Heterogeneous Information Networks Bi-clustering with Similarity Regularization

Xianchao Zhang, Haixin Li, Wenxin Liang(✉), Linlin Zong, and Xinyue Liu

Dalian University of Technology, Dalian, China
{xczhang,wxliang,xyliu}@dlut.edu.cn,
{lihaixin,linlinzong}@mail.dlut.edu.cn

Abstract. Clustering analysis of multi-typed objects in heterogeneous information network (HINs) is an important and challenging problem. Nonnegative Matrix Tri-Factorization (NMTF) is a popular bi-clustering algorithm on document data and relational data. However, few algorithms utilize this method for clustering in HINs. In this paper, we propose a novel bi-clustering algorithm, BMFClus, for HIN based on NMTF. BMFClus not only simultaneously generates clusters for two types of objects but also takes rich heterogeneous information into account by using a similarity regularization. Experiments on both synthetic and real-world datasets demonstrate that BMFClus outperforms the state-of-the-art methods.

Keywords: Heterogeneous information network · Nonnegative Matrix Tri-Factorization · Clustering

1 Introduction

Many real world data can be modeled using heterogeneous information networks (HINs) which consists of multiple types of objects. For example, a heterogeneous bibliographic network in Fig. 1(a) contains multi-typed objects including *authors*, *venues* (conferences or journals) and *terms*. A heterogeneous servers network in Fig. 1(d) contains *switches*, *email servers*, *database servers* and *web servers*. Clustering in HINs has attracted increasing attention in recent years. For instance, [1] finds that clustering analysis in heterogeneous bibliographic network helps generate more effective and comprehensive ranking result for *authors* and *venues*. In a heterogeneous servers network, if an attack script runs on some compromised web servers and the script reads data from database servers and sends out spam emails through the email servers, we call these servers compose an "attack sub-network". [2] proposes a framework to find such "attack sub-networks" with the help of clustering in HINs.

However, most existing studies on HIN clustering have some limitations. Many studies [3–5] deal with HINs as homogeneous networks, i.e., a network

This work was supported by National Science Foundation of China (No. 61272374,61300190) and 863 Project (No. 2015AA015463).

M. Chau et al. (Eds.): PAISI 2016, LNCS 9650, pp. 19–30, 2016.
DOI: 10.1007/978-3-319-31863-9_2

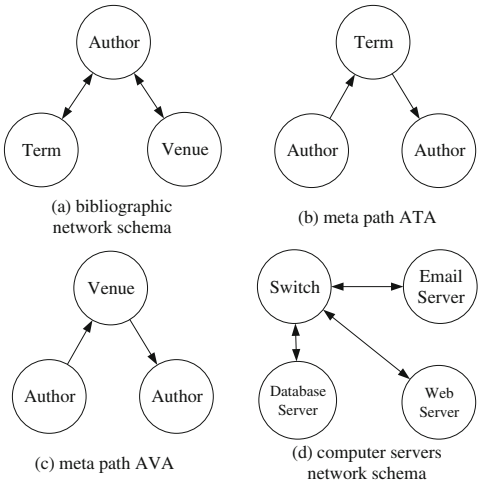

Fig. 1. Example of network schemas and meta paths

consisting of single type of objects. All the types (author type, venue type or term type) are treated in the same way in these algorithms. Therefore, these algorithms fail to use rich heterogeneous information during the clustering process. Some studies, RankClus [1] and GenClus [6] for example, distinguish different types in HINs. They assign a certain type in HIN as target type and other types as attribute types. Then, they focus on clustering target type and only generate clusters for that type of objects. HIN clustering aims at finding K partitions for multi-typed objects, so that objects in the same partition should be more similar (have more connections) to each other than to those in other partitions. For example, in a heterogeneous bibliographic network, we can tell which venues belong to "Information Security" and which venues belong to "Information Retrieval" by using clustering analysis in venues. While, if we are interested in which authors are authorities in "Information Security" and which authors are authorities in "Information Retrieval", we have to cluster the authors. Therefore, a better way to analysis heterogeneous bibliographic networks is to generate clusters for venues and authors simultaneously. Unlike RankClus [1] and GenClus [6], we try to find partitions for more than one type of objects. In this case, bi-clustering is a feasible technique because it generates clusters for two types of objects. In recent years, bi-clustering based on Nonnegative Matrix Tri-Factorization (NMTF) [7,8] attracts increasing attention because of their mathematical elegance and encouraging empirical results on document data and relational data. Nevertheless, few algorithms utilize NMTF for the HIN clustering. The main challenge of applying NMTF to HIN is how to incorporate rich heterogeneous information into the clustering process.

