# Semi-supervised Meta-path-based Algorithm for Community Detection in Heterogeneous Information Networks

Limin Chen(✉), Yan Zhang, and Liu Yang

School of Computer and Information Technology,
Mudanjiang Normal University, Mudanjiang 157012, China
chenlimin_clm@126.com

**Abstract.** Similarity between target objects is mostly computed based on meta-paths for semantic-based community detection algorithm in heterogeneous information networks. The meta-path-based similarity between target objects is very efficient, but the semantics of meta-path-based similarity is not integrity and can not truly express relations of target objects. And now there is lack of better expression for similarity semantics in heterogeneous information networks. The complex topology structure is usually neglected in semantic-based community detection algorithms for heterogeneous information networks. To effectively improve accuracy of community detection algorithm in heterogeneous information networks, a semi-supervised meta-path-based algorithm for community detection is proposed in this paper. First, spectral method is used to analyse the topology structure of heterogeneous information networks to select representative objects. Then the similarity of target objects is adjusted by representatives in every cluster. Last, NMF method is used to detect communities. Through experiments in simulation datasets and real datasets, the experimental results showed the proposed algorithm is effective.

**Keywords:** Heterogeneous information networks · Community detection · Semi-supervised · Meta-path · Topology structure

## 1 Introduction

Heterogeneous information networks are very common. The real world is reflected truly by heterogeneous information networks. Analyzing heterogeneous information networks can better understand the hidden structure of networks and the roles represented by each community data [1, 2].

Community detection algorithms in heterogeneous information networks are the research foundation, and community detection algorithms based on semantic similarity in heterogeneous information networks are mainstream methods [3, 4]. Among them, most of the semantic similarity measures are based on meta-path computation. The typical meta-path-based similarity measures are PathCount, PathSim [5] and JoinSim [6]. However, the similarity semantic of target objects based on meta-path is incomplete, and it can not truly reflect the relations of target objects. Now, there is no more accurate method to express the similarity semantics of target objects. Affected by it, the


W. Ni et al. (Eds.): WISA 2019, LNCS 11817, pp. 506–511, 2019.
https://doi.org/10.1007/978-3-030-30952-7_50

accuracy of the results of community detection for heterogeneous information networks is not high. Semantic-based community detection algorithms for heterogeneous information networks often neglect the complex topology of networks, Analyzing heterogeneous information networks from the view of topological structure can grasp the global distribution of data [7, 8]. Semi-supervised method can improve the accuracy of community detection [9], so the accuracy of community detection can be effectively improved in heterogeneous information networks while similarity adjusted semi-supervised from global structure.

Spectral clustering can capture the global distribution of datasets. In addition, NMF (Non-negative Matrix Factorization) method and spectral clustering method have supported by reliable theory [9]. Integrating the two methods to analyze communities of heterogeneous information networks, SMpC (Semi-supervised meta-path-based algorithm for community detection in heterogeneous information networks) is proposed in this paper. First, the SRC algorithm [8] is used to obtain the global distribution of the target dataset, and the representative objects in each cluster are selected to construct the prior information. Then, the meta-path-based similarity of target objects is computed, and the similarity of target objects is adjusted by prior information. And last, the target objects are partitioned by NMF algorithm to find reasonable communities.

## 2 Similarity Based on Meta-path

**Definition 1:** Given a dataset $X = \{X_t\}_{t=1}^{T}$ with $T$ types, where $X_t = \left\{x_1^{(t)}, x_2^{(t)}, \ldots, x_{nt}^{(t)}\right\}$ is a dataset belonging to the $t$-th type, a weighted graph $G = <V, E, W>$ on $X$ is called an information network; if $V = X$, the $E$ is a binary relation on $V$ and $W$: $E \rightarrow R^{+}$. Such an information network is called a heterogeneous information network when $T > 1$ [10].

