2, and v_3 are denoted as $\{\widehat{v}_1, \widehat{v}_2, \widehat{v}_3\}$. Before merging process, they connect with neighbor vertices $\widehat{v}_1, \widehat{v}_2, \widehat{v}_3$, respectively. Therefore $\Lambda(v_1) = \Lambda(v_2) = \Lambda(v_3) = \{\widehat{v}_1, \widehat{v}_2, \widehat{v}_3\}$; we regard v_1, v_2, and v_3 as redundant vertices candidates. In network G_2, vertices set and relation set are $V = \{v_4, v_5, \widehat{v}_4, \widehat{v}_5, \widehat{v}_6\}$ and $R = \{r_{10}, r_{11}, r_{12}, r_{13}\}$. The neighbor vertices set is $\{\widehat{v}_4, \widehat{v}_5, \widehat{v}_6\}$. Both v_4 and v_5 have two relations but structure of them is different; namely, $\Lambda(v_4) = \{\widehat{v}_5, \widehat{v}_6\}$, $\Lambda(v_5) = \{\widehat{v}_4, \widehat{v}_6\}$ and $\Lambda(v_1) \neq \Lambda(v_2)$. Thus, they are not redundant.

4.1.3. Structure Error. After merging process, the number of relations between neighbor $\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_J$ and character vertices $v_1, v_2, \ldots, v_I$ has been changed, but the number of relations between character vertices and their neighbor vertices and the number of neighbors remain unchanged. According to this principle, we use these numbers to define structure error which is the validation criteria of structure merging.

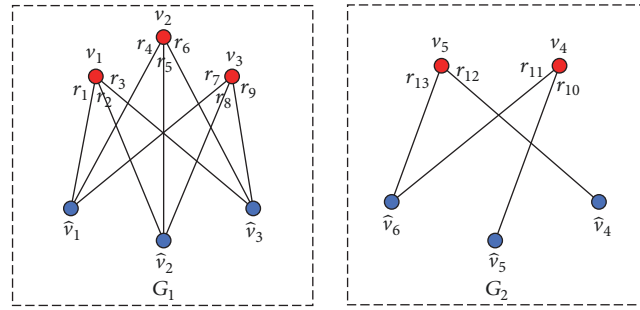

Figure 2: Redundant vertices candidate.

Definition 4 (structure error). In network $G = \langle V, R, \Phi \rangle$, the character vertices subset is denoted as $\{v_1, v_2, \ldots, v_I\} \subseteq V$, and the neighbor vertices subsets before and after merging are $\{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_J\} \subseteq V$ and $\{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_{J'}\} \subseteq V$, $\forall v_x, v_y \in \{v_1, v_2, \ldots, v_I\}$; then

$$\varepsilon_\xi\left(v_x, v_y\right) = \sum_{j=1}^{J} \delta\left(\widehat{v}_j\right) - \sum_{j'=1}^{J'} \delta'\left(\widehat{v}_{j'}\right) - \frac{\delta\left(v_x\right) + \delta\left(v_y\right)}{2}. \quad (1)$$

In (1), $\widehat{v}_j \in \{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_J\}$, $\widehat{v}_{j'} \in \{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_{J'}\}$, and the numbers of neighbor vertices before and after merging are denoted as J and J'. $\delta(v_x)$ and $\delta(v_y)$ severally represent the number of relations between v_x and v_y and their neighbors. ξ represents the type label of social networks and Ξ is the set of type labels, $\xi \in \Xi$. The structure error of v_x and v_y is denoted as $\varepsilon_\xi(v_x, v_y)$.

Based on this notion, we can recognize redundant vertices candidate from social networks according to structure error. If $\varepsilon_\xi(v_x, v_y) \neq 0$, we can regard v_x and v_y as a vertex pair with uniqueness, whereas they are redundant vertices candidates.

4.2. Algorithm. We designed redundant vertices candidates screening method in social networks according to the above notion, which is shown in Algorithms 1 and 2. Firstly, we arbitrarily select two character vertices v_x and v_y from networks and then calculate the number of relations between character vertices and their neighbors. We denote it as *preRelations*. Secondly, based upon merging principle we calculate the number of the relations between them after correct merging, and it is denoted as *postRelations*. Lastly, we calculate structure error of each vertex pairs and put the vertices which have zero value of structure error into redundant vertices candidates set.

5. Character Uniqueness Measure Based on Activity Path Similarity

The temporal attributes of semantic relations are composed by start time and end time of activities. We can use them to construct heterogeneous temporal social networks which consisted several different types of subnetworks. Each of the subnetworks contains only one type of relations. Vertices similarity is therefore decided by activity relations between

Input: **social network** $G = \langle V, R, \Phi \rangle$
Output: **candidate redundant vertices set H**
(01) **Initialize** list $L_1, L_2 \leftarrow V$
(02) **for** $i \leftarrow 1$ **to** $|V|$ **do**
(03) **for** $j \leftarrow 1$ **to** $|V|$ **do**
(04) **if** the name of $L_1[i] \neq$ the name of $L_2[j]$ **then**
(05) **if** $\varepsilon_\xi(L_1[i], L_2[j]) = 0$ **then**
(06) **add** $L_1[i], L_2[j]$ **into** H
(07) **end if**
(08) **end if**
(09) **end for**
(10) **end for**
(11) **return** H

Algorithm 1: Candidate redundant vertices screening.

Input: Two character vertices v_x and v_y
Output: structure error value ε
(01) **Initialize** list $L_3 \leftarrow v_x$ and v_y in G_ξ
(02) **for** $m \leftarrow 1$ **to** $|L_3|$ **do**
(03) $L_4 \leftarrow$ relations of person node $L_3[m]$
(04) **for** $n \leftarrow 1$ **to** $|L_4|$ **do**
(05) $preRelations \leftarrow preRelations + 1$
(06) **end for**
(07) **end for**
(08) **Initialize** list $L_5 \leftarrow$ relations in G_ξ which shared by character vertices v_x and v_y
(09) **for** $m \leftarrow 1$ **to** $|L_5|$ **do**
(10) $postRelations \leftarrow postRelations + 1$
(11) $\varepsilon \leftarrow (|L_3|/(|L_3| - 1)) * preRelations - (preRelations - postRelations)$
(12) **end for**
(13) **return** ε

Algorithm 2: $\varepsilon_\xi(v_x, v_y)$.

character vertices and entity vertices in different subnetworks. As differences of temporal attributes cause differences of relation path, we introduce activity path to describe these network structure. Based on this notion, we quantitatively measure similarity of character vertices by calculating temporal weight of activity paths. After combining all results in each subnetwork, character uniqueness can be measured precisely.

5.1. Transaction Activity Network (TAN). Let $\mathscr{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ and $\mathscr{B} = \{\beta_1, \beta_2, \ldots, \beta_n\}$ be label sets of vertices types and relation types, $\alpha \in \mathscr{A} \cap \beta \in \mathscr{B}$. Nonempty definite sets V_α and $\widehat{V}_\alpha$ denote character vertices set and entity vertices set, respectively. Nonempty relation set is denoted by R_β. Let T_β be temporal attributes set of activities in G. The mapping from relations to vertices and temporal attributes is denoted by $\Phi_\beta : R_\beta \to \{\langle v_i, \widehat{v}_j, \tau_k \rangle \mid v_i \in V_\alpha \cap \widehat{v}_j \in \widehat{V}_\alpha \cap \tau_k \in T_\beta\}$, and its inverse mapping is ${\Phi_\beta}^{-1}$. The mappings of vertices types and relation types are denoted by $\Omega_\alpha : V_\alpha \to \mathscr{A}$ and $\Theta_\beta : R_\beta \to \mathscr{B}$ severally.