To address the problem, in this paper we propose a novel bi-clustering algorithm, BMFClus (HIN Bi-Clustering based on Matrix tri-Factorization), which simultaneously generates clusters for two types of objects, and takes advantage

of rich heterogeneous information during the clustering process. To achieve this goal, the NMTF is adopted as a basic bi-clustering method of BMFClus to cluster two types of objects in HIN. Furthermore, a similarity regularization term is introduced to the objective function of NMTF. The similarity regularization term enables BMFClus to utilize rich heterogeneous information in HINs, which leads to an improvement over the basic NMTF method. Our contributions are summarized as follows:

1. We propose a bi-clustering algorithm for HINs based on NMTF.
2. We incorporate rich heterogeneous information into the bi-clustering process by a similarity regularization term.
3. Experiments on both synthetic and real-world datasets demonstrate the effectiveness of the proposed algorithm in comparison with the state-of-the-art algorithms.

The rest of this paper is organized as follows: Sect. 2 introduces the problem statement and related work. Section 3 describes the details of the proposed algorithm. Section 4 reports the performance of the proposed algorithm comparing with the state-of-the-art algorithms. Finally, we concludes the paper and outlines the future work.

2 Problem Statement and Related Work

A graph $G = (V, E)$, where $V = \bigcup_{i=1}^{t} X_i$, and $X_1 = \{x_{11}, ..., x_{1n_1}\}, ..., X_t = \{x_{t1}, ..., x_{tn_t}\}$ denote the t different types of nodes. E is the set of links between any two data objects in V. If $t = 1$, G is a **homogeneous information network**. If $t > 1$, G is a **heterogeneous information network**.

As described in [9], a graph $S = (T, R)$ is called **network schema**, if S is an undirected connected graph defined over object types T, with edges as relations from R. A network schema provides a meta structure description of a heterogeneous information network.

For example, Fig. 1 (a) is a network schema of a HIN, specifically, a heterogeneous bibliographic network. It contains three types of objects including *authors*, *venues* and *terms*. For this HIN, $T = \{Author(A), Venue(V), Term(T)\}$, $R = \{A - V, V - A, A - T, T - A\}$,$t = 3$, X_1 denotes the objects of author type, X_2 denotes the objects of venue type, X_3 denotes the objects of term type. We define the **meta path** as a path in network schema which connects two types, following the definition of meta path in [9]. In Fig. 1, (b) and (c) are two meta paths selected from (a). Meta path (b) composed by relations $A - T$ and $T - A$, and is denoted as ATA. ATA encodes the semantic that whether two authors are interested in the same term, e.g. both two authors like "kmeans". Meta path (c), i.e. AVA, composed by relations $A - V$ and $V - A$, denotes the semantic that whether two authors are interested in the same venue, e.g. two co-authors publish a paper in "SIGKDD".

Given a meta path, we can use **PathCount** to measure the similarity between a pair of objects [9]. The PathCount of x_{1i}, x_{1j} is the number of path from x_{1i}

to x_{1j} following a certain meta path. For instance, in Fig. 1, two authors can be connected via "author-term-author" (ATA) path if they use a same term in their papers, and $PathCount(x_{1i}, x_{1j})$ under path ATA is the number of common terms used by author x_{1i} and x_{1j}. Meta path "author-venue-author" (AVA) denotes a relation between authors via venues (i.e., publishing in the same venues), and $PathCount(x_{1i}, x_{1j})$ under path AVA is the number of common venues attended by author x_{1i} and x_{1j}. Given a meta path, the higher value of $PathCount(x_{1i}, x_{1j})$, x_{1i} and x_{1j} are considered to be more similar. Since meta path encodes the relationship between different types, it captures rich heterogeneous information of a HIN [9–11].

