Meta-Path based Nonnegative Matrix Factorization for Clustering on Multi-type Relational Data

Yangyang Zhao*, Zhengya Sun*, Changsheng Xu†, Hongwei Hao*
*IDMTech, Institute of Automation, Chinese Academy of Sciences
Email: {yangyang.zhao,zhengya.sun,hongwei.hao}@ia.ac.cn
†NLPR, Institute of Automation, Chinese Academy of Sciences
Email: csxu@nlpr.ia.ac.cn

***Abstract*—Clustering on multi-type relational data has attracted increasing interest due to its great practical and theoretical importance. One of the most popular solutions is nonnegative matrix factorization. However, previous work on nonnegative matrix factorization typically copes with multi-type relations individually, and ignores the underlying semantics conveyed by the relation propagation. Additionally, these approaches may suffer from data sparsity as most of the relations between object pairs are unknown. In this paper we propose a novel Meta-Path based Nonnegative Matrix Factorization (MPNMF) framework, which enriches potentially useful similarity semantics for the improved clustering performance. We begin with constructing meta-paths, i.e., paths that connects object types via a sequence of relations, which are appropriately weighted according to certain propagation decay rules. Based on the weighted meta-paths, we are promised to characterize the strength of pairwise interactions among the objects. Together with the attributes in the bag-of-word form, we cluster the objects of target type by collective nonnegative matrix factorization. Experiments on real world datasets demonstrate the effectiveness of our method.**

I. Introduction

With the rapid development of social network, online shopping sites, genomic medicine systems and etc., multi-type relational data have become increasingly ubiquitous. And clustering on multi-type relational data has attracted increasing interest due to its great practical and theoretical importance. The uniqueness of multi-type relational data is that it goes beyond the i.i.d. hypothesis, in the sense that there may exist interactions among individual data points.

In many cases, such multi-type relational data sets consist of multiple types of objects and multiple types of relations as well. ***Note that***, attributes are often regarded as object types for convenience. It is intuitive that the similarity semantics conveyed by meta-paths (a single relation is a special case of the meta-paths, whose path length is 1) is really informative and beneficial in multi-type relational data clustering. For instance, two persons can be highly possibly grouped into the same cluster according to the fact that there are books they both read, films they both watch, which can be implied by the meta-path $People \xrightarrow{read} Book \xrightarrow{read^{-1}} People$ and $People \xrightarrow{watch} Film \xrightarrow{watch^{-1}} People$. However, previous clustering methods on multi-type relational data ignore the underlying semantics conveyed by the relation propagation, so that some similarity between object pairs can not be appropriately characterized just via the relations (i.e. $People \xrightarrow{read} Book$ and $People \xrightarrow{watch} Film$).

As one of the most popular clustering algorithm series, Nonnegative Matrix Factorization (NMF) based clustering methods can enhance clustering with various constraints. The existing methods can be roughly categorized into attribute-view methods and multi-view methods.

NMF [11] with the sum of squared error cost function is proved to be equivalent to a relaxed K-means clustering, which affirmed its clustering capabilities theoretically [12]. Thanks to the clustering capability, NMF has attracted a lot of attentions in machine learning and data mining fields. And most of these traditional NMF based clustering methods are designed for one-side clustering, i.e., clustering the data points or the attributes. To take advantage of the duality and interdependence between data point and attribute clusters, several co-clustering approaches have been proposed to do two-side clustering simultaneously [9], [10], [17]. These existing NMF based co-clustering methods are mostly the two-factor or three-factor NMF with orthonormal constraints or graph regularization on one-side or two-side. However, all these methods can be regarded as the simplification of possible multi-type relational data model, which omit possible relations except for the attributes, so they unsurprisingly show mediocre clustering performances on multi-type relational data.

To get a better clustering result by exploiting information from multiple views (the attribute view and different relation-type views), various multi-view clustering approaches have been proposed. The existing work on the multi-view clustering can be divided into co-regularized spectral clustering approaches and subspace NMF approaches. The former series first construct graphs of multiple views from the data points, with edges between them representing the similarities, and then seek the optimal representation under the assumption that the clustering results of data in each view are similar or the representation derived from each view is close to each other [20]. While the latter series factorize the information in each view, and assume the information in multiple views will be shared or mutually promoted via the constraint that the representations of instances in each view are similar or identical [21], [22], [23]. Although these multi-view clustering methods can handle multi-type relational data, they only utilize the relations but ignore the rich similarity semantics conveyed by the meta-paths. In other words, they face a dilemma if the relations are sparse and barren.

To address the limitations of attribute-view NMFs and existing multi-view clustering methods, we propose a novel meta-path based nonnegative matrix factorization (MPNMF)

978-1-4799-1959-8/15/$31.00 ©2015 IEEE

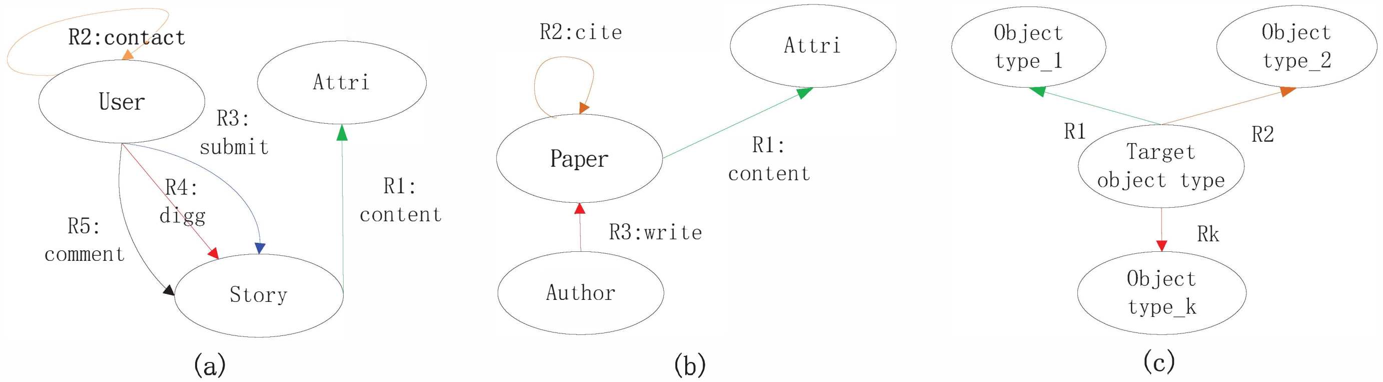

Fig. 1: (a)The structure of Digg dataset;(b)The structure of Cora1 dataset;(c)The structure of star-structured multi-type relational data.