Definition 5 (transaction activity network). A transaction activity network (TAN for short) contains activity information and temporal attributes. It is denoted by $G = \langle V_\alpha, \widehat{V}_\alpha, R_\beta, T_\beta, \Phi_\beta, \Omega_\beta, \Theta_\beta \rangle$, $\alpha \in \mathscr{A} \cap \beta \in \mathscr{B}$.

Property 1 (heterogeneity). In a TAN denoted by $G = \langle V_\alpha, \widehat{V}_\alpha, R_\beta, T_\beta, \Phi_\beta, \Omega_\beta, \Theta_\beta \rangle$, there is $|\mathscr{A}| \geq 2 \cap |\mathscr{B}| \geq 1$.

Property 2 (temporality). In a TAN denoted by $G = \langle V_\alpha, \widehat{V}_\alpha, R_\beta, T_\beta, \Phi_\beta, \Omega_\beta, \Theta_\beta \rangle$, there is $\tau_k = \langle \tau_k^S, \tau_k^E \rangle \cap \tau_k \in T_\beta$, τ_k denotes time attribute of r_k, and τ_k^S and τ_k^E denote severally start time and end time of r_k.

A large-scale TAN always contains several types of social activities. We can divide them into two or more subnetworks. Each of them contains one type of transaction activities. Let $\mathscr{G}_1, \mathscr{G}_2, \ldots, \mathscr{G}_{|\mathscr{B}|}$ be subnetworks with different types of activities. If its relations have temporal attributes and the set is $T_G = \{T_\beta \mid \beta \in \mathscr{B}\}$, we denote G as $G = \{\mathscr{G}_\beta \mid \mathscr{G}_\beta = \langle V_\alpha, \widehat{V}_\alpha,$

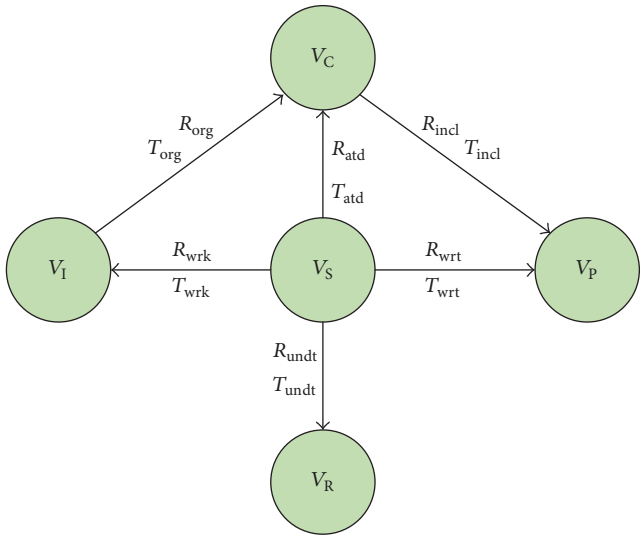

Figure 3: A heterogeneous TAN.

$R_\beta, T_\beta, \Phi_\beta, \Omega_\beta, \Theta_\beta\rangle \cap \alpha \in \mathscr{A} \cap \beta \in \mathscr{B}\}$, and the sets of vertices, relations, and types are denoted, respectively, by $V_G = \{V_\alpha \mid \alpha \in \mathscr{A}\} \cup \{\widehat{V}_\alpha \mid \alpha \in \mathscr{A}\}$, $R_G = \{R_\beta \mid \beta \in \mathscr{B}\}$, and $\Phi_G = \{\Phi_\beta \mid \beta \in \mathscr{B}\}$. It indicates that G consists of subnetworks $\mathscr{G}_1, \mathscr{G}_2, \ldots, \mathscr{G}_{|\mathscr{B}|}$; thus we denote it as $G = \{\mathscr{G}_\beta \mid \beta \in \mathscr{B}\}$ simply. Thus, it can be seen that a large-scale TAN contains multitype vertices and relations, and differences of vertex types lead to differences of relation types [32]. In real world, a TAN always contains several types of social activity; namely, there are different types of relations and vertices in a network.

Figure 3 shows a heterogeneous academic TAN. It contains two types of vertices: scholar vertices V_S and entity vertices V_I, V_C, V_P, V_R which represent *institution*, *conference*, *publication,* and *research project*. Due to differences of academic activities, there are different relations between vertices, such as *write relation* between scholars and papers and *participation relation* between scholars and conferences. We use R_{org}, R_{atd}, R_{incl}, R_{wrk}, R_{wrt}, and R_{undt} to denote six types of relations (organize, attend, included, work at, write, and undertake) and T_{org}, T_{atd}, T_{incl}, T_{wrt}, T_{wrt}, and T_{undt} denote temporal attribute set.

5.2. Transaction Activity Path (TAP). In a TAN, transaction activity paths (TAPs for short) are relative to topology of it. We regard character vertices and entity vertices, respectively, as master vertices and their neighbor vertices, and then we can describe TAPs. A TAP is a path which goes through a pair of character vertices and one entity vertices and the relations between them. From one master vertex to another, there is one or more TAPs through their common neighbors, and they contain semantics and temporal attributes of original transaction records.

Let character and neighbor vertices be $v_x, v_y \in V_\alpha$ and $\widehat{v}_z \in \widehat{V}_\alpha$, respectively, $X_1 = \{1, 2, 3, \ldots, \chi_1, \ldots, |X_1|\}$ and $X_2 = \{1, 2, 3, \ldots, \chi_2, \ldots, |X_2|\}$ are two label sets of relations in relation sets R_{xz} and R_{yz}, and

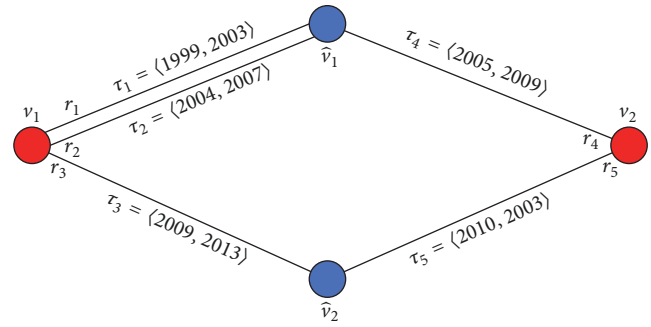

Figure 4: An instance of TAPs.

$$\begin{aligned} R_{xz} &= \left\{r_{\chi_1}\chi_1 \in X_1\right\} \\ &= \Phi_\beta^{-1}\left\{\left\langle v_x, \widehat{v}_z, \tau_{\chi_1}\right\rangle v_x \in V_\alpha \cap \widehat{v}_z \in \widehat{V}_\alpha \cap \tau_{\chi_1} \in T_\beta\right\}, \\ R_{yz} &= \left\{r_{\chi_2}\chi_2 \in X_2\right\} \\ &= \Phi_\beta^{-1}\left\{\left\langle v_y, \widehat{v}_z, \tau_{\chi_2}\right\rangle v_y \in V_\alpha \cap \widehat{v}_z \in \widehat{V}_\alpha \cap \tau_{\chi_2} \in T_\beta\right\}. \end{aligned} \tag{2}$$

Definition 3 (transaction activity path). In a TAN $\mathscr{G}_\beta = \langle V_\alpha, \widehat{V}_\alpha, R_\beta, T_\beta, \Phi_\beta, \Omega_\beta, \Theta_\beta\rangle$, let v_x or v_y be start vertex; a path which begin at v_x, and go through neighbor vertex $\widehat{v}_z$ and then end at v_y is called transaction activity path. It is denoted by $p(v_x\widehat{v}_z v_y)_{\chi_1\chi_2}$. The set of TAPs between v_x and v_y is denoted by $\mathscr{P}_{xy} = \{p(v_x\widehat{v}_z v_y)_{\chi_1\chi_2} \mid z \in Z \cap \chi_1 \in X_1 \cap \chi_2 \in X_2\}$.

