International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.107-118
http://dx.doi.org/10.14257/ijdta.2016.9.8.12

A Three-Phase Algorithm for Clustering Multi-typed Objects in Star-Structured Heterogeneous Data

Huang Yue

Library of Beijing Language and Culture University, Beijing, P. R. China, 100083
huang.yuet@blcu.edu.cn

Abstract

Heterogeneous networks, composed of multiple types of objects and relationships, are ubiquitous in real life. Although many methods have been proposed for community detection in homogeneous networks which contain only one type of objects and one type of relationships between these objects, effective direct clustering objects of all types in heterogeneous networks without heterogeneous-to-homogeneous transformation remains a challenge. To achieve this goal, we propose a three-phase method for clustering star-structured heterogeneous data based on diffusion path. By adopting the principle that central objects are more important than attribute objects, we firstly assess the proximity of central objects in terms of their connected objects of all types, then based on which we cluster central objects, and thirdly we detect attribute objects groups according to their associated central objects. Finally, experiments on a real-world data set show the effectiveness and efficiency of the proposed methods.

***Keywords**: Clustering; heterogeneous data; multi-typed objects; star-structured*

1. Introduction

With the advent of Web 2.0, there has been a growing attention on data science in recent years. Clustering, which explores hidden groups consisting of similar objects without pre-defined classes, is a widely used data mining technique. For example, detecting research community from bibliographic data is one of the important applications of clustering, which helps researchers discover useful knowledge from high-volume of scientific papers, such as authors with similar research interest or papers with common topic.

Data sets with interrelated objects are often represented as homogeneous networks, for the reason that there is only one type of nodes and one type of relations in these networks [1]. For example, bibliographic data are often represented as author graphs in which the authors form nodes and the relations between authors form edges, or paper graphs in which nodes represent papers and edges represent citation relations. A great many algorithms proposed for homogeneous networks clustering are suitable for this kind of bibliographic data but with prior processing work. One needs to first extract a co-author network and then apply traditional graph clustering methods. However, such extraction is an information reduction process [2], since some valuable information such as paper title or venue is lost in the constructed co-author network. Whereas in real-world settings, heterogeneous networks [1,3] which involve distinct types of objects or links are ubiquitous. In fact, mining heterogeneous networks has become a new and promising frontier in data mining research. [4] The greater the number of types a data set contains, the more heterogeneous it is. [5] Because of the complexity of heterogeneous networks, interests of objects of different types are not the same. Usually, the objects of interest are called central objects or target objects, and the objects of other types are called attribute objects. [6] A star-structure [1] is a special and common case where the central type is related to several attribute types.

ISSN: 2005-4270 IJDTA

Generally, methods for clustering on heterogeneous networks can be classified into three classes based on their principles of clustering [5]: heterogeneous-transformed homogeneous network clustering, simultaneous clustering of objects of each type, and target-object clustering based on attribute objects. One straightforward way of dealing with heterogeneous data is to first project a heterogeneous network onto a series of homogeneous networks, one for each mode [3], or to construct a homogeneous network with relationships that are a combination of each dimension [3] and then group the objects using existing graph partitioning methods. In fact, this way is not a fundamental solution to heterogeneous clustering, and due to information loss during transformation it is necessary to directly clustering heterogeneous data. From about two decades ago, some methods try to clustering bi-typed relational data which is called two-way clustering [7], co-clustering [8] or bi-clustering [9], and they are only suitable for data sets with two types of objects. While, more recently some methods are developed for data sets with heterogeneity of no less than three, which are called data sets with “real heterogeneity” [5]. These methods cluster objects of one type (target objects or central objects) based on the objects of other types that are linked with them (attribute objects). Direct-link-based methods [10-11], SimRank group methods [12-16], meta-path-based methods [17-18] and ranking-based methods [6, 19-21] are from this group. However, attribute information is usually ignored by these methods. In sum, when dealing with heterogeneous data especially with more than two types of heterogeneous data (high-order heterogeneous data [22]), traditional clustering methods which only focus on one particular type of objects of interest no longer work. Compared with clustering methods for homogeneous networks, only a small proportion of methods clustering on heterogeneous networks are developed for data sets with “real heterogeneity”. Besides, both attributes and links are critical in clustering objects, and clustering from perspective of either attribute or link as existing methods do is inaccurate. Therefore, it is still necessary to propose novel algorithms considering both attributes and links for clustering high-order heterogeneous data without heterogeneous-to-homogeneous transformation.

