4-2014

Community discovery in social networks via heterogeneous link association and fusion

Lei MENG

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

Community Discovery in Social Networks via Heterogeneous Link Association and Fusion

Lei Meng * Ah-Hwee Tan *

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 prior, do not address the weighting problem for fusing heterogeneous types of links and have a heavy computational cost. In this paper, we explore the feasibility of a newly proposed heterogeneous data clustering algorithm, called Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART), for discovering communities in heterogeneous social networks. Different from existing algorithms, GHF-ART performs real-time matching of patterns and one-pass learning which guarantee its low computational cost. With a vigilance parameter to restrain the intra-cluster similarity, GHF-ART does not need the number of clusters a prior. To achieve a better fusion of multiple types of links, GHF-ART employs a weighting function to incrementally assess the importance of all the feature channels. Extensive experiments have been conducted to analyze the performance of GHF-ART on two heterogeneous social network data sets and the promising results comparing with existing methods demonstrate the effectiveness and efficiency of GHF-ART.

1 Introduction

Clustering [4] for discovering communities in social networks [1], aiming at identifying groups of users with common interests and behavior, has been an important task for the understanding of collective social behavior and associative mining such as social link prediction and recommendation [2, 3]. However, with the popularity of social websites such as Facebook, users may communicate and interact with each other easily and diversely, such as posting blogs and tagging documents. The availability of those social media data, on one hand, facilitates the extraction of rich link information among users for further analysis. On the other hand, new challenges have been identified for traditional clustering techniques on community discovery from heterogeneous social networks, such as the scalability to large social networks, techniques for link representation and methods for the fusion of heterogeneous types of links.

In the recent years, many works have been done on the clustering of heterogeneous data. Existing methods may be considered in four categories: multi-view clustering approach [5, 12, 13, 17], spectral clustering approach [11, 14, 21, 23], matrix factorization approach [6, 15], aggregation approach [8, 9]. However, they have several limitations for clustering heterogeneous social links in practice. Firstly, existing algorithms typically involve iterative optimization which does not scale well to big data sets. Secondly, most of them need the number of clusters a prior, 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 similarity measure and may not yield optimal performance.

In this paper, we explore the feasibility of Generalized Heterogeneous Fusion Adaptive Resonance Theory (GHF-ART) for identifying user groups in the heterogeneous social networks. GHF-ART [10], extended from Fusion ART [16], has been proposed for clustering web multimedia data through the fusion of an arbitrary rich level of heterogeneous data resources such as images, articles and surrounding text. For clustering data patterns of social networks, we develop a set of specific feature representation and learning rules 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 different from 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 prior. Third, GHF-ART globally and locally evaluates the similarity between patterns across and in each feature channel. Fourth, the obtained similarities from all the feature channels are fused by a weighting function, termed *Robustness Measure* (RM), which adaptively tunes the weights of different feature channels.

*School of Computer Engineering, Nanyang Technological University, Singapore, Email: {meng0027, asahtan}@ntu.edu.sg

We analyze the performance of GHF-ART on two public social network data sets, namely the YouTube data set [8] and the BlogCatalog data set [7], in terms of parameter selection, clustering performance comparison, performance in *Robustness Measure* and time cost. From our experimental results, we analyze the parameter selection methods for GHF-ART and show that GHF-ART outperforms and is much faster than many existing heterogeneous data clustering algorithms.

The remainder of paper is summarized as follows. Section 2 reviews existing works on the problem of heterogeneous data clustering. Section 3 formulates the problem of community discovery in the heterogeneous social networks. The technical details of GHF-ART are described in Section 4. Section 5 presents the analysis of experimental results and the last section concludes our work.

2 Related Work

Our work on identifying social groups of users via heterogeneous social links is related to the problem of heterogeneous data clustering. Considering different model formulation, existing approaches can be categorized into four categories: 1) **The multi-view clustering approach** [5, 12, 13, 17] considers to use two clustering models for two types of independent features. Subsequently, the learnt parameters of them are further refined by learning from each other iteratively. However, this approach is restricted to two types of links. 2) **The spectral clustering approach** [11, 14, 21, 23] typically models each feature modality as a graph and uses different unified objective function to identify an overall best cut of the graphs, which is typically an embedding vector and needs traditional clustering algorithms to obtain the final results. 3) **The Matrix factorization approach** [6, 15] factorizes a similarity matrix into two or three matrices and minimize the reconstruction error objective, which identifies the cluster membership of patterns by finding a cluster indicator matrix that contains the projection values of each data pattern to a pre-defined number of clusters. 4) **The aggregation approach** [8, 9] follows the idea of first obtaining the relational vectors [8] or similarities [9] between patterns for each type of features and then integrating them to produce the final results.