framework. Instead of exploiting information from multiple relations individually, we incorporate the meta-paths with NMF to take advantage of potentially useful similarity semantics for the improved clustering performance. We begin with constructing meta-paths, which are appropriately weighted and selected according to certain propagation decay rules. Based on the weighted meta-paths, we characterize the strength of pairwise interactions among the objects. Together with the attributes in the bag-of-word form, we cluster the objects of the target type by collective nonnegative matrix factorization, with the coefficient matrix constrained to be a cluster indicator matrix. Experiments on real world datasets demonstrate the effectiveness of our method.

The contributions of this paper are as follows: (1) we propose a novel framework which incorporates the meta-paths into NMF, and provides an effective solution for clustering on arbitrary multi-type relational data. (2) we design a general meta-path based similarity measure to characterize the strength of pairwise interactions between the objects of the target type. (3) the proposed NMF objective function of the framework provides the double semantic integration in another sense, which also advances the clustering performance.

The rest of the paper is organized as follows. Related work is briefly reviewed in Section II. In Section III we formulate the problem and explain our basic idea. The details of our proposed MPNMF framework are presented in Section IV. Then we report and discuss the experimental results in Section V. Finally, the conclusion is given in Section VI.

II. Related Work

In this section, we briefly review some of the research literatures related to multi-type relational data clustering.

A. *Meta-path based similarity measures*

Meta-path is a sequence of relations between entity types, which constructs a synthetic relation between its starting entity type and ending entity type [4]. The length of a meta-path is defined as the number of the serially connected relations. In order to capture richer semantic meanings of the structure, many studies have raised many similarity search methods, which have been widely used in information retrieval and data mining. However, most of these studies focus on homogeneous networks or bipartite networks, such as personalized PageRank (P-PageRank) [5] and SimRank [6]. These similarity measures disregard the subtle semantic differences between different meta-paths, taking all of them as one type. Additionally, they are biased to either highly visible or highly concentrated objects. Because of these drawbacks, it is not wise to employ these methods to calculate the similarities in heterogeneous networks. As in many scenarios, finding similar entities in networks is to find similar peers, Sun et al. proposed a symmetric-meta-path-based similarity search measure PathSim to find similar peer entities that not only are strongly connected, but also share comparable visibility [4]. To measure the relatedness between different-typed objects, Shi et al. proposed HeteSim which searches the strength of both asymmetry and symmetry meta-paths [7]. It is claimed that these two methods show superiority in dealing with the clustering problem and top-k similarity problem of heterogeneous information networks when compared with the above proposed homogeneous network methods [4], [7].

B. *Related Clustering Methods*

The attribute-view NMF methods usually perform well on data without relations. Most of these traditional NMF based clustering methods are designed for one-side clustering, i.e. , clustering the data points or the attributes. In fact, the two-factor NMF with orthonormal constraints enforced on both of the coefficient matrices, is a basic NMF based co-clustering method [12]. But the two-factor NMF is a rather poor low-rank matrix approximation. To solve the poor approximation for better co-clustering, Ding et al. proposed ONMTF [9], a three-factor NMF to ensure accurate low rank matrix approximation.

To take advantage of the duality and interdependence between data and attribute clusters, several co-clustering approaches have been proposed to do two-side clustering simultaneously [9], [10], [17]. e.g., we can simultaneously partition documents and words for document clustering. Utilizing two-side NMF clustering methods, we can co-cluster the objects and their attributes simultaneously and improve the clustering performance to some extent. To preserve the manifolds (i.e. , the smoothness between points) in the resulting matrix factor space, Cai et al. proposed Graph regularized Nonnegative Matrix Factorization (GNMF), which is able to easily obtain co-clustering results [8], but it just considers about the manifold on one side. Unlike GNMF, Gu et al. introduced both the similarity graph of data points and that of attributes in co-clustering tasks, referred to as Dual Regularized Co-Clustering (DRCC) [10]. More and more approaches are inclined to take advantages of the manifolds on both data side and attribute side, Shang et al. proposed a Dual regularized NMTF (D-NMTF) model which is almost the same as DRCC, but the three matrix factors are all constrained to be nonnegative [17]. Wang et al. proposed a fast DNMTF method (LP-FNMTF) by constraining factors to be cluster indicator matrices [18].

As one of the multi-view methods, co-regularized spectral clustering approaches and subspace NMF approaches consider the information from multiple views and seek consistency between them. Co-reguSC is a spectral clustering framework that co-regularizes the clustering hypotheses and looks for clusterings that are consistent across the multiple views [20]. ColNMF, as a collective matrix factorization method [21], shares the coefficient matrix but employs different basis matrices across views [22]. MultiNMF attempts to regularize the coefficient matrices learnt from factorizations of different views towards a common consensus, by factorizing individual matrices of all the views and minimizing the differences between coefficient matrices and the consensus matrix [23]. To sum up, all these multi-view clustering methods tend to characterize different views of the data to find a common consensus, which is inclined to suffer from a loss of potentially useful similarity semantics and sparsity. In contrast, our method characterize the multi-type relational data in an information-fusion way, which mines effective similarity semantics and alleviates the sparsity.

III. Notations and Problem Formulation

Given a certain multi-type relational dataset, it may contain one or more object types and various original inter-type relations and intra-type relations, which can be represented as a directed graph $\mathbb{G} = (\mathbb{Q};\mathbb{Z})$. $\mathbb{Q}$ is the set of vertices, including t types of objects with the corresponding object set $\mathbb{Q}_1 = \{q_{11}, \cdots, q_{1n_1}\}$,...,$\mathbb{Q}_t = \{q_{t1}, \cdots, q_{tn_t}\}$. $\mathbb{Z} \subseteq \mathbb{Q} \times \mathbb{Q}$ is the set of edges between the vertices in $\mathbb{Q}$, including s types of relations $\mathbb{Z} = \{Z_1, \cdots, Z_s\}$, where $t \geq 1$ and $s \geq 1$. The size of the relations are different as different relations involve different object types.