Therefore, owing to these limitations of existing method, our main focus in this paper is to directly clustering high-order heterogeneous data without heterogeneous-to-homogeneous transformation as well as considering attribute information. We put forward an algorithm called StarClusDP (star-structured heterogeneous data clustering based on diffusion path) to deal with star-structured heterogeneous data and experimental results show its promises for clustering different types of objects effectively. By focusing on multi-typed objects, our work differs from previous studies and could lead a more meaningful partition.

2. Our Method

In this section, we first give basic concepts to serve the purpose of our research. Then, we propose a framework composing three phases to clustering star-structured heterogeneous data. Lastly we drop out a solution to each sub-problem sequentially.

2.1. Concepts

Definition 1 (Heterogeneous Data) We denote a heterogeneous data set D with n central objects and r types of attribute objects as $D=(C,A,R)$, where $C=\{C_i\}_{i=1}^{n}$ is a set of central objects; $A=\{A_k\}_{k=1}^{r}$ is a set of attribute objects, where $A_k=\{a_p^k\}_{p=1}^{n_k}$ is a set of attribute objects of the k th type, and $n_k=|A_k|$; $R=\{r_l\}_{l=1}^{n_R}$, where $r_l=<r_l.from, r_l.to>$, $r_l.from \in (C \cup A)$, $r_l.to \in (C \cup A)$, and $n_R=|R|$.

With this definition, it is clear to see the relations of central objects and attribute objects, which could help us quantify the role of different attribute objects in connecting central objects. In this paper, we deal with a special case of heterogeneous data where the central objects are connected to each other only via attribute objects. We call this as star-structured heterogeneous data, which is illustrated in Figure 1.

Definition 2 (Star-structured Heterogeneous Data) For a given heterogeneous data set $D=(C,A,R)$, it is referred to as star-structured heterogeneous data, if $\forall r_l =< r_l.from, r_l.to >$ it meets the conditions: $r_l.from \in C$ and $r_l.to \in A$, or $r_l.from \in A$ and $r_l.to \in C$.

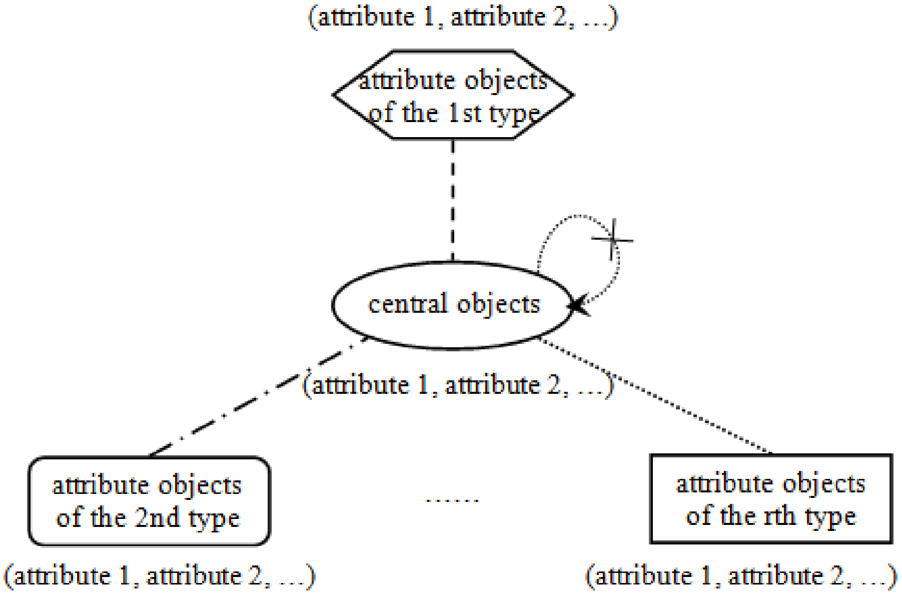
(attribute 1, attribute 2, ...)
attribute objects of the 1st type
central objects
(attribute 1, attribute 2, ...)
attribute objects of the 2nd type
......
attribute objects of the rth type
(attribute 1, attribute 2, ...)
(attribute 1, attribute 2, ...)