3 Problem Statement

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

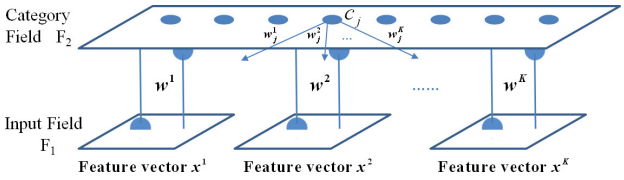

Figure 1: The architecture of GHF-ART for integrating K types of feature vectors.

Considering a set of users $\mathcal{U} = \{u_1, \ldots, u_N\}$ and their associated multiple types of links $\mathcal{L} = \{l_1, \ldots, l_K\}$, which, as described in section 1, may be the contact links, textual links or visual links. Each user u_n can be represented by a multi-channel input pattern $\mathcal{I} = \{\mathbf{x}^1, \ldots, \mathbf{x}^K\}$, where $\mathbf{x}^k$ is a feature vector extracted from the k-th link.

The community discovery task, in this work, is to identify a set of clusters $\mathcal{C} = \{c_1, \ldots, c_J\}$ 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 $\forall p, q$ such that $u_p, u_p \in \mathcal{U}$, we have $S_{u_N,u_p} > S_{u_N,u_q}$, where S_{u_N,u_p} denotes the overall similarity between u_N and u_p. Namely, users in the same cluster may consistently have a higher degree of similarity in terms of all types of links than those belonging to the other clusters.

4 GHF-ART for Clustering Heterogeneous Social Links

As shown in Fig. 1, GHF-ART consists of K independent feature channels in the input field designated to handle an arbitrarily rich level of heterogeneous links and a category field. The clustering process of GHF-ART processes one input pattern at a time, which comprises four key steps: 1) **Category choice**: select a best-matching cluster, called a winner, across all the feature channels; 2) **Template matching**: Evaluate if the degree of similarity between the input pattern $\mathcal{I}$ and the winner satisfies a threshold, called the vigilance criteria, for each feature channel; 3) **Resonance and Reset**: If the vigilance criteria is violated, a reset occurs so that a new winner is selected from the rest of the clusters in the category field; Otherwise, a resonance occurs which leads to the learning of winner from the input pattern for all the feature channels. 4) **Network Expansion**: If no cluster meets the vigilance criteria, a new cluster is generated to encode the new pattern. The dynamics of GHF-ART is summarized as follows.

Input vectors: Let $\mathcal{I} = \{\mathbf{x}^k|_{k=1}^K\}$ denote the multi-channel input pattern, where $\mathbf{x}^k$ is the feature vector for the k-th feature channel. Note that the min-max normalization should be employed to make sure that the

input values are in the interval [0, 1]. The complement coding [18] is used to normalize the input feature vector through which $\mathbf{x}^k$ is concatenated with its complement vector $\bar{\mathbf{x}}^k$ in the input field such that $\bar{\mathbf{x}}_i^k = 1 - \mathbf{x}_i^k$.

Weight vectors: Let $\{\mathbf{w}_j^k|_{k=1}^K\}$ denote the weight vectors associated with the j-th cluster c_j in the category field F_2.

Parameters: The GHF-ART's dynamics is determined by choice parameter $\alpha > 0$, learning parameter $\beta \in [0,1]$, contribution parameters $\gamma^k \in [0,1]$ for $k = 1,\ldots,K$ and vigilance parameters $\rho \in [0,1]$.

4.1 Heterogeneous Link Representation

4.1.1 Density-based Features for Relational Links Relational links, such as contact and co-subscription links, uses the number of interactions as the strength of connection between users. Considering a set of users $\mathcal{U} = \{u_1,\ldots,u_N\}$, each user u_n is represented by a feature vector $\mathbf{FD}_n = [f_{n,1},\ldots,f_{n,N}]$, wherein $f_{n,N}$ reflects the density of interactions between the user u_n and the N-th user in the user set $\mathcal{U}$.

4.1.2 Text-similarity Features for Articles Text-similarity features are used to represent the articles of users with long paragraphs such as blogs. Considering the word list $\mathcal{G} = \{g_1,\ldots,g_M\}$ of all the M distinct keywords from the articles of a set of users $\mathcal{U} = \{u_1,\ldots,u_N\}$, the feature vector of u_n can be represented by $\mathbf{FA}_n = [f_{n,1},\ldots,f_{n,M}]$, where $f_{n,M}$ indicates the importance of keyword g_M to represent the user u_n, which can be valued by the term frequency-inverse document frequency (tf-idf).