Several approaches have been proposed to find K partitions for the multi-typed objects in a HIN. SpectralBiclustering [12] is proposed to bi-clustering two types of objects using spectral clustering. RankClus [1] combines ranking and clustering techniques to analysis two types of objects in HIN. PathSelClus [11] utilizes rich heterogeneous information encoded by meta path. While, PathSelClus only generates cluster for a single type in HIN. Our work is different from theirs, as we focus on simultaneously generating clusters for two types of objects. In addition, we also propose how to incorporate rich heterogeneous information into the NMTF clustering process.

3 Proposed Algorithm

3.1 NMTF

We give a brief review of Nonnegative Matrix Tri-Factorization (NMTF) [8] which is an effective bi-clustering method. Given a data matrix $M \in \mathbb{R}_+^{m \times n}$, the objective function of the NMTF is

$$\min_{F \geq 0, G \geq 0, S \geq 0} \left\| M - FSG^T \right\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm, $F \in \mathbb{R}_+^{m \times k}$, $S \in \mathbb{R}_+^{k \times k}$ and $G \in \mathbb{R}_+^{n \times k}$. S provides additional degrees of freedom such that the low-rank matrix representation remains accurate, while F gives m row cluster assignment vectors and G gives n column cluster assignment vectors. Equation (1) can be computed using the following update rules [8].

$$G_{jk} \leftarrow G_{jk} \sqrt{\frac{(M^T FS)_{jk}}{(GG^T M^T FS)_{jk}}}, \tag{2}$$

$$F_{ik} \leftarrow F_{ik} \sqrt{\frac{(MGS^T)_{ik}}{(FF^T MGS^T)_{ik}}}, \tag{3}$$

$$S_{ik} \leftarrow S_{ik} \sqrt{\frac{(F^T MG)_{ik}}{(F^T FSG^T G)_{ik}}}. \tag{4}$$

Although the objective function in Eq. (1) is not convex in all variables together, it is proved that the above update rules will find a local minimum of Eq. (1). Using NMTF for data clustering has following merits:

1. We can obtain the clusters of rows and columns simultaneously of a data matrix. Actually, it is also proved that NMTF is equivalent to do kernel K-means clustering on both columns and rows [8].
2. NMTF conducts a knowledge transformation between the row feature space and the column feature space [13]. It means that the quality of the row clustering and the column clustering are mutually enhanced during the update iteration.

The performance of NMTF in document data or relation data has been well studied. However, to the best of our knowledge, we are the first to apply NMTF to HIN clustering.

3.2 BMFClus

Although NMTF can be used to generate clusters for two types on a HIN, it fails to take advantage of rich heterogeneous information captured by meta path. Therefore, we propose BMFClus which not only inherits advantages of NMTF, but also takes into account rich heterogeneous information of HIN.

First, we use NMTF to model a HIN. Two types are selected from T. Then, a nonnegative edges weight matrix M is constructed, where $M_{i,j}$ is the number of the links between two nodes. For example, in Fig. 1, we choose *author* type and *venue* type. The $M_{i,j}$ denotes how many papers of author i published by venue j. If the topic of venue j is "data mining", and the value of $M_{i,j}$ is high, then author i has a high probability to be labeled as "data mining", and vice versa. According to Eq. (1), F gives the cluster assignment vectors of authors and G gives the cluster assignment vectors of venues. If $k = 3$ and $F_{i,*} = [0.9, 0.1, 0]$, then author i will be assigned to cluster 1. If $G_{*,j} = [0.05, 0.95, 0.05]^T$, then venue j will be assigned to cluster 2.

Next, we describe how to incorporate rich heterogeneous information into NMTF. As mentioned before, meta paths encode rich heterogeneous information of HIN. Given a meta path, a nonnegative similarity matrix is constructed using PathCount. For example, in Fig. 1, if we use meta path $p = ATA$ (author-term-author), a nonnegative similarity matrix $W^{(F)}$ between each authors can be constructed using $W^{(F)}_{i,j} = PathCount(x_{1i}, x_{1j})$. And $W^{(F)}_{i,j}$ is the number of path from object x_{1i} to object x_{1j} following p.

