5-2019

Community discovery in heterogeneous social networks

Lei MENG
Nanyang Technological University

Ah-hwee TAN
Singapore Management University, ahtan@smu.edu.sg

Donald C. WUNSCH
Missouri University of Science and Technology

Citation

MENG, Lei; TAN, Ah-hwee; and WUNSCH, Donald C.. Community discovery in heterogeneous social networks. (2019). *Adaptive Resonance Theory in Social Media Data Clustering*. 137-154.

Available at: https://ink.library.smu.edu.sg/sis_research/6063

Chapter 6
Community Discovery in Heterogeneous Social Networks

Abstract Discovering social communities of web users through clustering analysis of heterogeneous link associations has drawn much attention. However, existing approaches typically require the number of clusters a priori, do not address the weighting problem for fusing heterogeneous types of links, and have a heavy computational cost. This chapter studies the commonly used social links of users and explores the feasibility of the proposed heterogeneous data co-clustering algorithm GHF-ART, as introduced in Sect. 3.6, for discovering user communities in social networks. Contrary to the existing algorithms proposed for this task, GHF-ART performs real-time matching of patterns and one-pass learning, which guarantees its low computational cost. With a vigilance parameter to restrain the intra-cluster similarity, GHF-ART does not need the number of clusters a priori. To achieve a better fusion of multiple types of links, GHF-ART employs a weighting algorithm, called *robustness measure* (RM), to incrementally assess the importance of all the feature channels for the representation of data objects of the same class. Extensive experiments have been conducted on two social network datasets to analyze the performance of GHF-ART. The promising results compare GHF-ART with existing methods and demonstrate the effectiveness and efficiency of GHF-ART. The content of this chapter is summarized and extended from [11] (Copyright ©2014 Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved).

6.1 Introduction

Clustering [17] for discovering communities of users in social networks [19] has been an important task for understanding collective social behavior [21] and associative mining such as social link prediction and recommendation [6, 20]. However, with the popularity of social websites such as Facebook, users may communicate and interact with each other easily and diversely, such as by posting blogs and tagging documents. The availability of that social media data enables the extraction of rich link information among users for further analysis. Alternatively, new challenges have risen for traditional clustering techniques attempting to perform community discovery of social users from heterogeneous social networks in which the users

L. Meng et al., *Adaptive Resonance Theory in Social Media Data Clustering*, Advanced Information and Knowledge Processing,
https://doi.org/10.1007/978-3-030-02985-2_6

are associated by multiple but different types of social links, such as the scalability for large social networks, techniques for link representation, and methods for fusing heterogeneous types of links.

In recent years, many works have been created on the clustering of heterogeneous data. The existing methods may be considered in four categories: multi-view clustering approach [1, 4, 7, 9], spectral clustering approach [10, 12, 15, 23], matrix factorization approach [5, 14] and aggregation approach [2, 13]. However, they all have several limitations for clustering heterogeneous social network data in practice. Firstly, existing algorithms typically involve iterative optimization which does not scale well for big datasets. Secondly, most of them need the number of clusters a priori, which is hard to decide in practice. Thirdly, most of those algorithms do not consider the weighting problem when fusing multiple types of links. Since different types of links have their own meanings and levels of feature values, equal or empirical weights for them may bias their importance in the similarity measure and may not yield a satisfactory performance.

This study explores the feasibility of Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART) for identifying user groups in heterogeneous social networks. As discussed in Sect. 3.6 and Chap. 5, GHF-ART can process social media data that is represented with an arbitrary rich level of heterogeneous data resources such as images, articles and surrounding text. For clustering data patterns of social networks, a set of specific feature representation and learning rules have been developed for GHF-ART to handle various heterogeneous types of social links, including relational links, textual links in articles and textual links in short text.

GHF-ART has several key properties that differ from the existing approaches. First, GHF-ART performs online and one-pass learning so that the clustering process can be done in just a single round of pattern presentation. Second, GHF-ART does not need the number of clusters a priori. Third, GHF-ART employs a weighting function, termed *robustness measure* (RM), which adaptively tunes the weights for different feature channels according to their importance in pattern representation, to achieve a satisfactory level of overall similarity across all the feature channels. Additionally, GHF-ART not only globally considers the overall similarity across all the feature channels, but it also locally evaluates the similarity obtained from each channel. This helps to handle cases when users share some common interests but behave differently in some other aspects.

The performance of GHF-ART was analyzed on two public social network datasets, namely the YouTube dataset [13] and the BlogCatalog dataset [16], through the parameter sensitivity analysis, the clustering performance comparison, the effectiveness evaluation of *robustness measure* and the time cost comparison. The experimental results show that GHF-ART outperforms and is much faster than many existing heterogeneous data clustering algorithms.

6.2 Problem Statement and Formulation

The community discovery problem in heterogeneous social networks is to identify a set of social user groups by evaluating different types of links between users, such that members in the same group interact with each other more frequently and share more common interests than those outside the group.

Considering a set of users $\mathscr{U} = \{u_1, \ldots, u_N\}$ and their associated multiple types of links $\mathscr{L} = \{l_1, \ldots, l_K\}$, such as contact links and subscription links, each user u_n therefore can be represented by a multi-channel input pattern $\mathscr{I} = \{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$, where $\mathbf{x}^k$ is a feature vector extracted from the k-th link.

Consequently, the community discovery task is to identify a set of clusters $\mathscr{C} = \{c_1, \ldots, c_J\}$ according to the similarities among the user patterns evaluated within and across different types of links. As a result, given a user $u_N \in c_J$ and two users $u_p \in c_J$ and $u_q \notin c_J$, for $\{p, q | u_p, u_q \in \mathscr{U}\}$, $S_{u_N,u_p} > S_{u_N,u_q}$, where S_{u_N,u_p} denotes the overall similarity between u_N and u_p. Namely, users in a cluster may consistently be similar in terms of all types of links than those belonging to other clusters.

