TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 14/18 pp386–395
DOI: 10.26599/TST.2020.9010049
Volume 27, Number 2, April 2022

Nonnegative Matrix Tri-Factorization Based Clustering in a Heterogeneous Information Network with Star Network Schema

Juncheng Hu, Yongheng Xing, Mo Han, Feng Wang*, Kuo Zhao, and Xilong Che

Abstract: Heterogeneous Information Networks (HINs) contain multiple types of nodes and edges; therefore, they can preserve the semantic information and structure information. Cluster analysis using an HIN has obvious advantages over a transformation into a homogenous information network, which can promote the clustering results of different types of nodes. In our study, we applied a Nonnegative Matrix Tri-Factorization (NMTF) in a cluster analysis of multiple metapaths in HIN. Unlike the parameter estimation method of the probability distribution in previous studies, NMTF can obtain several dependent latent variables simultaneously, and each latent variable in NMTF is associated with the cluster of the corresponding node in the HIN. The method is suited to co-clustering leveraging multiple metapaths in HIN, because NMTF is employed for multiple nonnegative matrix factorizations simultaneously in our study. Experimental results on the real dataset show that the validity and correctness of our method, and the clustering result are better than that of the existing similar clustering algorithm.

Key words: heterogeneous information network; data mining; clustering; nonnegative matrix tri-factorization

1 Introduction

With the rapidly growing number of mobile phones and smart devices accessing the Internet[1], massive amounts of data are produced every moment. In the era of big data, intelligent and effective technologies are required to expedite data processing to deal with explosive volumes of data.

Effective data representation enables a mining algorithm to obtain more accurate results. A network is an important representation of the complex relationship between data objects, such as social networks, communication networks, and traffic systems[2]. The information network analysis has drawn extensive attention currently, a great number of researches focused on the study of data analysis enabling technologies, comprising network representation learning[3, 4], knowledge graphs[5], and Heterogeneous Information Networks (HINs)[6].

Nodes represent data entities, and the edges between nodes represent relationships between data entities in the information network. If the type of node and the type of edge are unique in an information network, then the network is called a homogeneous information network; examples of such a network are an author collaboration network, a paper citation network, and a friendship network. The topological structure of the paper citation network is shown in Fig. 1, which forms a star network schema. Otherwise, it is an HIN with different types of nodes and links, such as a medical information network or a bibliographic information network.

As illustrated in Fig. 2, the bibliographic information network is a typical HIN. Each paper is written by a group of authors, contains a variety of terms, and is published in a venue (a conference or a journal). The bibliographic HIN consists of four types of nodes: authors, venues, papers, and terms. The edges between

• Juncheng Hu, Yongheng Xing, Mo Han, Feng Wang, and Xilong Che are with the College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: {jchu19, xingyh18, hanmo15, wangfeng12}@mails.jlu.edu.cn; chexilong@jlu.edu.cn.

• Kuo Zhao is with the School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China. E-mail: zhaokuo@jnu.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2020-09-20; accepted: 2020-10-09

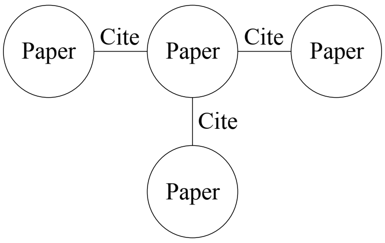


Fig. 1 Topological structure of the paper citation network.

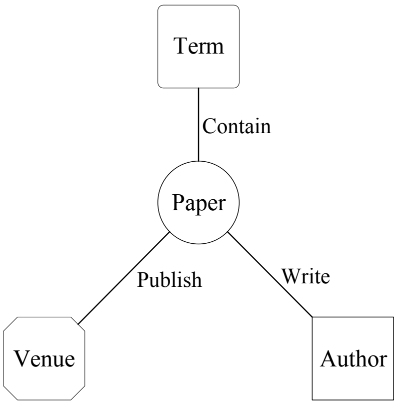


Fig. 2 Topological structure of bibliographic information network.

the authors and papers have "write" and "written" relationships; the edges between papers and conferences have "published" and "include" relationships; and the edges between papers and terms have "contain" and "contained" relationships.

The metapath[6] is an important concept in HIN, which expresses the characteristic information and semantic relationship between nodes. Figures 3 and 4 show two different metapaths in the bibliographic network. These metapaths are Author-Paper-Author (APA) and Author-Paper-Venue-Paper-Author (APVPA), which represent two authors collaborating to publish a paper and two authors publishing a paper in the same venue, respectively.

Unlike the widely studied homogeneous information networks, HINs contain comprehensive structural information and rich semantic information, which also provides new opportunities and challenges for data mining.

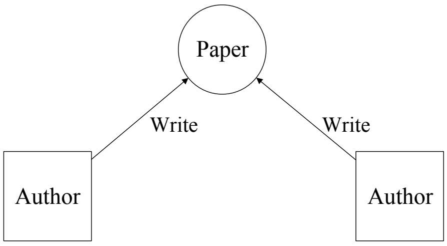


Fig. 3 Metapath (APA) in a bibliographic information network.

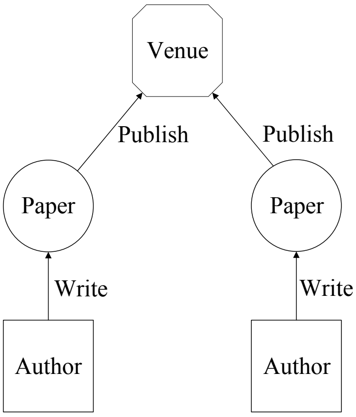


Fig. 4 Metapath (APVPA) in a bibliographic information network.