Our goal is to encourage two objects (x_{1i} and x_{1j}) who have a high similarity ($W^{(F)}_{i,j}$) to have similar cluster assignment vectors ($F_i \approx F_j$). To achieve this goal, we introduce the following similarity regularization term.

$$O_1 = \frac{1}{2} \sum_{i,j} \|F_i - F_j\|_2^2 W^{(F)}_{i,j}. \tag{5}$$

The regularization term Eq. (5) is a cost function. It is obvious that if x_{1i} and x_{1j} have a high similarity value with respect to $W_{i,j}^{(F)}$, we should make $\|F_i - F_j\|_2^2$ small to reduce the punishment of Eq. (5). Minimizing O_1 will smooth the cluster distributions between a object and its similar objects. Define diagonal matrix $D_{i,i}^{(F)} = \sum_j W_{i,j}^{(F)}$. Then, we construct the consistent Laplacian matrix $L_F = D^{(F)} - W^{(F)}$. Now, we rewrite the regularization term into trace form:

$$\begin{aligned} O_1 &= \frac{1}{2}\sum_{i,j} \|F_i - F_j\|_2^2 W_{i,j}^{(F)} \\ &= \sum_i F_i D_{i,j}^{(F)} F_i{}^T - \sum_{i,j} F_i W_{i,j}^{(F)} F_j{}^T \\ &= Tr(F^T L_F F). \end{aligned} \tag{6}$$

Similar with the construction of O_1, we construct another objective function for venue type:

$$O_2 = Tr(G^T L_G G). \tag{7}$$

Now, we define our **BMFClus** by adding the regularization terms Eqs. (6) and (7) to Eq. (1):

$$\min_{F\geq 0, G\geq 0, S\geq 0} \left\|M - FSG^T\right\|_F^2 + \lambda(Tr(F^T L_F F) + Tr(G^T L_G G)), \tag{8}$$

where the first term represents the reconstruction error for nonnegative edges weight matrix M, and the second term represents the similarity regularization. λ is a trade-off parameter. This parameter is not much sensitive and we set it to be 0.1 in our experiments.

Algorithm 1. *BMFClus*

Require: M, $W^{(F)}$ and $W^{(G)}$ constructed using HIN,
trade-off parameters λ
number of clusters K
Ensure: the cluster assignment vectors F and G
1: Normalize M, $W^{(F)}$ and $W^{(G)}$;
2: Initialize F, S and G ;
3: **repeat**
4: Fixing other factors, update S by Eq. (11);
5: Fixing other factors, update F by Eq. (10);
6: Fixing other factors, update G by Eq. (9);
7: **until** Eq. (8) converges.

BMFClus can be computed using the following update rules [14]:

$$G_{jk} \leftarrow G_{jk}\sqrt{\frac{[\lambda L_G^- G + A^+ + GB^-]_{jk}}{[\lambda L_G^+ G + A^- + GB^+]_{jk}}}, \tag{9}$$

$$F_{ik} \leftarrow F_{ik}\sqrt{\frac{[\lambda L_F^- F + P^+ + FQ^-]_{ik}}{[\lambda L_F^+ F + P^- + FQ^+]_{ik}}}, \tag{10}$$

$$S = (F^T F)^{-1} F^T M G (G^T G)^{-1}, \tag{11}$$

where $L_G = L_G^+ - L_G^-$, $A = M^T FS = A^+ - A^-$, $B = S^T F^T FS = B^+ - B^-$, $P = MGS^T$, $Q = SG^T GS^T$, $A_{ij}^+ = (|A_{ij}| + A_{ij})/2$, $A_{ij}^- = (|A_{ij}| - A_{ij})/2$ [15]. The detailed process of the BMFClus is summarized in Algorithm 1.

4 Experiments

In this section, we conduct a series of experiments to show the effectiveness of BMFClus on both synthetic and real datasets.

