Semi-supervised clustering on heterogeneous information networks

Chen Luo[1], Wei Pang[2], and Zhe Wang[1*]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China
rackingroll@163.com wz2000@jlu.edu.com
[2] School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, AB24 3UE, UK
pang.wei@abdn.ac.uk

Abstract. Semi-supervised clustering on information networks combines both the labeled and unlabeled data sets with an aim to improve the clustering performance. However, the existing semi-supervised clustering methods are all designed for homogeneous networks and do not deal with heterogeneous ones. In this work, we propose a semi-supervised clustering approach to analyze heterogeneous information networks, which include multi-typed objects and links and may contain more useful semantic information. The major challenge in the clustering task here is how to handle multi-relations and diverse semantic meanings in heterogeneous networks. In order to deal with this challenge, we introduce the concept of *relation-path* to measure the similarity between two data objects of the same type. Thereafter, we make use of the labeled information to extract different weights for all *relation-path*s. Finally, we propose SemiRPClus, a complete framework for semi-supervised learning in heterogeneous networks. Experimental results demonstrate the distinct advantages in effectiveness and efficiency of our framework in comparison with the baseline and some state-of-the-art approaches.

Keywords: Heterogeneous information network, Semi-supervised clustering.

1 Introduction

The real world is interconnected: objects and inter-connections between these objects constitute various information networks. Clustering methods in information networks [1] become more and more popular in recent years. One can discover much interesting knowledge from the information networks by using appropriate clustering methods, and the clustering result can also be used in many fields such as information retrieval [2] and recommendation systems [3]. In particular, the real world information networks are often heterogeneous [4], which means in these networks objects and links between these objects may belong to different types. In order to handle the multi-relational data in heterogeneous networks, semi-supervised learning methods [5] can be an appropriate tool. In this paper,

we focus on the semi-supervised clustering task in heterogeneous information networks.

Up till now many semi-supervised clustering algorithms have been proposed for information networks [6, 7, 8]. Some of these algorithms consider the labeled information as constraints for clustering tasks [6]. These constraints can guide the clustering process to achieve better results. Others focus on semi-supervised learning on graphs [7], which uses a small portion of labeled objects to label all the other objects in the same network by propagating the labeled information. The semi-supervised algorithm, proposed in [8], integrates both the constraint-based learning and distance-function learning methods. All the above-mentioned link-based clustering methods are specifically designed for homogeneous information networks, in which all the links in the network are assumed to be of the same type [1]. However, most of the real-world networks are heterogeneous ones [4]. In KDD-2012, Sun *et.al* proposed PathSelClus [9], a user guided clustering method in heterogeneous information networks. PathSelClus integrates both the meta-path selection and clustering processes. The experimental results produced by PathSelClus also showed that more meaningful results could be obtained by considering the clustering task on the heterogeneous information networks instead of the homogeneous ones. However, in PathSelCLus, the number of clusters needs to be pre-specified at the beginning of the algorithm, which is not realistic in many real-world problems.

In this paper, we will investigate semi-supervised clustering [5] in heterogeneous information networks, and intend to develop a clustering algorithm that does not need to pre-specify the number of clusters. In a heterogeneous information network, two objects may be connected via different relation paths or sequences of relations [9]. These different relation paths have different semantic meanings. For example, in the academic community network, two authors can be connected via either the co-author relationship or the co-institution relationship, but these two relations have very different meanings. Sun *et.al* proposes the concept of 'meta-path' [10] to indicate the relation sequence. In this research we propose a similar definition —'*relation-path*', which is specifically for our clustering task. Correspondingly, we also propose a topological measure for our relation-path, which is different from existing path topological measures [10, 11]. By using a logistic regression approach, we evaluate each weight of the relation-path. Finally, SemiRPClus, a novel framework for clustering in heterogeneous information network, is presented. Experiments on DBLP showed the distinct advantages in effectiveness and efficiency of SemiRPClus in comparison with some clustering methods on information networks.