Supposing we aim to cluster objects of a certain object type $\mathbb{Q}_{target} = \{q_{target1}, \cdots, q_n\}$ in multi-type relational data set, R $(n \times n)$ denotes the joint similarity matrix, X $(m \times n)$ denotes the bag-of-word attribute matrix, F $(m \times c1)$ denotes the coefficient matrix of attributes, G $(n \times c2)$ denotes the coefficient matrix of objects, where n is the number of objects and m is the number of attributes. The clustering task is to group the n objects into $c2$ clusters, accompanied by grouping the m attributes into $c1$ clusters. Inspired by [18], we denote the set of all cluster indicator matrices as Ψ, where each matrix has one and only one element equal to 1 in each row to indicate the cluster membership, with the rest elements as 0.

The two-factor NMF with orthonormal constraints on the both coefficient matrices, is a NMF based co-clustering method [12]. The objective is:

$$\begin{aligned} &J_1 = \|X - FG^T\|_F^2 \\ &s.t. \quad F \geq 0, G \geq 0, F^T F = I, G^T G = I, \end{aligned} \tag{1}$$

where the resulting matrix factors F, G in terms of attributes and data points respectively can be regarded as soft labels [9], [12]. However, this brings about an over-constrained problem as orthogonal factor approximations for X may not exist, leading to a rather poor low-rank matrix approximation.

To solve the poor approximation for better co-clustering, Ding et al. proposed ONMTF [9] :

$$\begin{aligned} &J_2 = \|X - FSG^T\|_F^2 \\ &s.t. \quad F \geq 0, G \geq 0, F^T F = I, G^T G = I, \end{aligned} \tag{2}$$

where S provides increased degrees of freedom to ensure accurate low rank matrix approximation, F gives row clusters and G gives column clusters. Moreover, the nonnegative constraint on S can be relaxed, which can be regarded as a Semi-Nonnegative Matrix Tri-Factorization method [16].

To preserve the manifolds (i.e. , the smoothness between points) in the resulting matrix factor space, Cai et al. proposed Graph regularized Nonnegative Matrix Factorization (GNMF) [8]:

$$\begin{aligned} &J_3 = \|X - FG^T\|_F^2 + \lambda_G \sum_{i=1}^{n} \sum_{j=1}^{n} W^G{}_{i,j} \|G_{i,*} - G_{j,*}\|^2 \\ &= \|X - FG^T\|_F^2 + \lambda_G tr(G^T L_G G) \\ &s.t. \quad F \geq 0, G \geq 0 \end{aligned} \tag{3}$$

where $L_G = D_G - W_G$ is known as the graph Laplacian matrix with D_G being a diagonal matrix whose diagonal elements $(D_G)_{i,i} = \sum_j (W_G)_{i,j}$. This method can easily obtain co-clusterig results, but just considers the manifold on one side.

The major shortcoming of these attribute-view methods is that they omit all the relations except for the attributes, while the existing multi-view methods only utilize the relations but ignore the rich similarity semantics conveyed by the meta-paths. Therefore, in this paper, we introduce an effective solution for clustering on arbitrary multi-type relational data. We first integrate the similarity semantics of different meta-paths into a joint similarity matrix, and then making the most of the joint similarity matrix together with the attributes in bag-of-word form via NMF.

IV. Details of MPNMF

In this section, we will detail the proposed MPNMF framework. We first construct the weighted meta-paths, and filter out the ones whose weights are less than the threshold. Then we propose new functions to characterize the strength of pairwise interactions among the objects. Finally we get a weighted combination of the similarity semantics, and utilize a graph Laplacian regularized collective NMF to cluster the objects of the target type.

A. Meta-Path Constructing and Weighting

We confine our meta-paths to the ones that start and end with the target object type for enriching similarity semantics. It also offers the following benefits. (1) It provides a strong interaction between objects, and is easier to integrate more related relations together. E.g., we can incorporate the relation $User \xrightarrow{contact} User$ by $Story \xrightarrow{comment^{-1}} User \xrightarrow{contact} User \xrightarrow{comment} Story$ in Fig. 1(a). (2) It can characterize arbitrary multi-type relational data (e.g., Fig. 1(a)) which goes beyond the star structure as illustrated in Fig. 1(c). (3) It is practicable to integrate all the similarity semantics into a more powerful joint similarity matrix.

W.r.t. (**w**ith **r**espect **t**o) the meta-paths, each of them may have different degree of impact in cluster consistency. So it is essential to assign weights for them. However, the existing weighting methods usually compute the weights with the label information of the sampled data points, which turns the unsupervised problem to be a semi-supervised problem. To keep our method as a complete unsupervised algorithm, we assign weights to the meta-paths according to certain propagation decay rules, which means that the weights decays with the length of the meta-path. In this paper, to facilitate the experiments, we set all the weights as follows,

$$Weight_k = \varrho^N \tag{4}$$

where ϱ is set to be 0.8 according to cross-validation, and N is the length of the meta-path.

B. Meta-Path based Similarity Measure

As the selected meta-paths are all inner-typed, we propose a meta-path-based similarity measure g-PathSim (general Path Similarity) which characterizes the strength of pairwise interactions among the objects connected along a certain meta-path. The basic idea is that similar objects are not only strongly connected but also have few connections with others. Given an arbitrary meta-path $P : P_1 \times P_2 \times \cdots \times P_n$ of length $n(n{>}1)$, it can be decomposed into two sub-meta-paths that share one common object type, just like $P_L : P_1 \times \cdots \times P_m$ and $P_R : P_{m+1} \times \cdots \times P_n$ where the object types that P_m end with and P_{m+1} start with are identical. The $g\text{-}PathSim$ is defined as follows,

$$\begin{aligned} & g\text{-}PathSim(o_i \xrightarrow{P} o_j) \\ & = \frac{2 * (|o_i \xrightarrow{P} o_j| + |o_j \xrightarrow{P} o_i|)}{|O(o_i \mid P_L)| + |O(o_j \mid P_L)| + |I(o_i \mid P_R)| + |I(o_j \mid P_R)|} \end{aligned} \tag{5}$$

where $|\ |$ is a counter function, $|o_i \xrightarrow{P} o_j|$ is the number of meta-path instances from object o_i to o_j along the meta-path P_{LR}, $|O(o_i \mid P_L)|$ is the weighted out-degree of object o_i along the meta-path P_L, and $|I(o_i \mid P_R)|$ is the weighted in-degree of object o_i along the meta-path P_R. Note that if o_i and o_j are the same object, $g\text{-}PathSim(o_i, o_j \mid P_{LR})$ is directly set to 1.

