Automatic K Community Mining in Heterogeneous Networks Using Convergence Aware Dirichlet Process Mixture Model

Renuga Devi. R and Hemalatha. M*
Department of Computer Science, Karpagam University, Coimbatore, India.

Abstract

Network Community mining development has emerged as a fascinating research area in many fields. Several researches focused on mining the hidden communities in homogeneous social networks. But in real-world, most of the social networks are heterogeneous. Each node in a network contains particular kind of relationship and it plays a different role in a particular task. Mining such communities are challenging task. Finding evolutionary communities in heterogeneous networks can help the researchers to understand the structural information of the networks. To solve this heterogeneous networks problem, earlier researchers used Dirichlet Process (DP) mixture models which are promising candidates for clustering applications where the number of clusters is unknown. Because of some computational considerations the existing models are unfortunately unsuitable for large scale data mining applications. This paper presents Convergence aware Dirichlet Process Mixture Model (CADPM) to solve the above mentioned problem. CADPM is proposed to routinely handle millions of data-cases. A spectral construction of the networks and its scalability problems are studied. Experiments are carried out with the real world large networks. Results show that the efficiency of our proposed algorithm and recommend its generalization in solving troubles with complex relationships.

1. Introduction

The fast and growing development of online social networks, vast heterogeneous information networks thus derived are omnipresent that contains differing kinds of objects. A network which is containing totally different reasonably objects is completely different from traditional homogeneous networks. A social network is represented as a graph. The nodes are representing individuals, and the relationships between the nodes are represented as edges. In a classical social network, a different relationship exists between the nodes or between individuals. Like normal friendships to organizational relationships. Each relation plays a different role in different process. Most of the existing algorithms on social network analysis assume that there is only one single social network and is usually representing the comparatively homogenous connection. But In real social networks contains different types of relations. Each relationship can be treated as a network. Such kinds of networks are called as multi relational social network or heterogeneous social network. Finding unrefined process communities from these heterogeneous networks can profit the users of those online databases higher understanding the structures of the complicated networks and their evolution beside time. Also, such information can facilitate users observe predictions on the longer term trends of the community. In distinction to community outlined in an exceedingly same network, that may be a set of objects from one sort, a community in an exceedingly heterogeneous network ought to be heterogeneous itself.

However, most existing strategies solely study the community evolution in same networks. The normal network evolution analysis on same networks, that is barely able to track one sort of objects' evolution, cannot properly model the evolution of a community that really contains multiple styles of objects. Different recent works on evolution study on heterogeneous networks, like in [1], have thought of the interaction of communities among differing kinds, however, their community definition remains single type and cannot replicate the construct of multi type communities like analysis areas. It ought to contain two properties like the quantity of communities in every time stamp ought to be versatile and mechanically learned and, the communities in adjacent timestamps ought to be consistent [2]. To find a community with some specific properties, first try identify that relation plays a very important role in such a community. Moreover, such relation may not exist expressly; first discover such a hidden relation before finding the community on such a relation network. Such issues may be designed mathematically as relation choice and extraction in multi-relational social network analysis.

Mixture model is usually used methodology in clustering. It is generally hard for human to identify the correct cluster range within the mixture model. Dirichlet process Mixture Model may be a distinctive way to resolve the problem, wherever the cluster range is measured as numerable unlimited, and also the distribution of part weights follows a Dirichlet Process with a base distribution. A Dirichlet process Mixture model (DPM) is employed to determine the natural cluster range and contemplate historical impacts from net-clusters of previous time windows at the same time. Net-clusters with the most effective cluster range and best in line with the historical net-clusters are generated. Specific evolution structure will so be obtained from the previous dependency among completely different net-clusters between adjacent timestamps [1].