4.1 Dataset

We give a brief description of the datasets used in our experiments as follows:

1. **SynData**: We generate a synthetic HIN following the properties of real word HIN. A HIN is composed with several bipartite networks. We apply the method described in [1] to generate 3 synthetic bipartite networks and construct a synthetic HIN, SynData. SynData contains 3 clusters and 3 types, denoted as A, B and C. The number of objects: $N_a = \{1000, 1200, 1300\}$ for type A, $N_b = \{3000, 3200, 3500\}$ for type B, $N_c = \{1000, 1200, 1300\}$ for type C.
2. **DBLP4** [16] is a real-world HIN extracted from DBLP bibliography dataset in four research areas: database (DB), data mining (DM), information retrieval (IR) and artificial intelligence (AI). DBLP4 contains 3 types of objects: 4236 authors, 20 venues and 11771 unique terms. Each author object links with several venues and terms. The link weight of author-venue pair is the number of papers the author publishes in the venue. The author-term sub-network contains all the terms appeared in the abstract of papers of each author with stopwords removed. All the venue objects and author objects are labeled.
3. **Flickr** [9,17]: Flickr is a HIN contains three types of objects: image, user and tag. Each image object links with several tags and one user. Image objects are labeled.

The statistics of the datasets are summarized in Table 1. The network schemas of the datasets are shown in Fig. 2.

Table 1. Statistics of the datasets

dataset	# object	# type	# link	# cluster
SynData	16700	3	135000	3
DBLP4	16027	3	735710	4
Flickr	4076	3	14396	8

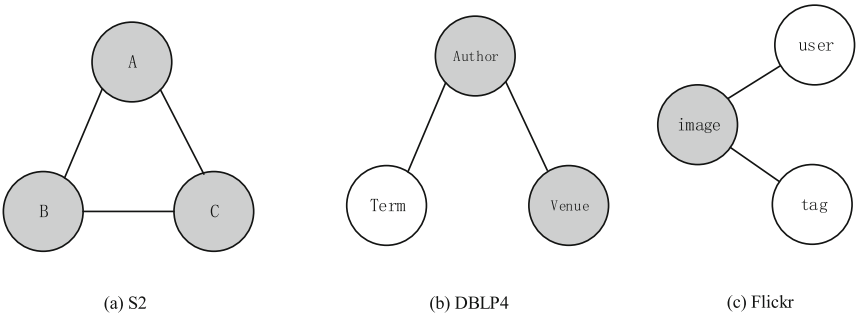

Fig. 2. Network schemas of the datasets. (a):SynData; (b):DBLP4; (c): Flickr. The labeled object types are in grey.

4.2 Baselines

We compare the proposed method with the following state-of-the-art algorithms:

- **SpectralBiclustering (SBC)** [12] is a well known spectral clustering based bi-clustering algorithm. Give a $m \times n$ matrix, SBC generates clusters for m rows and n columns.
- **GreedyCoClustering** [9] is an information-theoretic greedy bi-clustering algorithm. It use a greedy KL-divergence based bi-clustering method to cluster a $m \times n$ matrix.
- **NMTF** [8] is the basic form of the proposed method.
- **RankClus** [1] is a rank-based algorithm which integrates ranking and clustering together for a heterogeneous bibliographic network. RankClus treats a $m \times n$ matrix as a bipartite graph. RankClus generates clusters for m rows, and applies ranking for n columns. Then, it generates a better cluster structure for m rows based on the ranking distribution on n columns. After several this iteration, quality of clustering and ranking are mutually enhanced.

4.3 Evaluation Metrics

The clustering performance is evaluated by comparing the ground truth labels with the predicted labels. Two popular metrics, i.e., *accuracy* (ACC) and *normalized mutual information* (NMI), are used to measure the clustering performance [1,18,19].

Given an object v_i of a certain type T_a $(1 \leq a \leq t)$, let c_i and r_i be the predicted label and the ground truth label of v_i, respectively. The ACC of type T_a is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n_a} \delta\left(c_i, map\left(r_i\right)\right)}{n_a}. \tag{12}$$

where $\delta(x, y)$ equals one if $x = y$ and equals zero otherwise. $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the ground truth labels. Kuhn-Munkres algorithm [20] is used for finding the best mapping.