Figure 1. Model of Star-Structured Heterogeneous Data

2.2. The General Framework

The objective of clustering is to assign objects that are similar to the same cluster while objects that are not similar to different clusters. A fundamental while critical concept in clustering is the measurement of similarity between two objects. Thus, we first calculate pair-wise similarity between central objects with the use of their connected attribute objects, since in star-structured heterogeneous data central objects are not directly related to each other but through attribute objects. With the result of pair-wise proximity of central objects, in the second step, we apply a hierarchical clustering scheme to group central objects, which yields a cluster assignment of central objects. In the third step, since attribute objects depend on their related central objects, we first measure proximity of attribute objects based on the assignment of their related central objects and then cluster attribute objects. Therefore, we propose a three-phase general framework as follows for clustering objects of all types in star-structured heterogeneous data:

- Similarity measurement between central objects.
- Central object clustering based on pair-wise similarity between central objects.
- Attribute object clustering in terms of clustering results of central objects.

2.3. Method for Similarity Measurement Between Central Objects

We first represent each central object with its related attribute objects so that all information is prepared well when calculating similarity between central objects.

Definition 3 (Representation of Central Objects in Star-structured Heterogeneous Data) For a given star-structured heterogeneous data set $D=(C,A,R)$, each central object is represented by its related attribute objects, that is, $C=\{C_i.A\}_{i=1}^{n}$, where $C_i.A=\{C_i.A_k\}_{k=1}^{r}$ is the set of attribute objects that are related to the i th central object C_i, and $C_i.A_k$ is the set of the k th attribute objects that are related to the i th central object C_i.

As discussed above, central objects are the objects of most interest. However, they are not directly related to each other in the case of star-structured heterogeneous data, but they are indirectly associated with each other via attribute objects. Thus, it is reasonable that each central object in star-structured heterogeneous data is represented by its connected attribute objects. Besides, a specific central object could interrelate with different central objects by diffusing through different kinds of attribute objects. Our idea is to combine structure and attribute information to assess similarity between central objects. Hence, several concepts are proposed as follows.

Definition 4 (Central Object Diffusion Set) For a given star-structured heterogeneous data set $D=(C,A,R)$ and a particular central object C_i, reachable central objects of C_i is defined by diffusing through r types of attribute objects as central object diffusion set of C_i, that is, $C_i.CODS=\{C_i \to A_k\}_{k=1}^{r}$, where $C_i \to A_k=\{C_i \to A_{kp}\}_{p=1}^{n_k}$ denotes the central object set through the k th type of attribute objects, $C_i \to A_{kp}$ denotes the central object set through A_k on the p th value, and $n_k=|A_k|$.

Notice that the central object diffusion set of a central object C_i includes itself, for the reason that one object has the strongest relation with itself. Since a central object could spread to several central objects, the diffusion paths of different central objects might intersect, which indicates some common character between central objects. Hence, similarity of central object diffusion path is developed as follows.

Definition 5 (Similarity of Central Object Diffusion Path) For a given star-structured heterogeneous data set $D=(C,A,R)$, the diffusion path similarity of any two central objects C_i and C_j by the k th type of attribute objects, $sim_{path}(C_i,C_j\mid k)$, is defined as ratio of size of their related attribute objects intersection and that of their union, which is depicted as follows:

$$sim_{path}(C_i,C_j\mid k)=\frac{|C_i.A_k \cap C_j.A_k|}{|C_i.A_k \cup C_j.A_k|} \tag{1}$$

where $C_i.A_k$ is the set of the k th type of attribute objects that are related to the i th central object C_i, and $C_j.A_k$ is the set of the k th type of attribute objects that are related to the j th central object C_j.

As discussed above, central object diffusion set of any two central objects can be obtained by diffusing through attribute objects. It is possible that these two sets overlap which indicates that they share something in common. Hence, the similarity of central object diffusion set is defined as follows.