Property 1. Let $|\mathscr{P}_{xy}|$ be the number of TAPs in set $\mathscr{P}_{xy}$, $|\mathscr{P}_{xy}| = |X_1| * |X_2|$.

Instance 1. Figure 4 shows that, in a TAN $\mathscr{G}_\beta = \langle V_{\text{person}}, \widehat{V}_{\text{club}}, R_\beta, T_\beta, \Phi_\beta, \Omega_\beta, \Theta_\beta\rangle$, the sets of character vertices and their neighbor vertices are $V_\beta = \{v_1, v_2\}$ and $\widehat{V}_\beta = \{\widehat{v}_1, \widehat{v}_2\}$; $R_\beta = \{r_1, r_2, r_3, r_4, r_5\}$ and $T_\beta = \{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}$ are the sets of relations and temporal attributes. We can find that $R_{11} = \{r_1, r_2\}$, $R_{12} = \{r_3\}$, $R_{21} = \{r_4\}$, $R_{22} = \{r_5\}$. Evidently though, there are two activity paths from vertex v_1 to v_2 through neighbor vertex $\widehat{v}_1$, and we denote them by $p(v_1\widehat{v}_1 v_2)_{14}$ and $p(v_1\widehat{v}_1 v_2)_{24}$. Similarly, we denote the path through neighbor $\widehat{v}_2$ by $p(v_1\widehat{v}_2 v_2)_{35}$.

5.3. Character Uniqueness Measure. Owing to temporal attributes of relations, we can define and calculate temporal weight of relations and TAPs, which reflect temporal characteristics of transaction activity networks. Based on temporal weight we can calculate TAP similarity to measure similarity degree of character vertices pairs. The similarity threshold is a filter to screen out unique vertices so that we can get redundant vertices set.

5.3.1. Temporal Weight Calculation. In a transaction activity network, temporal weights of relations are decided by start time and end time, while temporal weights of TAPs are decided by the former. Based on time attribute $\tau_k = \langle \tau_k^S, \tau_k^E\rangle$,

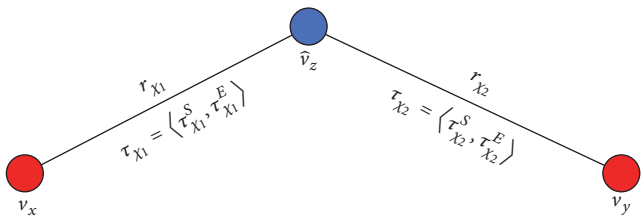

Figure 5: The first type of TAP.

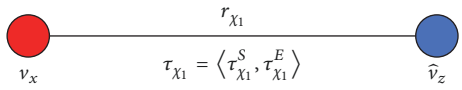

Figure 6: The second type of TAP.

we can use the following equation to calculate temporal weight of r_k:

$$W_k^z = \left(\text{Now} + 1 - \tau_k^S\right) * \left(\tau_k^E + 1 - \tau_k^S\right). \tag{3}$$

Now denotes current data, k denotes label of relations, and z is label of neighbor vertex $\widehat{v}_z$. The following equation is the temporal weight of TAPs:

$$W\left(p\left(v_x \widehat{v}_z v_y\right)_{\chi_1\chi_2}\right) = W_{\chi_1}^z * W_{\chi_2}^z. \tag{4}$$

The temporal weight of relations reflects the start time and end time, as well as the duration of relations. Apparently, the temporal weight of TAPs contains all of this information since TAPs consisted of two relations. The weight is decided by the temporal attributes of relations.

5.3.2. Transaction Activity Path Similarity. In a transaction activity network $\mathcal{G}_\beta$, let character vertices and entity vertex be $v_x, v_y \in V_\alpha$ and $\widehat{v}_z \in \widehat{V}_\alpha$ which is the neighbor of v_x and v_y. The TAP sets are denoted by $\mathcal{P}_{xy} = \{p(v_x \widehat{v}_z v_y)_{\chi_1\chi_2} \mid z \in Z \cap \chi_1 \in \mathrm{X}_1 \cap \chi_2 \in \mathrm{X}_2\}$, $\mathcal{P}_{xx} = \{p(v_x \widehat{v}_z v_x)_{\chi_1\chi_1} \mid z \in Z \cap \chi_1 \in \mathrm{X}_1 \cap \chi_1 \in \mathrm{X}_1\}$, and $\mathcal{P}_{yy} = \{p(v_y \widehat{v}_z v_y)_{\chi_2\chi_2} \mid z \in Z \cap \chi_2 \in \mathrm{X}_2 \cap \chi_2 \in \mathrm{X}_2\}$. $\mathcal{P}_{xy}$, $\mathcal{P}_{xx}$, and $\mathcal{P}_{yy}$ represent three types of paths, respectively. They have three different structures.

Figure 5 shows the first type of TAP between v_x and v_y. These paths begin from v_x then go through relations r_{χ_1}, neighbor $\widehat{v}_z$, and relation r_{χ_2} and end to v_y. In a network, all of the TAPs between different two vertices are this type. The second and third types are showed in Figures 6 and 7. Both of them begin from one vertex (v_x or v_y) and end to the same vertex, and they are through the same relation twice.

Definition 2 (SimTAP). SimTAP is the similarity between two vertices v_x and v_y. It is decided by structure and temporal weight of TAPs between v_x and v_y. The definition formula of SimTAP is as follows:

$$\text{SimTAP}_\beta\left(v_x, v_y\right) = \frac{2 * W\left(\mathcal{P}_{xy}\right)}{W\left(\mathcal{P}_{xx}\right) + W\left(\mathcal{P}_{yy}\right)}. \tag{5}$$

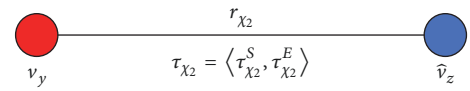

Figure 7: The third type of TAP.

In this formula, $W(\mathcal{P}_{xy})$, $W(\mathcal{P}_{xx})$, and $W(\mathcal{P}_{yy})$ denote the temporal weight sums of these three types of TAPs. We use the following formulas to calculate these weights:

$$\begin{aligned} W\left(\mathcal{P}_{xy}\right) &= \sum_{z=1}^{|Z|} \sum_{\chi_1=1}^{|\mathrm{X}_1|} \sum_{\chi_2=1}^{|\mathrm{X}_2|} W_{\chi_1}^z * W_{\chi_2}^z, \\ W\left(\mathcal{P}_{xx}\right) &= \sum_{z=1}^{|Z|} \sum_{\chi_1=1}^{|\mathrm{X}_1|} \sum_{\chi_1=1}^{|\mathrm{X}_1|} W_{\chi_1}^z * W_{\chi_1}^z, \\ W\left(\mathcal{P}_{yy}\right) &= \sum_{z=1}^{|Z|} \sum_{\chi_2=1}^{|\mathrm{X}_2|} \sum_{\chi_2=1}^{|\mathrm{X}_2|} W_{\chi_2}^z * W_{\chi_2}^z. \end{aligned} \tag{6}$$

$W_{\chi_1}^z$ and $W_{\chi_1}^z$ are weights of relations between v_x and v_y.