6.3 GHF-ART for Clustering Heterogeneous Social Links

GHF-ART is designed for clustering composite data objects which are represented by multiple types of features. As introduced in Sect. 3.6, GHF-ART consists of a set of independent feature channels, which can handle an arbitrarily rich level of heterogeneous links. To fit the dynamic changing of social network data, GHF-ART can process input data objects one at a time, during which each of them is either identified as a novel template/prototype, which incurs the generation of a new cluster, or categorized into an existing cluster of similar patterns. In this way, the category space of GHF-ART is incrementally partitioned into regions of clusters.

The following sub-sections will illustrate the key procedures of GHF-ART for clustering social network data, in terms of the representation of commonly used social links, the heterogeneous link fusion for pattern similarity measure, the learning strategies for cluster template generalization, the weighting algorithm for heterogeneous links and algorithm time complexity. The pseudo code of GHF-ART for clustering heterogeneous social links is presented in Algorithm 6.1.

6.3.1 Heterogeneous Link Representation

In GHF-ART, each social user with multi-modal links is represented by a multi-channel input data object $\mathbf{I} = \{\mathbf{x}^k|_{k=1}^K\}$, where $\mathbf{x}^k$ is the feature vector for the k-th feature channel. When presented to GHF-ART, $\mathbf{I}$ undergoes two normalization procedures. First, *min-max normalization* is employed to guarantee that the input

Algorithm 6.1 GHF-ART

```
Input: Input patterns $\mathscr{I}_n = \{\mathbf{x}^k|_{k=1}^K\}$, $\alpha$, $\beta$ and $\rho$.
Present $\mathscr{I}_1 = \{\mathbf{x}^k|_{k=1}^K\}$ to the input field.
Set $J = 1$. Create a node $c_J$ such that $\mathbf{w}_J^k = \mathbf{x}^k$ for $k = 1, \dots, K$.
set $n = 2$.
repeat
    Present $\mathscr{I}_n$ to the input field.
    For $\forall c_j$ $(j = 1, \dots, J)$, calculate the choice function $T(c_j, \mathscr{I}_n)$ according to Eq. (6.2).
    Identify the winner cluster $c_{j^*}$ so that $j^* = \arg\max_{j:c_j \in F_2} T(c_j, \mathscr{I}_n)$. If $j^* = 0$, go to 11.
    Calculate the match function $M(c_{j^*}, \mathbf{x}^k)$ for $k = 1, \dots, K$ according to Eq. (6.3).
    If $\exists k$ such that $M(c_{j^*}, \mathbf{x}^k) < \rho^k$, set $T(c_{j^*}, \mathscr{I}_n) = 0$, $j^* = 0$, go to 7.
    If $j^* \neq 0$, update $\mathbf{w}_{j^*}^k$ for $k = 1, \dots, K$ according to Eqs. (6.4) and (6.5) respectively, and
    update $\gamma$ according to Eqs. (6.6)–(6.7).
    If $j^* = 0$, set $J = J + 1$, create a new node $c_J$ such that $\mathbf{w}_{J+1}^k = \mathbf{x}^k$ for $k = 1, \dots, K$,
    update $\gamma$ according to Eq. (6.8).
    $n = n + 1$.
until All the input patterns are presented.
Output: Cluster Assignment Array $\{A_n|_{n=1}^N\}$.
```

values are in the interval of [0, 1]. Second, For the feature channels using the learning function of Fuzzy ART, *complement coding* [3] normalizes the input feature vector by concatenating $\mathbf{x}^k$ with its complement vector $\bar{\mathbf{x}}^k$ such that $\bar{\mathbf{x}}^k = 1 - \mathbf{x}^k$.

To fit GHF-ART with the social network data, the commonly used social links were divided into three categories, and the respective representation methods were developed accordingly, as discussed below.

6.3.1.1 Density-Based Features for Relational Links

Relational links, such as contact and co-subscription links, use the number of interactions as the strength of the connection between users. Considering a set of users $\mathscr{U} = \{u_1, \dots, u_N\}$, the density-based feature vector of the n-th user u_n is represented by $\mathbf{x}^k = [f_{n,1}, \dots, f_{n,N}]$, wherein $f_{n,i}$ reflects the density of interactions between the user u_n and the ith user u_N.

6.3.1.2 Text-Similarity Features for Articles

Text-similarity features are used to represent the articles of users with long paragraphs such as blogs. Considering a set of users $\mathscr{U} = \{u_1, \dots, u_N\}$ and the word list $\mathscr{G} = \{g_1, \dots, g_M\}$ of all of the M distinct keywords from their articles, the text-similarity feature vector of the n-th user u_n is represented by $\mathbf{x}^k = [f_{n,1}, \dots, f_{n,M}]$, where $f_{n,i}$ indicates the importance of keyword g_i to represent the user u_n, which can be computed by term frequency-inverse document frequency (tf-idf).

6.3.1.3 Tag-Similarity Features for Short Text

Tag-similarity features are used to represent short text, such as tags and comments. Short text from textual articles is unique because the short text consists of a small amount of semantically meaningful words and many noisy ones. Given a set of users $\mathscr{U} = \{u_1, \ldots, u_N\}$ and the corresponding word list $\mathscr{G} = \{g_1, \ldots, g_H\}$ of all the H distinct words, the tag-similarity feature vector of the n-th user u_n is expressed by $\mathbf{x}^k = [f_{n,1}, \ldots, f_{n,H}]$. Following the representation method for meta-information of Probabilistic ART as introduced in Sect. 3.5, given $\mathscr{G}_n$, the word list of the user u_n, the value of its i-th feature $f_{n,i}$ $(i = 1, \ldots, H)$ is given by

$$f_{n,i} = \begin{cases} 1, & if \ g_i \in \mathscr{G}_n \\ 0, & otherwise \end{cases}. \tag{6.1}$$

6.3.2 *Heterogeneous Link Fusion for Pattern Similarity Measure*

GHF-ART selects the best-matching cluster from the input pattern and evaluates the fitness between them through a two-way similarity measure: a bottom-up measure to select the winning cluster by globally considering the overall similarity across all the feature channels and a top-down measure to locally evaluate if the similarity for each feature channel meets the vigilance criteria, defined as

$$T(c_j, \mathbf{I}) = \sum_{k=1}^{K} \gamma^k \frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{\alpha + |\mathbf{w}_j^k|}, \tag{6.2}$$

$$M(c_{j^*}, \mathbf{x}^k) = \frac{|\mathbf{x}^k \wedge \mathbf{w}_{j^*}^k|}{|\mathbf{x}^k|}. \tag{6.3}$$

More discussions on the similarity measure of ART variants can be found in Sects. 3.1.2, 3.6.2 and 5.3.2.