4.1.3 Tag-similarity Features for Short Text Tag-similarity features are used to represent short text, such as tags and comments. The key difference of short text from article is that short text consists of few but meaningful words. Considering a set of user $\mathcal{U} = \{u_1,\ldots,u_N\}$ and the corresponding word list $\mathcal{G} = \{g_1,\ldots,g_H\}$ of all the H distinct tags, the feature vector of user u_n can be expressed by $\mathbf{FS}_n = [f_{n,1},\ldots,f_{n,H}]$. Following the representation method for meta-information in [10], $f_{n,h}$ for $h = 1,...,H$ is given by

$$f_{S,h} = \begin{cases} 1, & if\ g_h \in \mathcal{G}_n \\ 0, & otherwise \end{cases}. \tag{4.1}$$

4.2 Heterogeneous Link Fusion for Similarity Measure

GHF-ART measures the similarity between the input pattern and each cluster in the category field through a two-way similarity measures: a bottom-up measure to select a winner cluster by considering the overall similarity across all the feature channels; and a top-down measure to evaluate if the similarity for each feature channel meets the vigilance criteria threshold.

4.2.1 Bottom-Up Similarity for Category Choice In the first step, a choice function is employed to evaluate the overall similarity between the input pattern and the template weight of each cluster in the category field, which is defined by

$$T(c_j,\mathcal{I}) = \sum_{k=1}^{K} \gamma^k \frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{\alpha + |\mathbf{w}_j^k|}, \tag{4.2}$$

where the fuzzy AND operation $\wedge$ is defined by $(\mathbf{p} \wedge \mathbf{q})_i \equiv \min(\mathbf{p}_i, \mathbf{q}_i)$, and the ℓ_1 norm $|.|$ is defined by $|\mathbf{p}| \equiv \sum_i \mathbf{p}_i$. The choice function evaluates the proportion of intersection between the feature vectors of the input pattern and the prototypes of the winner across all the feature channels so that the winner cluster with the best matching feature distribution in the category field is identified.

4.2.2 Top-Down Similarity for Template Matching After identifying the winner cluster, a match function is used to evaluate if the selected winner matches the input pattern in terms of each feature channel. For the k-th feature channel, the match function is defined by

$$M(c_{j^*},\mathbf{x}^k) = \frac{|\mathbf{x}^k \wedge \mathbf{w}_{j^*}^k|}{|\mathbf{x}^k|}. \tag{4.3}$$

If the match function values for all the K feature channels satisfies the vigilance criteria defined by $M(c_{j^*},\mathbf{x}^k) > \rho$ for $k = 1,...K$, a resonance occurs so that the input pattern is categorized into the winner cluster. Otherwise, a reset occurs to select a new winner from the rest of the clusters in the category field.

4.3 Learning from Heterogeneous Links

4.3.1 Learning from Density-based and Text-similarity Features Assuming the k-th feature channel is for the density-based features, the learning function for the k-th feature channel of the winner cluster c_{j^*} is defined by

$$\hat{\mathbf{w}}_{j^*}^k = \beta(\mathbf{x}^k \wedge \mathbf{w}_{j^*}^k) + (1-\beta)\mathbf{w}_{j^*}^k. \tag{4.4}$$

We can observe that the values of the new weight vector will not be greater than the old ones so that this learning function may incrementally identify the key features by preserving the key features which have stably high values while depressing the features which are unstable in values.

4.3.2 Learning from Tag-similarity Features Assuming the k-th feature channel is for the tag-sinilarity features of short text, given the k-th feature vector $\mathbf{x}^k = [x_1^k, \ldots, x_H^k]$ of the input pattern $\mathcal{I}$, the winner 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

$$\hat{w}_{j^*,h}^k = \begin{cases} \eta w_{j^*,h}^k & if\ x_h^k = 0 \\ \eta(w_{j^*,h}^k + \frac{1}{L}) & otherwise \end{cases}, \tag{4.5}$$

where $\eta = \frac{L}{L+1}$. (4.5) 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 high frequency to occur in the cluster are given high weights while those of the noisy words are incrementally decreased.

4.4 Adaptive Weighting of Heterogeneous Links GHF-ART employs the *Robustness Measure* (*R-M*) to adaptively tune γ for different feature channels, which evaluates the importance of different feature channels by considering the intra-cluster scatters.