The clustering problem on HIN is frequently investigated in published studies[7]. With the HIN expressing through the metapath, semistructured information can be transformed into structured information for mining. Instead of converting it into a homogeneous network, HIN can represent multiple types of data information directly, which can reduce the loss of semantic information or structural information. Thus, HIN makes the clustering result more accurate.

However, the existing HIN clustering algorithms generally have disadvantages in versatility and co-clustering with multiple metapaths simultaneously. With regard to the deficiencies, our study proposes a multipath-based heterogeneous network clustering algorithm: the HIN-Nonnegative Matrix Tri-Factorization (HIN-NMTF) algorithm. HIN-NMTF has the following advantages.

First, the HIN-NMTF organizes HINs into a star network structure, with a central-type node connecting subordinate-type nodes based on multiple metapaths. The network is structured, and the algorithm has a strong migration capability. Second, different data matrixes are constructed for every metapath for co-clustering analysis. A nonnegative matrix tri-factor decomposition algorithm is used for multiple nonnegative data matrices at the same time, and soft clustering of all types of nodes in a heterogeneous network is achieved. The clustering results of different metapaths promote each other, thus the method gives us a global clustering perspective of HIN. Finally, the experimental result on a real dataset is better than that obtained by a similar existing clustering algorithm.

The rest of this paper is organized as follows: Related work is presented in Section 2. The problem is formally defined and an algorithm is detailed in Section 3. The experiment and validation results are described in Section 4. Finally, we conclude our paper and suggest avenues for future work in Section 5.

2 Related Work

Clustering of HINs has attracted much attention in recent years. Some scholars have conducted clustering research from the perspective of a parameter estimation method on the basis of network ranking and a probability model. RankClus[6] applies to bi-type HINs and combines ranking functions and clustering. In RankClus, each target object is randomly assigned an initial cluster label, and the conditional ranking distribution is calculated in the current cluster. The Expectation Maximization (EM) algorithm is used to estimate the prior probability, while the prior and posterior probabilities are iterated continuously, and the probability that each conference belongs to each class is obtained.

The NetClus algorithm[7] is an improvement over the RankClus algorithm. RankClus can be clustered only on a bi-type information network, whereas NetClus implements clustering on a star scheme information network, which is capable of ranking and clustering in a star scheme HIN. The PathSelClus algorithm[8] uses a probabilistic model to model the link relationships and assigns different weights to different metapaths. PathSelClus can select the correct metapath for user-guided clustering tasks and use metapaths to represent the relationship between two information objects.

In addition, researchers have gradually modeled the interconnected, multitype networked data as HINs and designed structural analysis methods by leveraging the wealth of object and relationship information in the network. Things2Vec[9] models the function sequence relationships that are generated by the interaction of things as HIN and produce the latent semantic representations from the IoT. Reference [10] used HIN to achieve good results for Android malware detection[11].

Nodes and edges in HINs have rich semantic information. Therefore, clustering based only on edges or nodes is not comprehensive. The process of how to effectively combine the different information to obtain a global clustering result in HIN is also important.

Nonnegative Matrix Factorization (NMF) is used as a relatively novel paradigm for dimensionality reduction[12]. NMTF[13] is a 3-factor decomposition algorithm of a nonnegative data matrix. Two dependent latent variables are associated with the different object clusters in NMTF, leading NMTF suited to co-clustering. Hu et al.[14] utilized NMTF to cluster users and POIs simultaneously and thereby discovered the potential preference of users.

Unlike the parameter estimation method of the probability distribution in existing work, several two-dimensional data matrices are constructed for every metapath for co-clustering analysis in our study. And then we leverage NMTF to obtain several dependent latent variables simultaneously, which is associated with the cluster for the corresponding object in the HIN. It reflects the clustering results of the nodes and also the ambiguity of the clusters to which the nodes belong, making the clustering results that originally had no obvious clustering boundaries more interpretable.

3 Problem Statement and Algorithm Design

Definition 1. Heterogeneous Information Network. An HIN is a directed graph with multiple types of nodes or multiple types of links. It can be denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of links. In addition, HIN should be associated with a node-type mapping function $\phi = \mathcal{V} \rightarrow \mathcal{A}$ and a link-type mapping function $\psi = \mathcal{E} \rightarrow \mathcal{R}$. Each node $v \in \mathcal{V}$ belongs to a particular node type $\phi(v) \in \mathcal{A}$. $\mathcal{A}$ is a set that consists of different types of entities, and each link $l \in \mathcal{E}$ belongs to a particular link type $\psi(l) \in \mathcal{R}$. $\mathcal{R}$ is a set consisting of different relations between different types of entities.

An HIN is an abstraction of the real world, and it focuses on entities and relationships between entities. We use a network schema to represent the nodes and relationship types in an HIN, denoted by $\mathcal{T}_G = (\mathcal{A}, \mathcal{R})$. As shown in Fig. 2, the bibliographic information network can be considered as an HIN. In Fig. 2, node type $\mathcal{A}$ includes paper, term, venue, and author, while the relation type set $\mathcal{R}$ includes "contain", "publish in", and "write".