The rest of this paper is organized as follows: in Section 2 some important definitions used in this paper are introduced. The proposed framework, named SemiRPClus, is described in Section 3. In Section 4, we present a series of experiments on DBLP, which demonstrated the effectiveness and efficiency of SemiRPClus. Finally, we conclude our work in Section 5.

2 Problem Definition

As in [4], we use $G = \langle V, E, W \rangle$ to represent a heterogeneous network, where $V = \bigcup_{i=1}^{m} X_i$, and $X_1 = \{x_{11}, ..., x_{1n_1}\}, ..., X_m = \{x_{m1}, ..., x_{mn_m}\}$ denote the m different types of nodes. E is the set of links between any two data objects of V, and W is the set of weight values on the links. $T_G = \langle A, R \rangle$ denotes the network schema [10], which is a directed graph defined over object types A, with edges as relations from R. For more details about heterogeneous networks, please refer [4, 10]. First, the semi-supervised clustering in heterogeneous network is given below:

Definition 1 (Semi-supervised clustering in heterogeneous information network). *In a heterogeneous information network $G = \langle V, E, W \rangle$ following a network schema $T_G = \langle A, R \rangle$, suppose V' is a subset of V and $V' \subseteq V \in X_i$, where X_i is the target type for clustering, and each data object O in V' is labeled with a value γ indicating which cluster O should be in. Given a set of relation-path (see Definition 2), the learning task is to predict the labels for all the unlabeled objects $V - V'$.*

Second, we give the definition of *relation-path*, which can be considered as a special case of meta-path [10]:

Definition 2 (Relation-path). *Given a network schema $T_G = \langle A, R \rangle$, a relation-path RP is in the form of $A_t \overset{R_1}{\rightarrow} A_2 \overset{R_2}{\rightarrow} ...A_{l-1} \overset{R_{l-l}}{\rightarrow} A_t$, which defines a composite relation $RP = R_1 \circ R_2 \circ ... \circ R_{l-1}$ between two objects in the same target type A_t, and $\circ$ is the composition operator on relations.*

Different from the definition of meta-path [10], in *relation-path* the starting object and the end object of the relation-path must belong to the same target type. From Definition 2 we can see that a *relation-path* is always a meta-path, but not vice versa. More importantly, as the relation-path is defined for objects of the same target type, it will be more suitable for our clustering task. Third, we define a transform of our *relation-path* named '*inverse relation-path*' as follow:

Definition 3 (Inverse Relation-path). *Given a relation-path RP: $A_t \overset{R_1}{\rightarrow} A_2 \overset{R_2}{\rightarrow} ...A_{l-1} \overset{R_{l-l}}{\rightarrow} A_t$, RP^{-1} is the Inverse Relation-path of RP, if RP^{-1} is $A_t \overset{R_{l-1}{}^{-1}}{\rightarrow} A_{l-1} \overset{R_{l-2}{}^{-1}}{\rightarrow} ...A_1 \overset{R_l}{\rightarrow} A_t$, where R^{-1} is the inverse relation of R.*

After defining all the above concepts, we introduce a typical heterogeneous information network used in the experiments of our research: the DBLP network, which has been used as test cases in a number of papers [12, 4, 9].

Example 1 (The DBLP bibliographic network). DBLP, computer science bibliography database, is a representative of heterogeneous information networks. The DBLP schema is shown in Figure 1. There are four types of objects in the schema: Paper, Author, Term, and Conference. Links between Author and Paper are defined by the relation of "write" and "written by", denoted as "$write^{-1}$".

Relation between Term and Paper is "mention" and "mentioned by", denoted as "mention^{-1}". Relation between Paper and Conference is "publish" and "published by", denoted as "publish^{-1}". The "cite" relation exists between the papers in the schema. In this research we extract the "cite" relation from the Microsoft Academic Search API.

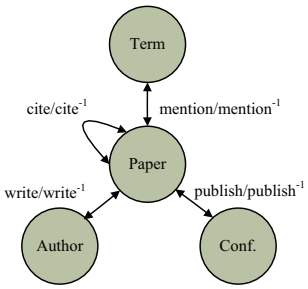


Fig. 1. DBLP schema

3 The SemiRPClus Framework

As mentioned in [4], there are two constraints which determine the clustering results on heterogeneous information networks: first, the clustering result should be consistent with the network structure; second, the clustering results should be consistent with the labeled information pre-assigned for some data objects. Our semi-supervised clustering process will follow these two constraints.