An information network $G = <V, E, W>$ on $X = \{X_t\}_{t=1}^{T}$ with $T$ types is given, $X_1$ is the target dataset and $X_t$ ($t \neq 1$) is the attribute dataset, $x_n^{(t)} \in X_t$ is the $n$-th object of $X_t$. The meta-path is a concatenation of multiple nodes or node types linked by edge types between object $x_u^{(1)}$ and $x_v^{(1)}$, denoted by $P^{<u,v>}$. Meta-path count $S$ from object $x_u^{(1)}$ to $x_v^{(1)}$ is the total number of meta-paths between $x_u^{(1)}$ and $x_v^{(1)}$. Let $P_s^{<u,v>}$ is the $s$-th meta-path between $x_u^{(1)}$ and $x_v^{(1)}$, the weight of $P_s^{<u,v>}$ is $\alpha_s$. The typical computing similarity method of target objects based on meta-path is as follows [5, 6]:

$$\begin{aligned}
\mathbf{PathCount}(u,v) &= \sum_s^S \alpha_s P_s^{<u,v>} \\
\mathbf{PathSim}(u,v) &= \sum_s^S \alpha_s \left(2P_s^{<u,v>} / \left(P_s^{<u,u>} + P_s^{<v,v>}\right)\right) \\
\mathbf{JoinSim}(u,v) &= \sum_s^S \alpha_s \left(2P_s^{<u,v>} / \left(P_s^{<u,u>} \cdot P_s^{<v,v>}\right)^{1/2}\right)
\end{aligned}$$

## 3 Multi-type Data Partitioning Based on Topological Structure

From the view of topological structure, SRC algorithm solves the problem of collaborative clustering heterogeneous data well. Given $G = <V, E, W>$ on $X = \{X_t\}_{t=1}^{T}$ with $T$ types and $M$ relation matrices, where $\left\{W^{(pq)} \in R^{n_p \times n_q}\right\}_{1 \le p,q \le T}$ is the relation matrix between $X_p$ and $X_q$. $\beta_{pq}$ is the weight of $W^{(pq)}$, where $\sum \beta_{pq} = 1, \beta_{pq} > 0$ . Let

$$L = \sum\nolimits_{1 \le p,q \le T} \beta_{pq} \left\| W^{(pq)} - C^{(p)} A^{(pq)} \left(C^{(q)}\right)' \right\| \tag{1}$$

where $\|\cdot\|$ refers Frobenius norm, $C^{(p)} \in \{0,1\}^{n_p \times k_p}$, $A^{(pq)} \in R^{k_p \times k_q}$, $\left(C^{(p)}\right)' C^{(p)} = I$ and $\sum_{i=1}^{k_i} C_{ij}^{(p)} = 1$, $I$ is unit matrix, $0 \le i \le n_i$. While $L$ is the smallest, $C$ is the best indicator matrix. Equation (1) can also be expressed as:

$$L = \sum\nolimits_{1 \le p,q \le M} \beta_{pq} \mathrm{tr} \left\{ \left( W^{(pq)} - C^{(p)} A^{(pq)} \left(C^{(q)}\right)' \right)' \left( W^{(pq)} - C^{(p)} A^{(pq)} \left(C^{(q)}\right)' \right) \right\} \tag{2}$$

where, tr denotes the trace of matrix. While $\partial L / \partial A^{(pq)} = 0$, $L$ is minimum. i.e.

$$\max_{(C(p))' C(p) = I} \mathrm{tr} \left( \left(C^{(p)}\right)' \Phi^{(p)} C^{(p)} \right) \tag{3}$$

where $\Phi^{(p)}$ is

$$\begin{aligned} \Phi^{(p)} = & \beta_{pp} \left( A^{(pp)} \left(A^{(pp)}\right)' \right) + \sum\nolimits_{p < i \le M} \beta_{pi} \left( A^{(pi)} C^{(i)} \left(C^{(i)}\right)' \left(A^{(pi)}\right)' \right) \\ & + \sum\nolimits_{1 < i \le p} \beta_{pi} \left( \left(A^{(ip)}\right)' C^{(i)} \left(C^{(i)}\right)' A^{(ip)} \right) \end{aligned} \tag{4}$$

The local maximum of Eq. (3) is easy obtained by iterative algorithm. First, given $M - 1$ indicator matrix $C^{(i)}$, then determining the optimal indicator matrix $C^{(p)}$, where $i \ne p, 0 \le i, p \le M$. The algorithm is as follows:

**Algorithm 1**: SRC for multi-type data

1) input matrices $\{W^{(pq)} \in R^{n_p \times n_q}\}_{1 \le p,q \le T}$, weight $\beta^{(pq)}$, cluster number $\{k_i\}_{1 \le i \le K}$;
2) initialize $\{C^{(p)}\}_{1 \le p \le T}$ with orthogonal normal matrices;
3) repeat
4) for p=1 to T do
5) { compute Eq.(4) $\Phi^{(p)}$; update $C^{(p)}$ with eigenvectors of $\Phi^{(p)}$;}
6) until convergence
7) partition indicator data $C^{(1)}$ of target objects by k-means algorithm.
8) output cluster indicator matrices $\{C^{(p)}\}_{1 \le p \le T}$.