6.3.3 *Learning from Heterogeneous Links*

6.3.3.1 Learning from Density-Based and Text-Similarity Features

The density-based features and textual features for articles use a distribution to represent the characteristics of a user. Therefore, GHF-ART should be able to learn the generalized distribution of similar patterns in the same cluster so that the users with similar feature distribution can be identified.

To this end, the learning function of Fuzzy ART is used, as illustrated in Sect. 3.1. Assuming the k-th feature channel is for density-based features, the corresponding learning function of the winning cluster c_{j*} is therefore defined by

$$\hat{\mathbf{w}}_{j*}^{k} = \beta(\mathbf{x}^{k} \wedge \mathbf{w}_{j*}^{k}) + (1-\beta)\mathbf{w}_{j*}^{k}, \tag{6.4}$$

It was observed that the updated weight values will not be larger than the old ones, so this learning function may incrementally identify the key features by preserving the key features that have stably high values while depressing the features which are unstable in values.

6.3.3.2 Learning from Tag-Similarity Features

The learning function is used for the meta-information, as in Probabilistic ART (Sect. 3.5.2), to model the cluster prototypes for tag-similarity features. Given the feature vector $\mathbf{x}^{k} = [x_1^{k}, \ldots, x_H^{k}]$ of the input pattern $\mathbf{I}$ which encodes short text, the winning cluster c_{j*} with L users and the corresponding weight vector $\mathbf{w}_{j*}^{k} = [w_{j*,1}^{k}, \ldots, w_{j*,H}^{k}]$ of c_{j*} for the k-th feature channel, the learning function for $w_{j*,h}^{k}$ is defined by

$$\hat{w}_{j*,h}^{k} = \begin{cases} \eta w_{j*,h}^{k} & if\ x_h^{k} = 0 \\ \eta(w_{j*,h}^{k} + \frac{1}{L}) & otherwise \end{cases}, \tag{6.5}$$

where $\eta = \frac{L}{L+1}$.

Equation (6.5) models the cluster prototype for the tag-similarity features by the probabilistic distribution of tag occurrences. Thus, the similarity between tag-similarity features can be considered as the number of common words. During each round of learning, the keywords with a high-frequency occurrence in the cluster are given high weights while those of the noisy words are incrementally decreased.

6.3.4 Adaptive Weighting of Heterogeneous Links

GHF-ART employs the *robustness measure* (RM) to adaptively tune the contribution parameter γ^{k} for different feature channels in the choice function (Eq. 6.2), which evaluates the importance of different feature channels by considering the intra-cluster scatters.

As illustrated in Sect. 3.6.3, the *robustness measure* initially gives equal weights to all the feature channels, and then subsequently updates them after the assignment of each input data object according to two scenarios:

- **Resonance in the existing cluster**: Given an existing cluster c_j with L data objects, when a new data object $\mathbf{I}_{L+1}$ is assigned to this cluster, the intra-cluster scatter, called *Difference*, is first computed using

$$\hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(|\mathbf{w}_j^k|D_j^k + |\mathbf{w}_j^k - \hat{\mathbf{w}}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|). \tag{6.6}$$

  where $\eta = \frac{L}{L+1}$.
  Subsequently, the contribution value γ^k is obtained by normalizing a *Robustness* R^k using that of all the feature channels, defined as

$$\gamma^k = \frac{R^k}{\sum_{k=1}^K R^k} = \frac{\exp(-\frac{1}{J}\sum_j D_j^k)}{\sum_{k=1}^K \exp(-\frac{1}{J}\sum_j D_j^k)}. \tag{6.7}$$

  where J is the number of clusters.
- **Generation of new cluster**: When generating a new cluster, the *Difference* of the other clusters remains unchanged. Therefore, the addition of a new cluster just introduces a proportion change to the *Robustness*, which is defined as

$$\hat{\gamma}^k = \frac{\hat{R}^k}{\sum_{k=1}^K \hat{R}^k} = \frac{(R^k)^{\frac{J}{J+1}}}{\sum_{k=1}^K (R^k)^{\frac{J}{J+1}}}, \tag{6.8}$$

6.3.5 Computational Complexity Analysis

The time complexity of GHF-ART with the *robustness measure* has been demonstrated to be $O(n_i n_c n_f)$ in Sect. 3.6.4, where n_i is the number of input patterns, n_c is the number of clusters, and n_f is the total number of features. In comparison with other community detectionalgorithms, the time complexity of LMF [14] is $O(tn_i n_c(n_c + n_f))$, PMM [13] is $O(n_i^3 + tn_c n_i n_f))$, SRC [10] is $O(tn_i^3 + n_c n_i n_f))$ and NMF [5] is $O(tn_c n_i n_f)$, where t is the number of iterations. As observed, GHF-ART has a much lower time complexity.

6.4 Experiments

This section presents an experimental analysis of GHF-ART on the detection of social user communities. Specifically, experiments are conducted on two social network datasets, including the YouTube dataset and the BlogCatalog dataset, in terms of parameter selection, clustering performance comparison, robustness measure for heterogeneous link association analysis and a case study on the discovered user communities.

6.4.1 YouTube Dataset

6.4.1.1 Data Description

The YouTube dataset[1] is a heterogeneous social network dataset, which is originally used to study the community detection problem via the heterogeneous interactions of users. This dataset contains 15, 088 users from the YouTube website and involves five types of relational links, including contact network, co-contact network, co-subscription network, co-subscribed network and favorite network.