In this section, we first introduce in detail the proposed framework, SemiRPClus, which includes two components: (1) the linear regression based topological measure and (2) the relation extraction model. Then we present the overall clustering framework.

3.1 Linear Regression Based Topological Measure

Topological features, also called structural features, are connectivity properties extracted from a network for some pairs of objects. Many topological features have been proposed for the homogeneous networks, and see more details in [13]. There are also some topological features proposed for heterogeneous networks, and we redefine them based on *relation-path* as below:

- **Path Count Measure [12].** Given a *relation-path*, denoted as RP, the Path Count can be calculated as the number of path instances of RP between two objects, say $x_{t,i}$ and $x_{t,j}$, denoted as $S^{PC}_{RP}(x_{t,i}, x_{t,j})$, where $x_{t,i}, x_{t,j} \in X_t$ and X_t is the target type.
- **Random Walk Measure [12].** Random walk measure following a *relation-path* RP is defined as $S^{RW}_{RP}(x_{t,i}, x_{t,j}) = \frac{S^{PC}_{RP}(x_{t,i}, x_{t,j})}{S^{PC}_{RP}(x_{t,i}, :)}$. Here, $S^{PC}_{RP}(x_{t,i}, :)$ denotes the path count value following RP starting from $x_{(t,i)}$.

- **PathSim Measure [10].** Given a *relation-path* RP, PathSim between two objects $x_{t,i}, x_{t,j}$ is defined as :

$$S^{PS}_{RP}(x_{t,i}, x_{t,j}) = \frac{2 * S^{PC}_{RP}(x_{t,i}, x_{t,j})}{S^{PC}_{RP}(x_{t,i}, x_{t,i}) + S^{PC}_{RP^{-1}}(x_{t,j}, x_{t,j})}$$

  Here, S^{PC}_{RP} is a path count measure. In the above RP^{-1} denotes the *inverse path-relation* of RP (see Definition 3);
- **HeteSim Measure [11]:** Given a *relation-path* $RP = R_1 \circ R_2 \circ ... \circ R_l$ (as in Definition 2), HeteSim [11] is defined as follows:
  $S^{HS}_{RP(R_1 \circ R_2 \circ ... \circ R_l)}(x_{t,i}, x_{t,j}) =$

$$\frac{\sum_{p=1}^{|O(x_{t,i}|R_1)|} \sum_{q=1}^{|I(x_{t,j}|R_l)|} S^{HS}_{RP(R_2 \circ R_3 \circ ... \circ R_{l-1})}(O_p(x_{t,i}|R_1), I_q(x_{t,j}|R_l))}{S^{PC}_{RP}(x_{t,i}, :) + S^{PC}_{RP^{-1}}(:, x_{t,j})}$$

  In the above, $x_{t,i}, x_{t,j} \in X_t$, $O(x_{t,i}|R_1)$ is the set of out-neighbors of $x_{t,i}$ based on relation R_1, and $I(x_{t,j}|R_l)$ is the set of in-neighbors of $x_{t,j}$ based on relation R_l.

All the above-described topological measures only focus on the topological structure of the networks. However, in the semi-supervised clustering process, different labeled information will lead to different similarity measures and different clustering results [9]. As a result, we propose a linear regression based measure which considers the small amount of labeled information. We use the labeled information as guidance, and propose a linearly combined measure, which is defined as follows:

$$S^{LS}_{RP}(x_{t,i}, x_{t,j}) = \sum_{d=1}^{m} \alpha_d s_d(x_{t,i}, x_{t,j}) \tag{1}$$

