Spectral Clustering for Large-Scale Social Networks via a Pre-Coarsening Sampling Based NystrÖm Method

Ying Kang[1,2(✉)], Bo Yu[1], Weiping Wang[1], and Dan Meng[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{kangying,yubo,wangweiping,mengdan}@iie.ac.cn
[2] University of Chinese Academy of Sciences,
No.89 Minzhuang Road, Haidian District, Beijing, China

Abstract. Spectral clustering has exhibited a superior performance in analyzing the cluster structure of network. However, the exponentially computational complexity limits its application in analyzing large-scale social networks. To tackle this problem, many low-rank matrix approximating algorithms are proposed, of which the NystrÖm method is an approach with proved lower approximate errors. Currently, most existing sampling techniques for NystrÖm method are designed on affinity matrices, which are time-consuming to compute by some similarity metrics. Moreover, the social networks are often built on link relations, in which there is no information to construct an affinity matrix for the approximate computing of NystrÖm method for spectral clustering except for the degrees of nodes. This paper proposes a spectral clustering algorithm for large-scale social networks via a pre-coarsening sampling based NystrÖm method. By virtue of a novel triangle-based coarsening policy, the proposed algorithm first shrinks the social network into a smaller weighted network, and then does an efficient sampling for NystrÖm method to approximate the eigen-decomposition of matrix of spectral clustering. Experimental results on real large-scale social networks demonstrate that the proposed algorithm outperforms the state-of-the-art spectral clustering algorithms, which are realized by the existing sampling techniques based NystrÖm methods.

Keywords: Spectral clustering · NystrÖm method · Pre-coarsening sampling

1 Introduction

Spectral clustering is one of the most popular methods for analyzing the cluster structure of networks. In comparison with other classical clustering algorithms such as k-means or linkage algorithm, spectral clustering often yields superior performances. However, spectral clustering is unable to extend its application in large-scale networks [1]. The key reason is that, the exponential increment of time consumption and space occupation restricts the scalability of spectral clustering, especially facing up to the daily exploding social networks. To break the limitation, a number of low-rank matrix approximating algorithms for spectral clustering have been proposed.

The NystrÖm method is a widely used and efficient low-rank matrix approximating technique to speed up spectral clustering, which is able to reduce the computing

T. Cao et al. (Eds.): PAKDD 2015, Part II, LNAI 9078, pp. 106–118, 2015.
DOI: 10.1007/978-3-319-18032-8_9

and memory burdens enormously [2]. Lots of sampling techniques such as uniform sampling [3], weighted sampling [4], random walking sampling [5], k-means sampling [6] etc. have been designed to lower the error of approximation for spectral clustering as much as possible. In common, these techniques select the interpolation points for NystrÖm method with some probabilities, which are computed in view of the degrees of nodes or the weights of edges. Generally, the weights come from the similarity between pairwise nodes in the networks, and the similarity computation by a metric is time-consuming. Even though based on node degrees, there is no valuable information for the judicious selection of columns and rows for a low-rank approximate matrix. In addition, the social networks are often built on the link structure simply, as well lack of enough useful information for the construction of affinity matrix (or similarity matrix). Hence, there are many obstacles for the NystrÖm method to generate a low-rank matrix, which is used for the approximation of spectral clustering efficiently to mine the cluster property of social networks with large scale nodes.

Apart from link relations, if there is no other resource such as the attribute or feature of nodes in the social networks to construct an affinity matrix, can we extract available information to motivate the approximate computing of NystrÖm method for spectral clustering? As is known to us, the triangle has a strong cluster property due to its any two vertices' connection [7]. If we traverse the triangles in social networks and shrink the encountered triangle into a single node (multi-node), the accumulated edge weights appear, which reflect the local aggregating property and can be extracted to cut down the blindness of further study; moreover, along with the networks become smaller and smaller, the basic cluster structure of original networks has still been kept up. Therefore, In light of what is discussed, we propose a spectral clustering algorithm for large-scale social networks via a pre-coarsening sampling based NystrÖm method, which embeds a new triangle-based coarsening process to make the low-rank matrix approximation of spectral clustering much more efficient and targeted.

2 Related Work

Due to the outstanding capability of identifying the objective clusters in the sampling space with arbitrary shape efficiently and the convergence to the global optimal results [8], spectral clustering has been widely applied to many research fields such as machine learning, computer vision, and bioinformatics etc. In general, spectral clustering makes use of the eigenvectors of an affinity matrix which is derived from the data points, to analyze the aggregating characteristic of the original data implicitly and group the similar points into one cluster [9]. Both the procedures of the construction of affinity matrix and the eigen-decomposition of matrix to obtain eigenvectors are complex, which induces spectral clustering unsuitable for the problems with large data sets. Thus, a number of heuristic methods and approximation algorithms are designed to alleviate the computational burdens of spectral clustering.