Considering a cluster c_j with L users, each of which is denoted by $\mathcal{I}_l = \{\mathbf{x}_l^1, \ldots, \mathbf{x}_l^K\}$ for $l = 1, \ldots, L$ and the corresponding weight vectors for the K feature channels denoted by $\mathcal{W}_j = \{\mathbf{w}_j^1, \ldots, \mathbf{w}_j^K\}$, the *Difference* for the k-th feature channel of c_j is defined by

$$D_j^k = \frac{\frac{1}{L}\sum_l |\mathbf{w}_j^k - \mathbf{x}_l^k|}{|\mathbf{w}_j^k|}. \tag{4.6}$$

Considering all the clusters, the contribution parameter for the k-th feature channel γ^k is defined by

$$\gamma^k = \frac{\exp(-\frac{1}{J}\sum_j D_j^k)}{\sum_{k=1}^K \exp(-\frac{1}{J}\sum_j D_j^k)}. \tag{4.7}$$

The respective incremental update equations for the contribution parameters are further derived for the following two cases:

- **Resonance in existing cluster**: Assuming the input pattern $\mathcal{I}_{L+1} = \{\mathbf{x}_{L+1}^1, \ldots, \mathbf{x}_{L+1}^K\}$ is assigned to an existing cluster c_j. For the k-th feature channel, the corresponding update equations for the density-based and text-similarity features and tag-similarity features are defined by (4.8) and (4.9) respectively:

  (4.8)
  $$\hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(|\mathbf{w}_j^k| D_j^k + |\mathbf{w}_j^k - \hat{\mathbf{w}}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|)$$

  (4.9)
  $$\hat{D}_j^k = \frac{\eta}{|\hat{\mathbf{w}}_j^k|}(\eta D_j^k + |\hat{\mathbf{w}}_j^k - \eta\mathbf{w}_j^k| + \frac{1}{L}|\hat{\mathbf{w}}_j^k - \mathbf{x}_{L+1}^k|).$$

  After the update for all feature channels, the new contribution parameter can then be obtained by calculating (4.7).
- **Generation of new cluster**: When generating a new cluster, the differences of other clusters remain unchanged. Therefore, it just introduces a proportionally change of the *Difference*. Considering the *robustness* R^k $(k = 1, ..., K)$ for all of the feature channels, the update equation for the k-th feature channel is derived as:

  $$\hat{\gamma}^k = \frac{\hat{R}^k}{\sum_{k=1}^K \hat{R}^k} = \frac{(R^k)^\eta}{\sum_{k=1}^K (R^k)^\eta}. \tag{4.10}$$

Algorithm 1 GHF-ART

Input: Input patterns $\mathcal{I}_n = \{\mathbf{x}^k|_{k=1}^K\}$, α, β and ρ.

1: set $n = 1$.
2: **repeat**
3: Present $\mathcal{I}_n = \{\mathbf{x}^k|_{k=1}^K\}$ into the input field.
4: For $\forall c_j$, calculate the choice function $T(c_j, \mathcal{I}_n)$ in (4.2).
5: Identify the winner cluster c_{j^*} so that $j^* = \arg\max_{j:c_j \in F_2} T(c_j, \mathcal{I}_n)$.
6: Calculate the match function $M(c_{j^*}, \mathbf{x}^k)$ for $k = 1, \ldots, K$ in (4.3).
7: If $\exists k$ such that $M(c_{j^*}, \mathbf{x}^k) < \rho^k$, set $T(c_{j^*}, \mathcal{I}_n) = 0$, go to 5.
8: If c_{j^*} exists, Update $\mathbf{w}_{j^*}^k$ for $k = 1, \ldots, K$ according to (4.4) and (4.5) respectively and update γ according to (4.7)-(4.9).
9: If no cluster meets the vigilance criteria, Create a new node c_{J+1} such that $\mathbf{w}_{J+1}^k = \mathbf{x}^k$ for $k = 1, \ldots, K$, update γ according to (4.10)
10: $n = n + 1$.
11: **until** All the input patterns are presented.

Output: Cluster Assignment Array $\{A_n|_{n=1}^N\}$.

4.5 Time Complexity Comparison The time complexity of GHF-ART with *Robustness Measure* has been demonstrated to be $O(n_i n_c n_f)$ in [10], where n_i is the number of input patterns, n_c is the number of clusters and n_f denotes the number of features across all of the feature channels.

In comparison with existing heterogeneous data clustering algorithms, the time complexity of LMF [6] is $O(tn_i n_c(n_c + n_f))$, PMM [8] is $O(n_i^3 + tn_c n_i n_f))$, SRC [14] is $O(tn_i^3 + n_c n_i n_f))$ and NMF [15] is $O(tn_c n_i n_f)$, where t is the number of iteration. We can observe that GHF-ART has a much lower time complexity.

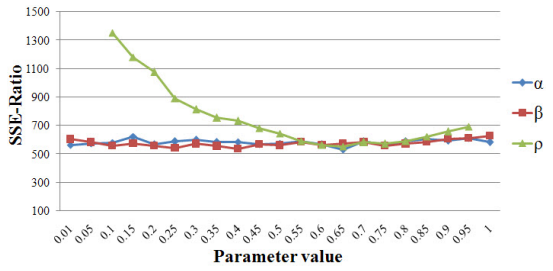

Figure 2: The clustering performance of GHF-ART on the YouTube data set in terms of *SSE-Ratio* by varying the values of α, β and γ respectively.