Generally, a transaction activity network contains several subnetworks. In order to measure the similarity of all characters, we need to add all similarity values in each subnetwork and then calculate arithmetic mean. Let a TAN be $G = \{\mathcal{G}_\beta \mid \beta \in \mathcal{B}\}$, we calculate similarity of vertices pair v_x and v_y in $\mathcal{G}_\beta$; then we get the TAPs similarity set $\{\text{SimTAP}_\beta(v_x, v_y) \mid \beta \in \mathcal{B}\}$. After that we calculate the arithmetic mean in G. The formula is as follows:

$$\text{SimTAP}\left(v_x, v_y\right) = \frac{1}{|\mathcal{B}|} \sum_{\beta=1}^{|\mathcal{B}|} \text{SimTAP}_\beta\left(v_x, v_y\right). \tag{7}$$

In the formula, $|\mathcal{B}|$ is the number of subnetworks in G and $\text{SimTAP}(v_x, v_y)$ is the TAP similarity of v_x and v_y.

5.3.3. Character Uniqueness Measure. $\text{SimTAP}(v_x, v_y)$ can measure uniqueness of characters quantitatively in TANs. The larger its value is, the greater the similarity between character vertices is, and vice versa. According to this idea, we proposed uniqueness measurement of characters: after $\text{SimTAP}(v_x, v_y)$ calculating, we set character uniqueness threshold θ based on features of networks and data-analytic requirements to screen out the results. If $\text{SimTAP}(v_x, v_y) < \theta$, we regard v_x and v_y as unique characters, while if $\text{SimTAP}(v_x, v_y) \geq \theta$, vertices v_x and v_y have high similarity, which indicates that we need to merge these vertices and their shared relations.

Instance 2. In a transaction activity network $G = \{\mathcal{G}_\mathrm{S}, \mathcal{G}_\mathrm{W}, \mathcal{G}_\mathrm{R}, \mathcal{G}_\mathrm{C}\}$, the type sets of vertices and relations are, respectively, denoted by $\mathcal{A}$ and $\mathcal{B} = \{\text{Study}, \text{Work}, \text{Research}, \text{Coauthor}\}$. There are 10 character vertices in this network and the values of similarity of them are shown in Table 1.

In this table, columns v_x and v_y are name of characters, $\mathcal{G}_\mathrm{S}$, $\mathcal{G}_\mathrm{W}$, $\mathcal{G}_\mathrm{P}$, and $\mathcal{G}_\mathrm{C}$ indicate the similarity of vertices pairs in these four subnetworks, and G denote the similarity in G. After setting the threshold $\theta = 0.70$, we can

Table 1: The results of transaction activity similarity.

v_x	v_y	$\mathcal{G}_S$	$\mathcal{G}_W$	$\mathcal{G}_P$	$\mathcal{G}_C$	G
Long Chen	Jay Liu	0.0000	0.4235	1.0000	1.0000	0.6059
Hao Feng	Wei Zhang	0.8661	1.0000	1.0000	0.0000	0.7165
Jing Xu	Jiang Zhou	1.0000	1.0000	1.0000	0.0000	0.7500
Faye Wu	Fei X. Wu	1.0000	1.0000	1.0000	1.0000	1.0000
Faye Wu	Ron Xiao	0.5219	0.0000	1.0000	1.0000	0.6305
Faye Wu	Pan Zhang	0.5219	0.0000	1.0000	1.0000	0.6305
Fei X. Wu	Ron Xiao	0.5219	0.0000	1.0000	1.0000	0.6305
Fei X. Wu	Pan Zhang	0.5219	0.0000	1.0000	1.0000	0.6305
Ron Xiao	Pan Zhang	1.0000	0.0000	1.0000	1.0000	0.7500

find that there are four similarity values larger than θ: SimTAP($v_{\text{Hao Feng}}, v_{\text{Wei Zhang}}$), SimTAP($v_{\text{Jing Xu}}, v_{\text{Jiang Zhou}}$), SimTAP($v_{\text{Faye Wu}}, v_{\text{Fei X. Wu}}$), and SimTAP($v_{\text{Ron Xiao}}, v_{\text{Pan Zhang}}$). It indicates that these characters are remarkably similar, and they do not have uniqueness. Besides, the similarity of character *Long Chen* and *Jay Liu* is smaller than θ; namely, SimTAP($v_{\text{Long Chen}}, v_{\text{Jay Liu}}$) $< \theta$; thus they have uniqueness. This shows that we can screen out the character vertices which have uniqueness by calculating similarity SimTAP(v_x, v_y) and setting threshold θ.

5.4. Algorithm Design. We designed TAPs similarity algorithm based on the above-mentioned theories, which is shown in Algorithm 5. At first, we get the relation lists of vertices pair v_x and v_y from each subnetwork $\mathcal{G}_\beta$ and then calculate the temporal weight of transaction activity paths. Second, we calculate the transaction activity network similarity SimTAP$_\beta$ of v_x and v_y and then calculate arithmetic mean of similarity SimTAP in network G, shown in Algorithm 4. After traversing all vertices pairs in candidate redundant vertices set H and getting their similarity, we set threshold θ and compare it with each similarity. We regard the vertices whose similarity is larger than θ as redundant vertices and put them into redundant vertices set $\widetilde{\text{H}}$, shown in Algorithm 3. The vertices whose similarity is smaller than θ are regarded as unique vertices and they must remain in network.

5.5. Experiment and Analysis. The multimedia dataset for academic transaction networks building contains texts, images, and videos concerning proposals, papers, award certificates, and videos of academic conference. In the experiment of this paper, we extract academic activity transaction data from 724 proposals of *Natural Science Foundation of China (NSFC)* [33], which are texts in Chinese only, and then established a transaction database. After that, we import these data into graph database *Neo4J* and then construct transaction activity networks which contained 598 vertices. We mine academic relationship between scholars and then build academic networks. On this basis, we calculate structure error of character vertices and then give the visual presentation of network [34]. Based on the results of structure error calculation, we get vertices from redundant vertices set H and calculated SimTAP of each vertices pair.

Table 2: Structure error of vertices in G.

v_x	v_y	$\varepsilon_G(v_x, v_y)$
Faye Wu	Fei Wu	0.000
ShaoJia Zhu	ShaoNan Zhu	0.000
Ruifeng Fan	Chi Zhang	0.250
Xiaojie Liu	Chao Zhang	0.500
Kang Du	Lin Guo	0.500
Lijuan Liu	Guanghua Zhao	0.625
Yafei Hou	Yue Wang	0.750
Shengmin Fu	Yanni Peng	0.750
Zeo Zhou	Ze Zhong	0.750
Bin Du	Binbin Gao	0.875

5.5.1. Evaluation for Structure Error. Our academic transaction activity networks contain 589 scholars' academic transaction information. The first step was extracting academic activity data from transaction database and then importing them into graph database. We construct four types of activity networks; they were education experience network, work experience network, project cooperation network, and coauthor network. After that, we build academic network G based on them. Figure 3 shows this network.