Yan et al. [10] utilized a distortion-minimizing local transformation of data to speed up the approximating process of spectral clustering. However, the clustering results are prone to the local optimum and sensitive to the original selection because of k-means, or not steady because of RP tree. Mall et al. [11] employed a primal-dual

framework to infer the cluster affiliation for out-of-sample extensions, accelerating spectral clustering in the process of searching for eigenspace. Even if the similarity metric is simplified to the angular cosine between pairwise data points, too much time is cost to compute the affinity matrix.

Inspired by the original application in numerical approximate solution of integral equations [12], the NystrÖm method is introduced to generate a low-rank matrix approximation for the eigen-decomposition of spectral clustering, making spectral clustering capable of studying large-scale networks. Fowlkes et al. [13] simplified the spectral clustering problem by a small random sampling subset of data and then extend the computing results to the whole data set. However, the adopted sampling technique has not improved the computational accuracy. Zhang et al. [6] employed a k-means based sampling scheme to reduce the approximation error of NystrÖm method, but the constraint conditions that the kernel functions comply to make this scheme time-consuming. Belabbas et al. [14] proposed that the probability of choosing interpolation points for NystrÖm method was in proportion to the determinant of similarity matrix, and the Schur complement was used to analyze the NystrÖm reconstruction error. While the bigger the determinant is, the smaller the error is.

Unfortunately, the existing spectral clustering algorithms based on the NystrÖm method cannot be applied to large-scale social networks for the following two main reasons. Firstly, no matter what sampling scheme is adopted for the NystrÖm method, most low-rank matrix approximation algorithms for spectral clustering are based on an affinity matrix, which is too time-consuming to construct due to the high complexity of similarity computing. Secondly, the social networks are often built on link relations, while the node similarity information is insufficient and difficult to obtain. In order to solve the above problems, this paper proposes a spectral clustering algorithm for large-scale social networks via a pre-coarsening sampling based NystrÖm method, which avoids the time-consuming pairwise similarity computation and improves the performance of spectral clustering.

3 Our Approach

3.1 Background

The NystrÖm method is originally used to find the numerical approximations to eigenfunction problems, which are expressed by integral equations of the form [15] as:

$$\int K(x,y)\phi(y)p(y)dy = \lambda\phi(x) \tag{1}$$

where $p(*)$ denotes the probability density function, $K(*,*)$ denotes a kernel function, λ and $\phi(*)$ denote the eigenvalue and eigenvector of the kernel K based on the integral equation respectively. To approximate the integral on the left of Equation

(1), sample q interpolation points $\{x_1, x_2, \ldots, x_q\}$ drawn from $p(*)$, the approximate result by the empirical average is as follows:

$$\frac{1}{q}\sum_{j=1}^{q} K(x, x_j)\tilde{\phi}(x_j) \simeq \lambda\, \tilde{\phi}(x) \tag{2}$$

where $\tilde{\phi}(x)$ approximates $\phi(x)$ in Equation (1). In addition, choose x in Equation (2) from $\{x_1, x_2, \ldots, x_q\}$ as well to generate an eigen-decomposition $\overline{K}\overline{U} = q\overline{\Lambda}\overline{U}$, $\overline{K}$ denotes the positive semi-definite matrix with elements $\{K(x_i, x_j)|i, j = 1,2, \ldots, q\}$, $\overline{U}$ denotes the eigenvector matrix of $\overline{K}$ and $\overline{U} \in R^{q\times q}$ has orthonormal columns, $\overline{\Lambda} \in R^{q\times q}$ is a diagonal matrix whose non-zero elements are the eigenvalue of $\overline{K}$. Any eigenvector $\phi_i(x)$ and eigenvalue λ_i in Equation (1) can be estimated by $\overline{U}$ and $\overline{\Lambda}$.

$$\phi_i(x) \simeq \sqrt{q}\overline{U}_{ij}, \qquad \lambda_i \simeq \bar{\lambda}_{ii}/q \tag{3}$$

The eigenvector of any point x can be approximated by the eigenvectors of interpolation points in $\{x_1, x_2, \ldots, x_q\}$, because the k-th eigenvector at an unsampled point x can be computed by:

$$\tilde{\phi}_k(x) \simeq \frac{1}{q\bar{\lambda}_i}\sum_{j=1}^{q} K(x, x_j)\tilde{\phi}_k(x_j) \tag{4}$$

As is shown above, different interpolation points to be selected will lead to different approximating results.