Definition 6 (Similarity of Central Object Diffusion Set) For a given star-structured heterogeneous data set $D=(C,A,R)$, the diffusion set similarity of any two central objects C_i and C_j by the kth type of attribute objects, $sim_{set}(C_i,C_j\,|\,k)$, is defined as follows:

$$sim_{set}(C_i,C_j\,|\,k)=\frac{\left|(C_i\rightarrow A_k)\cap(C_j\rightarrow A_k)\right|}{\left|(C_i\rightarrow A_k)\cup(C_j\rightarrow A_k)\right|} \tag{2}$$

where $C_i\rightarrow A_k$ denotes diffusion central object set of C_i through the different values of A_k, and $C_j\rightarrow A_k$ denotes diffusion central object set of C_j through the different values of A_k.

It is obvious that the value range of $sim_{set}(C_i,C_j\,|\,k)$ is [0,1], and the bigger the value the more similar of the two.

In a star-structured heterogeneous data set $D=(C,A,R)=(C,A_1,A_2,...,A_r,R)$, any two central objects could communicate via at most r types of path, and one central object could spread to other central objects only through attribute objects, that is, $C_i-A_k-C_j$ ($k\in\{1,2,...,r\}$). However, attribute objects of different types play different roles in clustering of central objects. It is desirable that the role of attribute objects of distinct types should be automatically learnt through the calculation. Hence, similarity of central objects is proposed as follows.

Definition 7 (Diffusion Structure Similarity between Central Objects) For a given star-structured heterogeneous data set $D=(C,A,R)$ with r types of attribute objects, diffusion structure similarity between any two central objects C_i and C_j, $sim_{DS}(C_i,C_j)$, is defined as follows:

$$sim_{DS}(C_i,C_j)=\frac{\sum_{k=1}^{r}sim_{set}(C_i,C_j\,|\,k)^{AMP}}{r} \tag{3}$$

where $sim_{set}(C_i,C_j\,|\,k)$ denotes the similarity of central object diffusion set, and AMP denotes the amplifier factor which is calculated as follows:

$$AMP=1-sim_{path}(C_i,C_j\,|\,k) \tag{4}$$

where $sim_{path}(C_i,C_j\,|\,k)$ denotes the diffusion path similarity of any two central objects C_i and C_j by the kth type of attribute objects.

It is obvious that the value range of $sim_{DS}(C_i,C_j)$ is [0,1], and the bigger the value the more similar of the two.

Definition 8 (Attribute Similarity between Objects) For a given star-structured heterogeneous data set $D=(C,A,R)$, the attribute similarity between any two objects O_a and O_b, $sim_A(O_a,O_b)$, is defined as follows:

$$sim_A(O_a,O_b)=\begin{cases}1,O_a=O_b\\0,O_a\neq O_b\end{cases} \tag{5}$$

Definition 9 (Similarity between Central Objects) For a given star-structured heterogeneous data set $D=(C,A,R)$, the similarity between any two central objects C_i and C_j, $sim(C_i,C_j)$, is defined as follows:

$$sim(C_i, C_j) = \alpha sim_R(C_i, C_j) + \beta sim_A(C_i, C_j) \tag{6}$$

where $sim_R(C_i, C_j)$ denotes the relation similarity of C_i and C_j, $sim_A(C_i, C_j)$ denotes the attribute similarity of C_i and C_j, α denotes the relation factor, β denotes the attribute factor, and α β satisfy the following conditions: $\alpha + \beta = 1$, $\alpha, \beta \in R$, $\alpha \geqslant 0$, $\beta \geqslant 0$.

It is obvious that the value range of $sim(C_i, C_j)$ is [0,1], and the bigger the value the more similar of the two. In this paper, $sim_R(C_i, C_j)$ is calculated as $sim_{DS}(C_i, C_j)$.

Algorithm 1. Calculate similarity between central objects based on attribute objects in star-structured heterogeneous data.

Input: A star-structured heterogeneous data set $D = (C, A, R)$, with α denoting the weight of relation and β denoting the weight of attribute.