where S^{LS}_{RP} is the linear regression based measure of $x_{t,i}, x_{t,j} \in X_t$, and $s(x_{t,i}, x_{t,j}) = [s_1(x_{t,i}, x_{t,j}), s_2(x_{t,i}, x_{t,j}), ..., s_m(x_{t,i}, x_{t,j})]^T$ is the topological features following the given *relation-path* RP, and each feature is calculated using one of the formulae introduced at the beginning of this section, and m is the number of measures used in the framework. For example, $s_1(x_{t,i}, x_{t,j}) = S^{PC}_{RP}(x_{t,i}, x_{t,j})$, $s_2 = S^{RW}_{RP}(x_{t,i}, x_{t,j})$, $s_3 = S^{PS}_{RP}(x_{t,i}, x_{t,j})$, $s_4 = S^{HS}_{RP}(x_{t,i}, x_{t,j})$. Here, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_m]^T$ denotes the weights for all measures. An optimization algorithm can be used to solve the following approximation problem:

$$\boldsymbol{\alpha}^{opt} = \arg min_{\alpha} \left\| \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbb{R}(x_{t,i}, x_{t,j}) - \sum_{d=1}^{m} \alpha_d s_d(x_{t,i}, x_{t,j})) \right\| \tag{2}$$

In the above, n denotes the number of labeled objects. Matrix $\mathbb{R}$ is obtained from the pre-labeled information as follow:

$$\mathbb{R}(x_{t,i}, x_{t,j}) = \begin{cases} 1 & x_{t,i}, x_{t,j} \textit{ are labeled as the same label} \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

where $x_{t,i}, x_{t,j} \in X_t$ and X_t is the target type for clustering. Eq. (2) is actually a linear regression problem, which can be efficiently solved by many existing algorithms [14]. In this paper, we use the gradient descent method [14] to solve this problem.

3.2 Relationship Extraction Model

In our clustering task, given a set of *relation-path*s, each object within the target type may be connected via these *relation-path*s. By using these *relation-path*s, a heterogeneous information network can be reduced into a multi-relational network [15], in which the objects correspond to those of the target type in the original heterogeneous network and the different relations correspond to the given different relation-paths.

Similar to the model proposed in [15], the basic motivation of our relationship extraction model is as follows: different *relation-path*s correspond to different relation graphs[3], which can provide different clustering results. By combining these different clustering results through different weights of corresponding *relation-path*s, the final clustering result may be improved [15]. In this paper, logistic regression method [16] is used to handle the relation extraction problem.

The set of pre-labeled objects of the target type are regarded as the training set. Each two objects in the labeled set is regarded as a training pair, denoted as $x_{t,i}, x_{t,j} \in X_t$, and X_t is the target type for clustering. We first extract the topological features for these objects, and then build an extraction model to learn the weight values associated with these features. For each training pair $x_{t,i}, x_{t,j} \in X_t$, all the features are calculated by the linear regression based measure as in Eq. (1), denoted as $\boldsymbol{F} = [f_1, f_2, ..., f_d]$, where d is the number of *relation-paths* between the two objects. Here we denote the training set as $\boldsymbol{X}^{Train} = [X^1, X^2, ..., X^n]$, where X^i denotes the $i-th$ training pair, and n is the number of pairs in the training set. We define y^i as a label indicating whether these two objects are in the same cluster: if these two objects are in the same cluster, $y^i = 1$, and 0 otherwise, which is denoted as $\boldsymbol{Y}^{Train} = [y^1, y^2, ..., y^n]$, where $y^i \in \{0, 1\}$. The set of weights for all *relation-paths* is denoted as $\Lambda = [\lambda_1, \lambda_2, ..., \lambda_d]$. We use the Entropy maximization [17] method to calculate Λ: first, the conditional probability of the two objects in X^i belong to the same cluster can be modeled as:

$$p_\Lambda(y = y^i|X^i) = \frac{1}{Z(X^i)} exp(\sum_{i=1}^{d} f_i(X^i) * \lambda_i) \tag{4}$$

where Z is the normalization term calculated as $Z(X^i) = 1 + exp(\sum_{i=1}^{d} f_i(X^i) * \lambda_i)$. Second, we use the MLE (Maximum Likelihood Estimation) approach to derive Λ by maximizing the likelihood of all the training pairs:

[3] A **relation graph** is a homogeneous network reduced by the heterogeneous network using a typical relation-path

$$L(\Lambda) = \prod_{i}^{n} [p_\Lambda(y = y^i|X^i)]^{y^i} [p_\Lambda(y = y^i|X^i)]^{1-y^i} \tag{5}$$

Third, $\Lambda = [\lambda_1, \lambda_2, ..., \lambda_d]$ can be obtained as follows:

$$\Lambda^* = \arg max_{\Lambda=[\lambda_1,\lambda_2,...,\lambda_d]} L(\Lambda) \tag{6}$$

In this research we use the gradient descent [14] method to calculate Λ^*. Finally the combined affinity matrix $W^{combine}$ is defined by the following equation:

$$W^{combine} = \sum_{i=0}^{d} \lambda_i * W_i \tag{7}$$

Here W_i is the similarity matrix of $i-th$ homogeneous network reduced by $i-th$ *relation-path*. It is noted that the similarity of each two objects in W_i is calculated by the measure introduced in section 3.1. We can see from the Eq. (7) that the combined affinity matrix $W^{combine}$ is a linear combination of different relation graphs. After obtaining $W^{combine}$, we then perform clustering on $W^{combine}$ to obtain the finally result.

3.3 The Detailed Steps of SemiRPClus

After presenting the calculation method for each relevant variable, the detailed steps of SemiRPClus is given as follows:

Step 1 Given a heterogeneous information network $G = \langle V, E, W \rangle$, a set of *relation-paths*, the target type X_t for clustering and labeled information Y.

Step 2 Use Eq. (3) to calculate the relation matrix $\mathbb{R}$.

Step 3 Calculation of the linear based similarity measure

- **Step 3-a** Calculate each kind of relation-path measure using the measure methods introduced in Section 3.1.
- **Step 3-b** Use Eq. (2) to calculate the weight $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_m]^T$ for each measure.
- **Step 3-c** Calculate the linear based similarity S_{RP}^{LS} using Eq. (1).

Step 4 Use Eq. (6) to obtain the weight of each *relation-path*: $\Lambda = [\lambda_1, \lambda_2, ..., \lambda_d]$, then use Eq. (7) to calculate the combine affinity matrix $W^{combine}$.

Step 5 Cluster the relation matrix $W^{combine}$ to obtain the final clustering result.

It is pointed out that many useful clustering methods [15, 18] can be used in Step 5, no matter whether the cluster number is pre-assigned or not.

3.4 Complexity Analysis

In this section, we consider the time complexity of the SemiRPClus. At the beginning of our framework, all the traditional topological features are calculated, and the time complexity is $O(k_{path}n^2)$. Here k_{path} is the number of *relation-path*s selected for the framework, and n is the number of target objects. For the linear based measure calculation process, the time complexity is $O(t_1 \sum_m |T_m|)$. Here t_1 is the number of iterations, and m is the dimension of the training dataset. For the Relationship Selection Model, the time complexity is $O(t_2 \sum_m |T_m|)$. Here t_2 is the number of iterations of this model. Assuming that the time complexity of the clustering used in Step 5 is $O_{cluster}$, the overall time complexity of our framework is $O(k_{path}n^2) + O(t_1 \sum_m |T_m|) + O(t_2 \sum_m |T_m|) + O_{cluster}$.

4 Experimental Results

In this section, we use the DBLP dataset[2] as a test bed to evaluate both the effectiveness and efficiency of our approach compared with some existing methods.

4.1 Datasets

The DBLP dataset is used for the performance test. Following [9], we extract a sub network of DBLP, "four-area dataset", which contains 20 major conferences in four areas: Data Mining, Database, Information Retrieval and Machine Learning. Each area contains five top conferences. In this dataset, the term is extracted from the paper titles, and the paper citation relationship is obtained by the Microsoft academic search API[3]. We use the following three datasets for our experiments.

DataSet-1 top 100 authors in the DBLP within the 20 major conferences, and the corresponding papers published by these authors after 2007.

DataSet-2 top 500 authors in the DBLP within the 20 major conferences, and the corresponding papers.