Definition 2. Metapath. A metapath $\mathcal{P}$ is a path defined on the graph of network schema $T_G = (\mathcal{A}, \mathcal{R})$ and is denoted as $\mathcal{A}_1 \xrightarrow{\mathcal{R}_1} \mathcal{A}_2 \xrightarrow{\mathcal{R}_2} \cdots \xrightarrow{\mathcal{R}_i} \mathcal{A}_{i+1}$, $\mathcal{A}_i \in \mathcal{A}$, which defines a composite relation $\mathcal{R} = \mathcal{R}_1 \circ \mathcal{R}_2 \circ \cdots \circ \mathcal{R}_i$ between types $\mathcal{A}_1$ and $\mathcal{A}_{i+1}$, where $\circ$ denotes the composition operator on relations[15].

Examples of metapaths defined in the network schema in Fig. 2 include the "paper-contain-term" path, the "paper-publish in-venue" path, and the "author-write-paper" path. The metapath can be semantically explanatory, and when two nodes are connected via different paths, they hold different semantic information.

Definition 3. Constraint metapath. The constraint metapath is a metapath based on a specific constraint, which is expressed as

$$\mathcal{C}_{\mathcal{P}} = \mathcal{P}|_{C.\mathcal{P}}(\mathcal{A}_1\mathcal{A}_2\cdots\mathcal{A}_l),$$

where $\mathcal{C}$ represents the constraints on the metapath.

For example, $APA|_{P.L}$ = "DM" represents the author's cooperation in the field of data mining. $APCPA|_{P.L}$ = "DM" and C = "KDD" represent the author in the field of data mining and is a collaborative relationship with the conference, respectively.

In this paper, we choose only those metapaths with explicit semantic information. For example, Figs. 3 and 4 show two different metapaths in the bibliographic network, namely, APA and APVPA, which represent two authors collaborating to publish a paper and two authors that published a paper in the same venue, respectively.

In our proposed model, we use the semantic information expressed by these metapaths to provide an interpretation of the clustering results, and then we calculate the adjacency matrix based on different metapaths.

Definition 4. Weight matrix. Given nodes x_i, x_j and relationship $\mathcal{R}$, $x_i \xrightarrow{\mathcal{R}} x_j$, W_{ij} is leveraged to represent the weight between nodes x_i and x_j. Weights are used to measure the strength of the association between two nodes, and different applications use different measurement methods.

Definition 5. NMF. Given a nonnegative matrix $X = (x_1, x_2, \dots, x_n) \in \mathbf{R}^{m\times n}$ containing n column vectors, the objective of the NMF algorithm is to approximate the nonnegative matrix X by the product of two low-rank matrices U and V, thereby obtaining the compressed representation of original matrix. $U \in \mathbf{R}^{m\times k}$ is a basis matrix, $V \in \mathbf{R}^{n\times k}$ is an encoding matrix, and $k < \min(m, n)$ reduces the original matrix rank.

From the perspective of subspace analysis, NMF actually projects the original matrix X into a subspace consisting of a base vector (a column vector of U), and each sample has a linear representation in the subspace. The base vector U is the column vector, and the coordinates are the column vectors of the coding matrix V^{T}. With the use of the Frobenius norm () as the loss function, the objective function is

$$y_1 = \left\| X - UV^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{1}$$

where $U \geqslant 0$, $V \geqslant 0$, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm of the matrix.

Formula (1) is a convex optimization problem only for the factor matrix U or only for the factor matrix V, but it is a nonconvex problem for the factor matrices U and V at the same time. Therefore, obtaining a global optimal solution is unrealistic. To obtain the local minimum of the objective function y_1, an iterative multiplication update strategy similar to the EM algorithm is usually used[16], that is, only one factor matrix is updated at a time, other factor matrices are fixed, and a local optimal solution can be obtained by iterative updating. When the matrix converges or the number of iterations exceeds the threshold, the iteration stops. Equations (2) and (3) are iterative multiplication update formulas for NMF:

$$U_{ij} = U_{ij}\sqrt{\frac{(XV)_{ij}}{(UV^{\mathrm{T}}V)_{ij}}} \tag{2}$$

$$V_{ij} = V_{ij}\sqrt{\frac{(X^{\mathrm{T}}U)_{ij}}{(VU^{\mathrm{T}}U)_{ij}}} \tag{3}$$

where each data vector x can be approximated as a linear combination of the column vectors of the base matrix U, and the weight coefficient is an element value in the coding matrix V^{T}.

3.1 HIN with star network scheme

Grounded in the definition of the problems and related concepts proposed in the previous section, an HIN co-clustering algorithm based on a multiple metapath HIN-NMTF algorithm is proposed.