In this network, we calculate structure error of each vertices pair and screen out the vertices with 0 structure error. Table 2 shows partial results.

In Table 2, fields v_x and v_y denote two vertices and field denotes value of structure error of these vertices in G. We can find that the structure error of vertices pairs *Faye Wu* and *Fei Wu* and *Shaojia Zhu* and *Shaonan Zhu* equals zero. Therefore, these two vertices are regarded as redundant vertices candidates. We can find their structure features in Figure 8. Four highlighted character vertices are *Faye Wu*, *Fei Wu*, *Shaojia Zhu,* and *Shaonan Zhu*. These two highlighted subnetworks illustrate that the two vertex pairs have same neighbors, respectively.

In order to analyze our method deeply, we extract academic activity information from the database. Tables 3–6 show academic activity information of Faye Wu and Fei Wu.

We can find that *Faye Wu* and *Fei Wu* studied in the same school over the same period. Likewise, they have the same experience on the aspects of work, project, and publication. Namely, their experience of academy is selfsame. Thus, *Faye Wu* or *Fei Wu* is not unique, which is redundant information.

In Figure 8, vertices *Shaojia Zhu* and *Shaonan Zhu* own the same neighbors. Similarly, we extract their activity information.

From Tables 7–10 we can see that vertices *Shaojia Zhu* and *Shaonan Zhu* studied in the same universities and were employed by the same employer but the periods are different. That means their education and work experience is different. The difference between *Shaojia Zhu* and *Shaonan Zhu* is caused by the difference of temporal attributes. Therefore, both of them are unique and they do not contain redundant information.

The results indicate that character vertices which have the same neighbors may not contain the exact same social

Input: Candidate redundant vertices set H
Output: Redundant vertices set $\widetilde{\mathrm{H}}$
(01) **Initialize** list L_1, L_2 by candidate redundant vertices set H
(02) **for** $i \leftarrow 1$ **to** $|L_1|$ **do**
(03) **for** $j \leftarrow 1$ **to** $|L_2|$ **do**
(04) **if** the name of $L_1[i]$ = the name of $L_2[j]$ **then**
(05) **if** SimTAP$(L_1[i], L_2[j])$ >= θ **then**
(06) **if** $L_1[i] \in \widetilde{\mathrm{H}}_i$ or $L_2[j] \in \widetilde{\mathrm{H}}_i$ **then**
(07) Insert $L_1[i]$ or $L_2[j]$ into $\widetilde{\mathrm{H}}_i$
(08) **else**
(09) Insert $L_1[i]$, $L_2[j]$ into $\widetilde{\mathrm{H}}_{i+1}$
(10) **end if**
(11) **end if**
(12) **end if**
(13) **end for**
(14) **end for**
(15) **return** $\widetilde{\mathrm{H}}$

Algorithm 3: Redundant vertices screening in heterogeneous network G.

Input: Two character vertices v_x and v_y
Output: value of path similarity of v_x and v_y in $\mathcal{G}_\beta$
(01) **Initialize** list R_1 by relation list of v_x in $\mathcal{G}_\beta$
(02) **Initialize** list R_2 by relation list of v_y in $\mathcal{G}_\beta$
(03) **for** $k_1 \leftarrow 1$ **to** $|R_1|$ **do**
(04) **for** $k_2 \leftarrow 1$ **to** $|R_2|$ **do**
(05) **if** $\widehat{v}_z$ of $R_1[k_1] = \widehat{v}_z$ of $R_2[k_2]$ **then**
(06) $W(p(v_x\widehat{v}_z v_y)_{\chi_1\chi_2}) \leftarrow W^z_{\chi_1} * W^z_{\chi_2}$
(07) **end if**
(08) **end for**
(09) **end for**
(10) **for** $k_1 \leftarrow 1$ **to** $|R_1|$ **do**
(11) **for** $k_2 \leftarrow 1$ **to** $|R_1|$ **do**
(12) **if** $\widehat{v}_z$ of $R_1[k_1] = \widehat{v}_z$ of $R_1[k_2]$ **then**
(13) $W(p(v_x\widehat{v}_z v_x)_{\chi_1\chi_1}) \leftarrow W^z_{\chi_1} * W^z_{\chi_1}$
(14) **end if**
(15) **end for**
(16) **end for**
(17) **for** $k_1 \leftarrow 1$ **to** $|R_2|$ **do**
(18) **for** $k_2 \leftarrow 1$ **to** $|R_2|$ **do**
(19) **if** $\widehat{v}_z$ of $R_2[k_1] = \widehat{v}_z$ of $R_2[k_2]$ **then**
(20) $W(p(v_y\widehat{v}_z v_y)_{\chi_2\chi_2}) \leftarrow W^z_{\chi_2} * W^z_{\chi_2}$
(21) **end if**
(22) **end for**
(23) **end for**
(24) SimTAP$_\beta \leftarrow$
$(2 * W(p(v_x\widehat{v}_z v_y)_{\chi_1\chi_2}))/(W(p(v_x\widehat{v}_z v_x)_{\chi_1\chi_1}) + W(p(v_y\widehat{v}_z v_y)_{\chi_2\chi_2}))$
(25) **return** SimTAP$_\beta$

Algorithm 4: SimTAP$_\beta(v_x, v_y)$.

Input: Two character vertices v_x and v_y
Output: value of path similarity of v_x and v_y in G
(01) Initialize SimTAP $\leftarrow$ 0
(02) **for** $\beta \leftarrow 1$ **to** $|\mathscr{B}|$ **do**
(03) SimTAP $\leftarrow$ SimTAP + SimTAP$_\beta(v_x, v_y)$
(04) **end for**
(05) SimTAP $\leftarrow$ SimTAP/$|\mathscr{B}|$
(06) **return** SimTAP

Algorithm 5: SimTAP(v_x, v_y).