Dirichlet processes are often used in Bayesian non-parametric statistic. Here the word Nonparametric does not mean a parameter fewer models, relatively a model within which representations produce as a lot of knowledge square measure discovered. Bayesian statistic models have gained tidy quality within the field of machine learning as a result of the above-named flexibility, particularly in unattended learning. In a very Bayesian statistic model, the previous and posterior distributions do not seem to be constant distributions, however random processes. The fact that the Dirichlet distribution could be a likelihood distribution of non-negative numbers that add to at least one makes it a decent method candidate to model distributions of distributions functions. In addition, the non-parametric nature of this model makes it a perfect candidate for bunch issues wherever the distinct variety of clusters is unknown beforehand. As attracts from a Dirichlet process method square measure separate, a crucial use is as a previous likelihood in infinite mixture models. During this case, S is that the constant set of element distributions. The generative method is thus that a sample is drawn from a Dirichlet method, and for every datum successively a price is drawn from this statistical distribution and used because the element distribution for that datum. The fact that there's no limit to the amount of distinct parts which can be generated makes this type of model acceptable for the case once the amount of mixture parts is not well-defined earlier. For instance, the infinite mixture of Gaussians model [3].

In this paper, proposed Convergence aware Dirichlet Process Mixture Model (CADPM). The main idea of the proposed algorithm is to model the community mining problem as an optimization issues. Particularly, each relation is characterized as a graph. The obtained permutation can better meet up user’s desire. As a result, it leads to improved performance on community mining problem. The main contributions of this paper are we proposed the problem to find multi relational network communities in a heterogeneous network. CADPM algorithm was proposed to model the network communities. It robotically finds out the most excellent cluster number and remains reliability between neighboring nodes. We applied the proposed method on three real world datasets, the Amazon network data set, Gnutella data set and Stanford network and the experimental results show that the influence of our proposed model which is capable to use both heterogeneous network and instance information of the networks.

2. Related Works

Network Community finding and clustering in social networks studied for quite long-standing. Most of the prevailing studies try to divide a large network into over some comparatively small elements and mix similar nodes into constant clusters. The study of community detection downside was started with homogeneous networks, like spectral clustering methods [4, 5, 6], modularity measure based methods [7, 8], and probabilistic model based on the ways [9, 10, 11, 12], and later to bipartite networks [13, 14]. Recently several researchers focused on heterogeneous networks [15, 16].

In [1] authors analyzed the heterogeneous networks with star network schema. It is different from their previous analysis and alternative static community detection strategies. It principally concentrated on the model of dynamic evolution of the net-cluster-based multi relational communities. In general, new nodes can take part the network, whereas some nodes can leave, and so sequences of networks with totally different timestamps are often collected from dynamic evolving networks. Finding clusters on such network sequences will facilitate humans to highly perceive the evolution of communities. Some studies are devoted on unvaried networks, extended from static clustering ways, like in [17, 18, 19, 20, and 21].

In recent days studies [18] on heterogeneous networks are carried out by many researchers. The communities of the network are distinct based on every particular kind of objects, and also the number of clusters for each type of object is needs to be set and specified by the person. The community development studies are broad idea of community, which can also repeatedly choose the number of clusters. Main problem in community development is to deciding the exact number of clusters for every timestamp. The Existing Dirichlet Process Mixture Model based on generative model used to find the development of net-clusters. The Dirichlet Process [22, 23] model provides an easy way to insert priority for the clusters in mixture models, and it was supportive to make a decision the cluster number mechanically. Apart from these existing ways few alternative works have extended the Dirichlet process method into considering time information, like in [24] and [25]. Another DP-based extension [26, 27] are planned to model biological process clustering. The variations of their proposed model includes providing an exact solutions for net-cluster development in heterogeneous networks, developed a unique generative model for net-cluster development, which might model the evolution of identical cluster in several timestamps, whereas several existing works need identical clusters do not modification among completely different timestamps and did not claim a worldwide abstract thought of the model, greedy abstract thought at on every occasion stamp that is additional sensible for timely change the evolution.