6.4.1.2 Evaluation Measure

Since there are no ground truth labels of users in this dataset, the following five evaluation measures were adopted:

1. ***Cross-Dimension Network Validation*** (***CDNV***) [13]: It evaluates how well the cluster structure learned from one or more types of links that fits the network of the other type of links. A larger value indicates a better performance.
2. ***Average Density*** (***AD***): It measures the average probability of two users in the same cluster having a connection, defined by

$$AD = \frac{1}{J}\frac{1}{K}\Sigma_j \Sigma_k \frac{2e_j^k}{n_j(n_j-1)}, \tag{6.9}$$

   where e_j^k is the number of edges of the k-th link in cluster c_j, and n_j is the number of patterns in c_j.
3. ***Intra-cluster sum-of-squared error*** (***Intra-SSE***): It measures the weighted average of SSE within clusters across feature modalities, defined by

$$Intra\text{-}SSE = \Sigma_j \Sigma_{\mathbf{x}_i^k \in c_j} \Sigma_k \frac{n_j}{\Sigma_j n_j}(\mathbf{x}_i^k - \bar{\mathbf{x}}_j^k)^2, \tag{6.10}$$

   where $\mathbf{x}_i^k$ is the feature vector of the i-th pattern for the k-th link, and $\bar{\mathbf{x}}_j^k$ is the mean value of all the $\mathbf{x}_i^k \in c_j$.
4. ***Between-cluster SSE*** (***Between-SSE***): It measures the average distance between two cluster centers to evaluate how well the clusters are separated from each other, defined by

[1] http://socialcomputing.asu.edu/datasets/YouTube.

$$Between\text{-}SSE = \Sigma_j \Sigma_i \Sigma_k \frac{1}{J(J-1)} (\bar{\mathbf{x}}_j^k - \bar{\mathbf{x}}_i^k)^2. \tag{6.11}$$

5. **The ratio of Intra-SSE and Between-SSE** (***SSE-Ratio***): It gives a view of the overall performance, defined by

$$SSE\text{-}Ratio = \frac{Intra\text{-}SSE}{Between\text{-}SSE}. \tag{6.12}$$

6.4.1.3 Parameter Selection Analysis

$\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.6$ were initialized, and the change in the performance of GHF-ART in terms of *SSE-Ratio* was studied by varying one of them while fixing the others, as shown in Fig. 6.1. Despite some small fluctuations, the performance of GHF-ART is roughly robust to the change in the values of α and β. Regarding the vigilance parameter ρ, the performance is improved when ρ increases up to 0.65 and degrades when $\rho > 0.85$. The cluster structures generated under different values ρ were further analyzed, as shown in Fig. 6.2. It was observed that the increase of ρ leads to the generation of more clusters, which may contribute to the compactness of the clusters. At $\rho = 0.9$, a significant number of small clusters are generated, which degrades the performance in terms of recall.

To study the selection of ρ, the cluster structure was analyzed at $\rho = 0.5$ and 0.7, at which the best performance is obtained. When ρ increases from 0.5 to 0.7, the number of small clusters that contain less than 100 patterns increases. Therefore, it is assumed that when a suitable ρ is reached, the number of small clusters starts to increase. If this idea works, an interesting empirical way to select a reasonable value

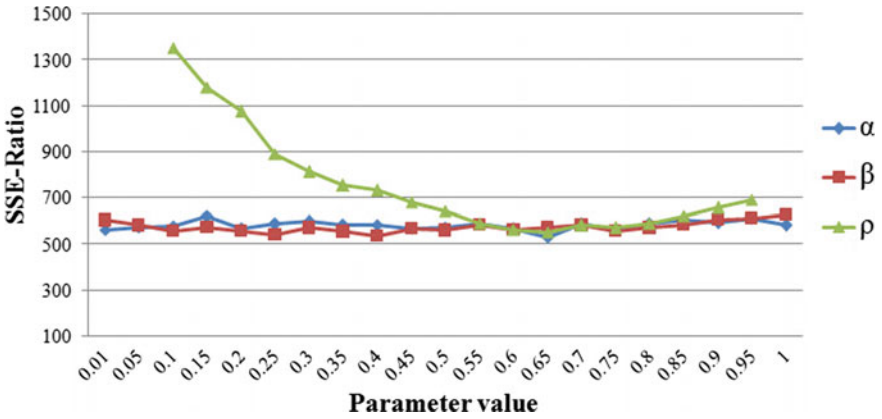

Fig. 6.1 The clustering performance of GHF-ART on the YouTube dataset in terms of *SSE-Ratio* by varying the values of α, β and ρ respectively

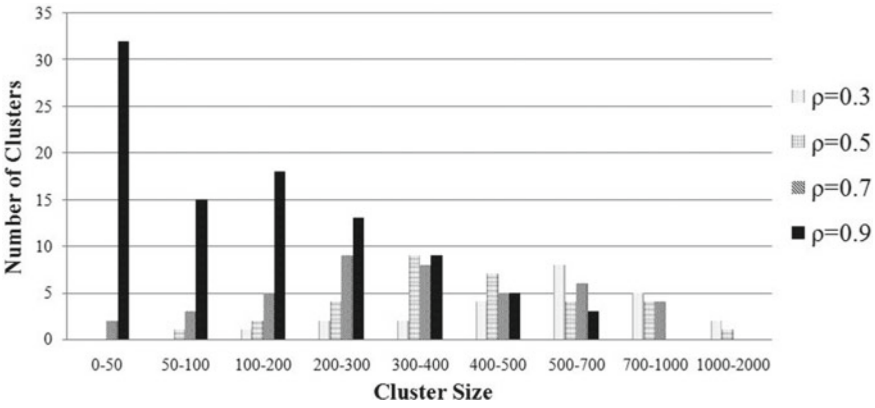

Fig. 6.2 The cluster structures generated by GHF-ART on the Youtube dataset in terms of different values of vigilance parameter ρ

of ρ is to tune the value of ρ until a small number of small clusters, less than 10% of the total number of clusters, are identified.