Given the clustering result of type T_a, let $n(i, j), i, j = 1, 2, ..., K$, denote the number of objects that predicted as label i and labeled as j in the ground truth. From $n(i, j)$, we define joint distribution $p(i, j) = \frac{n(i,j)}{n_a}$, row distribution $p_1(j) = \sum_{i=1}^{K} p(i, j)$ and column distribution $p_2(i) = \sum_{j=1}^{K} p(i, j)$. The NMI of type T_a is defined as follows:

$$NMI = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} p(i,j)\log(\frac{p(i,j)}{p_1(j)p_2(i)})}{\sqrt{\sum_{j=1}^{K} p_1(j)\log p_1(j)\sum_{i=1}^{K} p_2(i)\log p_2(i)}}. \tag{13}$$

NMI dose not require the mapping function between the predicted labels and ground truth labels.

Both metrics are in the range from 0 to 1 and a higher value indicates a better clustering performance in terms of the ground truth labels.

4.4 Settings

For DBLP4 dataset, *author* type and *venue* type are selected as the two types of objects we want to cluster. And the nonnegative edges weight matrix M is constructed by the link count between two objects. L_F and L_G is constructed using meta path ATA (author-term-author) and VAV (venue-author-venue), respectively. Meta path ATA means two authors share more common terms are more similar and meta path VAV means two venues share more common authors are more similar. For Flickr dataset, we choose *image* type and *tag* type as the types of objects we want to cluster. M is constructed by the links between image objects and tag objects. Since only image objects are labeled, the clustering performance are evaluated on image type. L_F and L_G is constructed using meta path IUI (image-user-image) and TIT (tag-image-tag), respectively. Meta path IUI means two images provided by same user are similar, and meta path TIT means two tags share more common images are more similar. For SynData dataset, we choose A type and B type as the clustering target. M is constructed by the links between A and B. L_F and L_G is constructed using meta path ACA and BCB, respectively. M also serves as the input data for SBC [12], GreedyCoClustering [9], NMTF [8] and RankClus [1]. Each result is the average of 10 runs.

4.5 Results

The results are shown in Table 2 and the best results are highlighted in boldface. On the DBLP4 and SynData dataset, as this dataset is nicely structured, all methods achieve outstanding performance. NMTF outperforms the other

Table 2. Cluster performance of different methods.

(a) ACC

dataset	DBLP4		Flickr	SynData	
type	author	venue	image	A	B
SBC [12]	0.8780	0.8500	**0.4669**	0.8803	0.7185
GreedyCoClustering [9]	0.7652	0.7668	0.3597	0.8207	0.6621
NMTF [8]	0.9308	0.9500	0.4205	0.9222	0.7850
RankClus [1]	0.6822	0.8500	0.4161	0.9316	0.7731
BMFClus	**0.9327**	**1.0000**	0.4317	**0.9438**	**0.8153**

(b) NMI

dataset	DBLP4		Flickr	SynData	
type	author	venue	image	A	B
SBC [12]	0.7279	0.8388	0.3997	0.7024	0.3518
GreedyCoClustering [9]	0.5812	0.6628	0.2419	0.6557	0.3021
NMTF [8]	0.7834	0.9058	0.4131	0.7328	0.3924
RankClus [1]	0.5931	0.8338	0.3934	0.7191	0.3889
BMFClus	**0.7901**	**1.0000**	**0.4242**	**0.7618**	**0.4137**

baselines on DBLP4 and SynData dataset. As expected, due to the rich heterogeneous information captured by similarity regularization term, BMFClus performs much better than NMTF on author type and venue type with respect to NMI and ACC. On the Flickr dataset, we observe that SBC [12] achieves the best accuracy on image type. While, BMFClus outperforms SBC with respect to NMI. BMFClus achieves the best results on DBLP4 and SynData dataset, and the second best results on Flickr dataset. Overall we conclude that BMFClus outperforms the base line methods.