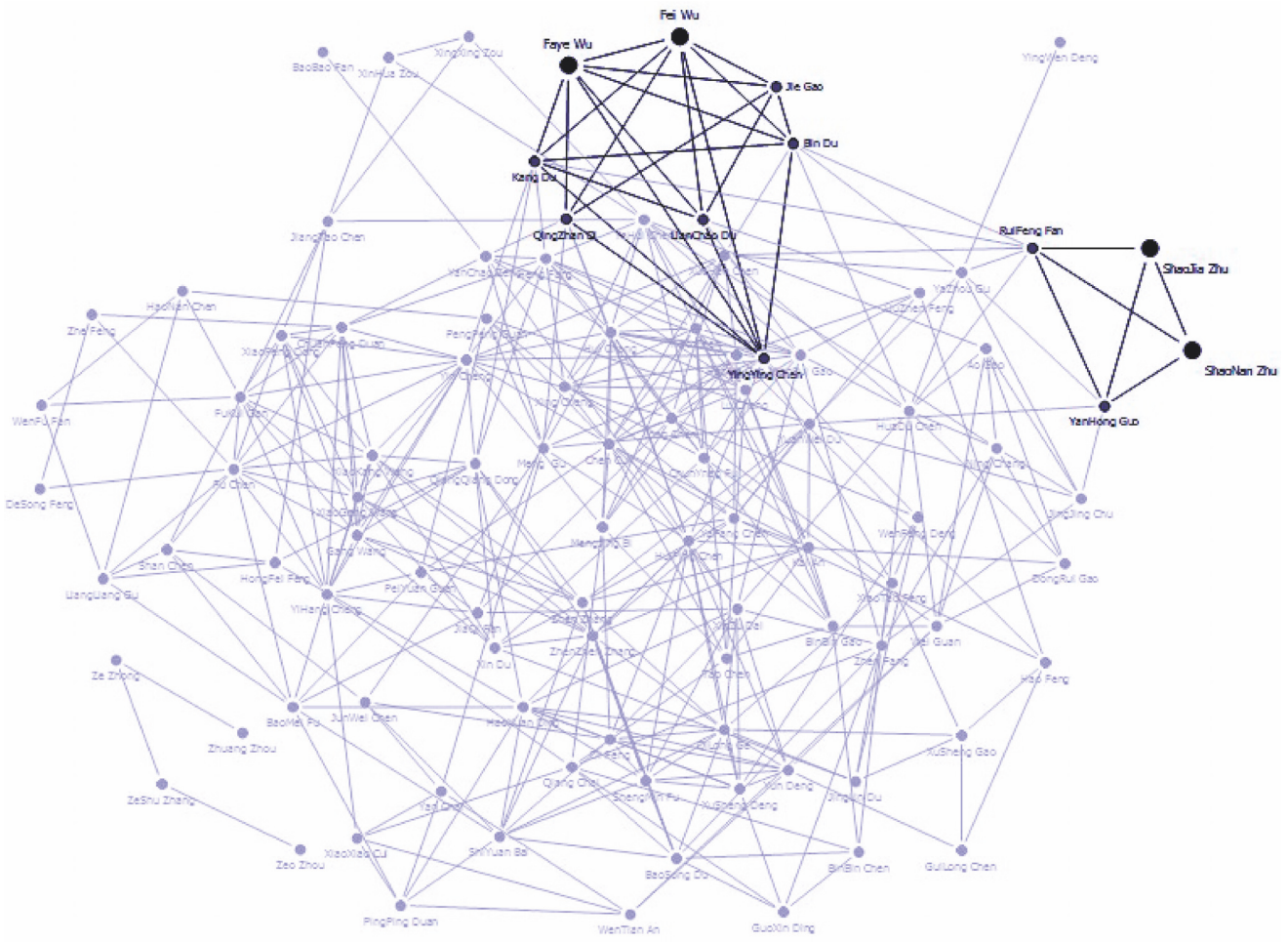

Figure 8: Visualization of academic network G.

Table 3: Education experience info of Faye Wu and Fei Wu.

Name	Institute	Start time	End time
Faye Wu	Hubei Minzu Univ.	1992	1996
Faye Wu	Guangzhou Normal Univ.	1997	1999
Faye Wu	Jinan Univ.	2000	2005
Fei Wu	Hubei Minzu Univ.	1992	1996
Fei Wu	Guangzhou Normal Univ.	1997	1999
Fei Wu	Jinan Univ.	2000	2000

Table 4: Work experience info of Faye Wu and Fei Wu.

Name	Employer	Start time	End time
Faye Wu	Central South Univ.	2010	2012
Fei Wu	Central South Univ.	2010	2012

Table 5: Project information of Faye Wu and Fei Wu.

Name	Project	Start time	End time
Faye Wu	Control Policy Research in Parabolic Distributed Parameter Systems	2010	2012
Fei Wu	Control Policy Research in Parabolic Distributed Parameter Systems	2010	2012

Table 6: Publication info of Faye Wu and Fei Wu.

Name	Publication	Start time	End time
Faye Wu	Adaptive Control Synchronization Approach Research of Unified Chaotic Systems	2000	2000
Faye Wu	Linear and Nonlinear Feedback Synchronization and Performance Research in Discrete Chaotic System	2004	2004
Fei Wu	Adaptive Control Synchronization Approach Research of Unified Chaotic Systems	2000	2000
Fei Wu	Linear and Nonlinear Feedback Synchronization and Performance Research in Discrete Chaotic System	2004	2004

Table 7: Education experience info of Shaojia Zhu and Shaonan Zhu.

Name	Institute	Start time	End time
ShaoJia Zhu	Changsh Univ. of Sci and Tech	1990	1994
ShaoJia Zhu	Central South Univ.	1996	1999
ShaoJia Zhu	Zejiang Univ.	2000	2004
ShaoNan Zhu	Changsh Univ. of Sci and Tech	2000	2004
ShaoNan Zhu	Central South Univ.	2004	2007
ShaoNan Zhu	Zejiang Univ.	2008	2012

Table 8: Work experience info of Shaojia Zhu and Shaonan Zhu.

Name	Employer	Start time	End time
ShaoJia Zhu	Hunan Univ. of Sci and Tech	2004	2014
ShaoNan Zhu	Hunan Univ. of Sci and Tech	2012	2014

Table 9: Project information of Shaojia Zhu and Shaonan Zhu.

Name	Project	Start time	End time
ShaoJia Zhu	Decimal Encryption Technology Research Based on AES	2012	2013
ShaoNan Zhu	Decimal Encryption Technology Research Based on AES	2012	2013

activity information. These vertices are redundant candidates and among them there are some vertices with uniqueness. But we cannot recognize them by structure error. On the contrary, we can only screen out vertices whose structure error is not zero. They exactly have uniqueness. Above all, we need to recognize character uniqueness ulteriorly.

5.5.2. TAPs Similarity Calculation. We first calculate the similarity of vertices pair in an academic network containing 589 characters. After setting θ as 0.70, we screen out the vertices whose SimTAP are higher than θ. The results are shown in Table 11.

In this table, we find that the value of similarity of vertices pair *Faye Wu* and *Fei Wu* is 1.0000, which indicates that their academic activity information is identical. That means the similarity between them has been maximized. The similarities of *Jia Gao* and *Di Feng*, *Xinhua Zou* and *Xingxing Zou*, and *Zhe Feng* and *Kang Du* are 0.7450, 0.7463, 0.8546, and 0.7500.

5.5.3. Regression Analysis. We chose the vertices whose similarity in subnetworks is zero and extracted their transaction information from database. It is shown in Tables 12, 13, and 14.

We can see from Table 12 that *Jia Gao* and *Di Feng* studied in three different universities. Likewise, in Table 13, *Xinghua Zou* and *Xingxing Zou* studied in different colleges as well. In Table 14, the publications of *Zhe Feng* and *Kang Du* are entirely different. These situations indicate that these three pairs of character are different in *education* and *publication*

Table 10: Publication info of Shaojia Zhu and Shaonan Zhu.

Name	Publication	Start time	End time
ShaoJia Zhu	Node Behavior Prediction and Optimized Routing Algorithm in Mobile Networks	2013	2013
ShaoNan Zhu	Node Behavior Prediction and Optimized Routing Algorithm in Mobile Networks	2013	2013

Table 11: The results of transaction activity similarity.

v_x	v_y	$\mathcal{G}_S$	$\mathcal{G}_W$	$\mathcal{G}_P$	$\mathcal{G}_C$	G
Faye Wu	Fei Wu	0.0000	0.4235	1.0000	1.0000	0.6059
ShaoJia Zhu	ShaoNan Zhu	0.8661	1.0000	1.0000	0.0000	0.7165
Jie Gao	Di Feng	1.0000	1.0000	1.0000	0.0000	0.7500
XinHua Zou	XingXing Zou	1.0000	1.0000	1.0000	1.0000	1.0000
Zhe Feng	Kang Du	0.5219	0.0000	1.0000	1.0000	0.6305

Table 12: Education info of Jia Gao and Di Feng.