DataSet-3 top 2000 authors in the DBLP within the 20 major conferences, and the corresponding papers published by them after 2007.

The ground truth used in our experiment is obtained from the "four-area dataset [9]". It is pointed out that the labeled information and the true clustering result are all obtained from the ground truth.

As in [12], we choose 10 types of *relation-path*s as bellow: $author - paper - author$, $author \rightarrow paper \rightarrow paper - author$, $author \rightarrow paper \leftarrow paper \rightarrow author$, $author \rightarrow paper \rightarrow conference \leftarrow paper \leftarrow author$, $author \rightarrow paper \leftarrow author \rightarrow paper \leftarrow author$, $author \rightarrow paper \rightarrow term \leftarrow paper \leftarrow author$, $author \rightarrow paper \rightarrow paper \rightarrow paper \leftarrow author$, $author \rightarrow paper \rightarrow paper \leftarrow paper \leftarrow author$, $author \rightarrow paper \leftarrow paper \leftarrow paper \leftarrow author$, $author \rightarrow$

[2] http://www.informatik.uni-trier.de/ ley/db/

[3] http://academic.research.microsoft.com/

$paper \leftarrow paper \rightarrow paper \leftarrow author$. For example, $author - paper - author$ denotes the co-author relationship, and $author \rightarrow paper \leftarrow paper \rightarrow paper \leftarrow author$ denotes two authors' papers are cited in the same paper.

4.2 Case study on Effectiveness

In this section, we study the effectiveness of our algorithm by comparing it with several existing methods on the three datasets given in Section 4.1. For the ease of comparison, we choose the hierarchical-cluster algorithm [19] to cluster the similarity matrix, and the cluster number is pre-assigned as the input of the cluster algorithm.

Three clustering methods are used in our experiment for comparison: PathSelClus [9], GNetMine [20] and LP [7]. The first two algorithms are proposed for heterogeneous networks, and they are regarded as the state-of-the-art clustering algorithms. LP is proposed for homogeneous network, thus we use two homogeneous networks reduced by two corresponding *relation-path*s. The two selected *relation-path*s have the highest weight in SemiRPClus.

Two evaluation methods are used for testing the clustering result: Accuracy [9], which is calculated as the percentage of target objects clustered into the correct clusters; and Normalized Mutual Information (NMI) [9], which is one of the most popular evaluation methods to evaluate the quality of clustering results.

The clustering results are presented in Table 1. In the table, performance under different percentage of labeled information (5%, 10% and 20%) in each cluster is tested. All the results are averaged for 10 times. In Table 1, results in bold indicate the best performance among all algorithms.

From Table 1 we see that the performance of SemiRPClus is comparable with PathSelClus and GNetMine, and better in some cases. From the result we can also see that SemiRPClus can have a better result evaluated by NMI in some cases. NMI considers not only the accuracy, but also the distribution of the objects within each cluster. From this perspective, SemiRPClus is more effective than the other three algorithms. The LP algorithm always performs worse than all other three heterogeneous clustering algorithms. This demonstrates that by considering the heterogeneous information better results can be obtained. On the other hand, we can see that mining heterogeneous networks can gain more useful information than homogeneous ones.

4.3 Case Study on Efficiency

In this section, we study the efficiency of SemiRPClus. We use the same hardware configuration to run SemiRPClus and the other three algorithms. Every algorithm is run 10 trials and the average performance is calculated. The CPU execution time for each algorithm is showed in Fig. 2. It is pointed out that we use LP-RP1 to represent LP algorithm in this section. In Fig. 2, we can see that SemiRPClus is more efficient than PathSelClus and GNetMine: it is three to four orders of magnitude faster than PathSelClus in all experiments.