How to extend the NystrÖm method to spectral clustering? Consider a $m \times n$ matrix A, which is partitioned as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{5}$$

without loss of generality, let $A_{11} \in R^{p\times q}$ denotes the submatrix which is generated by the intersection of q columns and p rows sampled in some manner, A_{12} and A_{21} denotes the submatrix consisting of elements with a sampled column label (exclusive) or sampled row label, respectively, and $A_{22} \in R^{(m-p)\times(n-q)}$ denotes the submatrix consisting of the remaining elements of A.

Suppose the eigen-decomposition of A_{11} is $A_{11} = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^{\mathrm{T}}$, the eigenvector matrix of A_{21} can be approximated by $A_{21}\widetilde{U}\widetilde{\Lambda}^{-1}$, thus the spectral analysis on the submatrix can be extended to the original matrix A, and the estimation of A is as follows:

$$\begin{aligned} \bar{A} &= \left[\widetilde{U}; A_{21}\widetilde{U}\widetilde{\Lambda}^{-1}\right]\widetilde{\Lambda}\left[\widetilde{U}; A_{21}\widetilde{U}\widetilde{\Lambda}^{-1}\right]^{T} \\ &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{21}{A_{11}}^{-1}A_{12} \end{bmatrix} \end{aligned} \tag{6}$$

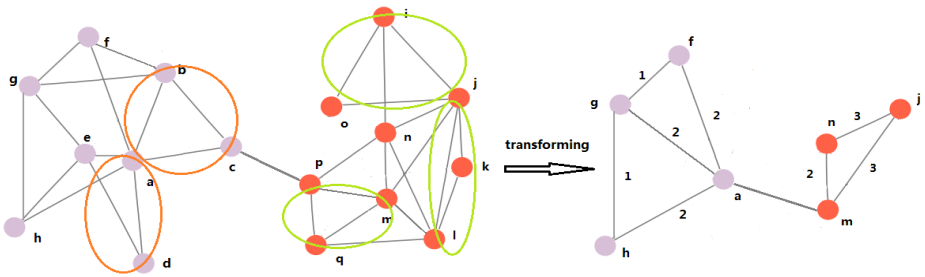
the corresponding approximate eigenvectors are $\overline{U} = \left[\widetilde{U}; A_{21}\widetilde{U}\widetilde{\Lambda}^{-1}\right]$.

In addition, some details on the normalization of eigenvectors, the transformation of matrix based on whether positive semi-definite or not etc. are not discussed here, which can be referred to [16] for further studies.

3.2 A Pre-Coarsening Sampling Based Nyström Method

From the viewpoint of topology, the social networks are built on link relations ubiquitously, which is the most direct resource for studying social networks. Without any help of other information like attribute or feature of nodes to serve for similarity computing in social networks, is it possible that the task of low-rank matrix approximation based on the Nyström method for spectral clustering can proceed? At least impossible for weighted sampling, random walking sampling or k-means sampling based Nyström methods, because there is no appropriate affinity matrix to match with them. Even though an endeavor is made to search for such a matrix, it is time-consuming to compute the matrix by some similarity metrics like Cosine [17].

How to mine some available information from the link structure of social networks for the approximate computing of Nyström method for spectral clustering is a challenge to our work. Let's start from an example in Figure 1. Figure 1(a) shows a link relation based network which contains many triangles. If we execute operations as follows: traverse the triangles in the network, and when encountering a triangle as labeled by one circle in Figure 1(a), shrink it into one single node (multi-node), the network will be transformed into a smaller weighted network gradually as is shown in Figure 1(b). During the transforming process, the edges adjacent to the shrunk vertices in a triangle are accumulated, thus the unit edge-weight values of original network become numerical values after transformation, which can be utilized as the foundation of further sampling of Nyström method. In addition, it is obvious that the generated network keeps the basic cluster topology structure of the original network.



(a) Before coarsening (b) After coarsening

Fig. 1. The transforming process of the link relation based network by shrinking triangles

As is familiar to us, the three vertices of a triangle are bound to belong to the same cluster, owing to the strong connection between any two of its inner vertices. Therefore, in virtue of the strong cluster property of triangles, the generated smaller weighted network by shrinking the traversed triangles into multi-nodes will not break down the original cluster structure. On the contrary, the weights of edges in the network appear after transformation, which reflect the joint strength of any two nodes. In addition, some connections between nodes in the transformed network are the indirect relations constructed over the shrunk vertices, which is beneficial for mining deeper associations and is an advantage that other methods lack of. So in summary, lots of