W.r.t. two objects of the same entity type, the g-PathSim of them can be calculated in matrix or vector manner as follows, where $\|*\|$ is the L2-norm function, L and R are the adjacency matrices corresponding to the left and the right meta-paths respectively.

$$\begin{aligned} & g\text{-}PathSim(o_i \xrightarrow{P} o_j) \\ & = \frac{2 * (L_{i,*}R_{*,j} + L_{j,*}R_{*,i})}{\|L_{i,*}\|^2 + \|L_{j,*}\|^2 + \|R_{*,i}\|^2 + \|R_{*,j}\|^2} \end{aligned} \tag{6}$$

W.r.t. a single relation, or referred as one-length meta-path, we add an imaginary object type E between the real object type and decompose the atomic relation into two relations as the authors did in [7]. The g-PathSim of one-length meta-path is calculated as follows, where P is the original adjacency matrix of the one-length meta-path.

$$\begin{aligned} & g\text{-}PathSim(o_i \xrightarrow{P} o_j) \\ & = \frac{2(P_{i,j} + P_{j,i})}{\|P_{i,*}\|^2 + \|P_{j,*}\|^2 + \|P_{i,*}\|^2 + \|P_{j,*}\|^2} \end{aligned} \tag{7}$$

PathSim [4] is a special case of g-PathSim with symmetry meta-paths. Compared with PathSim and HeteSim [7], the advantages of g-PathSim are as follows: (1) our method synthesizes the semantic propagation information from component sub-meta-paths in bi-direction. (2) the resulting similarity matrix maintains symmetry for any arbitrary meta-path, which is of vital significance in many cases. (3) the symmetry ensures us to search the similarities of the upper triangular matrix at half of the computational cost.

C. The NMF Objective Function for Semantic Integration

To make full use of the potentially useful similarity semantics contained in the similarity matrices of different meta-paths together with the attributes in bag-of-word form, we design a graph Laplacian regularized collective NMF objective function as follows

$$\begin{aligned} J_4 &= J_{attri} + J_{joint\text{-}simil} \\ &= \|X - FSG^T\|_F^2 + \lambda\|R - GUG^T\|_F^2 \\ &+ \lambda_F tr(F^T L_F F) + \lambda_G tr(G^T L_G G) \\ s.t.&\quad F \geq 0, G \geq 0 \end{aligned} \tag{8}$$

where R is the joint similarity matrix, J_{attri} and $J_{joint\text{-}simil}$ respectively correspond to the loss function of attribute matrix and that of the joint similarity matrix.

$$R = \sum^{k} Weight_k \cdot g\text{-}PathSim_k \tag{9}$$

Note that, to ensure the weighted adjacency matrix for graph Laplacian regularization reliable, we choose the top-K (where K is set to be 6 according to cross-validation) highest-weighted links for every node in the joint similarity matrix R and construct a filtered adjacency matrix $\widehat{R}$ to construct L_G.

This model simultaneously decomposes similarity semantics in attribute perspective and utilizes it as graph Laplacian regularization in relation perspective. Meanwhile we treat the attributes as an object type when constructing the joint similarity matrix R. In this way, the attributes also takes part in the graph Laplacian regularization, so that we can utilize them in relation perspective. Our regularization means that, the more effective links and similar attributes two nodes have, the closer the encoding vectors of them will be. As we integrate the similarity semantics by combining the weighted similarity matrices, this NMF objective function provides the double semantic integration in another sense.

Despite its mathematical elegance, Eq. (8) suffers from two problems that impede its practical use. First, similar to many other NMFs, the immediate outputs (G and F) are not cluster labels, which requires an additional post-processing step and often leads to non-unique solutions. Second and more importantly, Eq. (8) is usually solved by alternately iterative algorithms, and the intensive matrix multiplications are involved at each iteration step. As a result, it is infeasible to apply such algorithms to large-scale real world data due to the expensive computational cost. Inspired by [18], we solve the original clustering problem, i.e., the normal K-Means, instead of solving the relaxed clustering problems. Specifically, we constrain the coefficient matrices of the model to be cluster indicator matrices and minimize the following objective:

$$\begin{aligned} &J_5 = \|X - FSG^T\|_F^2 + \lambda\|R - GUG^T\|_F^2 \\ &+ \lambda_F tr(F^T L_F F) + \lambda_G tr(G^T L_G G) \\ &s.t. \quad F \in \Psi^{m\times c1}, G \in \Psi^{n\times c2} \end{aligned} \tag{10}$$

As the constraints of cluster indicator matrix enable us to split the whole NMF problem into parallelizable subproblems [18], the NMF objective function can be optimized by a scalable optimization algorithm.

D. Scalable Optimization Algorithm

We adopt an alternating projection method to learn the parameters S, U, $\boldsymbol{F}$ and $\boldsymbol{G}$. More specifically, each time we update one parameter and fix the others. This procedure will be repeated for several iterations until the termination condition is satisfied and the algorithm converges to an optimal solution. More importantly, we detail the derivation process, and prove the correctness and convergence of our method.

- $\boldsymbol{U}$pdating S and U

One straightforward way to learn the parameters is to set the gradient of J_5 with respect to S and U to 0 and solve the corresponding linear system or nonlinear system.

$$\begin{aligned} S &= (F^T F)^{-1} F^T X G (G^T G)^{-1} \\ U &= (G^T G)^{-1} G^T R G (G^T G)^{-1} \end{aligned} \tag{11}$$

- $\boldsymbol{U}$pdating F and G

Due to the inherent characteristics of cluster indicator matrices, the update of F and G is rather efficient and quite different from normal matrix update mechanisms. We would like to elaborate it with several lemmas.