With Digital Bibliography and Library Project (DBLP) as an experimental dataset used as an example, node objects are organized into a star network structure, which is shown in Fig. 5. APC represents the relationship in which authors published papers at a conference; APT represents the relationship in which terms are included in papers published by an author; and CPT represents the relationship in which terms are included in papers

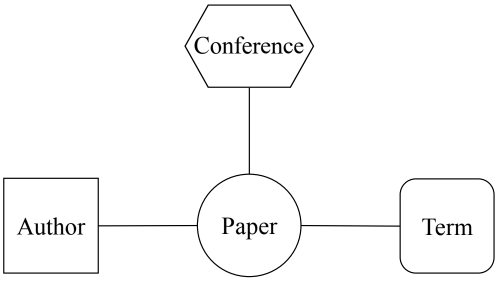


Fig. 5 DBLP star-scheme network structure.

published for a conference. The node types of the DBLP HIN graph $G = (V, E, W)$ are divided into central and attribute types. The central type set is paper, denoted by P, and the number of central types is $|P|$; the subordinate type set is conference, author, term, represented by C, A, T, and the quantity is $|C|, |A|, |T|$. $P = \{p_1, p_2, \ldots, p_{|P|}\}$, $C = \{c_1, c_2, \ldots, c_{|C|}\}$, $A = \{a_1, a_2, \ldots, a_{|A|}\}$, and $T = \{t_1, t_2, \ldots, t_{|T|}\}$. Only edges exist between the center-type and the attribute-type nodes in the star-scheme HIN. Let M be the edge weight set and $M_{ij} \in M$ be the weight value of the edge (x_i, x_j) between the objects, which is the number of edges between two nodes.

3.2 Co-clustering method

3.2.1 Nonnegative matrix tri-factor decomposition algorithm

The NMF algorithm captures only the relationship between two types of nodes. This paper introduces the NMTF[13] decomposition algorithm, which can capture the relationship between multiple node types.

To allow the factor matrices U and V to better simulate the approximation of the original data matrix, the correlation matrix S_l is introduced to obtain a nonnegative matrix three-decomposition algorithm. Meanwhile, the number of clusters of rows and columns of the original data matrix may be different (the dimensions of the rows and columns of the correlation matrix S_l may not be equal), which provides more flexibility for the clustering of the original matrix rows and columns.

The F-norm as a loss function is used to obtain the following objective function:

$$y_3 = \left\| X - USV^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{4}$$

where $U \in R^{m\times k} \geqslant 0, V \in R^{n\times l} \geqslant 0, S \in R^{k\times l} \geqslant 0$, and $k < \min(m, n)$.

Similar to the NMF algorithm, we call the factor matrix U the row clustering factor indication matrix, the factor matrix V the column cluster factor indication matrix, and S the correlation factor matrix. U reflects the relation between clusters of each row of the original data matrix X, and V reflects the relation between clusters of each column of the original data matrix X. S reflects the compressed representation of the original data matrix X; therefore, the original matrix can be compressed and represented as USV.

US contains the base of the original data matrix X column space. Therefore, if we cluster the column vectors, then we can use the columns of US as the base vector of the original data matrix X column space, and V is the encoding matrix of the original data matrix X. Then, if the maximum value of each column of X is on the same row, then the node objects corresponding to these columns belong to the same cluster.

Similarly, VS contains the base of the original data matrix row space. Therefore, if the maximum value of each row of U is in the same column, the node objects corresponding to these rows belong to the same cluster. Thus, the NMTF completes the cooperative clustering of the rows and columns of the original data matrix.

The NMTF multiplication iterative update formula based on the F-norm is as follows:

$$U_{ij} = U_{ij}\sqrt{\frac{(XVS^{\mathrm{T}})_{ij}}{(VV^{\mathrm{T}}S^{\mathrm{T}}US)_{ij}}} \tag{5}$$

$$V_{ij} = V_{ij}\sqrt{\frac{(X^{\mathrm{T}}US)_{ij}}{(UU^{\mathrm{T}}S^{\mathrm{T}}VS)_{ij}}} \tag{6}$$

$$S_{ij} = S_{ij}\sqrt{\frac{(U^{\mathrm{T}}XV)_{ij}}{(U^{\mathrm{T}}USV^{\mathrm{T}}V)_{ij}}} \tag{7}$$

3.2.2 Co-Clustering of star-scheme HIN based on multiple paths

Any HIN can be organized into a star scheme structure. Many entities and relationships are present in an HIN, and a star scheme structure depicts a variety of relationships around an entity. On the basis of the metapaths APC, CPT, and APT, co-clustering analysis of different types of objects and the same type of object construction relation matrix is performed at the same time. Co-clustering has been proved is better than traditional unilateral clustering[17, 18], but these tasks are only performed on one data matrix. However, heterogeneous information networks have multiple types of nodes and edges, and the clustering of different types of nodes should promote each other.

The multiple metapath based co-clustering algorithm for star HINs is proposed in this paper. Simultaneous clustering of multiple types of data objects is achieved, and the global perspective of clustering under HINs is given. Co-clustering is performed for multiple data matrices, and each data matrix can be locally optimized to obtain a local optimal solution of the entire HIN, thus obtaining the global clustering result of an HIN.

The M_{a-c} matrix is obtained based on the metapath APC, the M_{t-c} matrix is obtained based on the metapath TPC, and the M_{a-t} matrix is obtained based on the metapath APT. For M_{a-c}, M_{t-c}, and M_{a-t}, NMTF is performed at the same time to minimize the objective function in Eq. (8) to obtain the low-dimensional approximate representation of the original matrix, which is also the cluster matrix of the row and column of original matrix.

In this paper, the HIN-NMTF clustering algorithm is a soft clustering, so it does not restrict the orthogonality constraints to the clustering factor indication matrixes U, V, W. The element value in the matrixes U, V, W indicates the intensity of the node affiliated with the cluster. A large value corresponds to the great intensity of the node belonging to the cluster.