A new method was proposed to concurrently finding network communities and its topics in text augmented social networks [28]. It was developed based on the non parametric Bayesian approach along with the Dirichlet Process Mixture model (DPM) and its Hierarchical Dirichlet Process (HDP) mixture model. It was developed to mechanically decide the numbers of network communities and their topics. Because communities and its topics was improved by each extra by suggests that of community-topic allocation throughout parameter education procedure and the numbers of each communities and topics area unit allowed to grow infinitely once it has necessary, our model is termed mutual enhanced infinite community-topic model (MEI). In this model, we have a tendency to expressly distinguish community and topic from one another by modeling them via totally different latent variables. On the opposite hand, it is discovered that there are unit correlations between communities and topics. Users from constant community tend to have an interest in similar topics. Thus, we have a tendency to correlate community and topic along via community-topic distribution in our model. Moreover, most previous works for community detection or topic modeling need the numbers of latent categories, i.e. community or topic, to be laid out in advance [48].

There are intensive works learning the structural property of interactions between actors. One probabilistic approach is that the random block model [28], during which the links between nodes are generated trained on the latent cluster membership of nodes. Two nodes among constant cluster area unit treated as the same. That is, the connections among (A1, B1) and (A2, B2) have constant likelihood if A1 and A2, B1 and B2 belong to constant cluster, respectively. In classic block models, the quantity of clusters area unit fastened. The method proposed in [29] replaces the restriction by assignment a Chinese restaurant work method as a previous to get cluster membership for every node. The quantity of clusters will be mechanically determined by assignment correct previous. Mixed block model is additionally developed [30]. Long et al., [31] proposed a framework just like random block model to handle multi-mode networks with interactions and attributes. Typically, some MCMC technique is utilized to approximate the reasoning. With the event of topic models [32], it is additionally extended to model documents at intervals a social network [33], and also the author/document, or sender/receiver/email interactions [34, 35]. The model is often specific certainly variety of documents like Emails or papers. Another commit to model the structure is latent area model. Intuitively, latent area models map the social actors to a latent low-dimensional area such the actors whose positions square measure nearer to every different square measure a lot of seemingly to move with one another [36, 37].

Most of the existing latent space methods principally focused on one-mode networks. Other works try and address the matter in multi-mode networks. In [38] basically studied a two-mode network and maps each authors and words into an equivalent Euclidean space. Spectral relative agglomeration that is most associated with multi-mode network tries to find the latent structure supported multiple relative tables. Because the original downside of finding separate cluster assignment is NP-hard, spectral agglomeration relaxes the constraint to permit the membership vector to be continuous. The method of co-clustering [39, 40, 41], tries to handle the matter of agglomeration. Each words and documents taking advantage of bipartite at the similar time. In [42] extends the matter to a star-type network with multiple heterogeneous knowledge objects and proposes semi-definite programming to resolve the matter. [43] Proposes reinforce agglomeration for multiple heterogeneous knowledge objects.

A general spectral agglomeration framework is projected [44] to handle multi-type relative agglomeration with totally different forms of objects and attributes, associate degree an alternating improvement algorithmic program is given to seek out an area best. Temporal modification of social networks has been attracting increasing attentions. It has trial and error ascertained that some real-world networks are evolving [45] and a few practitioners try and investigate the network evolve and what may well be an affordable generative method to model the dynamics or the vital factors to see the cluster evolution. On the opposite hand, agglomeration to handle biological process knowledge is additionally developed. It is assumed that agglomeration results of the current state of affairs ought to be the same as the previous time stamps. Rather than taking multiple snapshots of the info and severally agglomeration objects, biological process agglomeration finds out a sequence of agglomeration with temporal smoothness. Latent house model with temporal modification is additionally developed [46], which aims to seek out associate degree embedding that is consistent with the trade-off between previous time stamp and current distance info extracted from the social network. [47] proposes a general framework to handle dynamic single-mode network by casting it as a graph coloring problem and a few greedy heuristics is developed to handle large-scale knowledge. All the same works are specializing in knowledge with attributes or single-mode network.