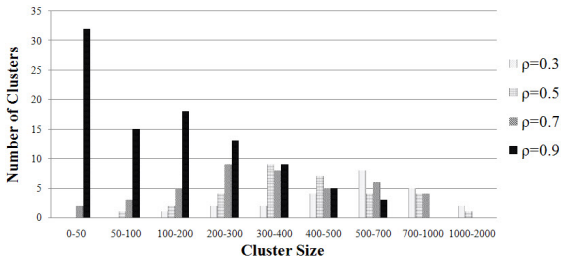

Figure 3: The cluster structures generated by GHF-ART on the Youtube data set in terms of different values of vigilance parameter ρ.

5 Experiments

5.1 YouTube Data Set

5.1.1 Data Description The YouTube data set [1] is a heterogeneous social network data set, which is originally used to study the community detection problem via heterogeneous interactions of users. This data set contains 15,088 users from YouTube website and involves five different types of links, including contact network, co-contact network, co-subscription network, co-subscribed network and favorite network. Detailed descriptions can be found in [8].

5.1.2 Evaluation Measure Since there is no ground truth labels of users in this data set, we adopt five evaluation measures to evaluate cluster quality: 1) *Cross-Dimension Network Validation* (*CDNV*) [8]; 2) *Average Density* (*AD*) measures the probability if two users have connection in the same cluster which is averaged by the number of clusters and feature modalities; 3) *Intra-cluster sum-of-squared error* (*Intra-SSE*) measures the weighted average of *SSE* within clusters across feature modalities; 4) *Between-cluster SSE* (*Between-SSE*) measures the average distance between two cluster centers in a clustering to evaluate how well-separated the clusters are from each other; and 5) *SSE-Ratio* = *Intra-SSE*/*Between-SSE* takes both the precision and recall aspects of the clustering performance.

5.1.3 Parameter Selection Analysis We initialized $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.6$ and studied the change in performance of GHF-ART in terms of *SSE-Ratio* by varying one of them while fixing others, as shown in Fig. 2. We observe that 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, we find that the performance is improved when ρ increases up to 0.65 and degrades when $\rho > 0.85$. To study the reasons, we analyzed the cluster structures generated under different values ρ, which is shown in Fig. 3. We observe that the increase of ρ leads to the generation of more clusters, which may contribute to the compactness of 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, we analyzed the cluster structure at $\rho = 0.5$ and 0.7 in which the best performance is obtained. We observe that when ρ increases from 0.5 to 0.7, the number of small clusters with less than 100 patterns increases. Therefore, we may assume that when a suitable ρ is reached, the number of small clusters starts to increase. In this case, the number of small clusters is nearly 10% of the total number of clusters. If this idea works, an interesting empirical way to select a reasonable value of ρ may be tuning the value of ρ until a small amount of small clusters are identified.

5.1.4 Clustering Performance Comparison We compared the performance of GHF-ART with four existing heterogeneous data clustering algorithms as described in section 2, namely the Spectral Relational Clustering (SRC) [14], Linked Matrix Factorization (LMF) [6], Non-negative Matrix Factorization (NMF) [15] and Principal Modularity Maximization (PMM) [8]. Since SRC and PMM need K-means to obtain the final clusters, we also employed K-means with Euclidean distance metric as a baseline.

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

We obtained the clustering results of GHF-ART 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 number of clusters ranging from

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

Table 1: The clustering Results of GHF-ART, K-means, SRC, LMF, NMF and PMM under the best setting of 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 data set.

	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

20 to 100. The best performance of each algorithm for each evaluation measure is reported in Table 1. We observe that the best performance of each algorithm is 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*, in which the result of GHF-ART is still competitive to the best one.

5.1.5 Correlation Analysis of Heterogeneous Networks We first ran GHF-ART under $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.65$ and showed the track of contribution parameters for each type of links during clustering in Fig. 4. We observed that the weights for all types of features begin with 0.2. The sudden change at $n = 1500$ is due to the the incrementally presenting of new patterns. After $n = 12000$, the weight values of all types of features become stable.

We further analyzed the probability that pairs of connected patterns fall into the same cluster to study how each type of relational network affects the clustering results, which is shown in Fig. 5. We observe that the order of relational networks is consistent with the results shown in Fig. 4, which demonstrates the performance of *Robustness Measure*. Contact network achieves much higher probability than other relational networks. This may be due to that the contact network is much sparser than the other four networks. As thus, the links of contact network are more representative and less links of patterns exist between clusters.