Table 1. Cluster Accuracy and NMI for Three Dataset

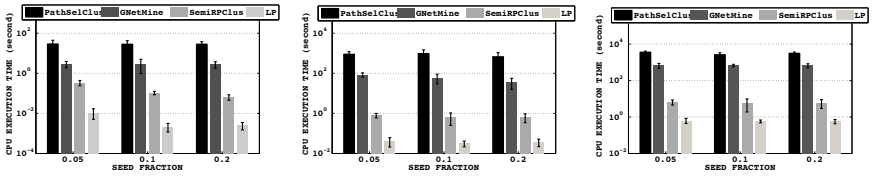
Labeled	Evaluation	SemiRPClus	LP-RP1[4]	LP-RP5[5]	PathSelClus	GNetMine
Dataset-1						
5%	NMI	.048±.015	.078±.012	.020±.002	**.457±.095**	.387±.089
5%	Accuracy	.380±.021	.350±.036	.280±.016	**.570±.040**	.520±.073
10%	NMI	.318±.032	.056±.021	.031±.015	**.523±.026**	.408±.127
10%	Accuracy	.510±.024	.320±.041	.390±.012	**.710±.096**	.550±.048
20%	NMI	**.696±.032**	.069±.009	.036±.009	.541±.081	.488±.057
20%	Accuracy	.680±.042	.320±.023	.320±.084	**.730±.070**	.620±.058
Dataset-2						
5%	NMI	.621±.103	.014±.007	.004±.008	.609±.045	**.677±.042**
5%	Accuracy	.720±.087	.306±.085	.272±.052	.786±.116	**.884±.034**
10%	NMI	**.698±.038**	.023±.010	.006±.005	.646±.073	.664±.117
10%	Accuracy	.736±.085	.316±.034	.284±.094	**.830±.042**	.664±.042
20%	NMI	**.774±.034**	.026±.006	.013±.024	.718±.049	.702±.039
20%	Accuracy	.862±.046	.356±.088	.282±.044	.854±.0350	**.900±.028**
Dataset-3						
5%	NMI	**.798±.046**	.001±.006	.007±.002	.652±.089	.621±.045
5%	Accuracy	.750±.048	.254±.012	.207±.034	**.872±.015**	.862±.065
10%	NMI	**.759±.095**	.003±.014	.004±.002	.664±.015	.632±.015
10%	Accuracy	.784±.014	.254±.074	.271±.045	**.880±.034**	.868±.034
20%	NMI	**.868±.015**	.002±.001	.004±.002	.697±.095	.676±.024
20%	Accuracy	.800±.031	.261±.049	.275±.041	**.897±.012**	.889±.025

4.4 Case Study on Relation Path Weight

In this section, we study the learned weights for different *relation-path*s obtained by SemiRPClus compared with PathSelClus. As the ranking of the *relation-path*s showed in Table 2, the ranking learned by SemiRPClus is fundamentally the same as the ranking learned by the PathSelClus. From the result we can see the *relation-path* $author - paper - author$ always has a more trusted weight than other relation paths. This is consistent with human intuition: two authors having the co-author relationship means that they are very likely to have a very similar research interest. On the other hand, the *relation-path* $author \rightarrow paper \rightarrow term \leftarrow paper \leftarrow author$ has the lowest weight in both algorithms. This is consistent with real-world scenarios: it is not rare that two papers from different areas can have the same term. For example, the words "optimization" and "method" appear in many papers from different areas.

[4] $author - paper - author$

[5] $author \rightarrow paper \leftarrow author \rightarrow paper \leftarrow author$



(a) Efficiency in DataSet-1 (b) Efficiency in DataSet-2 (c) Efficiency in DataSet-3

Fig. 2. Running time of SemiRPClus compared with the other three algorithms.

Table 2. *Relation-Paths* Weight Comparison

Rank	**PathSelClus**	**PathSelClus**
1	$A - P - A^6$	$A - P - A$
2	$A \to P \leftarrow A \to P \leftarrow A$	$A \to P \leftarrow A \to P \leftarrow A$
3	$A \to P \leftarrow P \leftarrow P \leftarrow A$	$A \to P \leftarrow P \to P \leftarrow A$
4	$A \to P \to P - A$	$A \to P \leftarrow P \to A$
5	$A \to P \leftarrow P \to A$	$A \to P \to T \leftarrow P \leftarrow A$
6	$A \to P \to P \to P \leftarrow A$	$A \to P \to P - A$
7	$A \to P \to T \leftarrow P \leftarrow A$	$A \to P \to P \to P \leftarrow A$
8	$A \to P \leftarrow P \to P \leftarrow A$	$A \to P \to C \leftarrow P \leftarrow A$
9	$A \to P \to C \leftarrow P \leftarrow A$	$A \to P \leftarrow P \leftarrow P \leftarrow A$
10	$A \to P \to T \leftarrow P \leftarrow A$	$A \to P \to T \leftarrow P \leftarrow A$