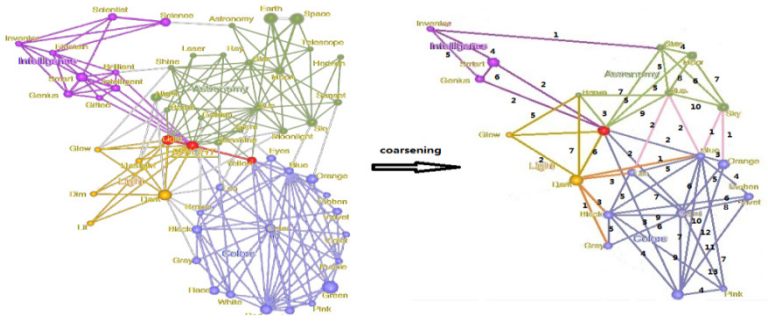
information about edge weight, cluster and connection etc. has been mined from the transformation of link relation based network, which is able to be utilized for the further sampling of NystrÖm method explicitly.

Extending the principle to the social networks, where triangles are the basic structural elements, we propose a pre-coarsening sampling based Nyström method. We first give the formulized definition of the novel triangle-based coarsening problem:

Definition 1. (Triangle-based Coarsening Problem)
Input: *undirected, link relation based social network* $N = (V, E, W)$, $v \in V$ *denote nodes of* N, $e \in E$ *denote edges,* Δ *denote triangles, edge weight* $\omega \in W$ *are unit.*
Output: *undirected, coarsened weighted network* $N' = (V', E', W')$, *what* $v' \in V'$ *and* $e' \in E'$ *of* N' *denote is similar to* N, $\omega' \in W'$ *are numerical.*

Process: *traverse* Δ *of* N *in an order (ascending or descending) of node degrees, {*v *has not been shrunk or to be a multi-node just only one time}; when encounter a* Δ, *shrink it into a single multi-node* $v'_{multi-node}$; *accumulate the edges adjacent to shrunk nodes* v_{shrunk} *and reweight the corresponding remained edges* $\omega'_{e-remaining}$.

The certain order of node degrees in Definition 1 is a necessary condition, which is used to avoid the indeterminacy of generated network by coarsening the random traversed triangles. And if traversing and shrinking triangles in the network without any constraint, the coarsening will make the original cluster structure disappear sometimes, thus to prevent this phenomenon from happening, another condition as '*to be a multi-node just only one time*' must be added. Figure 2 shows an example of the generation of a smaller weighted network by triangle-based coarsening.



(a) A network of word association [18] (b) Generated weighted network by coarsening

Fig. 2. The generation of a smaller weighted network

We put the triangle-based coarsening as the preprocessing for sampling of Nyström Method, so the proposed pre-coarsening sampling based Nyström method is:

Definition 2. (Pre-Coarsening Sampling based Nyström Method)
Input: *weight matrix* $\tilde{A}$ *constructed on the weighted network which is generated by the triangle-based coarsening.*
Output: *low-rank approximating matrix* $\bar{A}$.

Process: *in view of* $\tilde{A}$ *and the probability distribution* $p_i = \left|A^{(i)}\right|^2 / \|A\|_F^2$, *we sample* p *rows and* q *columns to generate a low-rank approximating matrix* $\bar{A} \in R^{p \times q}$.

3.3 Spectral Clustering for Large-Scale Social Networks

Subsequently, the key problem is how to estimate the cluster affiliation of out-of-sample nodes. Depending on the obtained low-rank approximating matrix $\bar{A}$, the eigen-decomposition of $\bar{A}$ is able to extended to the out-of-samples in $\tilde{A}$, making the approximation of spectral clustering more efficient for large-scale social networks built on link relations. We define the out-of-sample extension problem as follows:

Definition 3. (Out-of-Sample Extension Problem)
Input: *low-rank approximating matrix* $\bar{A}$.
Output: *approximate eigenvector matrix* $\tilde{U}$.

Process: *compute the eigenvector matrix* $\bar{U}$ *by* $\bar{A} = \bar{U}\bar{\Lambda}\bar{U}^T$, *and according to Eq.(4), approximate* $\tilde{U}$ *by* $\tilde{U} = \left[\bar{U}; A_{out-of-sample}\bar{U}\bar{\Lambda}^{-1}\right]$.

In contrast to the process of pre-coarsening sampling of Nyström method, except for the extension of eigenvectors in $\bar{U}$, it is necessary to use k-means to group all of the approximate eigenvectors in $\tilde{U}$ into k clusters, and then unfold the multi-nodes and classify the unfolded nodes into the cluster which the corresponding multi-node belongs to. By now, we have obtained the ultimate spectral clustering results from large-scale social networks. The implementation of integral process of coarsening, sampling and clustering of our spectral clustering is depicted in Algorithm 1.