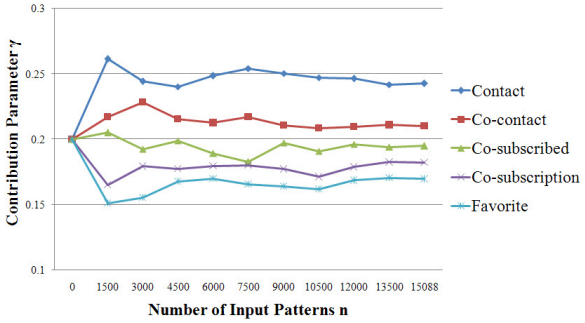

Figure 4: Track of contribution parameters for five types of links during clustering with the increase in the number of input patterns.

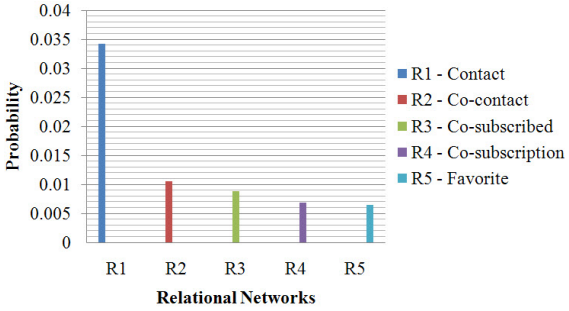

Figure 5: The probability that pairs of patterns fall into the same cluster if connected in each of the five relational networks.

5.2 BlogCatalog Data Set

5.2.1 Data Description The BlogCatalog data set [2] is crawled in [7] 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.

We extracted three types of links, including a friendship network and two textual similarity networks in terms of blog content and tags. By filtering infrequent words from tags and blogs, we obtained 66,418 users, 6,666 tags and 17,824 words from blogs. As suggested in [7], we used the most frequent category in the blogs of a user as the class label and got 147 class labels.

5.2.2 Evaluation Measure With the ground truth labels, we used *Average Precision*, *Cluster Entropy* and *Class Entropy* [22], *Purity* [19] and *Rand Index* [20]

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

Table 2: The clustering Results of GHF-ART, K-means, SRC, LMF, NMF and PMM under the best setting of 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 data set 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

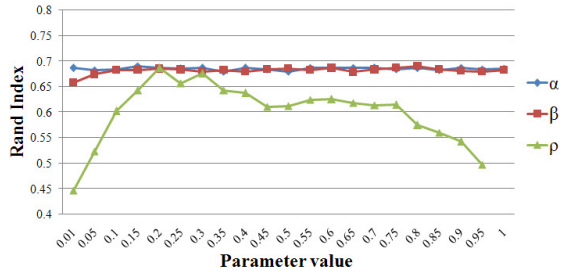

Figure 6: The clustering performance of GHF-ART on the BlogCatalog data set in terms of Rand Index by varying the values of α, β and γ respectively.

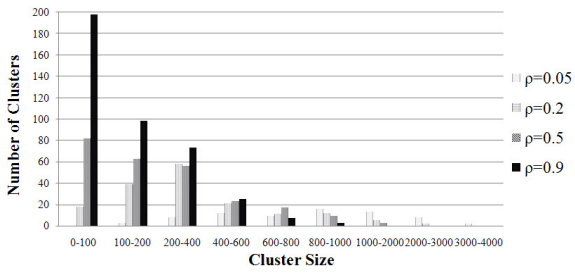

Figure 7: The cluster structures generated by GHF-ART on the BlogCatalog data set in terms of different values of vigilance parameter ρ.

as 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. In our experiments, *Average Precision* is defined by the weighted sum of precision of all the clusters.

5.2.3 Parameter Selection Analysis We studied the influence of parameters to the performance of GHF-ART on the BlogCatalog data set with initial settings $\alpha = 0.01$, $\beta = 0.6$ and $\rho = 0.2$, as shown in Fig. 6. We observed that, consistent with those in Fig. 2, 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 by the same reason for that in Fig. 2.

To validate our findings to select a suitable ρ in section 5.1.3, we analyzed the cluster structures corresponding to the four key points of ρ, as shown in Fig. 7. At $\rho = 0.2$, we observe that nearly 20 small clusters with less than 100 patterns are generated. Interestingly, we find that the number of small clusters is also around 10% of the total number of clusters, which fits the findings that we observed on the YouTube data set in section 5.1.3. This demonstrates the feasibility of the empirical way to select ρ: Run GHF-ART several times by tuning the value of ρ until 10% of the identified clusters are small clusters having less than 100 patterns.

5.2.4 Clustering Performance Comparison We compared the performance of GHF-ART on the BlogCatalog data set with the same set of algorithms compared in the YouTube data set under the same parameter settings as mentioned in section 5.1.4, except the number of clusters. We varied the value of ρ from 0.1 to 0.4 with an interval of 0.05 and the number of clusters from 150-200 with an interval of 5. The best performance for each algorithm with the number of clusters is shown in Table 2. We observe that GHF-ART obtained much better performance (at least 4% improvement) than other algorithms in terms of *Average Precision*, *Cluster Entropy* and *Purity*. This indicates that GHF-ART may well identify similar patterns 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, we find that GHF-ART identifies a similar number of clusters to other algorithms, which demonstrates the effectiveness of GHF-ART.