Lemma 1: Given the optimization problem

$$\min_{G\in\Psi^{n\times c}} \|X - FG^T\|_F^2$$

the solution of G can be obtained by

$$G_{i,j} = \begin{cases} 1 & j = \arg\min_k \|X_{*,i} - F_{*,k}\|_F^2 \\ 0 & otherwise \end{cases} \tag{12}$$

Meanwhile, the solution of F with respect to $\min\limits_{F\in\Psi^{m\times c}} \|X - FG^T\|_F^2$ can be obtained by

$$F_{h,v} = \begin{cases} 1 & v = \arg\min_e \|X_{h,*} - (G^T)_{e,*}\|_F^2 \\ 0 & otherwise \end{cases} \tag{13}$$

Proof: W.r.t. the i-th row of G, $G_{i,*}(1 \leq i \leq n) \in \Psi^{1\times c}$ is a cluster indicator vector, in which only one element is 1 and the others are 0. The solution of $G_{i,*}$ lies in finding which column is 1. Assuming the k-th ($1 \leq k \leq c$) column of $G_{i,*}$ is 1, then $F{G_{i,*}}^T = F_{*,k}$ and we get the k-th column of F. So we just need to calculate $\|X_{*,i} - F_{*,k}\|_F^2$ for c times, and choose the k that leads to the minimal value. We repeat the above step n times and obtain the solution of G. The optimization of F is similar. ■

Lemma 2: Let Ω be the set of cluster indicator matrices that optimize the following objective

$$\min_{C\in\Psi} tr(C^T L_C C) \tag{14}$$

where $L_C = I - A$ is the normalized graph Laplacian of C, $A = {D_C}^{-\frac{1}{2}} W_C {D_C}^{-\frac{1}{2}}$ and $(D_C)_{i,i} = \sum_j (W_C)_{i,j}$. If the adjacency matrix W_C is symmetric, then there exists a solution in Ω such that the following objective is optimized

$$\min_{C\in\Psi, Q^T Q = I} \|C - BQ\|_F^2 \tag{15}$$

where $B = H\Gamma^{\frac{1}{2}}$. $H, \Gamma = eigh(A, c)$ is the eigen-decomposition ($A = H\Gamma H^T$) of A, and $\Gamma \in \Lambda^{c\times c}$ is a real diagonal matrix with c largest eigenvalues as its diagonal elements. $Q = U_C {V_C}^T$, in which U_C, V_C are calculated by the Singular Value Decomposition ($B^T C = U_C \Lambda_C {V_C}^T$) of $B^T C$.

Proof: From the proof of Proposition 1 in [18], we can obtain that

$$\min_{C\in\Psi} tr(C^T L_C C) = \min_{C\in\Psi, Q^T Q = I} \|CC^T - BQ(BQ)^T\|_F^2$$

Suppose C approximates BQ, then CC^T approximates $BQ(BQ)^T$. Note that it offers one possible solution although $CP \neq C$ approximating BQ when $PP^T = I$ and $CP \in \Psi$ can also make CC^T approximate $BQ(BQ)^T$. Hence, the solution derived from Eq. (15) reasonably leads to the minimization of the objective Eq. (14), which completes the proof of Lemma 2. ■

According to Lemma 2, we can transform the objective in Eq. (10) as follows:

$$\begin{aligned} &J_6 = \|X - FSG^T\|_F^2 + \lambda\|R - \underline{G}UG^T\|_F^2 \\ &+ \lambda_F \|F - B_F Q_F\|_F^2 + \lambda_G \|G - B_G Q_G\|_F^2 \\ &s.t. \quad F \in \Psi^{m\times c1}, G \in \Psi^{n\times c2}, {Q_F}^T Q_F = I, {Q_G}^T Q_G = I \end{aligned} \tag{16}$$

where B_F, B_G, Q_F, Q_G are calculated by Lemma2.

Since Eq. (16) in terms of G cannot be solved directly, we adopt an alternative approach inspired by [19], which simplifies this nonlinear problem by updating the left G with the underline in one iteration step, with the right G adapted accordingly. Through the experiments, we verify the viability of the update of G, which is able to converge to a local optimum fast. Thanks to G is a cluster indicator matrix, Eq. (16) can be decoupled into the following incremental subproblems for each $1 \leq i \leq n$:

$$\begin{aligned} &\min_{G_{i,*}\in\Psi^{1\times c2}} \|X_{*,i} - FS{G_{i,*}}^T\|_F{}^2 + \lambda\|R_{*,i} - \underline{G}U{G_{i,*}}^T\|_F{}^2 \\ &+ \lambda_G \|G_{i,*} - (B_G Q_G)_{i,*}\|_F^2 \end{aligned} \tag{17}$$

and the solution can be obtained by

$$G_{i,j} = \begin{cases} 1 & j = \arg\min_k \{\|X_{*,i} - (FS)_{*,k}\|_F^2 \\ & \quad + \lambda\|R_{*,i} - (\underline{G}U)_{*,k}\|_F^2 - 2\lambda_G (B_G Q_G)_{i,k}\} \\ 0 & otherwise \end{cases} \tag{18}$$

Analogously, we obtain the solution of F as follows:

$$F_{h,v} = \begin{cases} 1 & v = \arg\min_{e}\{\|X_{h,*} - (SG^T)_{e,*}\|_F^2 \\ & \quad - 2\lambda_F (B_F Q_F)_{h,e}\} \\ 0 & otherwise \end{cases} \tag{19}$$

E. Correctness and Convergence Analysis

- *C*orrectness Analysis

Theorem 1: The update algorithms for G and F according to Eq. (18) and (19) monotonically decrease the objective in Eq. (16).