Algorithm 1. Spectral clustering

Input: Adjacency matrix of social network $A \in R^{n \times n}$, n is the number of nodes, m is the number of sampled columns, r-rank approximation, m and $r<<n$.
Output: k clusters of A.

```
1  Begin
2    Ã= weight matrix generated from A by the triangle-based coarsening;
3    S =indices of m columns sampled by probability p_i = |A^(i)|^2 / ||A||_F^2;
4    Ā = Ã(:S);
5    Ā = ŪΛ̄Ū^T;
6    U_r = SmallestEigenVectors(Ū, r);
7    U_os = sqrt(m/n) C U_r Σ_r^-1; // U_os is the approximating eigenvectors of out-of-sample
                                    nodes, Σ_r is the diagonal eigenvalue matrix
8    Y = NormalizeRows[U_r, U_os];
9    K̄ = ClusterRows(Y);        // group the approximate eigenvector matrix into k
                                   clusters by k-means
10   K = Classify(Unfold(Ã)) by reference to K̄;
11 End
```

3.4 Performance Analysis

By virtue of the triangle-based coarsening, valuable prior information about cluster is extracted from the link structure of social networks, which is useful for the further sampling of NystrÖm method. What's more, due to the intrinsic strong cluster property of triangle, the basic cluster structure of original network is maintained after coarsening preprocessing. Therefore, via the pre-coarsening sampling based NystrÖm method, not only the complexity of spectral clustering can be reduced, but also the computational accuracy of the low-rank matrix approximation of spectral clustering can be promoted greatly.

Lemma 1 (Running Time). *The worst case time complexity of our algorithm is less than* $\Theta(e^{3/2}) + \Theta(m^3) + \Theta(nm^2) + \Theta(nkt)$*, and the worst case space complexity is* $\Theta(mn)$*, where e denotes the number of edges in network* A*, n denotes the number of columns in matrix* $\tilde{A}$*,* m *denotes the number of sampled columns from* $\tilde{A}$*,* k *denotes the number of clusters,* t *denotes the iteration times of k-means.*

Proof. The process of traversing all of the triangles in the network takes $\Theta(e^{3/2})$ time (refer to compact-forward algorithm), but for the triangle-based coarsening it is unnecessary to do the traversing, because the triangles adjacent to the shrunk nodes become unavailable, so the time complexity of generating matrix $\tilde{A}$ is much less than $\Theta(e^{3/2})$; from the eigen-decomposition of matrix $\bar{A}$ to the extension to out-of-sample nodes in Algorithm 1, the consuming time is $\Theta(m^3) + \Theta(nm^2)$; and the execution of k-means on Y only needs $\Theta(nkt)$ operations (neglect the time of normalizing Y). Therefore, the *time complexity of our algorithm is less than* $\Theta(e^{3/2}) + \Theta(m^3) + \Theta(nm^2) + \Theta(nkt)$ in the worst case. In the process of approximate computing of spectral clustering based on NystrÖm method, the maximum scale of the matrices which need to be stored is $m \times n$, so the memory usage of our algorithm is $\Theta(mn)$ in the worst case.

4 Experiments

4.1 Dataset and Experiment Setup

Our experiments are designed on the dataset of large-scale social networks which are collected from Stanford University's SNAP networks [19]. The details of each social network are listed in Table 1. Our spectral clustering algorithm for large-scale social networks is realized by a pre-coarsening sampling based NystrÖm method. To test our algorithm's performance, we compare the pre-coarsening sampling with uniform sampling [20], weighted sampling [21], k-means sampling [6] and incremental sampling [22]. Our pre-coarsening sampling can proceed explicitly based on the link relations of social networks, but the weighted sampling, k-means sampling and incremental sampling need to search for some available information by other approaches to compute the similarity matrices firstly. Therefore, we design two sub-experiments for different testing tasks. One is to use all the sampling techniques to analyze the social networks explicitly, while the other is to add the similarity matrix computations for the latter three sampling techniques.

Table 1. Large-scale social networks

Dataset	No. of nodes	No. of edges	No. of clusters	No. of triangles
Youtube	1,134,890	2,987,624	8,385	3,056,386
Orkut	3,072,441	117,185,08	6,288,363	627,584,181
LiveJournal	3,997,962	34,681,189	287,512	177,820,130
Twitter	12,309,718	91,765,139	355,179	519,402,625
Friendster	65,608,366	1,806,067,135	957,154	4,173,724,142