3. Dirichlet Process Mixture Model

The Existing Mixture model is often used in clustering. This model assumes observation Oi is generated from K fixed number of different statistical methods. $(clusters)\{\phi_k\}_{K=1}^{K}$ with different component weights $\prod_K$. By maximizing the log-likelihood of all the observations, both the component weights and the parameters for each cluster are obtained, and a soft clustering can be achieved accordingly.

A mixture model can be formalized as

$$O_i \sim \sum_{K=1}^{K} \Pi_K P(O_i | Z_i = K \quad \text{Oi} \quad)(1)$$

Where Zi denotes the hidden cluster label associated with object Oi. However, it is usually difficult for people to specify the correct cluster number K in the mixture model. Dirichlet Process Mixture Model is a typical way to solve the problem, where the cluster number is considered as countable infinite, and the distribution of component weights follows a Dirichlet Process (an extension of Dirichlet Distribution to infinite space) with a base distribution $G0$.

Define the DPM model as

$$O_i | \phi_i f(\phi_i)$$

$$O_i | G \sim G \rightarrow (2)$$

$$G \sim DP(G0, \infty)$$

Where

$\emptyset_1$is the parameter of the cluster associated with Oi and it follows the distribution of G. The distribution G is$^{\propto}$ $G0$.$^{\propto}$ is the concentration parameter.This model is equivalent to the following infinite mixture models, with the cluster number K goes to infinity:

$$O_i | Z_i \{\phi_K\}_{K=1}^{K} \sim f(\theta_{zi})$$

$$Z_i | \pi \sim Drichlet(\pi_1, \ldots\ldots\ldots \pi_k)$$

$$\phi_k \sim G0$$

$$\pi \sim Dirichlet(\alpha | K, \ldots\ldots \alpha | K)$$

Where Z_i stands for the latent class label of the observation O_i.In this model, given the cluster number K, the parameters for all the clusters are drawn from the same prior distribution $G0$, and the component weights are strained from a Dirichlet Distribution as former.

4.The Proposed Convergence Aware Dirichlet Process Mixture Model (CADPM)

We propose a slightly different model for q that allows families over T to be nested. L goes to inanity but we tie the parameters of all models after a specific level T. In particular, we impose the condition that for all components with indexes i>T the variational distributions for the social network-length $q_{vi}(vi)$ and the variational distributions for the components $q_{\eta i}(\eta i)$ are equal to their corresponding priors, i.e. $q_{vi}(vi\phi_i^{v}) = pv(vi | \alpha) and$ $q_{\eta i}(vi_i\phi_i^{\eta}) = p_{\eta}(\eta_i\lambda)$. We define the free energy F as the limit $F = \lim_{L \rightarrow \infty} F_L \; whereF_L$ is the free energy defined by q and a truncated DP mixture at level L. Using the parameter tying assumption for i>T the free energy reads

$$F = \sum_{i=1}^{T} \left\{ E_{qv_i} [\log \frac{q_{vi}(vi; \phi_i^{v})}{pv(vi | \alpha}] + E_{q\eta_i} [\log \frac{q_{\eta i}(\eta i; \phi_i^{\eta})}{p\eta(\eta i | \lambda}] \right\} +$$

$$\sum_{i=1}^{T} E_q [\log \frac{q_{zin}(z_n)}{pz(z_n | V) px(xi | n_{zn}} \rightarrow (1)$$