Output: Similarity matrix $SimMatrix(C)$ with each element denoting similarity between corresponding two central objects.

```
Step 1. for each central object $C_i \in C$ do
        $C_i.A = \{C_i.A_k\}_{k=1}^{r}$
        $C_i.CODS = \{C_i \rightarrow A_k\}_{k=1}^{r}$
Step 2. for $i \leftarrow 1$ to $n$ do
        for $j \leftarrow 1$ to $n$ do
            calculate $sim_{path}(C_i, C_j \mid k)$
            calculate $sim_{set}(C_i, C_j \mid k)$
            calculate $sim_{DS}(C_i, C_j)$
            calculate $sim_A(C_i, C_j)$
            calculate $sim(C_i, C_j)$
Step 3. return $SimMatrix(C)$
```

In Algorithm 1, the parameter α controls the role of relation structure on similarity between central objects, while β controls the role of attributes of central objects on similarity between central objects.

2.4. Method for Central Object Clustering

Hierarchical clustering is a widely used clustering technique, which can identify hierarchical structure of clusters. With pair-wise similarity between central objects, it is reasonable to apply hierarchical clustering to group them. In this paper, we define the similarity between two central object clusters as follows.

Definition 10 (Similarity between Central Object Clusters) For a given star-structured heterogeneous data set $D = (C, A, R)$ and similarity matrix of central objects $SimMatrix(C)$, the similarity between any two central object cluster

$Clus_i = \{C_i \mid i = 1,2,...,n_i\}$ and $Clus_j = \{C_j \mid j = 1,2,...,n_j\}$, $sim(Clus_i, Clus_j)$, is defined as follows:

$$sim(Clus_i, Clus_j) = \frac{\sum_{i=1}^{n_i}\sum_{j=1}^{n_j} sim(C_i, C_j)}{n_i \ n_j} \tag{7}$$

where $sim(C_i, C_j)$ denotes the similarity between any two central objects C_i and C_j, n_i denotes the number of central objects $Clus_i$ contains, and n_j denotes the number of central objects $Clus_j$ contains.

Algorithm 2. Cluster central objects using hierarchical clustering technique.

Input: For a given star-structured heterogeneous data set $D = (C, A, R)$ and similarity matrix of central objects $SimMatrix(C)$, with ε denoting the threshold of similarity between two central object clusters and $n_{CentClus}$ denoting the number of central object clusters.

Output: Central objects clustering result $Clus(C)$.

```
Step 1. for each central object $C_i \in C$ do
            $Clus_i \leftarrow C_i$
Step 2. for $i \leftarrow 1$ to $n_{CentClus}$ do
            for $j \leftarrow 1$ to $n_{CentClus}$ do
                calculate $sim(Clus_i, Clus_j)$
Step 3. get $Clus_i$ and $Clus_j$ with max $sim_{max}(Clus_i, Clus_j)$
Step 4. combine $Clus_i$ and $Clus_j$
Step 5. repeat Step 2 to Step 4 until $sim_{max}(Clus_i, Clus_j) < \varepsilon$
Step 6. return $Clus(C)$
```

2.5. Method for Attribute Object Clustering

It is obvious that the clustering result of central objects is of guidance in attribute object clustering, since attribute objects depend on central objects. Therefore, we group attribute objects according to the cluster assignment of their related central objects, without computing similarity between attribute objects. To achieve this goal, we first define the nearest neighbour of attribute objects in star-structured heterogeneous data, and then assess similarity between attribute objects according to the intersection of their neighbours. The fact that two attribute objects share a lot of common neighbours makes them a good cluster.

Definition 11 (Nearest Neighbour of Attribute Objects in Star-structured Data) For a given star-structured heterogeneous data set $D = (C, A, R)$, the nearest neighbour of the pth attribute object of the kth attribute type, $A_{kp}.NN$, is defined as its directly connected central objects, that is $A_{kp}.NN = \{C_i \mid (r_l.one = C_i \wedge r_l.other = A_{kp}) \vee (r_l.one = A_{kp} \wedge r_l.other = C_i)\}$.

Definition 12 (Similarity between Attribute Objects) For a given star-structured heterogeneous data set $D=(C,A,R)$, the similarity between any two attribute object A_{kp} and A_{kq} ($p \neq q$) of the k th attribute type, $sim(A_{kp}, A_{kq})$, is defined as follows:

$$sim(A_{kp}, A_{kq}) = \frac{\left|A_{kp}.NN.Clus \cap A_{kq}.NN.Clus\right|}{\left|A_{kp}.NN.Clus \cup A_{kq}.NN.Clus\right|} \tag{8}$$

where $A_{kp}.NN.Clus$ denotes the cluster number of $A_{kp}.NN$ contains and $A_{kq}.NN.Clus$ denotes the cluster number of $A_{kq}.NN$ contains.