5 Conclusions and Future Work

In this work, we explore the semi-supervised clustering analysis in heterogeneous information networks. Firstly, a similarity measure, which is more suitable for the semi-supervised clustering task, is proposed for measuring the similarity between objects in heterogeneous information networks. Secondly, a logistic regression model is used for extracting the relations. At last, an overall computational framework is proposed to perform semi-supervised clustering in heterogeneous information networks. Experimental results on the DBLP dataset demonstrate the effectiveness and efficiency of SemiRPClus.

In the future, we intend to apply SemiRPClus to more real-world clustering problems. In addition, another direction of our future research is to explore the potential of SemiRPClus on big data problems, such as massive social media and bioinformatics problems.

Acknowledgements

WP is supported by the partnership fund from dot.rural, RCUK Digital Economy research. This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61373051; the National Science and Technology Pillar Program (Grant No. 2013BAH07F05), Jilin Province Science and

Technology Development Program (Grant No. 20111020); Project of Science and Technology Innovation Platform of Computing and Software Science (985 engineering), and the Key Laboratory for Symbolic Computation and Knowledge Engineering, Ministry of Education, China.

References

[1] Fortunato, S.: Community detection in graphs. Physics Reports **486**(3) (2010) 75–174
[2] Lipka, N., Stein, B., Anderka, M.: Cluster-based one-class ensemble for classification problems in information retrieval. In: SIGIR'12, ACM (2012) 1041–1042
[3] Pham, M.C., Cao, Y., Klamma, R., Jarke, M.: A clustering approach for collaborative filtering recommendation using social network analysis. J. UCS **17**(4) (2011) 583–604
[4] Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: ICDT'09, ACM (2009) 565–576
[5] Zhu, X.: Semi-supervised learning literature survey. Computer Science, University of Wisconsin-Madison **2** (2006) 3
[6] Basu, S., Banerjee, A., Mooney, R.J.: Semi-supervised clustering by seeding. In: ICML. Volume 2. (2002) 27–34
[7] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in neural information processing systems **16**(16) (2004) 321–328
[8] Bilenko, M., Basu, S., Mooney, R.J.: Integrating constraints and metric learning in semi-supervised clustering. In: ICML, ACM (2004) 11
[9] 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: KDD'12, ACM (2012) 1348–1356
[10] Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. VLDB11 (2011)
[11] Shi, C., Kong, X., Yu, P.S., Xie, S., Wu, B.: Relevance search in heterogeneous networks. In: ICDT'12, ACM (2012) 180–191
[12] Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C., Han, J.: Co-author relationship prediction in heterogeneous bibliographic networks. In: ASONAM'11, IEEE (2011) 121–128
[13] Lü, L., Zhou, T.: Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications **390**(6) (2011) 1150–1170
[14] Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. Volume 821. Wiley (2012)
[15] Cai, D., Shao, Z., He, X., Yan, X., Han, J.: Mining hidden community in heterogeneous social networks. In: LinkKDD, ACM (2005) 58–65
[16] Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression. Wiley. com (2013)
[17] Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational linguistics **22**(1) (1996) 39–71

[6] A: author, P: paper, C: conference, T: term

[18] Frey, B.J., Dueck, D.: Clustering by passing messages between data points. science **315**(5814) (2007) 972–976
[19] Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. The Computer Journal **26**(4) (1983) 354–359
[20] Ji, M., Sun, Y., Danilevsky, M., Han, J., Gao, J.: Graph regularized transductive classification on heterogeneous information networks. In: Machine Learning and Knowledge Discovery in Databases. Springer (2010) 570–586