6.4.1.4 Clustering Performance Comparison

The performance of GHF-ART was compared with four existing heterogeneous data clustering algorithms, namely the Spectral Relational Clustering (SRC) [10], Linked Matrix Factorization (LMF) [14], Non-negative Matrix Factorization (NMF) [5] and Principal Modularity Maximization (PMM) [13]. Since SRC and PMM need K-means to obtain the final clusters, K-means with Euclidean distance was also employed as a baseline.

To make a fair comparison, since GHF-ART needs to perform min-max normalization, the normalized data was applied as the input to the other algorithms. For GHF-ART, $\alpha = 0.01$ and $\beta = 0.6$ were fixed. For K-means,the feature vectors of the five types of links were concatenated. For SRC, the same weight values from GHF-ART were used. The number of iterations for K-means, SRC, LMF, NMF and PMM was set to 50.

The clustering results of GHF-ART were obtained with different values of ρ ranging from 0.3 to 0.9 and those of K-means, SRC, LMF, NMF and PMM with different pre-defined numbers of clusters ranging from 20 to 100. The best performance of each algorithm for each evaluation measure is reported in Table 6.1 and was typically achieved with 34–41 clusters. GHF-ART usually achieves the best performance with $\rho = 0.65$ which is more consistent than other algorithms. GHF-ART outperforms other algorithms in terms of all the evaluation measures except *between-SSE*, but the result of GHF-ART is still comparable to the best one.

Table 6.1 The clustering performance of GHF-ART, K-means, SRC, LMF, NMF and PMM under the best setting of a pre-defined number of clusters ("k") ($\rho = 0.6$ and 0.65 when $k = 35$ and 37 respectively for GHF-ART) in terms of *CDNV*, *Average Density* (*AD*), *Intra-SSE*, *Between-SSE* and *SSE-Ratio* on the YouTube dataset

	CDNV		*AD*		*Intra-SSE*		*Between-SSE*		*SSE-Ratio*	
	Value	k	Value	k	Value	k	Value	k	Value	k
K-means	0.2446	43	0.0572	40	7372.4	41	9.366	40	774.14	41
SRC	0.2613	37	0.0691	35	6593.6	36	10.249	35	652.34	36
LMF	0.2467	39	0.0584	38	6821.3	41	9.874	37	694.72	40
NMF	0.2741	36	0.0766	35	6249.5	36	**10.746**	34	591.57	35
PMM	0.2536	36	0.0628	37	6625.8	37	9.627	34	702.25	35
GHF-ART	**0.2852**	37	**0.0834**	37	**5788.6**	37	10.579	35	**563.18**	37

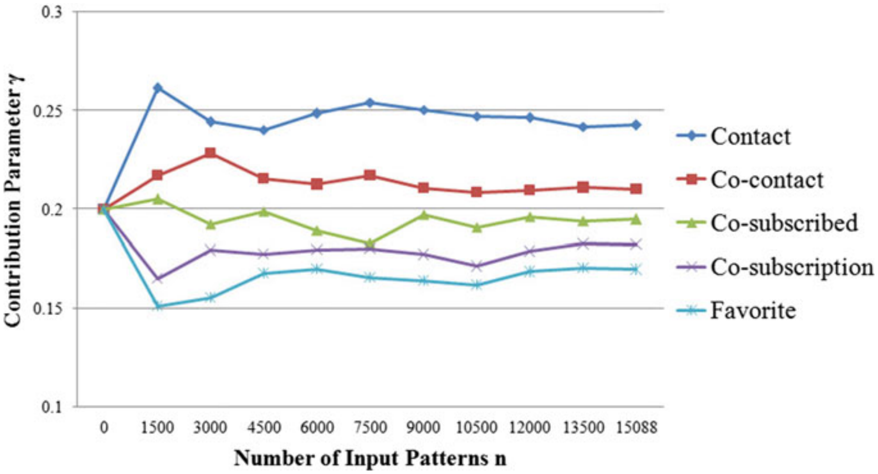

Fig. 6.3 Trace of contribution parameters for five types of links during clustering with an increase in the number of input patterns

6.4.1.5 Correlation Analysis of Heterogeneous Networks

GHF-ART was first run under $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.65$ and showed the trace of contribution parameters for each type of link during clustering in Fig. 6.3.The weights for all types of features begin with 0.2. The initial fluctuation at $n = 1,500$ is due to the incremental generation of new clusters. After $n = 12,000$, the weight values for all types of features become stable.

The probability of pairs of connected patterns falling into the same cluster was further analyzed to determine how each type of relational network affects the clustering results, as shown in Fig. 6.4. It was observed that the order of relational networks is consistent with the results shown in Fig. 6.3. This demonstrates the validity of

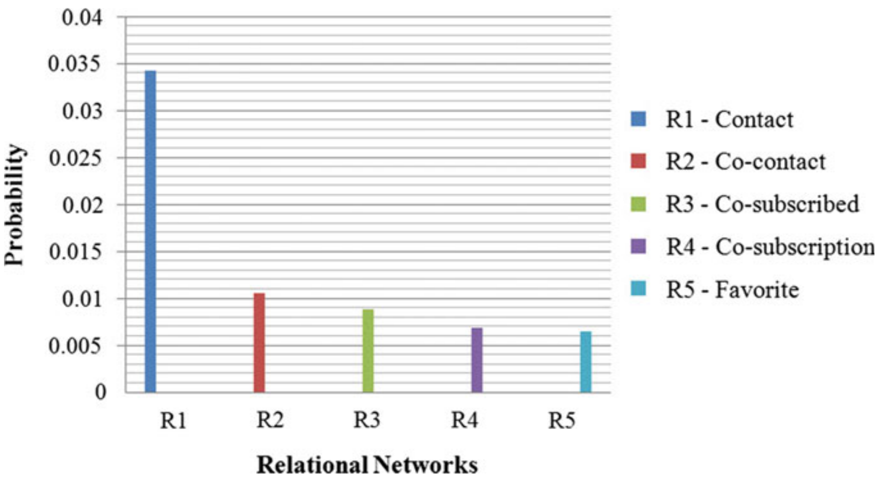

Fig. 6.4 The probability that pairs of patterns falling into the same cluster are connected in each of the five relational networks

robustness measure. Among all types of links, the contact network achieves a much higher probability than other relational networks. This may be due to the contact network being much sparser than the other four networks. As such, it is expected that the links of the contact network are more representative.