For example, the element value $U(i, j)$ in the matrix U indicates that the strength of the i-th author belongs to the j-th cluster, and the larger $U(i, j)$ means a high likelihood that the author i belongs to the j-th cluster. Thus, the purpose of the optimization algorithm is to minimize the objective function:

$$y = \left\|M_{a-\mathrm{c}} - US_1V^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \left\|M_{t-c} - WS_2V^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \left\|M_{a-t} - US_3W^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \tag{8}$$

where $U \in R^{m\times K_1}$, $V \in R^{n\times K_2}$, $W \in R^{w\times K_3} \geqslant 0$.

Introducing the correlation matrix S_l provides flexibility for U, V, W, so the number of rows and columns of the matrix can be different, that is, the dimensions of U, V, W can be different. However, the number of clusters for rows and columns is generally the same.

The correlation matrix S_l represents the strength of the association between the cluster indication matrices, or it can be said to be the correlation between the hidden variables of the rows and columns of the matrix. In this paper, S_l is the strength between the cluster of author and the cluster of conference, the strength of the cluster of conference and the cluster of keyword, and the association between authors. Meanwhile, the elements in the matrix S_l reflect the size of clusters in the row space and the column space.

3.3 Co-clustering HIN-NMTF algorithm in HIN based on multiple metapaths

In summary, the objective function of HIN-NMTF, which is the clustering algorithm of the star scheme HIN based on multiple metapaths in this paper is as follows:

$$y = \left\|M_{a-\mathrm{c}} - US_1V^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \left\|M_{t-c} - WS_2V^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \left\|M_{a-t} - US_3W^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \tag{9}$$

s.t. $S_1 \in R^{K_1\times K_2}$, $S_2 \in R^{K_3\times K_2}$, $S_3 \in R^{K_1\times K_3} \geqslant 0$, $U, V, W \geqslant 0$.

For Formula (9), the multiplication iterative updating algorithm is obtained according to the NMTF[16] discussed in the previous section:

$$U \leftarrow U \circ \sqrt{\frac{M_{a-\mathrm{c}}VS_1^{\mathrm{T}} + M_{a-t}WS_3^{\mathrm{T}}}{US_1V^{\mathrm{T}}VS_1^{\mathrm{T}} + US_3W^{\mathrm{T}}WS_3^{\mathrm{T}}}} \tag{10}$$

$$V \leftarrow V \circ \sqrt{\frac{M_{a-c}^{\mathrm{T}}US_1 + M_{t-c}^{\mathrm{T}}\mathrm{W}S_2}{US_1^{\mathrm{T}}U^{\mathrm{T}}VS_1 + VS_2W^{\mathrm{T}}WS_2^{\mathrm{T}}}} \tag{11}$$

$$W \leftarrow W \circ \sqrt{\frac{M_{t-c}VS_2^{T} + M_{a-t}^{\mathrm{T}}US_3}{WS_2V^{\mathrm{T}}VS_2^{T} + WS_3^{\mathrm{T}}U^{\mathrm{T}}US_3}} \tag{12}$$

$$S_1 \leftarrow S_1 \circ \sqrt{\frac{U^{\mathrm{T}}M_{a-c}V}{U^{\mathrm{T}}US_1V^{\mathrm{T}}V}} \tag{13}$$

$$S_2 \leftarrow S_2 \circ \sqrt{\frac{W^{\mathrm{T}}M_{t-c}V}{W^{\mathrm{T}}WS_2V^{\mathrm{T}}V}} \tag{14}$$

$$S_3 \leftarrow S_3 \circ \sqrt{\frac{U^{\mathrm{T}}M_{a-t}W}{U^{\mathrm{T}}US_3W^{\mathrm{T}}W}} \tag{15}$$

On the basis of the above iterative update algorithm, the procedure of the HIN-NMTF algorithm is as shown in Algorithm 1. The algorithm flowchart is shown in Fig. 6. In each step of the iterative update algorithm, only one factor matrix is updated at a time, and the other factor matrices are fixed. When the factor matrix

Algorithm 1 Personalized recommendation model

```
Input:
  Author and conference weight matrix $M_{a-c}$;
  Term and conference weight matrix $M_{t-c}$;
  Author and term weight matrix $M_{a-t}$.
Output:
  Author cluster indication matrix $U$;
  Conference cluster indication matrix $V$;
  Term cluster indication matrix $W$.

  Initialize $[U, V, W, S_1, S_2, S_3] \geqslant 0$
  while When the convergence conditions are not met do
    Fix other variables and update the cluster indication matrix $U$ using Formula (10)
    Fix other variables and update the cluster indication matrix $V$ using Formula (11)
    Fix other variables and to update the cluster indication matrix $W$ using Formula (12)
    Fix other variables and update the author-conference clustering associative matrix $S_1$ using Formula (13)
    Fix other variables and update the term-conference clustering associative matrix $S_2$ using Formula (14)
    Fix other variables and update the author-term clustering associative matrix $S_3$ using Formula (15)
  end while
```

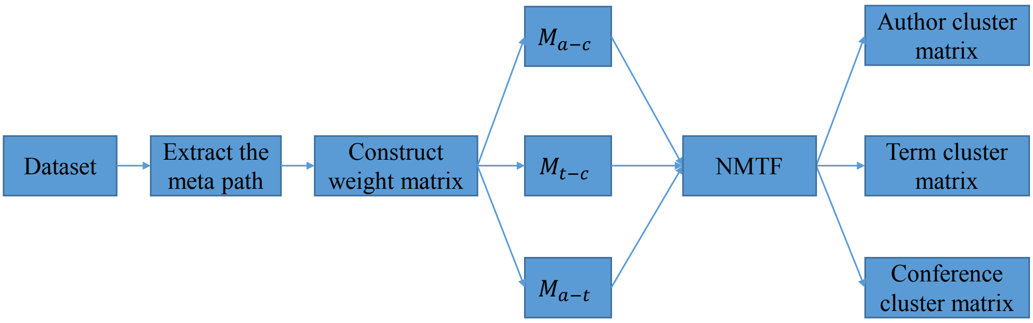


Fig. 6 Algorithm flowchart of HIN-NMTF.

converges or the number of iterations exceeds a given threshold, the algorithm stops.