Algorithm 3. Cluster attribute objects based on clustering result of central objects.

Input: For a given star-structured heterogeneous data set $D=(C,A,R)$ and clustering assignment of central objects $Clus(C)$, with η denoting the threshold of similarity between attribute objects, r denoting the number of attribute object types, and n_k denoting the number of objects of the k th attribute type.

Output: Attribute objects clustering result $Clus(A_k)$ ($k=1,2,...,r$).

```
Step 1. for each attribute object $A_{kp} \in A_k$ ( $k=1,2,...,r$ ) do
            calculate $A_{kp}.NN$
Step 2. for $k \leftarrow 1$ to $r$ do
            for $p \leftarrow 1$ to $n_k$ do
              if $p == 1$ then $Clus(A_{k1}) \leftarrow A_{k1}$
              else
                calculate $sim(A_{k(p+1)}, A_{kp})$
                if $sim(A_{k(p+1)}, A_{kp}) \geqslant \eta$ then $AttrClus(A_{kp}) \leftarrow A_{k(p+1)}$
                else new $AttrClus(A_{k(p+1)}) \leftarrow A_{k(p+1)}$
              $p++$
Step 3. return $Clus(A_k)$( $k=1,2,...,r$ )
```

3. Experimental Results

As mentioned above, bibliographic data is an ideal data source to validate clustering methods for heterogeneous data due to its inherent multi-typed structure. In this paper, we use Chinese academic papers from CNKI (China National Knowledge Infrastructure) as our data set because CNKI is an authoritative full-text website (http://www.cnki.net/) whose databases include almost all of Chinese published academic papers.

Analysing academic papers written by authors from the same social academic unit, such as a school of a university, can help us get the unit's overall academic situation. From China Academic Journals Full-text Database of CNKI, we firstly take 2872 papers affiliated with Dongling School of Economics and Management (DSEM), University of Science and Technology, Beijing, as an example (searched on March 19, 2014). Then, to illustrate StarClusDP, we extract 205 papers with distinct title, author and source, which contain eight most frequent sources as the experimental data. Brief statistics about this data set are shown in Table 1, in which heterogeneity stands for the number of distinct types of objects.

Table 1. Brief Statistics on the Data Set for Starclusdp from CNKI

Heterogeneity	Paper number	Author number	Source number	Total object number
3	205	298	8	511

To be specific, distribution of journals from which these 205 papers come is as follows: 39 papers are from Computer Engineering and Applications (CEA), 32 papers are from Industrial Engineering Journal (IEJ), 30 papers are from China Economist (CE), 25 papers are from Productivity Research (PR), 24 papers are from Metallurgical Financial Accounting (MFA), 20 papers are from Securities & Futures of China (SFC), 18 papers are from Computer Integrated Manufacturing Systems (CIMS), 17 papers are from Finance and Accounting Monthly (FAM).

We implement our method in Java and run it on the data set to test its effectiveness.

(1) Clustering result of central objects

StarClusDP divides papers (central objects of this data set) into 3 groups with $\alpha = 1.000$ and $\varepsilon = 0.003$ (Table 2). After careful analysis, we find that StarClusDP groups papers from SFC, FAM and MFA into the first cluster, papers from IEJ, CEA and CIMS into the second cluster, and papers from CE and PR into the third cluster. The clustering result of these papers are in accord with human cognitive on the periodical range, which shows that StarClusDP can effectively assign similar papers into the same cluster. Besides, from the result it can be drawn that authors from DSEM focus their study on three main aspects: financial, computer engineering, and economics.

Table 2. Clustering Result of Papers

Cluster id	Number of papers	Main focus of the cluster	Sources
1	61	Securities, futures, financial accounting theory and practice.	From SFC, FAM and MFA.
2	89	Industrial engineering, algorithms, computer aided design and manufacturing.	From IEJ, CEA and CIMS.
3	55	Economic theory, economic reform, development of major problems, accounting, and tax.	From CE and PR.