6.4.2 BlogCatalog Dataset

6.4.2.1 Data Description

The BlogCatalog dataset[2] is crawled in [16] and used for discovering the overlapping social groups of users. It consists of the raw data of 88, 784 users, each of which involves the friendship to other users and the published blogs. Each blog of a user is described by several pre-defined categories, user-generated tags and six snippets of blog content.

Three types of links were extracted, including a friendship network and two textual similarity networks in terms of blog content and tags. By filtering infrequent words from tags and blogs, 66, 418 users, 6, 666 tags and 17, 824 words from blogs were obtained. As suggested in [16], the most frequent category in the blogs of a user was used as the class label and a total of 147 class labels were obtained.

[2] http://dmml.asu.edu/users/xufei/datasets.html#Blogcatalog.

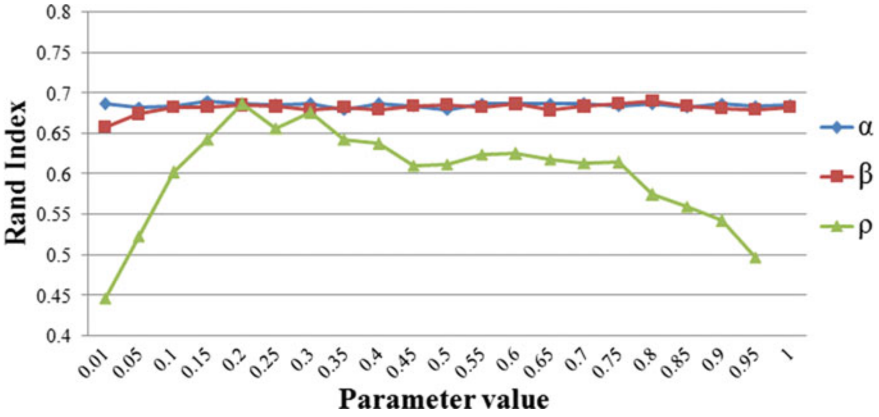

Fig. 6.5 The clustering performance of GHF-ART on the BlogCatalog dataset in terms of rand index by varying the values of α, β and ρ respectively

6.4.2.2 Evaluation Measure

With the ground truth labels, *Average Precision* (*AP*), *Cluster Entropy* and *Class Entropy* [8], *Purity* [22] and *Rand Index* [18] were used as the clustering evaluation measures. *Average Precision*, *Cluster Entropy* and *Purity* evaluate the intra-cluster compactness. *Class Entropy* evaluates how well the classes are represented by the minimum number of clusters. *Rand Index* considers both cases.

6.4.2.3 Parameter Selection Analysis

The influence of parameters on the performance of GHF-ART was studied for the BlogCatalog dataset with the initial settings of $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.2$, as shown in Fig. 6.5. It was observed that, consistent with those in Fig. 6.1, the performance of GHF-ART is robust to the change in the choice and learning parameters. As expected, the performance of GHF-ART varies a lot due to the change in ρ. This curve may also be explained in the same way as that in Fig. 6.1.

To validate the findings and to select a suitable ρ in Sect. 6.4.1.3, the cluster structures corresponding to the four key points of ρ were analyzed, as shown in Fig. 6.6. It was observed that, at $\rho = 0.2$, nearly 20 small clusters with less than 100 patterns were generated. Interestingly, the number of small clusters was also around 10% of the total number of clusters, which fits the findings from the the YouTube dataset. This demonstrates the feasibility of the proposed empirical way of selecting a suitable value of ρ.

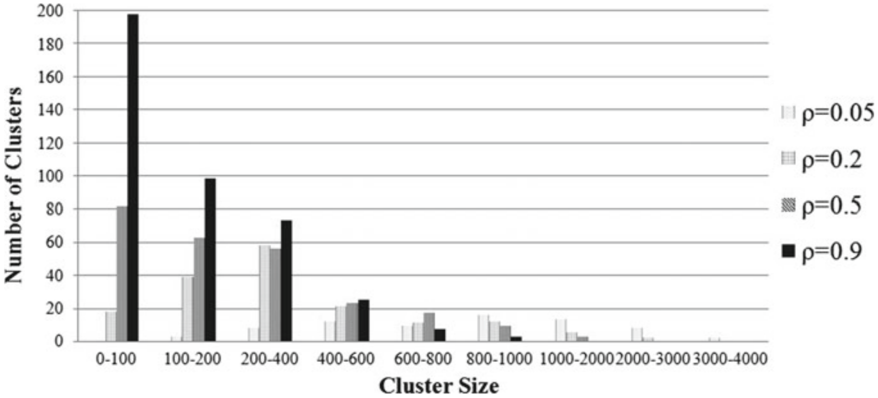

Fig. 6.6 The cluster structures generated by GHF-ART on the BlogCatalog dataset in terms of different values of vigilance parameter ρ

6.4.2.4 Clustering Performance Comparison

The performance of GHF-ART was compared with the same set of algorithms compared in the YouTube dataset, under the same parameter settings mentioned in Sect. 6.4.1.4, except the number of clusters. The value of ρ was varied from 0.1 to 0.4 with an interval of 0.05, and the number of clusters was varied from 150-200 with an interval of 5.

The best performance for each algorithm with the number of clusters is shown in Table 6.2. GHF-ART obtained a much better performance (at least a 4% improvement) than the other algorithms in terms of *Average Precision*, *Cluster Entropy* and *Purity*. This indicates that GHF-ART may identify similar patterns well and produce more compact clusters. Competitive performance is obtained by SRC and NMF in terms of *Class Entropy*. Considering the number of clusters under the best settings, it was found that GHF-ART identifies a similar number of clusters to other algorithms, which demonstrates the effectiveness of GHF-ART.