# 4 Semi-supervised Meta-path-based Algorithm

## 4.1 Constructing Prior Information

An information network with a star network schema is selected to analyse the global distribution of target objects in this paper in order to reduce computation complexity. There only exist relation matrices $\{W^{(1q)} \in R^{n_1 \times n_q}\}_{1<q\le T}$ in a star network schema. Then Eq. (1) is expressed as:

$$L = \sum\nolimits_{1<q\le T} \beta_q \left\| W^{(1q)} - C^{(1)} A^{(1q)} \left( C^{(q)} \right)' \right\|$$

where $\beta_q$ is the weight of $W^{(1q)}$, $\sum \beta_q = 1$ and $\beta_q > 0$.

Target indicator matrix $C^{(1)}$ is got by partitioning target dataset $X_1$ into $K_1$ clusters using Algorithm 1. To select representative objects in the $k$-th cluster, where $1 \le k \le K_1$, given threshold $\delta$, first, select a seed $c_u^{(1)}$ in the $k$-th indicator cluster randomly, and $x_u^{(1)}$ is as the representative object corresponding to indicator object $c_u^{(1)}$. Then compute $dis = \left\| c_u^{(1)} - c_v^{(1)} \right\|$ between object $c_u^{(1)}$ and $c_v^{(1)}$ in the same indicator cluster. if $dis > \delta, c_v^{(1)}$ is selected as the next seed, and $x_v^{(1)}$ is as the representative object corresponding to $c_v^{(1)}$, repeat the step until no seed exists. Selecting the object whose distance from the seed is more than $\delta$ as the representative object, it can make representative objects associated in the topology structure to adjust the bias or incompleteness of the semantic similarity based on meta-path. Let $Z \in R^{n_1 \times n_1}$ is the priori information relation matrix of $X_1$. Given $z_{uv} \in Z, z_{uv} = 1$, if both $x_u^{(1)}$ and $x_v^{(1)}$ are represent objects and belong to the same cluster, otherwise $z_{uv} = 0$.

## 4.2 SMpC Algorithm

Semi-supervised meta-path-based algorithm for community detection proposed in this paper is as follow:

**Algorithm 2: SMpC** Algorithm

1) input cluster number $K_1$, threshold $\delta$, weight $a$, weight $b$;
2) compute the similarity matrix $H$ of target objects based on meta-path and regularized $H$;
3) partition $X_1$ into $K_1$ clusters by algorithm 1;
4) select representative objects, construct priori information matrix $Z \in R^{n_1 \times n_1}$ and regularized $Z$;
5) construct relation matrix $aH+bZ$;
6) decompose $aH+bZ$ by NMF method;
7) allocate target objects into clusters $\{Y_k\}_{1\le k\le K1}$ by indicator matrix;
8) output clusters $\{Y_k\}_{1\le k\le K1}$ of target objects.

# 5 Experiment

## 5.1 Experiment Data and Parameter Analysis

$S_s$ is the small test dataset as in the literature [10] and extracted from the DBLP dataset contains 4 areas related to data mining. $S_l$ is the large test dataset and extracted from the Chinese DBLP dataset, which are sharing resources released by Institute of automation, Chinese Academy of Sciences. $S_l$ includes 34 computer science journals, 2,671 papers, 4,576 authors and 4,962 terms.

The object similarity matrix $H$ is computed respectively by PathCount, PathSim and JoinSim in this experiment. $a$ is the parameter of similarity matrix $H$, and $b$ is the parameter of the priori information matrix $Z$, and $a + b = 1$. The dataset $S_s$ is used in this experiment to analyse parameters. $b$ ranges from 0.1 to 1, While $0.4 \leq b \leq 0.6$, the accuracy of communities is higher as shown in Fig. 1. $b = 0.5$ is used in the next experiment.