Proof: Because $B = H\Gamma^{\frac{1}{2}}$, $A = H\Gamma H^T$, $\Gamma = \Gamma^T$, ${Q_G}^T Q_G = I$, so

$$\begin{aligned} &(B_G Q_G)_{i,*} {(B_G Q_G)_{i,*}}^T \\ &= ({B_G}_{i,*} Q_G)({B_G}_{i,*} Q_G)^T \\ &= {B_G}_{i,*} (Q_G {Q_G}^T) {{B_G}_{i,*}}^T \\ &= {B_G}_{i,*} {{B_G}_{i,*}}^T \\ &= H_{i,*}\Gamma^{\frac{1}{2}}\Gamma^{\frac{1}{2}} {H_{i,*}}^T \\ &= H_{i,*}\Gamma {H_{i,*}}^T \\ &= A_{i,i} \end{aligned} \tag{20}$$

And because A is computed from the adjacency matrix W_c, so $A_{i,i}$ is a constant w.r.t a certain multi-type relational data set.

Supposing the k-th ($1 \leq k \leq c$) column of $G_{i,*}$ is 1, then

$$\begin{aligned} &\|G_{i,*} - (B_G Q_G)_{i,*}\|_F^2 \\ &= \sum_{z=1}^{c} (G_{i,z} - (B_G Q_G)_{i,z})^2 \\ &= \sum_{z=1}^{c} ((G_{i,z})^2 - 2G_{i,z}(B_G Q_G)_{i,z} + {(B_G Q_G)_{i,z}}^2) \\ &= (G_{i,k})^2 - 2(B_G Q_G)_{i,k} + \sum_{z=1}^{c} {(B_G Q_G)_{i,z}}^2 \\ &= 1 - 2(B_G Q_G)_{i,k} + tr((B_G Q_G)_{i,*} {(B_G Q_G)_{i,*}}^T) \\ &= 1 - 2(B_G Q_G)_{i,k} + tr(A_{i,i}) \\ &= -2(B_G Q_G)_{i,k} + 1 + A_{i,i} \end{aligned} \tag{21}$$

thus

$$\begin{aligned} &\arg\min_k \|G_{i,*} - (B_G Q_G)_{i,*}\|_F^2 \\ &= \arg\min_k (-2(B_G Q_G)_{i,k} + constant) \\ &= \arg\min_k (-2(B_G Q_G)_{i,k}) \end{aligned} \tag{22}$$

and the third item in Eq. (18) is correct.

According to Lemma 1, we can get the following equation:

$$\begin{aligned} &\min_{G_{i,*} \in \Psi^{1\times c2}} {\|X_{*,i} - FS{G_{i,*}}^T\|_F}^2 \\ &= \arg\min_k \|X_{*,i} - (FS)_{*,k}\|_F^2 \end{aligned} \tag{23}$$

$$\begin{aligned} &\min_{G_{i,*} \in \Psi^{1\times c2}} {\|R_{*,i} - GU{G_{i,*}}^T\|_F}^2 \\ &= \arg\min_k \|R_{*,i} - (GU)_{*,k}\|_F^2 \end{aligned} \tag{24}$$

Thus, Eq. (18) is correct. Since the selection criteria of k is to make the left side of the equation minimized, the update of G according to Eq. (18) monotonically decreases the objective in Eq. (16). Analogously, we can deduce a similar conclusion for F. ■

- *C*onvergence Analysis

Theorem 2: The objective function of MPNMF monotonically decreases, and converges to local minimum.

Proof: Based on Theorem 1, the objective function in Eq. (16) monotonically decreases at each update step for G and F. In addition, owing to the property of the equation solution, the update of S and U according to Eq. (11) can also make the objective monotonically decrease. Since the objective function is lower bounded by 0, the algorithm will converge to local minimum. Note that multiple local optima may exist, so there is no guarantee that the objective will converge to the global optimum. ■

V. Experiments.

To test the main idea of our method and learn more about the characterizations of MPNMF, we designed a series of comparative experiments with a number of baselines.

A. Datasets and Evaluation Scheme

To facilitate independent replication of the experiments, we employ three real world multi-type relational datasets to evaluate the compared methods. The characteristics of the datasets are presented in Table I.

The dataset Cora1 [2] and Cora2 [3] contain research papers from the computer science community. We adopt the whole Cora2 and the subset EC, OS, NW and DB of Cora1. In Cora2, there is only two type of original relations, and the original adjacency matrix M_{PP} describes the relation $Paper \xrightarrow{cite} Paper$, M_{PAttri} describes the relation $Paper \xrightarrow{content} Attri$. We first characterize Cora2 by the meta-paths of $M_{PP}, M_{PP} \times {M_{PP}}^T, {M_{PP}}^T \times M_{PP}, M_{PP} \times M_{PP}$ and augment them by $M_{PAttri} \times {M_{PAttri}}^T$. And then we do meta-path weighting and selection, relation strength learning via g-PathSim and finally get a weighted combination of the selected meta-paths. Meanwhile, the subsets of Cora1 have one more original adjacency matrix M_{PA} corresponding to the relation $Paper \xrightarrow{write^{-1}} Author$, which constitutes a simple heterogeneous information network together with $Paper \xrightarrow{cite} Paper$. We augment the candidate meta-paths by $M_{PA} \times {M_{PA}}^T$ for Cora1.

The Digg [1] dataset we utilize in this paper consists of stories, users and their actions (*submit*, *digg*, *comment*) w.r.t. the stories, as well as the explicit *contact* relation among these users. In addition, the attribute of the Digg stories is made up of keywords extracted from the story titles. In this paper we choose 1000 stories of five topics (i.e., pc games, space, pets/animals, linux/unix, political news) as the objects of target entity type, 200 for each topic, as well as the related users. As described in Fig.1, $R_k (k = 1, \cdots, 5)$ are original adjacency matrices of the five original relations. The candidate meta-paths of this network are quite various, including symmetric meta-path such as $R_1 \times {R_1}^T, {R_3}^T \times R_3, {R_3}^T \times R_2 \times R_2 \times R_3$, as well as many asymmetric ones such as ${R_3}^T \times R_2, {R_3}^T \times$

$R_2 \times R_3, {R_3}^T \times R_2 \times {R_2}^T \times R_4$. We traverse the meta-paths with the length constraint 4 according to cross-validation.

To evaluate the clustering results, we adopt three widely used standard metrics: cluster purity [9], normalized mutual information (NMI) [8] and clustering precision [13].