4 Experimental Evaluation

4.1 Experiment setup

The DBLP is a collection of bibliographic information on major computer science journals and proceedings, which can be used to build an HIN with a star-scheme as shown in Fig. 5. In this experiment, we use a subset of the DBLP records that belongs to four areas, namely, database, data mining, information retrieval, and artificial intelligence, and each area contains five top conferences[19].

We selected the bibliographic dataset with authors who published more than two articles and whose keyword frequency was greater than five times as experimental data, and it contains 28 569 papers that were used as axes, linking 3251 author nodes, 3397 term nodes, and 20 conference nodes. Tables 1 and 2 show a statistics of the data stored in the graph database.

4.2 Quantitative results

This paper uses accuracy to measure the clustering results. Each cluster is mapped to a primitive category.

Table 1 Statistics of the DBLP.

Type	Number	Type	Number
Papers (P)	28 569	Links (A-P)	37 317
Authors (A)	3251	Links (P-T)	158 605
Terms (T)	3397	Links (C-P)	28 569
Conferences (C)	20	Cluster (K)	4

Table 2 Twenty selected conferences in four areas.

Area	Conference				
Database (DB)	EDBT	ICDE	PODS	SIGMOD	VLDB
Data Mining (DM)	ICDM	KDD	PAKDD	SDM	PKDD
Information Retrieval (IR)	CIKM	ECIR	WSDM	SIGIR	WWW
Artificial Intelligence (AI)	AAAI	CVPR	ECML	ICML	IJCAI

By using the Kuhn-Munkres algorithm[20], given an object x_i, its class tag cl_i, and real class tag tl_i, the accuracy of the clustering is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(cl_i, map(tl_i))}{n} \tag{16}$$

where n is the number of all objects of that type, $delta(x, y) = 1$, if $x = y$, otherwise $delta(x, y) = 0$. $map(tl_i)$ is a mapping function obtained by the Kuhn-Munkres algorithm. The accuracy rate is calculated as the proportion of correctly clustered objects.

The selected conferences are in four similar fields; thus, no strict boundary is set for the clusters of each conference, author, and term. Therefore, the soft clustering method used in this paper can reflect the ambiguity and intensity of the clusters to which the node belongs and has better physical meaning and interpretability.

The element value in the cluster matrix shows the intensity of the subordinate-type object that belongs to the cluster. A great value corresponds to a great probability that the subordinate-type node belongs to the cluster. If we want to achieve hard clustering, then we can post-process the cluster matrix and use the maximum value for the cluster to which the node belongs. Table 3 shows the results of the conference cluster matrix V in which the bold font is the maximum value in the row.

The cluster matrix V of the conference is normalized, as shown in Table 4, the bold font is the maximum value in the row. The results show that the AAAI, CVPR, ECML, ICML, and IJCAI conferences are classified in Cluster 3; the CIKM, ECIR, WSDM, SIGIR, and WWW conferences are classified in Cluster 2; the EDBT, ICDE, PODS, SIGMOD, and VLDB conferences are classified in Cluster 4; the ICDM, KDD, PAKDD, SDM, and PKDD conferences are classified in Cluster 1. The use of Formula (16) obtains a 100% accuracy rate for the HIN-NMTF algorithm in conference clustering.

The method used in this paper belongs to soft

Table 3 Conference cluster matrix results.

Conference	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AAAI	1.12	9.56	**58.9**	3.88
CVPR	5.92	5.27	**20.6**	0.00
ECML	3.35	0.36	**16.5**	0.00
ICML	5.49	0.37	**39.1**	0.00
IJCAI	0.00	11.4	**67.9**	6.38
CIKM	10.8	**23.7**	6.05	18.4
ECIR	1.27	**16.4**	11.2	0.00
WSDM	0.20	**0.39**	0.13	0.00
SIGIR	1.88	**93.7**	1.81	0.93
WWW	8.42	**20.1**	4.26	2.37
EDBT	2.88	2.01	1.83	**15.5**
ICDE	16.4	2.81	9.15	**72.1**
PODS	1.55	0.19	3.60	**20.6**
SIGMOD	11.0	2.88	5.16	**64.7**
VLDB	14.9	2.00	6.64	**78.6**
ICDM	**28.4**	1.37	5.32	0.00
KDD	**39.5**	0.15	6.67	4.46
PAKDD	**30.3**	1.07	6.25	0.00
SDM	**14.6**	0.00	2.25	0.77
PKDD	**19.5**	0.28	4.92	2.19

Table 4 Conference cluster matrix normalized results.