6.4.2.5 Case Study

The communities identified by GHF-ART were further studied. First, details of the five biggest clusters discovered are listed, as shown in Table 6.3. Those clusters are well-formed to reveal the user communities since more than $1,000$ patterns are grouped with a reasonable level of precision. Additionally, most of the top tags discovered by the cluster weight values are semantically related to their corresponding classes. Interestingly, the clusters ranked 1 and 4 belong to the class "Personal". This may be because, according to the organized statistics, "Personal" is much larger than the other classes. However, in the top-5 tags, only "life" is shared by them. To gain insight of the relationship between these two clusters, their tag clouds are

Table 6.2 The clustering performance of GHF-ART, K-means, SRC, LMF, NMF and PMM under the best setting of a pre-defined number of clusters ("k") ($\rho = 0.15$, 0.2 and 0.25 when $k = 158$, 166 and 174 respectively for GHF-ART) on the BlogCatalog dataset in terms of *Average Precision* (*AP*), *Cluster Entropy* ($H_{cluster}$), *Class Entropy* (H_{class}), *Purity* and *Rand Index*(*RI*)

	AP		$H_{cluster}$		H_{class}		*Purity*		*RI*	
	Value	*k*	Value	*k*	Value	*k*	Value	*k*	Value	*k*
K-means	0.6492	185	0.5892	185	0.5815	165	0.6582	185	0.5662	170
SRC	0.7062	175	0.5163	175	0.4974	160	0.7167	175	0.6481	170
LMF	0.6626	175	0.5492	175	0.5517	155	0.6682	175	0.6038	165
NMF	0.7429	175	0.4836	175	0.4883	155	0.7791	175	0.6759	165
PMM	0.6951	170	0.5247	170	0.5169	165	0.6974	170	0.6103	165
GHF-ART	**0.7884**	174	**0.4695**	174	**0.4865**	158	**0.8136**	174	**0.6867**	166

Table 6.3 The five biggest clusters identified by GHF-ART with class labels, top tags, cluster size and *Precision*

Cluster rank	Class label	Top tags	Cluster size	*Precision*
1	Personal	Music, life, art, movies, Culture	2692	0.7442
2	Blogging	News, blog, blogging, SEO, Marketing	2064	0.8166
3	Health	Health, food, beauty, weight, diet	1428	0.7693
4	Personal	Life, love, travel, family, friends	1253	0.6871
5	Entertainment	Music, movies, news, celebrity, funny	1165	0.6528

plotted below. As shown in Fig. 6.7, the two clusters share many key tags such as "love", "travel", "personal" and "film". Furthermore, when looking into the large number of smaller tags in the clouds, it was found that such tags in Fig. 6.7a are more related to "music" and enjoying "life", such as "game", "rap" and "sport", while those in Fig. 6.7b are more related to "family" life, such as "kids", "parenting" and "wedding". Therefore, although the shared key tags indicate their strong relations to the same class "Personal", they are separated into two communities due to the differences in the sub-key tags.

6.4.2.6 Time Cost Analysis

To evaluate the efficiency of GHF-ART on big data, the time cost of GHF-ART, K-means, SRC, LMF, NMF and PMM with the increased number of input patterns was further analyzed. To make a fair comparison, the number of clusters was set to $k = 166$ for K-means, SRC, LMF, NMF and PMM and $\rho = 0.2$ for GHF-ART so that the numbers of the generated clusters for all the algorithms were the same. Figure 6.8

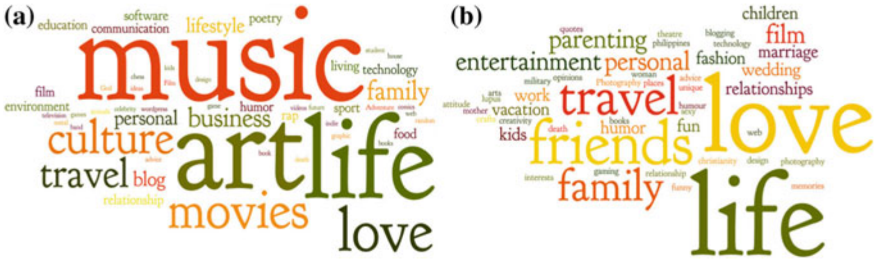

Fig. 6.7 The tag clouds generated for the **a** 1st and **b** 4th biggest clusters. A larger font of tag indicates a higher weight in the cluster

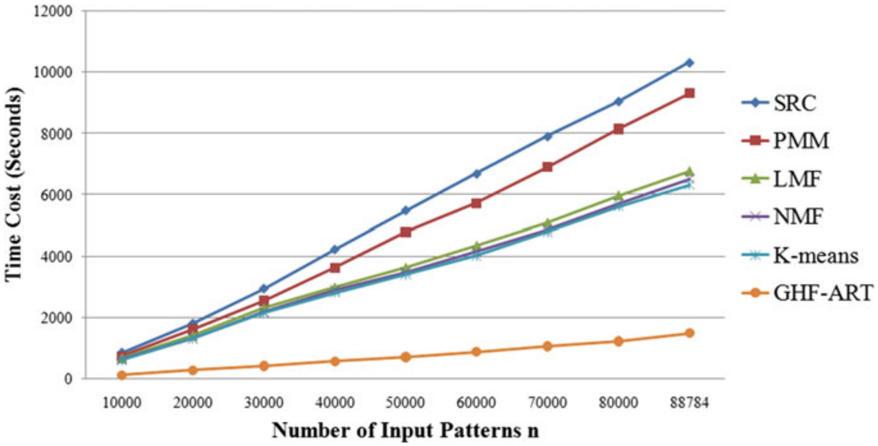

Fig. 6.8 Time cost of GHF-ART, K-means, SRC, LMF, NMF and PMM on the BlogCatalog Dataset with the increase in the number of input patterns

shows that GHF-ART runs much faster than the other algorithms. Additionally, the other algorithms incur a great increase in the time cost with the increase in the number of input patterns, but GHF-ART maintains a relatively small increase. This demonstrates the scalability of GHF-ART for big data.