B. Baselines and Parameters Setting

The compared approaches include the state-of-art NMF based clustering methods and multi-view clustering methods, and we also take K-Means and SVD-initialized NMF (SVD-NMF) [15] as baselines. For co-clustering methods, including Orthogonal NMTF (ONMTF) [9], Dual Regularized Co-clustering (DRCC) [10], Locality preserved fast NMTF (LP-FNMTF) [18] and our method, the number of attribute clusters is set to be the same as that of data clusters, i.e., $c1 = c2$. To compare these algorithms fairly, we run them under different parameter settings and select the best average result for comparison. For graph regularized methods, including GNMF, DRCC and our method, we construct K-nearest neighbor graph following [10], where the K is set by searching the grid of $\{1, 2, ..., 10\}$ and the regularization parameters (i.e., λ_F and λ_G in Eq. (16)) are set by searching the grid of $\{0.1, 1, 10, 100, 500, 1000\}$. In our method there is another parameter λ need to be set, which denotes the relative weight of structure-content information NMTF and set with the grid of $\{0.01, 0.1, 1, 3, 5, 10, 100\}$. The multi-view clustering methods include Co-reguSC [20], ColNMF [22], MultiNMF [23] and our method.

C. Performance and Result Analysis

Under each parameter setting of each method mentioned above, we independently repeat the experiments for 10 times and report the best average results. We choose the dataset EC and NW to examine the effectiveness of g-PathSim and the 10-times best average results is reported in Table II and

TABLE I: The characteristics of the datasets

dataset	cluster	objects of target entity type	attribute	original adjacency matrix
EC	3	648(295,120,233)	1551	$M_{\boldsymbol{P}\boldsymbol{P}}, M_{\boldsymbol{P}A}, M_{\boldsymbol{P}Attri}$
OS	4	1667(582,641,308,136)	2326	$M_{\boldsymbol{P}\boldsymbol{P}}, M_{\boldsymbol{P}A}, M_{\boldsymbol{P}Attri}$
NW	4	912(397,90,262,163)	1675	$M_{\boldsymbol{P}\boldsymbol{P}}, M_{\boldsymbol{P}A}, M_{\boldsymbol{P}Attri}$
DB	7	766(206,96,89,90,100,97,88)	1358	$M_{\boldsymbol{P}\boldsymbol{P}}, M_{\boldsymbol{P}A}, M_{\boldsymbol{P}Attri}$
Coar2	7	2708(298,418,818,426,217,180,351)	1433	$M_{\boldsymbol{P}\boldsymbol{P}}, M_{\boldsymbol{P}Attri}$
Digg	5	1000(200,200,200,200,200)	1805	R_1, R_2, R_3, R_4, R_5

TABLE II: The results of LP-FNMTF based on different similarity semantics (AS : Attribute based Similarity)

data set	Metrics	AS	AS+originalrelation	AS+PathSim	AS+HeteSim	g-PathSim
EC	Purity	0.6080	0.6125	0.6197	0.6361	**0.6409**
	NMI	0.1153	0.1167	0.1387	0.2463	**0.2881**
	Precision	0.4386	0.4426	0.4973	0.5029	**0.5349**
NW	Purity	0.4923	0.4879	0.4988	0.5092	**0.5226**
	NMI	0.0849	0.0821	0.1301	0.1432	**0.1625**
	Precision	0.3466	0.3353	0.3547	0.4082	**0.4452**

TABLE III: The results of MPNMF based on different similarity semantics (AS : Attribute based Similarity)

data set	Metrics	AS	AS+originalrelation	AS+PathSim	AS+HeteSim	g-PathSim
EC	Purity	0.6558	0.6742	0.7654	0.7731	**0.7910**
	NMI	0.2154	0.2459	0.5246	0.5093	**0.5318**
	Precision	0.4654	0.4983	0.6763	0.6679	**0.6801**
NW	Purity	0.5342	0.5488	0.5474	0.5652	**0.5756**
	NMI	0.1329	0.1787	0.1857	0.1973	**0.2156**
	Precision	0.3866	0.3895	0.4042	0.4128	**0.4373**

TABLE IV: The results of the compared methods

data set	Metrics	K-means	SVD-NMF	ONMTF	GNMF	DRCC	LP-FNMTF	Co-reguSC	ColNMF	MultiNMF	MPNMF
EC	Purity	0.5061	0.6697	0.6490	0.6429	0.6543	0.6080	0.7258	0.6952	0.6837	**0.7910**
	NMI	0.1345	0.2130	0.1861	0.1731	0.2359	0.1153	0.4223	0.3519	0.3155	**0.5318**
	Precision	0.3810	0.5324	0.5034	0.4791	0.4683	0.4386	0.6014	0.5824	0.5489	**0.6801**
OS	Purity	0.4841	0.5494	0.5428	0.6112	0.5550	0.5795	0.6305	0.6159	0.6325	**0.6490**
	NMI	0.1148	0.1212	0.1433	0.1690	0.1426	0.1519	0.2215	0.1820	0.2137	**0.3059**
	Precision	0.3105	0.3837	0.3981	0.4439	0.3847	0.3995	0.4078	0.4224	0.4569	**0.4988**
NW	Purity	0.5252	0.5405	0.5716	0.5635	0.5383	0.4923	0.5579	0.5382	0.5729	**0.5756**
	NMI	0.1038	0.1222	0.1391	0.1793	0.1338	0.0849	0.1838	0.1425	0.2058	**0.2156**
	Precision	0.3448	0.3811	0.3921	0.4338	0.3557	0.3466	0.3922	0.4016	0.4153	**0.4373**
DB	Purity	0.3015	0.3133	0.3368	0.3498	0.3172	0.3093	0.3718	0.3425	0.3647	**0.4412**
	NMI	0.1028	0.0854	0.0854	0.1081	0.0848	0.0718	0.1925	0.1153	0.1588	**0.2321**
	Precision	0.1671	0.1906	0.2028	0.2166	0.1917	0.1968	0.2349	0.2105	0.2292	**0.2713**
Cora2	Purity	0.3604	0.3962	0.4251	0.4782	0.3040	0.4106	0.5637	0.5516	0.5128	**0.6026**
	NMI	0.1400	0.1392	0.1665	0.2354	0.0240	0.1560	0.3750	0.3812	0.2983	**0.4013**
	Precision	0.2287	0.2764	0.3052	0.2660	0.1806	0.1965	0.5423	0.5238	0.4395	**0.7456**
Digg	Purity	0.3900	0.4520	0.4679	0.5240	0.5180	0.4700	0.4976	0.4185	0.5520	**0.5800**
	NMI	0.3706	0.2126	0.2289	0.3352	0.3154	0.2587	0.3257	0.2315	0.3825	**0.4274**
	Precision	0.2571	0.3453	0.3353	0.3937	0.2990	0.3523	0.3582	0.3329	0.4210	**0.4640**

Table III. The results show that g-PathSim exceeds PathSim and HeteSim in expressing the similarity semantic information for relational data cltering, and verify our assumption that the integration of multiply-mined relation information with the content information will improve the clustering performance by enhancing the credibility of the adjacency matrix.