(2) Clustering results of attribute objects

Based on clustering result of papers, StarClusDP ($\eta = 0.900$) detects three clusters of venues: SFC, FAM and MFA; IEJ, CEA and CIMS; CE and PR, which is consistent with scope of these journals. Besides, it gets three clusters of authors, which identifies authors do research on financial (95), computer engineering (113), and economics (90) respectively as an author cluster.

(3) Impact of parameters on clustering results

Parameters in StarClusDP have impacts on both number of central object clusters and number of attribute object clusters. When α is fixed, the number of paper clusters varies according to ε (Table 3). It is shown that with the decrease of ε , which stands for threshold of similarity between paper clusters, the number of paper clusters decreases. This is consistent with the principle of agglomerative hierarchical clustering. While, with different clustering results of central objects, StarClusDP gets different attribute object clusters. It is demonstrated that with the decrease of number of paper clusters, number of venue clusters decreases (Table 3). And by checking scopes of these journals, we find that the merging of journals with the decrease of paper cluster number is consistent with human cognitive. Except clustering result of central objects, we also find that clustering results of attribute objects using StarClusDP ($\alpha = 1.0$ and $\varepsilon = 0.003$) under different values of η (0.1 to 1.0 with 0.1 interval) are exactly the same. It indicates that η has minor effect on clustering result of attribute objects, which is of great help in determining the value of η.

Table 3. Impact of Parameters on Clustering Result of Papers

Id	α	ε (.000)	Number of paper clusters
1	1.0	[0.064,0.500]	8
2	1.0	[0.040,0.063)	7
3	1.0	(0.020,0.040)	6
4	1.0	(0.005,0.020]	5
5	1.0	(0.003,0.005]	4
6	1.0	(0.002,0.003]	3
7	1.0	[0.001,0.002]	2
8	1.0	0.000	1

Table 4. Impact of Clustering Result of Papers on Venue Clustering Result

Number of paper clusters	Number of venue clusters	Composition of venue clusters							
8	8	CEA	CIMS	IEJ	CE	PR	FAM	SFC	MFA
7	7	V1		V2	V3	V4	V5	V6	V7
6	6	V1		V2	V3	V4	V5	V6	6
5	5	V1			V2	V3	V4		V5
4	4	V1			V2	V3	V4		
3	3	V1			V2		V3		
2	2	V1			V2				
1	1	V1							

(4) Efficiency analysis

Last, we test time consumption of each main step of StarClusDP. To eliminate errors, we run StarClusDP on this data set 20 times and take the mediate value as running time. The total procedure takes 3.185s, in which Algorithm 1 takes 0.178s, Algorithm 2 takes 2.521s, and Algorithm 3 takes 0.486s. It can be seen that the clustering of central objects takes almost 80% of the whole time. And clustering of attribute objects only takes 15% time of it, which means it is effective exploiting information from clustering result of central objects.

4. Conclusions

In this paper, we present a three-phase method of utilizing information from all kinds of objects and relations in star-structured heterogeneous data to cluster objects of all types directly. Our method, as demonstrated comprehensively, can effectively detect clusters of objects of all types in the data. Firstly, we introduce a new representation of heterogeneous data with central objects and attribute objects, which lay a foundation for efficiently solving the issue of clustering on star-structured heterogeneous data. Then, we put forward a general framework for clustering all kinds of objects in star-structured heterogeneous data, which includes three phrases: similarity measurement between central objects, central objects clustering and attribute objects clustering. Then, we propose a novel approach for star-structured heterogeneous data clustering based on diffusion path (StarClusDP). Each central object is represented by its connected attribute objects, thus making it possible to efficiently obtain all the necessary information for computing similarity between central objects based on their diffusion path. Then we use hierarchical clustering to accurately group them. At last, we propose a method for clustering attribute objects based on the clustering results of their related central objects, which groups attribute objects quickly according to the cluster assignment of their neighbour central objects. In the experiments, we focus our method on the application scenario of scientific papers. However, as a matter of fact, the proposed method StarClusDP is not confined to scientific papers. To be specific, it can be widely used in other scenarios where there is more than one type of objects. For example, in text mining, where there are documents, words, and authors. However, it is worthwhile that efficiency of our method on larger real-world data sets needs to be further explored.