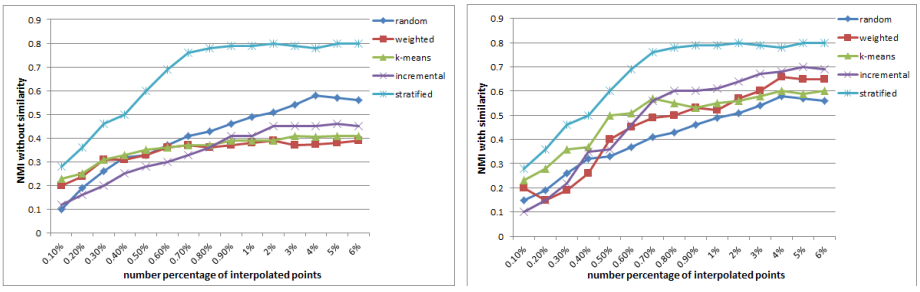
All social networks in Table 1 contain ground-truth clusters, so we can utilize the normalized mutual information (NMI) [23] to evaluate the clustering performance of different clustering algorithms, which are based on different sampling techniques of NystrÖm methods. In general, the larger NMI is, the better the clustering results are.

We perform all the experiments on a Linux machine with 4Core 2.6GHz CPU and 8G main memory. The implementations of all algorithms are in Java. Moreover, we repeat to run each algorithm 30 times to obtain an average result of NMI, making the analytical results more accurate.

4.2 Experimental Results and Analysis

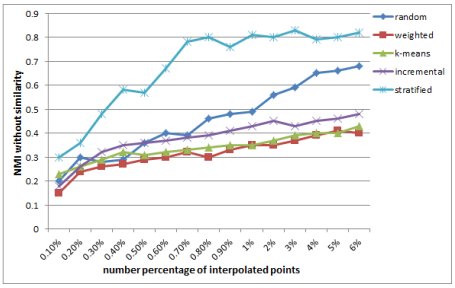
The spectral clustering results corresponding to different social networks are displayed in Figure 3, along with the running time of different algorithms in Figure 4.

From Figure 3(a) we can observe that, our spectral clustering algorithm outperforms other algorithms in analyzing the social networks, in which there is no useful information except for the link structure. Meanwhile, the algorithm adopting a uniform sampling technique has done a little better job than the algorithms which are based on the weighted sampling, k-means sampling, and incremental sampling.

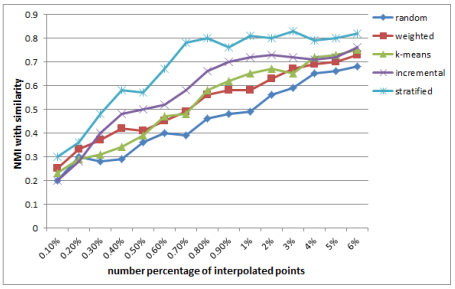


(1) Youtube (1) Youtube

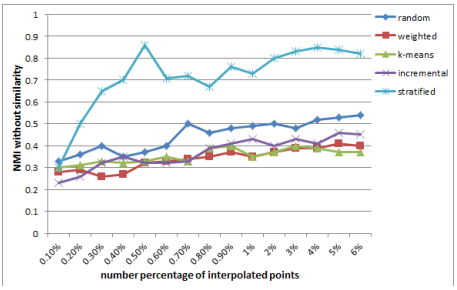
Fig. 3. The comparison of computing accuracy on social networks