Name	Institute	Start time	End time	Degree
Jie Gao	Qingdao University	2005	2008	Ph.D.
Jie Gao	Heilongjiang University	2002	2005	M.Sc.
Jie Gao	Zhengzhou University	1998	2002	B.Sc.
Di Feng	Peking University	1999	2004	Ph.D.
Di Feng	Nanjing Agriculture University	1996	1999	M.Sc.
Di Feng	Guangxi University	1992	1996	B.Sc.

Table 13: Education info of Xinhua Zou and Xingxing Zou.

Name	Institute	Start time	End time	Degree
XinHua Zou	Central South University	2002	2006	Ph.D.
XinHua Zou	Hunan University of Medicine	1999	2001	M.Sc.
XinHua Zou	Chang'an University	1995	1999	B.Sc.
XingXing Zou	Peking University	2005	2008	Ph.D.
XingXing Zou	Hunan University	2002	2005	M.Sc.
XingXing Zou	Ocean University of China	1998	2002	B.Sc.

Table 14: Publication info of Zhe Feng and Kang Du.

Name	Institute	Start time	End time
Zhe Feng	Lyapunov Exponent Algorithm Design and Implementation	1992	1996
Kang Du	Cell Image Separation Algorithm Based on Contour-stripped	1997	1999

activities. It leads to differences of academic relationship between them. However, high similarity of other types of academic activities leads to high value of SimTAP of these characters. It is even higher than threshold θ so that these characters cannot be screened out from networks. This situation adverse impact character uniqueness identification. This problem can be solved by calculating TAPs similarity and screening out redundant character vertices from social networks.

Based on experiment results in Section 4, we calculated similarity of candidate redundant vertices in H. The results of structure error calculation are shown in Table 15.

We got the redundant vertices set $H = \{v_{\text{Faye Wu}}, v_{\text{Fei Wu}}, v_{\text{Shaojia Zhu}}, v_{\text{Shaonan Zhu}}\}$ after structure error calculation and then calculated the similarity of these four vertices. θ was set as 0.80; the results are shown in Table 16.

5.5.4. Redundant Vertices Merging. After vertices screening we merged redundant vertices $v_{\text{Faye Wu}}$ and $v_{\text{Fei Wu}}$ and their relations; then we got academic network G' without redundant information. In Figure 9, we can find that $v_{\text{Fei Wu}}$ and its relations were removed by vertices merging, but $v_{\text{Faye Wu}}$ is saved. Likewise, we can save $v_{\text{Faye Wu}}$ and remove $v_{\text{Fei Wu}}$ in the process. Compared to the network before merging, the

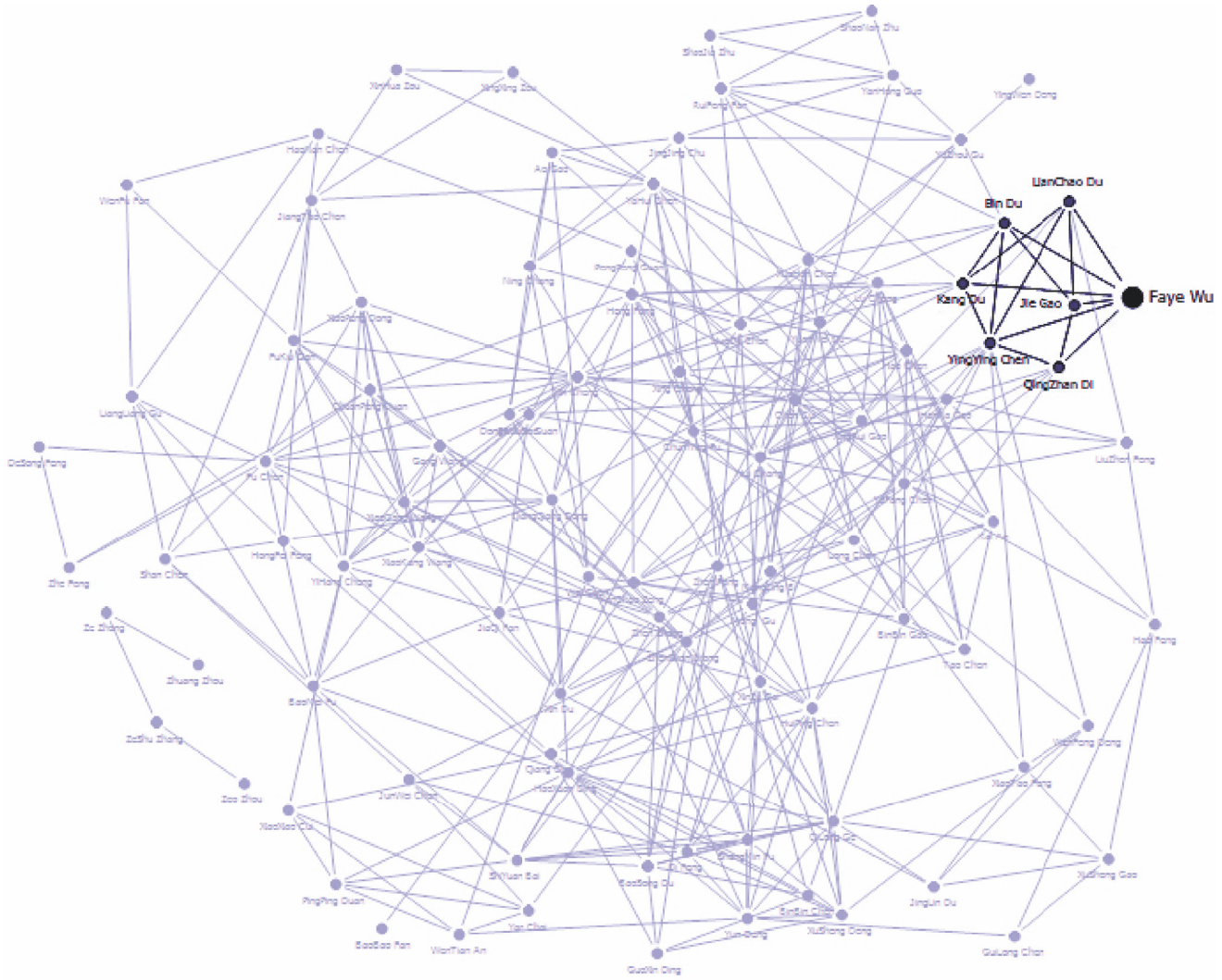

Figure 9: Academic network G' after vertices merging.

Table 15: The results of structure error calculation.

v_x	v_y	$\varepsilon_G(v_x, v_y)$
Faye Wu	Fei Wu	0.0000
Shaojia Zhu	Shaonan Zhu	0.0000

Table 16: The vertices are screened out by θ.

v_x	v_y	$\mathcal{G}_S$	$\mathcal{G}_W$	$\mathcal{G}_P$	$\mathcal{G}_C$	G
Faye Wu	Fei Wu	1.0000	1.0000	1.0000	1.0000	1.0000

relations between $v_{\text{Faye Wu}}$ and neighbors were not changed. It indicated that the vertices merging was correct.

The analyzing above indicates that we can promote accuracy of vertices uniqueness identifying based on structure error calculation and transaction activity paths similarity. Similarity threshold setting implements vertices screening, which is the basis of redundant vertices merging. Therefore, our solution realized character correction in social networks from multimedia datasets.