6.5 Discussion

This chapter discusses the task of community discovery in social networks using Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART) for the fusion and analysis of heterogeneous types of social links. Specifically, it considers three types of commonly used social links, namely the relational links, the textual links in articles and the textual links in short text. For each type of social link, GHF-ART incorporates specific feature representations and the corresponding pattern matching and learning strategies. Through the two-way similarity measure with choice and match functions, GHF-ART is able to identify user communities by globally and locally evaluating the similarity between users within and across all types of social links.

Compared with existing work [5, 10, 13, 14] for clustering heterogeneous social networks, GHF-ART has an advantage in four aspects:

1. **Scalability for big data**: GHF-ART employs an incremental and fast learning method which results in a linear time complexity so that GHF-ART is capable of efficiently learning from a large social network.
2. **Considering heterogeneity of links**: Different from existing approaches which consider more about the methods for integrating multiple links, GHF-ART is able to employ different representation and learning strategies for different types of links for a better similarity evaluation in terms of each link.
3. **Incorporating global and local similarity evaluation**: Existing co-clustering algorithms consider only the fusion of multiple links for an overall similarity measure so that two users may be grouped into the same cluster even if they are dissimilar in some of the links. In contrast, GHF-ART employs a two-way similarity measure wherein a bottom-up evaluation first globally considers the overall similarity to identify the most similar cluster, and then a top-down evaluation locally checks if the similarity for each link meets a threshold.
4. **Weighting algorithm for link fusion**: In order to make a better fusion of multiple links for the overall similarity measure of patterns, GHF-ART has a well-defined weighting algorithm which may adapt the weights for the features of each link by evaluating the intra-cluster scatter during the clustering process so that the features which are more prominent in representing the characteristics of a pattern in the same cluster will be assigned higher weight values.

The incremental clustering nature of GHF-ART makes it possible to process a very large social network that can hardly be processed by a single computer, and its *robustness measure* provides a good tool for link association analysis. Beyond the progress achieved in this study so far, there are several interesting directions worth further investigation. First, as GHF-ART uses feature vectors to represent social links, the dimension of those for relational networks are the number of users, which results in a high space complexity. Therefore, feature reduction techniques or hashing methods are preferred to reduce computer consumption. Second, as more types of communication methods emerge, there will be a future need to consider and utilize many more social links for user community or interest profiling. Thus, it is interesting to construct such a social network dataset and investigate the feasibility of GHF-ART for the associative mining tasks.

References

1. Bickel S, Scheffer T (2004) Multi-view clustering. In: ICDM, pp 19–26
2. Bisson G, Grimal C (2012) Co-clustering of multi-view datasets: a parallelizable approach. In: ICDM, pp 828–833
3. Carpenter GA, Grossberg S, Rosen DB (1991) Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Netw 4:759–771
4. Chaudhuri K, Kakade SM, Livescu K, Sridharan K (2009) Multi-view clustering via canonical correlation analysis. In: ICML, pp 129–136
5. Chen Y, Wang L, Dong M (2010) Non-negative matrix factorization for semisupervised heterogeneous data coclustering. TKDE 22(10):1459–1474
6. Dong Y, Tang J, Wu S, Tian J, Chawla NV, Rao J, Cao H (2012) Link prediction and recommendation across heterogeneous social networks. In: ICDM, pp 181–190
7. Drost I, Bickel S, Scheffer T (2006) Discovering communities in linked data by multi-view clustering. From data and information analysis to knowledge engineering. Springer, Berlin, pp 342–349
8. He J, Tan AH, Tan CL, Sung SY (2003) On quantitative evaluation of clustering systems. Clustering and information retrieval. Kluwer Academic Publishers, Netherlands, pp 105–133
9. Kumar AIII, Daumé H (2011) A co-training approach for multi-view spectral clustering. In: ICML, pp 393–400
10. Long B, Wu X, Zhang Z, Yu PS (2006) Spectral clustering for multi-type relational data. In: ICML, pp 585–592
11. Meng L, Tan AH (2014) Community discovery in social networks via heterogeneous link association and fusion. In: SIAM international conference on data mining (SDM), pp 803–811
12. Rege M, Dong M, Hua J (2008) Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In: Proceedings of international conference on world wide web, pp 317–326
13. Tang L, Wang X, Liu H (2009) Uncovering groups via heterogeneous interaction analysis. In: ICDM, pp 503–512
14. Tang W, Lu Z, Dhillon IS (2009) Clustering with multiple graphs. In: ICDM, pp 1016–1021
15. Wang X, Qian B, Ye J, Davidson I (2013) Multi-objective multi-view spectral clustering via Pareto optimization. In: SDM, pp 234–242
16. Wang X, Tang L, Gao H, Liu H (2010) Discovering overlapping groups in social media. In: ICDM, pp 569–578
17. Whang JJ, Sui X, Sun Y, Dhillon IS (2012) Scalable and memory-efficient clustering of large-scale social networks. In: ICDM, pp 705–714
18. Xu RII, Wunsch DC (2011) BARTMAP: a viable structure for biclustering. Neural Netw 24:709–716
19. Yang J, Leskovec J (2012) Defining and evaluating network communities based on ground-truth. In: SDM, pp 745–754
20. Yang Y, Chawla N, Sun Y, Han J (2012) Predicting links in multi-relational and heterogeneous networks. In: ICDM, pp 755–764
21. Zhang K, Lo D, Lim EP, Prasetyo PK (2013) Mining indirect antagonistic communities from social interactions. Knowl Inf Syst 35(3):553–583
22. Zhao Y, Karypis G (2001) Criterion functions for document clustering: experiments and analysis. Technical report, Department of Computer Science, University of Minnesota
23. Zhou D, Burges CJC (2007) Spectral clustering and transductive learning with multiple views. In: ICML, pp 1159–1166