5.2.5 Case Study We further studied the identified communities by GHF-ART. First, we listed the discovered five biggest clusters, as shown in Table 3. We ob-

Table 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

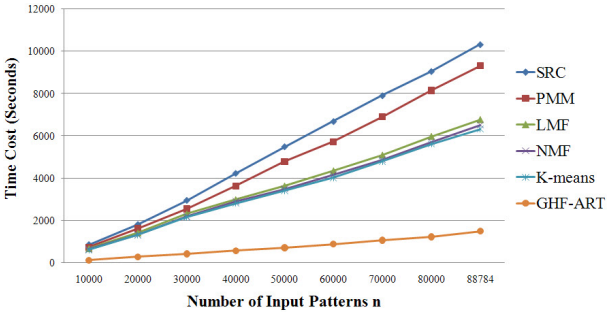

Figure 8: Time cost of GHF-ART, K-means, SRC, LMF, NMF and PMM on the BlogCatalog Dataset with the increase in the number of input patterns.

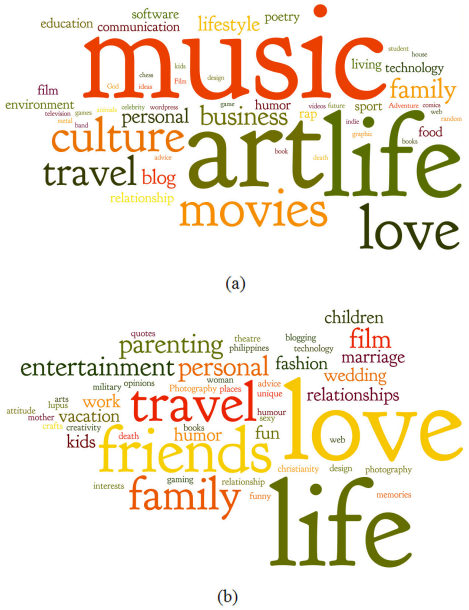

Figure 9: The tag clouds generated for the (a) 1^{st} and (b) 4^{th} biggest clusters. A larger font of tag indicates a higher weight in the cluster

serve that those clusters are well formed to reveal the user communities since more than 1000 patterns are grouped with a reasonable level of precision. We also observe that 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 our organized statistics, "Personal" is much larger than other classes. However, in the top 5 tags, only "life" is shared by them. To have an insight of the relation between these two clusters, we plot the tag clouds for them. As shown in Fig. 9, we observe that the two clusters share more key tags such as "love", "travel", "personal" and "film". Furthermore, when looking into the large amount of smaller tags in the clouds, we find that such tags in Fig. 9(a) are more related to "music" and enjoying "life", such as "game", "rap" and "sport", while those in Fig. 9(b) 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 difference in the potential trends of sub-key tags.

5.2.6 Time Cost Analysis To evaluate the efficiency of GHF-ART on big data, we further analyzed the time cost of GHF-ART, K-means, SRC, LMF, NMF and PMM with the increase in the number of input patterns. To make a fair comparison, we set the number of clusters $k = 166$ for K-means, SRC, LMF, NMF and PMM and set $\rho = 0.2$ for GHF-ART so that the numbers of the generated clusters for all the algorithms are the same. In Fig. 8, we can observe that GHF-ART runs much faster than the other algorithms. Whereas the other algorithms incur a great increase of time cost with the increase in the number of input patterns, GHF-ART maintains a relatively small increase. This demonstrates the scalability of GHF-ART to big data.

6 Conclusion

In this paper, we explored the feasibility of GHF-ART for the community discovery problem in the heterogeneous social networks. Comparing with existing heterogeneous data clustering algorithms [6, 8, 14, 15], as mentioned in section 2, for clustering heterogeneous social networks, GHF-ART has several advantages including: 1) Scalability to big data: GHF-ART performs real-time matching of patterns and one-pass learning which guarantee the low computational cost; 2) No need of the number of clusters a prior: GHF-ART employs a vigilance parameter to restrain the intra-cluster similarity so that clusters may be incrementally identified; 3) Considering heterogeneity of links: GHF-ART con-

siders different representation of learning functions for heterogeneous types of links, which is flexible and may produce better representation for heterogeneous links; 4) Incorporating global and local similarity evaluation; and 5) Weighting algorithm for heterogeneous link fusion.