T defines an implicit truncation level of the variational mixture, since there are no free parameters to optimize beyond level T. the free energy F is a function of T parameters $\{\phi_i^{v}; \phi_i^{v}\}_{i=1}^{T}$

and N distributions $\{q_{z_n}(z_n)\}_{n=1}^{N}$.Data-cases may now assign nonzero responsibility to components beyond level T and therefore each $q_{z_n}(z_n)$ must now have infinite support (which requires computing infinite sums in the various quantities of interest). An important implication of our setup is that the variational families are now nested with respect to T and as a result it is guaranteed that as we increase T there exist solutions that decrease F. This is an important result because it allows for optimization with adaptive T starting from T=1 for particular choices of models for q_{vi} and q_{ni} the free energy reads. From (1) we directly see that $q_{z_n}(z_n)$ that minimizes F is given by

$$q_{z_n}(z_n = i) = \frac{ep(S_{n,i})}{\sum_{j=1}^{\infty}\exp(s_{n,i}} \rightarrow (2)$$

Where

$$S_{n,i} = Eqv[\log\ P_z(Z_n = i\,|\,V)] + E_{q_{ni}}[\log\ p_x(x_n\,|\,\eta_i)] \rightarrow (3)$$

Minimization of F over $\phi_i^v and \phi_i^n$ can be carried out by direct differentiation of (1). Using q_{z_n} from (2) the free energy (1) reads

$$F = \sum_{i=1}^{T}\left\{E_{qv_i}[\log\frac{q_{vi}(vi;\phi_i^v)}{pv(vi\,|\,\alpha}] + E_{q\eta_i}[\log\frac{q_{\eta i}(\eta i;\phi_i^\eta)}{p\eta(\eta i\,|\,\lambda}]\right\} - \rightarrow (4)$$
$$\sum_{n=1}^{N}\log\sum_{i=1}^{\infty}\exp(S_{n,i})$$

Evaluation of F requires computing the infinite sum $\sum_{i=1}^{\infty}\exp(S_{n,i})$ in (4). The difficult part is $\sum_{i=T+1}^{\infty}\exp(S_{n,i})$ .Under the parameter tying assumption for i>T most terms of $(S_{n,i})$ in (3) factor out of the infinite sum as constants (since they do not depend on i except for the term

$$\sum_{j=T+1}^{i-1}E_{p_v}[\log(1-V)] = (i-1-T)E_{p_v}[\log(1-v)] .$$

From the above, the infinite sum can be shown to be

$$\sum_{i=T+1}^{\infty}\exp(S_{n,i}) = \frac{\exp(S_n T+1)}{1-\exp(E_{pv}[\log(1-v)]} \rightarrow (5)$$

Using the variatiotnal q (W) as an approximation to the true posterior $P(W\,|\,X,\theta)$, the required posterior over data labels can be approximated by $P(z_n\,|\,X,\theta \approx q_{z_n}(Z_n))$ .Although $q_{z_n}(z_n)$ has infinite support; in practice it suffices to use the individual $q_{z_n}(z_n = i)$ for the finite part $i \le T$, and the Cumulative $q_{z_n}(z_n > T)$ for the infinite part. As a final point, using this above parameter tying supposition for i>T, and the identity $\sum_{i=1}^{\infty}\pi_i(v) = 1$ the predictive density $P(x\,|\,X,\theta)$ can be approximated by

$$P(x\,|\,X,\theta) \approx \sum_{i=1}^{T}E_{qv}[\pi_i(v)]E_{q_{ni}}[\log - px(x_n\,|\,\eta_i) +$$
$$[1 - \sum_{i=1}^{T}E_{pv}[\pi_i(v)]E_{p\eta}[\log - px(x_n\,|\,\eta\ ) \rightarrow (6)$$

Note that each and every one quantity of concern, such as the free energy (4) and the predictive distribution (6), can be computed analytically even though they involve infinite sums. In table 1, given the parameters used by the proposed algorithm for better understanding of the algorithm. It helps the users to know the work flow of the algorithm.