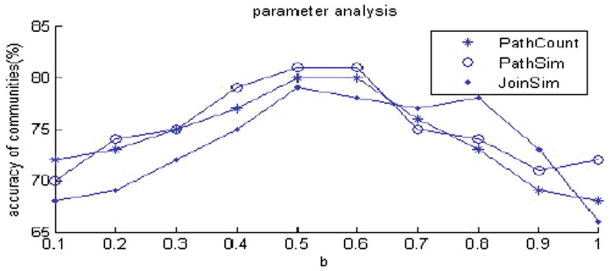


**Fig. 1.** Parameter analysis for papers in $S_s$

## 5.2 Accuracy Comparison of Community Detection

In this experiment, authors and papers as target objects are analyzed in datasets $S_s$ and $S_l$. Target object similarity is computed by using PathCount, PathSim, JoinSim respectively and all the similarity matrices are adjusted by SMpC. Communities is obtained in $S_s$ and $S_l$ by NMF method. Accuracy of communities by different methods is compared. The experimental results are shown in Table 1. It shows that SMpC can effectively improve the accuracy of communities in heterogeneous information networks.

**Table 1.** Comparision of community accuracy (%)

| Similarity | PathCount | SMpC | PathSim | SMpC | JoinSim | SMpC |
|---|---|---|---|---|---|---|
| Papers on $S_s$ | 73.91 | 80.87 | 71.54 | 81.72 | 72.83 | 79.65 |
| Authors on $S_s$ | 74.41 | 82.33 | 69.13 | 75.13 | 67.91 | 81.54 |
| Papers on $S_l$ | 70.84 | 76.36 | 68.28 | 72.89 | 72.93 | 79.92 |
| Authors on $S_l$ | 71.02 | 78.94 | 68.29 | 73.01 | 70.01 | 75.32 |

## 6 Conclusion

The incompleteness of semantic similarity is effectively adjusted by using priori information matrix. Therefore, SMpC algorithm in this paper can effectively improve the accuracy of communities in heterogeneous information networks, and SMpC algorithm does not need manual intervention, so that it improves the self-adaptability of algorithm. However, because of analysis of the topology of heterogeneous information networks, the computational complexity of the algorithm is increased. Reducing the complexity of the algorithm on ensuring the accuracy of communities will be solved in the future.

**Acknowledgments.** This paper is supported by Heilongjiang Natural Science Foundation: LH2019F051, F2016039; Science Project of Heilongjiang Education Department: 12521578.

## References

1. Sun, Y., Han, J.: Mining heterogeneous information networks: principles and methodologies. In: Proceedings of Mining Heterogeneous Information Networks: Principles and Methodologies, vol. 3, no. 2, pp. 1–159 (2012)
2. Li, Y., Li, C., Chen, W.: Research on influence ranking of chinese movie heterogeneous network based on PageRank algorithm. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 344–356. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_32
3. Shi, C., Li, Y., Zhang, J., et al.: A survey of heterogeneous information network analysis. IEEE Trans. Knowl. Data Eng. **29**(1), 17–37 (2016)
4. Sun, Y., Han, J., Yan, X., et al.: PathSim: meta path-based top-k similarity search in heterogeneous information networks. In: Proceedings of VLDB Endowment (2011)
5. Shi, Y., Chan, P.W., Zhuang, H., et al.: PReP: path-based relevance from a probabilistic perspective in heterogeneous information networks. In: KDD 2017, Canada, pp. 13–17 (2017)
6. Xiong, Y., Zhu, Y., Yu, P.S.: Top-k similarity join in heterogeneous information networks. IEEE Trans. Knowl. Data Eng. **27**(6), 1710–1723 (2015)
7. Yang, J., Chen, L., Zhang, J.: FctClus: a fast clustering algorithm for heterogeneous information networks. PLoS ONE **10**(6), e0130086 (2015)
8. Long, B., Zhang, Z.M., Wu, X., et al.: Spectral clustering for multi-type relational data. In: Proceedings of the 23rd International Conference on Machine learning, Pittsburgh, pp. 585–592 (2006)
9. Ma, X., Dong, D.: Evolutionary nonnegative matrix factorization algorithms for community detection in dynamic networks. IEEE Trans. Knowl. Data Eng. **29**(5), 1045–1058 (2017)
10. Sun, Y., Yu, Y., Han, J.: Ranking-based clustering of heterogeneous information networks with star network schema. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, pp. 797–806 (2009)