We have analyzed the performance of GHF-ART in terms of parameter selection, clustering performance comparison, performance in *Robustness Measure*, time cost and case studies. We show that, although GHF-ART needs to set three parameters, the performance of GHF-ART is robust to the choice and learning parameters. We have further illustrated an empirical way to select a suitable value for the vigilance parameter which has been demonstrated to be feasible on both the YouTube and BlogCatalog data sets. The experimental results also demonstrate the effectiveness of *Robustness Measure* and show that GHF-ART outperforms existing heterogeneous data clustering algorithms and has a much lower computational cost.

Although our work has so far obtained encouraging experimental results, there are several directions for further investigation. Firstly, 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 be employed to reduce computer consumption. Secondly, visual data such as images and videos are becoming more important in our social life and should also be considered as an important social link between users. Thus, identifying a social network data set with visual links and studying the feasibility of GHF-ART for effective fusion of visual links together with relational and textual links will be included into our future work.

References

[1] J. Yang and J. Leskovec, *Defining and Evaluating Network Communities based on Ground-truth*, In ICDM, pp. 745–754, 2012.

[2] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao and H. Cao, *Link Prediction and Recommendation across Heterogeneous Social Networks*, In ICDM, pp. 181–190, 2012.

[3] Y. Yang, N. Chawla, Y. Sun and J. Han, *Predicting Links in Multi-Relational and Heterogeneous Networks*, In ICDM, pp. 755–764, 2012.

[4] J. J. Whang, X. Sui, Y. Sun and I. S. Dhillon, *Scalable and Memory-Efficient Clustering of Large-Scale Social Networks*, In ICDM, pp. 705–714, 2012.

[5] I. Drost, S. Bickel and T. Scheffer, *Discovering Communities in Linked Data by Multi-View Clustering*, From Data and Information Analysis to Knowledge Engineering, pp. 342–349, 2006.

[6] W. Tang, Z. Lu and I. S. Dhillon, *Clustering with Multiple Graphs*, In ICDM, pp. 1016–1021, 2009.

[7] X. Wang, L. Tang, H. Gao and H. Liu, *Discovering Overlapping Groups in Social Media*, In ICDM, pp. 569–578, 2010.

[8] L. Tang, X. Wang and H. Liu, *Uncovering Groups via Heterogeneous Interaction Analysis*, In ICDM, pp. 503–512, 2009.

[9] G. Bisson and C. Grimal, *Co-clustering of Multi-View Datasets: a Parallelizable Approach*, In ICDM, pp. 828–833, 2012.

[10] L. Meng, A.-H. Tan and D. Xu, *Semi-supervised Heterogeneous Fusion for Multimedia Data Co-clustering*, IEEE Transactions on Knowledge and Data Engineering, 2013.

[11] D. Zhou and C. J. C. Burges, *Spectral Clustering and Transductive Learning with Multiple Views*, In ICML, pp. 1159–1166, 2007.

[12] K. Chaudhuri, S. M. Kakade, K. Livescu and K. Sridharan, *Multi-View Clustering via Canonical Correlation Analysis*, In ICML, pp. 129–136, 2009.

[13] A. Kumar and H. Daume III, *A Co-training Approach for Multi-View Spectral Clustering*, In ICML, pp. 393–400, 2011.

[14] B. Long, X. Wu, Z. Zhang and P. Yu, *Spectral Clustering for Multi-Type Relational Data*, In ICML, pp. 585-592, 2006.

[15] Y. Chen, L. Wang and M. Dong, *Non-negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering*, IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459–1474, 2010.

[16] A.-H. Tan, G. A. Carpenter and S. Grossberg, *Intelligence through Interaction: Towards a Unified Theory for Learning*, In LNCS, vol. 4491, pp. 1094–1103, 2007.

[17] S. Bickel and T. Scheffer, *Multi-View Clustering*, In ICDM, pp. 19–26, 2004.

[18] G. A. Carpenter, S. Grossberg and D. B. Rosen, *Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System*, Neural Networks, vol. 4, no. 6, pp. 759–771, 1991.

[19] Y. Zhao and G. Karypis, *Criterion functions for document clustering: experiments and analysis*, Technical Report, 2002.

[20] R. Xu and D. C. Wunsch II, *BARTMAP: A Viable Structure for Biclustering*, Neural Networks, vol. 24, no. 7, pp. 709–716, 2011.

[21] M. Rege, M. Dong and J. Hua, *Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering*, In WWW, pp. 317–326, 2008.

[22] J. He, A.-H. Tan, C.-L. Tan and S-Y. Sung, *On Quantitative Evaluation of Clustering Systems*, In Clustering and Information Retrieval, Kluwer Academic Publishers, pp. 105–133, 2003.

[23] X. Wang, B.Qian, J. Ye and I. Davidson, *Multi-Objective Multi-View Spectral Clustering via Pareto Optimization*, In SDM, pp. 234–242, 2013.