TABLE 1
Parameters Used For CADPM Algorithm

Q	Variational distribution
T	truncation level
P	Probability density function
vi	Length of a network community
ηi	Social users
Z_n	Component label
W	Latent variables
L	Level
S	Sum
F	Free energy
E	Expectation value
α, λ	Hyper parameters

5. Experimental Results and Analysis

We evaluate the effectiveness of the proposed CADPM algorithm using three large scale data sets. Those are Amazon data set, Gnutella data set, and Stanford data set. The Amazon data set was collected from Amazon website. This contains the information about the customers and their purchase information. If a particular product i is purchased continuously with the product j then the graph contains the directed edges from i to j. It consists of 262111 nodes and 1238477 edges among them. Average of co-efficient is 0.4198. The Gnutella data set consists of a series of snapshots of the peer-to-peer file sharing network information from the year of 2002 august. The nodes represent the hosts in the Gnutella network topology and edges between the nodes represent the connections between the hosts. Totally it consists of 10876 nodes, and 39994 edges. The average of co-efficient is 0.0062. And the Stanford data set consists of the pages information of the Stanford University (i.e. stanford.edu). Nodes represent the pages and edges represent hyperlinks between them. It has 281903 nodes and 2312497 edges. And the average co-efficient is 0.5976. The clustering results of the proposed CADPM method is compared with the previously proposed Dirichlet process modest. Accuracy, Precision and Recall parameters are used to evaluate the algorithms performance. In table 2, we have given the overall performance comparison results.

TABLE 2
CADPM Performance Comparison with Existing DP Method

Number of Data	Data set	Accuracy		Precision		Recall	
		Dirichlet Process	CADPM	Dirichlet Process	CADPM	Dirichlet Process	CADPM
1000	Amazo	87.1	95.5	0.8	0.96	0.87	0.96
	Gnutell	92.9	95.4	0.9	0.95	0.93	0.95
	Stanfor	90.9	95.2	0.9	0.95	0.91	0.95
2000	Amazo	88.5	94.1	0.8	0.94	0.89	0.94
	Gnutell	89.2	96.8	0.8	0.97	0.89	0.97
	Stanfor	90.6	95	0.9	0.95	0.91	0.95
3000	Amazo	88.5	96.8	0.8	0.97	0.89	0.97
	Gnutell	89.4	96.8	0.8	0.97	0.89	0.97
	Stanfor	90.4	95.3	0.9	0.95	0.9	0.95
4000	Amazo	87.3	94.4	0.8	0.94	0.87	0.94
	Gnutell	92.7	95.8	0.9	0.96	0.93	0.96
	Stanfor	92.3	95	0.9	0.95	0.92	0.95
5000	Amazo	87.6	94.5	0.8	0.95	0.88	0.95
	Gnutell	92.4	95.7	0.9	0.96	0.92	0.96
	Stanford	89.5	94.6	0.9	0.95	0.9	0.95

The results of the CADPM method are given in the following figures.