Acknowledgements

This paper is supported by Science Foundation of Beijing Language and Cultural University (supported by "the Fundamental Research Funds for the Central Universities") (grant no. 15YJ100005) and the National Natural Science Foundation of China (grant no. 71272161).

References

[1] Y. Sun, Y. Yu and J. Han, "Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema", Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, **(2009)**, pp. 797-806.

[2] Y. Sun and J. Han, "Mining Heterogeneous Information Networks: Principles and Methodologies", Morgan & Claypool Publishers, San Rafael, **(2012)**.

[3] L. Tang and H. Liu, "Community Detection and Mining in Social Media", Morgan & Claypool Publishers, San Rafael, **(2010)**.

[4] J. Han, "Mining Heterogeneous Information Networks: the Next Frontier", Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, **(2012)**, pp. 2-3.

[5] Y. Huang and X. Gao, "Clustering on Heterogeneous Networks", WIRES Data Min. Knowledge, vol. 4, no. 3, **(2014)**, pp. 213-233.

[6] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng and T. Wu, "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis", Proceedings of the 12th International Conference on Extending, Saint Petersburg, Russia, **(2009)**, pp.565-576.

[7] G. Getz, E. Levine and E. Domany, "Coupled Two-way Clustering Analysis of Gene Microarray Data", P. Natl. Acad. Sci. USA, vol. 97, no. 22, **(2000)**, pp. 12079-12084.

[8] I. S. Dhillon, "Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning", Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, **(2001)**, pp. 269-274.

[9] B. Long, Z. Zhang, X. Wu and P. S. Yu, "Spectral Clustering for Multi-type Relational Data", Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, **(2006)**, pp. 585-592.

[10] M. Kirsten and S. Wrobel, "Relational Distance-Based Clustering", Proceedings of the 8th International Conference on Inductive Logic Programming, Madison, Wisconsin, USA, **(1998)**, pp. 261-270.

[11] X. Yin, J. Han and P. S. Yu, "Cross-Relational Clustering with User's Guidance", Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, **(2005)**, pp. 344-353.

[12] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity", Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, **(2002)**, pp. 538-543.

[13] D. Fogaras and B. Rácz, "Scaling Link-Based Similarity Search", Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, **(2005)**, pp. 641-650.

[14] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao and W. Ma, "ReCoM: Reinforcement Clustering of Multi-Type Interrelated Data Objects", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, **(2003)**, pp. 274-281.

[15] X. Yin, J. Han and P. S. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links", Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, **(2006)**, pp. 427-438.

[16] P. Zhao, J. Han and Y. Sun, "P-Rank: A Comprehensive Structural Similarity Measure over Information Networks", Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, **(2009)**, pp. 553-562.

[17] Y. Sun, J. Han, X. Yan, P. S. Yu and T. Wu, "PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks", Proceedings of the 37th International Conference on Very Large Data Bases, Seattle, WA, **(2011)**, pp. 992-1003.

[18] C. Shi, X. Kong, P. S. Yu, S. Xie and B. Wu, "Relevance Search in Heterogeneous Networks", Proceedings of the 15th International Conference on Extending Database Technology, Berlin, Germany, **(2012)**, pp. 180-191.

[19] D. R. Cutting, D. R. Karger, J. O. Pedersen and J. W. Tukey, "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections", Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, **(1992)**, pp. 318-329.

[20] O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results", Computer Network, vol. 31, no. 11, **(1999)**, pp.1361-1374.

[21] L. Cao, X. Jin, Z. Yin, A. D. Pozo, J. Luo, J. Han and T. S. Huang, "RankCompete: Simultaneous Ranking and Clustering of Information Networks", Neurocomputing, Special Issue on Learning from Social Media Network, vol. 95, **(2012)**, pp.98-104.

[22] B. Gao, T. Liu, X. Zheng, Q. Cheng and W. Ma, "Consistent Bipartite Graph Co-partitioning for Star-Structured High-order Heterogeneous Data Co-clustering", Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, **(2005)**, pp. 41-50.

Author

Huang Yue, is a librarian at the Library of Beijing Language and Culture University. Her research interests include information management and data mining. Huang has a PhD in management science and engineering from University of Science and Technology, Beijing. Contact her at huang.yuet@blcu.edu.cn.