6. Conclusion and Future Work

In this paper, we introduce the framework of social network modeling via multimedia data. Then, we present the notion of structure error according to structure features of networks and vertices merging principles and then calculated structure error and screened out redundant vertices by using transaction information to build social networks. Besides, we designed algorithm of vertices similarity which can precisely measure character vertices uniqueness and created redundant vertices set. Finally, we removed redundant information in a network by merging redundant vertices in the set. Our solution improved the accuracy of character uniqueness recognition and solved character correction effectively in a network. At present, we set threshold empirical during experiment, but we do not implement intellectualized adaptive adjustment of threshold yet. In future work, we will compute the range

of threshold based upon large amount of network data and statistical techniques and then design adaptive adjustment algorithm.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61272150, 61379110, 61472450, 61402165, 61702560, S1651002, and M1450004), the Key Research Program of Hunan Province (2016JC2018), and Science and Technology Plan of Hunan Province Project (2018JJ2099 and 2018JJ3691).

References

[1] J. Sang, C. Xu, and R. Jain, "Social Multimedia Ming: From Special to General," in *2016 IEEE International Symposium on Multimedia (ISM)*, pp. 481–485, IEEE, San Jose, California, USA, 2016.

[2] J. Sang and C. Xu, "On Analyzing the 'Variety' of Big Social Multimedia," in *Proceedings of the 1st IEEE International Conference on Multimedia Big Data, BigMM 2015*, pp. 5–8, IEEE, China, April 2015.

[3] F. Amato, V. Moscato, A. Picariello, and F. Piccialli, "SOS: A multimedia recommender System for Online Social networks," *Future Generation Computer Systems*, 2017.

[4] Y. Wang, X. Lin, L. Wu, and W. Zhang, "Effective multi-query expansions: Robust landmark retrieval," in *Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015*, pp. 79–88, ACM, Australia, October 2015.

[5] L. Wu, X. Huang, C. Zhang, J. Shepherd, and Y. Wang, "An efficient framework of Bregman divergence optimization for co-ranking images and tags in a heterogeneous network," *Multimedia Tools and Applications*, vol. 74, no. 15, pp. 5635–5660, 2015.

[6] W.-Z. Nie, W.-J. Peng, X.-Y. Wang, Y.-L. Zhao, and Y.-T. Su, "Multimedia venue semantic modeling based on multimodal data," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 375–385, 2017.

[7] S. Pan, J. Wu, X. Zhu, G. Long, and C. Zhang, "Task Sensitive Feature Exploration and Learning for Multitask Graph Classification," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 744–758, 2017.

[8] S. Pan, J. Wu, X. Zhu, C. Zhang, and P. S. Yu, "Joint Structure Feature Exploration and Regularization for Multi-Task Graph Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 715–728, 2016.

[9] Y. Wang and L. Wu, "Multi-view spectral clustering via structured low-rank matrix factorization," https://arxiv.org/abs/1709.01212.

[10] Y. Wang, W. Zhang, L. Wu et al., "Multi-view spectral clustering via structured low-rank matrix factorization," https://arxiv.org/abs/1709.01212.

[11] Y. Wang, W. Zhang, L. Wu, X. Lin, and X. Zhao, "Unsupervised Metric Fusion over Multiview Data by Graph Random Walk-Based Cross-View Diffusion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 1, pp. 57–70, 2017.

[12] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3939–3949, 2015.

[13] S. Sagiroglu and D. Sinanc, "Big data: a review," in *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13)*, pp. 42–47, IEEE, San Diego, Calif, USA, May 2013.

[14] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in *Proceedings of the 6th International Conference on Contemporary Computing (IC3 '13)*, pp. 404–409, IEEE, Noida, India, August 2013.

[15] C. Ye, Z. Xiong, Y. Ding et al., "Joint fingerprinting and encryption in hybrid domains for multimedia sharing in social networks," *Journal of Visual Languages & Computing*, vol. 25, no. 6, pp. 658–666, 2014.

[16] L. Zhuhadar, R. Yang, and M. D. Lytras, "The impact of Social Multimedia Systems on cyberlearners," *Computers in Human Behavior*, vol. 29, no. 2, pp. 378–385, 2013.

[17] S. Zhao, Y. Gao, G. Ding, and T.-S. Chua, "Real-Time Multimedia Social Event Detection in Microblog," *IEEE Transactions on Cybernetics*, 2017.

[18] F. Laforest, N. Le Sommer, S. Frénot et al., "C3PO: A Spontaneous and Ephemeral Social Networking Framework for a Collaborative Creation and Publishing of Multimedia Contents," in *Proceedings of the International Conference on Selected Topics in Mobile and Wireless Networking, MoWNet 2014*, pp. 129–134, Italy, September 2014.

[19] F. Huang, X. Li, S. Zhang, J. Zhang, J. Chen, and Z. Zhai, "Overlapping community detection for multimedia social networks," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1881–1893, 2017.

[20] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proceedings of the IEEE Symposium on Security and Privacy (SP '08)*, pp. 111–125, IEEE, Oakland, Calif, USA, May 2008.

[21] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008*, pp. 990–998, Las Vegas, Nevada, USA, August 2008.

[22] R. Bekkerman and A. McCallum, "Disambiguating Web Appearances of People in a Social Network," in *WWW '05 Proceedings of the 14th international conference on World Wide Web*, pp. 463–470, ACM, Chiba, Japan, 2005.

[23] K. Liu and E. Terzi, "Towards identity anonymization on graphs," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data 2008, SIGMOD'08*, pp. 93–106, Canada, June 2008.

[24] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pp. 173–187, IEEE, Berkeley, Calif, USA, May 2009.

[25] X. Ding, L. Zhang, Z. Wan, and M. Gu, "De-anonymizing dynamic social networks," in *Proceedings of the 54th Annual IEEE Global Telecommunications Conference: "Energizing Global Communications", GLOBECOM 2011*, pp. 1–6, IEEE, Houston, Texas, USA, December 2011.

[26] M. Korayem and D. J. Crandall, "Crandall. De-anonymizing users across heterogeneous social computing platforms," in *The 7th international aaai conference on weblogs and social media (ICWSM 2013)*, pp. 899–908, Association for the Advancement of Artificial, Cambridge, Massachusetts, USA, 2013.

[27] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social networks as a side-channel," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS 2012*, pp. 628–637, ACM, Raleigh, NC, USA, October 2012.

[28] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[29] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," in *Proceedings of the VLDB'11*, ACM, Seattle, WA, USA, 2011.

[30] R. Hu, C. P. Yu, S.-F. Fung, S. Pan, H. Wang, and G. Long, "Universal network representation for heterogeneous information networks," in *Proceedings of the 2017 International Joint Conference on Neural Networks, IJCNN 2017*, pp. 388–395, IEEE, Anchorage, Alaska, USA, May 2017.

[31] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang, "Exploiting correlation consensus: Towards subspace clustering for multi-modal data," in *Proceedings of the 2014 ACM Conference on Multimedia*, pp. 981–984, ACM, Orlando, FL, USA, November 2014.

[32] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 73, no. 2, Article ID 026120, 2006.

[33] F. Huang, T. Shanmei, and X. Charles, "Extracting Academic Activity Transaction in Chinese Documents," in *Proceedings of the 8th International Conference on Intelligent Systems and Knowledge Engineering*, vol. 278, pp. 125–135, Springer, Shenzhen, China, November 2013.

[34] F. Huang, W. Xiao, and H. Zhang, "Visualization of clustered network graphs based on constrained optimization partition layout," *Lecture Notes in Electrical Engineering*, vol. 277, pp. 1381–1394, 2014.