Conference	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AAAI	0.02	0.13	**0.80**	0.05
CVPR	0.18	0.17	**0.65**	0.00
ECML	0.17	0.02	**0.81**	0.00
ICML	0.12	0.01	**0.87**	0.00
IJCAI	0.00	0.13	**0.79**	0.08
CIKM	0.18	**0.40**	0.10	0.32
ECIR	0.04	**0.57**	0.39	0.00
WSDM	0.28	**0.54**	0.18	0.00
SIGIR	0.02	**0.95**	0.02	0.01
WWW	0.24	**0.57**	0.12	0.07
EDBT	0.13	0.09	0.08	**0.70**
ICDE	0.16	0.03	0.09	**0.72**
PODS	0.06	0.01	0.14	**0.79**
SIGMOD	0.13	0.03	0.06	**0.78**
VLDB	0.14	0.02	0.07	**0.77**
ICDM	**0.81**	0.04	0.15	0.00
KDD	**0.78**	0.00	0.13	0.09
PAKDD	**0.80**	0.03	0.17	0.00
SDM	**0.82**	0.00	0.13	0.05
PKDD	**0.73**	0.01	0.18	0.08

clustering, which can not only determine the category of the conference but also quantify the strength of the category to which the conference belongs. For example, the WWW conference pays more attention to Web-related application issues, including data acquisition, information indexing, and retrieval methods, in which the information retrieval has a larger proportion. Therefore we can classify conferences in the WWW conference and information retrieval related cluster.

The CIKM conference focuses on information, knowledge management, and databases. From Table 4, we can see that the CIKM has almost the same probability of belonging to Clusters 2 and 4, which also shows that the papers included in the CIKM conference include both information retrieval-related fields and database areas. The main focus of the WSDM conference is search and DM. We can infer from Table 4 that the probability of WSDM publishing papers in the DM field is approximately 28%.

Our algorithm also embodies the ambiguity for the clustering of subordinate-type node authors and terms. The results of the clustering of authors are selected for display because of the large number of authors. For example, Jiawei Han covers two fields: DM and database fields. This phenomenon can be reflected in the author's clustering results (Tables 5 and 6).

4.3 Compared algorithms

In this paper, the NetClus algorithm[7], which is similar to the HIN-NMTF algorithm, is selected and compared to reflect the superiority of our algorithm. The NetClus algorithm is based on the RankClus algorithm[6] and is a combination of sorting and clustering algorithms to solve the global clustering problem in the HIN with a star network schema. The experiment was conducted in full accordance with the parameters required in the paper and the same experimental data were used.

This paper uses the clustering results of the conference to compare the accuracy of the algorithm because of

Table 5 Author cluster matrix results.

Name	DM	IR	AI	DB
Rakesh Agrawal	2.65×10^{-2}	4.20×10^{-45}	2.80×10^{-45}	$\mathbf{1.37\times10^{-1}}$
Jiawei Han	8.90×10^{-2}	2.80×10^{-45}	$\mathbf{2.62\times10^{-1}}$	1.40×10^{-45}
Christos Faloutsos	$\mathbf{1.50\times10^{-1}}$	1.40×10^{-45}	6.92×10^{-14}	5.61×10^{-45}
Michael Stonebraker	1.68×10^{-45}	2.16×10^{-43}	1.40×10^{-45}	$\mathbf{1.43\times10^{-1}}$
Jim Gray	0	1.40×10^{-45}	6.01×10^{-36}	$\mathbf{5.26\times10^{-2}}$

Table 6 Author cluster matrix normalized results.

Name	DM	IR	AI	DB
Rakesh Agrawal	0.16	0.00	0.00	**0.84**
Jiawei Han	0.25	0.00	**0.75**	0.00
Christos Faloutsos	**1.00**	0.00	0.00	0.00
Michael Stonebraker	0.00	0.00	0.00	**1.00**
Jim Gray	0.00	0.00	0.00	**1.00**

the large number of authors, terms, nodes, and fuzzy clustering clusters. Table 7 shows the clustering results of the NetClus algorithm for the conferences. With this result, Formula (16) derives a possible 85% accuracy rate for the NetClus algorithm. Therefore, the accuracy of the HIN-NMTF algorithm is higher than that of the NetClus algorithm.

5 Conclusion

Our research is based on an HIN of DBLP dataset. The node objects are organized into a star network structure. The paper node is used as a central node to connect the subordinate-type nodes. The number of paper nodes is used as the weight matrix of the subordinate-type nodes. The algorithm is based on the metapaths APC, CPT, and TPA.

One soft clustering algorithm is performed on the weight matrix of different types of objects. The algorithm is based on the multiple metapaths co-clustering algorithm for star HINs and produces simultaneous clustering of multiple types of data. The clustering results of different objects promote each other, giving a global perspective on clustering under HINs.

Furthermore, an NMTF algorithm is used for multiple nonnegative data matrices at the same time. The result can not only reflect the clustering results of the nodes but also reflect the ambiguity of the clusters to which the nodes belong, making the clustering results that originally had no obvious clustering boundaries more interpretable.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61701190), the Youth Science Foundation of Jilin Province of China (No. 20180520021JH), the National Key Research and Development Plan of China (No. 2017YFA0604500), the Key Scientific and Technological Research and Development Plan of Jilin Province of China (No. 20180201103GX), the China Postdoctoral Science Foundation (No. 2018M631873), the Project of Jilin Province Development and Reform Commission (No. 2019FGWTZC001), and the Key Technology Innovation Cooperation Project of Government and University for the Whole Industry Demonstration (No. SXGJSF2017-4).