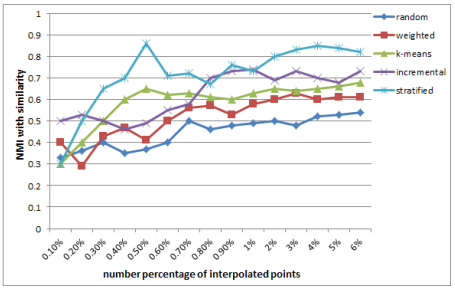
(2) Orkut



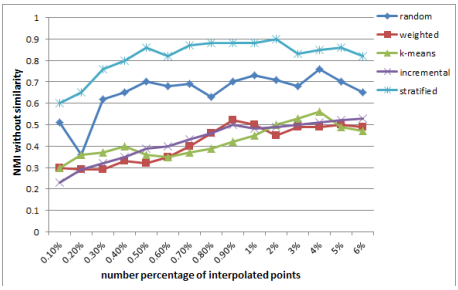
(2) Orkut



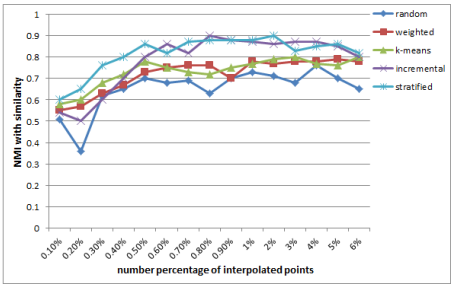
(3) LiveJournal



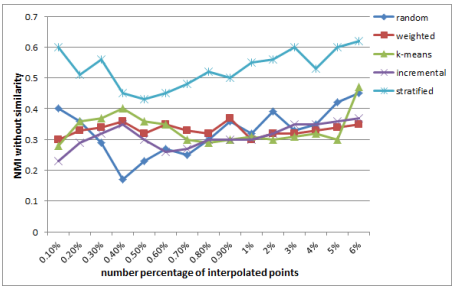
(3) LiveJournal



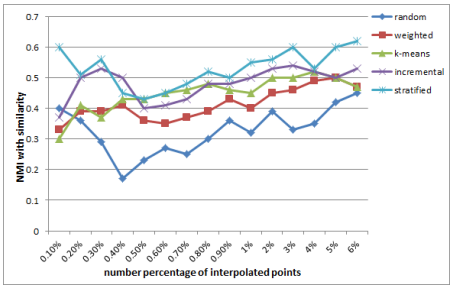
(4) Twitter



(4) Twitter



(5) Friendster

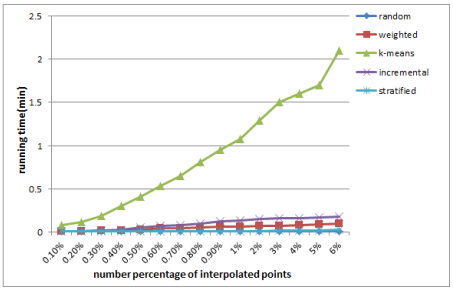


(5) Friendster

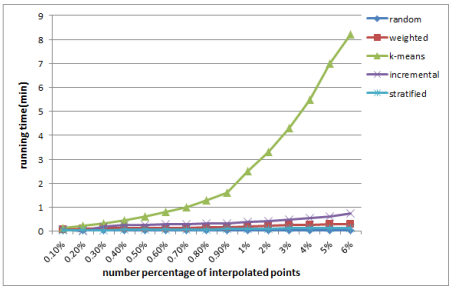
(a) Without any similarity matrix (b) Add similarity matrix for US, KS and CS

Fig. 3. (*continued*)

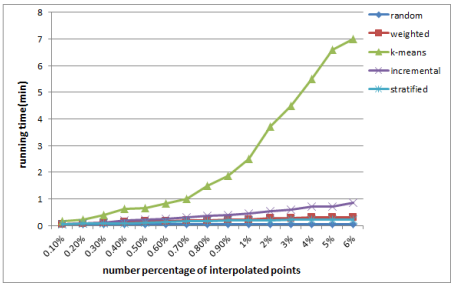
By contrast, we analyze the clustering results in Figure 3(b). There is no doubt that the accuracy of clustering results of our algorithm is superior to the others. And with the help of similarity matrix, the performances of the algorithms which are based on the weighted sampling, k-means sampling, and incremental sampling excel the ones of the uniform sampling based algorithm. Besides, when the sampling probabilities of different columns of matrix are identical, the temporarily adopted uniform random sampling among these columns will degrade the performance of our algorithm, so some exceptions appear in LiveJournal, Twitter and Friendster of Figure 3(b).



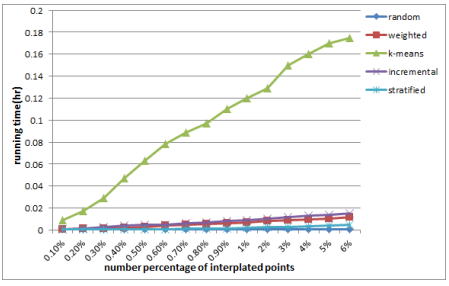
(a) Youtube



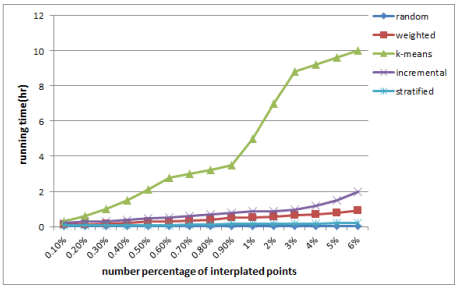
(b) Orkut



(c) LiveJournal



(d) Twitter



(e) Friendster

Fig. 4. The comparison of running time on social networks

Subsequently, let us compare the running time between different algorithms (note that we just do comparison in the case of adding similarity matrices). It is obvious that in Figure 4, the algorithm based on the k-means sampling technique spends much more time to tackle the clustering problems of large-scale social networks. The key reason consists in the inherent iterative computing complexity of k-means as an unsupervised method. Because of the relative easier sampling technique, the time consuming of the other algorithms is so small to be neglected in contrast to the k-means sampling based algorithm. As is shown in Figure 4, although our algorithm experiences a coarsening preprocessing before sampling, the running time of our algorithm is less than other algorithms except for the uniform sampling based algorithm. This is because that much more time needs to be cost to compute the similarity matrix for the algorithms which are based on the weighted sampling and incremental sampling.