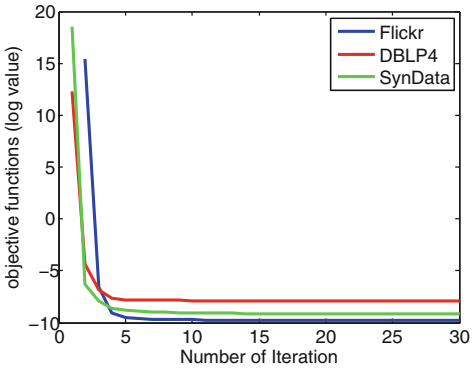

Fig. 3. Convergence of BMFClus

4.6 Algorithm Convergence

The update rules for minimizing the objective functions of BMFClus are essentially iterative We investigate the convergence of BMFClus. Figure 3 shows the convergence curve of the objective functions (in log values) on three datasets.

It is easy to see that the objective values of BMFClus falling fast at the first several iterations on each datasets.

5 Conclusion and Future Work

In this paper, we propose a bi-clustering algorithm (BMFClus) for HINs based on NMTF. Specifically, given a HIN, BMFClus simultaneously generates clusters for two types of objects. Besides, BMFClus takes rich heterogeneous information into account by using a similarity regularization. Experiments on both synthetic and real-world datasets demonstrate that BMFClus outperforms the state-of-the-art methods. For the future work, we will investigate how to extend BMFClus to arbitrary multi-typed heterogeneous information networks.

References

1. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 565–576. ACM (2009)
2. Gupta, M., Gao, J., Yan, X., Cam, H., Han, J.: Top-k interesting subgraph discovery in information networks. In: 2014 IEEE 30th International Conference on Data Engineering (ICDE), pp. 820–831. IEEE (2014)
3. Wang, N., Parthasarathy, S., Tan, K.-L., Tung, A.K.: Csv: visualizing and mining cohesive subgraphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 445–458. ACM (2008)
4. White, S., Smyth, P.: A spectral clustering approach to finding communities in graph. In: SDM, vol. 5, pp. 76–84. SIAM (2005)
5. Liu, X., Yu, S., Janssens, F., Glänzel, W., Moreau, Y., De Moor, B.: Weighted hybrid clustering by combining text mining and bibliometrics on a large-scale journal database. J. Am. Soc. Inform. Sci. Technol. **61**(6), 1105–1119 (2010)
6. Sun, Y., Aggarwal, C.C., Han, J.: Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. Proc. VLDB Endowment **5**(5), 394–405 (2012)
7. Pei, Y., Chakraborty, N., Sycara, K.: onnegative matrix tri-factorization with graph regularization for community detection in social networks. In: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI 2015, pp. 2083–2089. AAAI Press (2015)
8. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 126–135. ACM (2006)
9. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: VLDB 2011 (2011)

10. Sun, Y., Norick, B., Han, J., Yan, X., Yu, P.S., Yu, X.: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1348–1356. ACM (2012)
11. Yu, X., Sun, Y., Norick, B., Mao, T., Han, J.: User guided entity similarity search using meta-path selection in heterogeneous information networks. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2025–2029. ACM (2012)
12. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning: In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 269–274. ACM (2001)
13. Li, T., Ding, C., Zhang, Y., Shao, B.: Knowledge transformation from word space to document space. In: Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 187–194. ACM (2008)
14. Gu, Q., Zhou, J.: Co-clustering on manifolds. In: Proceedings of the 15th ACM SIGKDD International Cconference on Knowledge Discovery and Data Mining, pp. 359–368. ACM (2009)
15. Ding, C., Li, T., Jordan, M., et al.: Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 45–55 (2010)
16. Liu, J., Han, J.: Hinmf: A matrix factorization method for clustering in heterogeneous information networks. In: Proceedings of 2013 IJCAI Workshop on Heterogeneous Information Network Analysis (2013)
17. Liu, J., Wang, C., Gao, J., Gu, Q., Aggarwal, C., Kaplan, L., Han, J.: Gin: a clustering model for capturing dual heterogeneity in networked data. In: Proceedings of 2015 SIAM International Conference on Data Mining (2015)
18. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273. ACM (2003)
19. Cai, D., He, X., Han, J., Member, S.: Document clustering using locality preserving indexing. IEEE Trans. Knowl. Data Eng. **17**, 1624–1637 (2005)
20. Lovsz, L., Plummer, M.: Matching Theory. Annals of Discrete Mathematics, vol. 29 inria-00345669, version 3 - 21 November 2009 (1986)