Table 7 NetClus result.

Cluster	Conference				
Cluster1	PODS	SIGMOD	VLDB	ICDE	EDBT
Cluster2	KDD	ICDM			
Cluster3	SIGIR	ECIR	CIKM	WWW	WSDM
Cluster4	ICML	ECML	PKDD	AAAI	IJCAI
	SDM	PAKDD	CVPR		

References

[1] F. Wang, L. Hu, J. Zhou, and K. Zhao, A survey from the perspective of evolutionary process in the internet of things, *Int.J. Distrib. Sens. Netw.*, vol. 2015, p. 462752, 2015.

[2] C. Shi, Y. T. Li, J. W. Zhang, Y. Z. Sun, and P. S. Yu, A survey of heterogeneous information network analysis, *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 17–37, 2017.

[3] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[4] K. Yang, J. H. Zhu, and X. Guo, POI neural-rec model via graph embedding representation, *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 208–218, 2021.

[5] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, A review of relational machine learning for knowledge graphs, *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, 2016.

[6] Y. Z. Sun, J. W. Han, P. X. Zhao, Z. J. Yin, H. Cheng, and T. Y. Wu, RankClus: Integrating clustering with ranking for heterogeneous information network analysis, in *Proc. 12th Int. Conf. Extending Database Technology: Advances in Database Technology*, Saint Petersburg, Russia, 2009, pp. 565–576.

[7] Y. Z. Sun, Y. T. Yu, and J. W. Han, Ranking-based clustering of heterogeneous information networks with star network schema, in *Proc. 15th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Paris, France, 2009, pp. 797–806.

[8] Y. Z. Sun, B. Norick, J. W. Han, X. F. Yan, P. S. Yu, and X. Yu, PathSelClus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks, *ACM Trans. Knowl. Discov. Data*, vol. 7, no. 3, pp. 11, 2013.

[9] L. Hu, G. Wu, Y. H. Xing, and F. Wang, Things2Vec: Semantic modeling in the internet of things with graph representation learning, *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1939–1948, 2020.

[10] S. F. Hou, Y. F. Ye, Y. Q. Song, and M. Abdulhayoglu, HinDroid: An intelligent android malware detection system based on structured heterogeneous information network, in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 1507–1515.

[11] X. L. Zhang, I. Baggili, and F. Breitinger, Breaking into the vault: Privacy, security and forensic analysis of Android vault applications, *Comput. Secur.*, vol. 70, pp. 516–531, 2017.

[12] Y. X. Wang and Y. J. Zhang, Nonnegative matrix factorization: A comprehensive review, *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, 2013.

[13] J. Yoo and S. Choi, Probabilistic matrix tri-factorization, presented at 2009 IEEE Int. Conf. Acoustics, Speech and Signal Proc., Taipei, China, 2009, pp. 1553–1556.

[14] L. Hu, Y. H. Xing, Y. L. Gong, K. Zhao, and F. Wang,

Nonnegative matrix tri-factorization with user similarity for clustering in point-of-interest, *Neurocomputing*, vol. 363, pp. 58–65, 2019.

[15] Y. Z. Sun and J. W. Han. Meta-path-based search and mining in heterogeneous information networks, *Tsinghua Science and Technology*, vol. 18, no. 4, pp. 329–338, 2013.

[16] C. Ding, T. Li, W. Peng, and H. Park, Orthogonal nonnegative matrix t-factorizations for clustering, in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 2006, pp. 126–135.

[17] B. Long, Z. M. Zhang, and P. S. Yu, Co-clustering by block value decomposition, in *Proc. 11th ACM SIGKDD Int. Conf. Knowledge Discovery in Data Mining*, Chicago, IL, USA, 2005, pp. 635–640.

[18] I. S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, in *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2001, pp. 269–274.

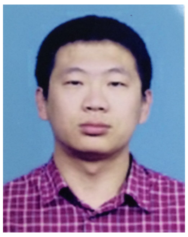
[19] H. B. Deng, J. W. Han, B. Zhao, Y. T. Yu, and C. X. Lin, Probabilistic topic models with biased propagation on heterogeneous information networks, in *Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 1271–1279.

[20] D. Cai, X. He, and J. Han, Document clustering using locality preserving indexing, *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, 2005.

Juncheng Hu received the MS degree from Jilin University in 2019, where he is currently pursuing the PhD degree. His research interests include data mining and machine learning.

Yongheng Xing received the MS degree from Jilin University in 2020, where he is currently pursuing the PhD degree. His research interests include data mining and machine learning.

Mo Han received the MS degree from Jilin University in 2018. Her research interests include data mining and machine learning.



Feng Wang received the MS and PhD degrees from Jilin University in 2012 and 2016. He is currently an associate professor in Jilin University. His research interests include computer networks, information security, Internet of Things, and cyber-physical systems.

Kuo Zhao received the BE degree from Jilin University in 2001, followed by the MS degree and PhD degree from the same university in 2004 and 2008. He is currently an associate professor in the School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai. His research interests are in operating systems, computer networks, and information security.

Xilong Che is an associate professor at the College of Computer Science and Technology, Jilin University, China. His current research areas are machine learning and parallel computing, including related theories, models, and algorithms of ANN, SVC/SVR, GA/ACO/PSO, and their combinations with parallel computing. He is a member of the IEEE.