5 Conclusion

This paper proposes a spectral clustering algorithm for large-scale social networks via a pre-coarsening sampling based NystrÖm method. By virtue of a new triangle-based coarsening policy, this algorithm first extracts a smaller weighted network from the link relation based social network, which reveals some useful prior information about cluster, and then executes an efficient sampling for the NystrÖm method to generate a low-rank matrix approximation for the eigen-decomposition of spectral clustering. Due to the cluster property of triangle, the process of coarsening maintains the original cluster topology structure. Moreover, the pre-coarsening sampling based NystrÖm method makes spectral clustering capable to analyze the social networks explicitly without any other available information except for link relations, promoting the computing accuracy of spectral clustering. Experimental results on real social networks demonstrate that our algorithm outperforms the state-of-the-art spectral clustering algorithms, which are based on other sampling techniques for the NystrÖm method.

Acknowledgement. This work is supported by the National Science Foundation of China under grant number 61402473, KeJiZhiCheng Project under grant number 2012BAH46B03, National HeGaoJi Key Project under grant number 2013ZX01039-002-001 -001, and "Strategic Priority Research Program" of the Chinese Academy of Sciences under grant number XDA06030200.

References

1. Ng, A.Y., Jordan, M.I., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm. Advances in Neural Information Processing Systems (2002)
2. Bellman, R., Bellman, R.E., et al.: Introduction to Matrix Analysis (1970)
3. Williams, C., Seeger, M.: Using the nyström method to speed up kernel machines. In: Proceedings of the 14th Annual Conference on Neural Information Processing Systems (2001)
4. Drineas, P., Mahoney, M.W.: On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-based Learning. The Journal of Machine Learning Research (2005)

5. Wang, X., Hamilton, H.J.: DBRS: A Density-based Spatial Clustering Method with Random Sampling. Springer, Heidelberg (2003)
6. Zhang, K., Tsang, I.W., Kwok, J.T.: Improved nyström low-rank approximation and error analysis. In: Proceedings of the International Conference on Machine Learning (2008)
7. Kumpula, J.M., Kivelä, M., Kaski, K., Saramäki, J.: Sequential Algorithm for Fast Clique Percolation. Phys. Rev. (2008)
8. Xiang, T., et al.: Spectral Clustering with Eigenvector Selection. Pattern Recognition(2008)
9. Von Luxburg, U.: A Tutorial on Spectral Clustering. Statistics and Computing (2007)
10. Yan, D., Huang, L., Jordan, M.I.: Fast approximate spectral clustering. In: Proceedings of the ACM SIGKDD (2009)
11. Mall, R., Langone, R., Suykens, J.A.K.: Kernel Spectral Clustering for Big Data Networks. Entropy (2013)
12. Sun, Z.H., Wei, X.H., Zhou, W.H.: A Nyström-based Subtractive Clustering Method. Wavelet Active Media Technology and Information Processing (ICWAMTIP) (2012)
13. Fowlkes, C., Belongie, S., Chung, F., et al.: Spectral Grouping Using the Nyström Method. IEEE Transactions on Pattern Analysis and Machine Intelligence (2004)
14. Belabbas, M.A., Wolfe, P.J.: Spectral methods in machine learning and new strategies for very large datasets. In: Proceedings of the National Academy of Sciences (2009)
15. Dick, J., Kritzer, P., et al.: Lattice-Nyström Method for Fredholm Integral Equations of the Second Kind with Convolution Type kernels. Journal of Complexity (2007)
16. Belongie, S., Fowlkes, C., Chung, F., et al.: Spectral Partitioning with Indefinite Kernels Using the Nyström Extension. European Conference on Computer Vision (2002)
17. Adamic, L.A., Adar, E.: You are What You Link. In: The 10th Annual International World Wide Web Conference, Hong Kong (2001)
18. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature (2005)
19. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: Proceedings of ACM SIGKDD (2012)
20. Bengio, Y., Paiement, J.F., et al.: Out-of-Sample Extensions for Lle, Isomap, Mds, Eigenmaps and Spectral Clustering. Advances in Neural Information Processing Systems (2004)
21. Choromanska, A., Jebara, T., Kim, H., et al.: Fast Spectral Clustering via the Nyström Method. Algorithmic Learning Theory (2013)
22. Xianchao, Z., Quanzeng, Y.: Clusterability analysis and incremental sampling for nyström extension based spectral clustering. In: International Conference on Data Mining (2011)
23. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing Community Structure Identification. Stat. Mech. (2005)