At the same time, a more careful examination on the results of MPNMF in TableIII and LP-FNMTF in TableII via item by item comparison, shows that our NMF objective function outperforms that of LP-FNMTF. It confirms that the double semantic integration provided by our NMF objective function can also advance the clustering performance.

The overall performance is presented in Table IV. It is evident that, with the results in Table II, MPNMF outperforms the other compared methods in both homogeneous information network (Cora2) and heterogeneous information networks(Cora1, Digg), sometimes very significantly, which demonstrate the advantage of our method in terms of clustering performance.

On dataset Cora1 and Cora2, the multi-view clustering methods perform better than the attribute-view methods, but on dataset Digg the situation reverses as the multi-view clustering method Co-reguSC and ColNMF are outperformed by some of the attribute-view methods. One of the most possible reasons is that there are five types of relations in Digg and some of them may adversely affect clustering while these multi-view clustering methods indiscriminately utilize information of relations. However, also as a multi-view clustering method, MPNMF weakens the negative effect of those disadvantageous relations by integrating relations with g-PathSim.

VI. Conclusion

In this paper, we propose a novel MPNMF framework which advances the multi-type relational data clustering by enriching potentially useful similarity semantics for the improved clustering performance. We integrate the meta-path based similarity semantics into a joint similarity matrix, and then make the most of the joint similarity matrix together with the attributes in bag-of-word form via NMF. The contributions of this paper are as follows: First, we propose a novel framework which incorporates the meta-paths into NMF, and provides an effective solution for clustering on arbitrary multi-type relational data. Second, we design a general meta-path based similarity measure to characterize the strength of pairwise interactions between the objects of the target type. Third, the proposed NMF objective function of the framework provides the double semantic integration in another sense, which also advances the clustering performance. Finally, experiments compared with existing state-of-art NMF based approaches demonstrate the effectiveness of our method.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments. This research work was funded by the National Natural Science Foundation of China under Grant No. 61303179.

References

[1] Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram and A. Kelliher, *Metafac: community discovery via relational hypergraph factorization*, ACM SIGKDD, 15 (2009), pp. 527–536.

[2] A. McCallum, K. Nigam, J. Rennie, and K.Seymore, *Automating the construction of internet portals with machine learning*,Kluwer Academic Publishers Hingham, Inf. Retr., Volume 3 Issue 2, (2000), pp. 127–163.

[3] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery, *Learning to extract symbolic knowledge from the world wide web*, In AAAI/IAAI, (1998), pp. 509–516.

[4] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu., *PathSim : Meta path-based top-k similarity search in heterogeneous information networks*, VLDB, (2011).

[5] G. Jeh and J. Widom, *Scaling personalized web search*, WWW, (2003), pp. 271–279.

[6] G. Jeh and J. Widom, *Simrank: a measure of structural-context similarity*, KDD, (2002), pp. 538–543.

[7] C. Shi, X. Kong, Y. Huang, P. S. Yu, and B. Wu, *HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks*, IEEE TKDE, DOI 10.1109/TKDE.2013.2297920,2013.

[8] D. Cai, X. He, J. Han, and T.S. Huang, *Graph regularized nonnegative matrix factorization for data representation*, IEEE TPAMI, 33(8):1548–1560, 2011.

[9] C. Ding, T. Li,W. Peng, and H. Park, *Orthogonal nonnegative matrix tri-factorizations for clustering*, SIGKDD, 2006.

[10] Q. Gu and J. Zhou, *Co-clustering on manifolds*, SIGKDD, 2009.

[11] D. Seung and L. Lee, *Algorithms for non-negative matrix factorization*, NIPS, 2001.

[12] C. Ding, X. He, and H.D. Simon, *On the equivalence of nonnegative matrix factorization and spectral clustering*, SDM, 2005.

[13] C.J. van Rijsbergen, *INFORMATION RETRIEVAL*, Second Edition, Butterworths, London, 1979.

[14] C. Ding, T. Li, and W. Peng, *Nonnegative matrix tri-factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method*, AAAI, 2006.

[15] C. Boutsidisa, and E. Gallopoulosb, *SVD based initialization: A head start for nonnegative matrix factorization*, Pattern Recognition,41(4):1350–1362, 2008.

[16] C. Ding, T. Li, and M.I. Jordan, *Convex and semi-nonnegative matrix factorizations*, IEEE TPAMI,32(1):45–55, 2010.

[17] F. Shang, L.C. Jiao, and F.Wang, *Graph dual regularization nonnegative matrix factorization for co-clustering*, Pattern Recognition, 2012.

[18] H.Wang, F. Nie, H. Huang, and F. Makedon, *Fast nonnegative matrix tri-factorization for large-scale data co-clustering*, IJCAI, 2011.

[19] M. Nickel, V. Tresp, and H. P. Kriegel, *A three-way model for collective learning on multi-relational data*, ICML, 2011.

[20] A. Kumar, P. Rai and H. Daum'e, *Co-regularized multi-view spectral clustering*, NIPS, 2011.

[21] A. Singh and G. Gordon, *Relational learning via collective matrix factorization*, KDD, 2008.

[22] M. Sachan and S. Srivastava, *Collective Matrix Factorization for Co-clustering*, WWW Companion,2013.

[23] J. Liu, C. Wang, J. Gao and J. Han, *Multi-view clustering via joint nonnegative matrix factorization*, SDM, 2013.