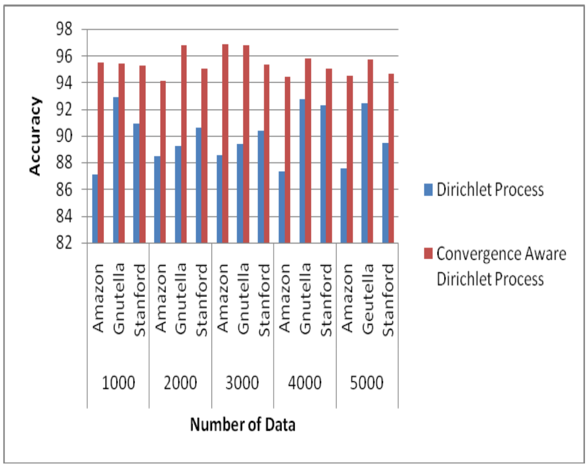

Fig 1: Accuracy Comparison

We obtain the accuracy value of the proposed CADPM algorithm. Accuracy Comparison is shown in figure 1. As we use large scale data set in the experiments, it is not possible to test all nodes and their edges in the network to get the exact accuracy value. To calculate the accuracy for every experiment we randomly choose 1000 nodes to test the algorithm performance. We tested 5000 nodes for each datasets. First we choose 1000 Amazon data set, for that the existing Dirichlet process model gives 87.10% accuracy, but the proposed CADPM gives 95.50% accuracy. In case of Gnutella data set, first we choose 1000 data, for this existing method gives 92.90% and the proposed CADPM method gives 95.40% accuracy. We also tested 1000 data from Stanford data set, the existing method gives 90.95% accuracy, and the proposed algorithm gives 95.25% accuracy. We tested our method up to 5000 data for each data set. The existing method gives 87.81% average accuracy with Amazon data set, but the proposed CADPM algorithm gives 95.07% average of accuracy. We tested the methods with Gnutella data sets, for 5000 data, the existing method gives 91.35% accuracy in average, but the proposed method gives 96.11% accuracy in average. For Stanford data set, the existing method gives 90.75% average accuracy and the proposed method gives 95.04% of accuracy in average.

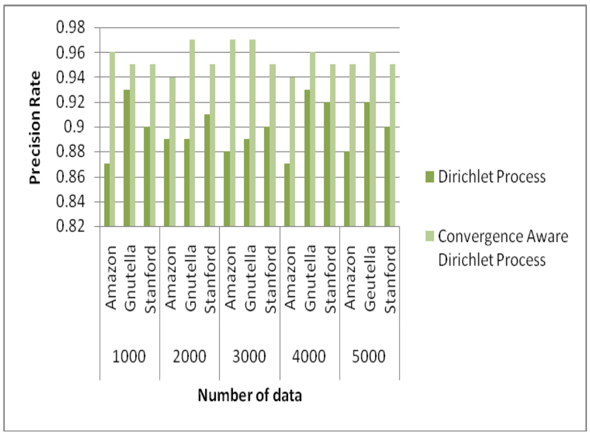

Fig 2: Precision rate comparison

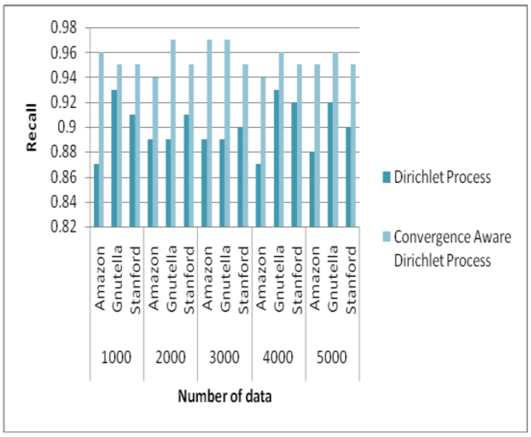

Fig 3: Recall rate comparison

Figure 2 shows the comparison results of precision rate among the existing method and the proposed CADPM method. Like accuracy calculation we choose 5000 data from each data set and we tested increasing number of 1000 data. First we tested 1000 data from Amazon data set, the existing Dirichlet process model gives 0.87 precision rates, but the proposed CADPM gives 0.96 precision rates. In case of Gnutella data set, first we choose 1000 data, for this, the existing method gives 0.93 precision, and the proposed CADPM algorithm gives 0.95 precision. We tested 1000 data from Stanford data set the existing method give 0.90 precision rates, but the proposed algorithms gives 0.95. In order to calculate the overall system performance in terms of precision we tested 5000 data from each data set, In Amazon data set the existing method gives the average of 0.89 precision rates, and the proposed CADPM algorithm gives 0.95 average of precision. We tested the methods with Gnutella data sets, for 5000 data, the existing method gives 0.91of precision rate in average, but the proposed method gives 0.96 in average. For Stanford data set, the existing method gives 0.91 in average and the proposed method gives 0.95 precision rates in average.

The comparison of the recall values for the proposed system and existing Dirichlet process model is given in figure 3. To evaluate the recall rate of the proposed algorithm we choose 5000 data from each data set, In Amazon data set the existing method gives an average of 0.88 recall rates, and the proposed CADPM algorithm gives 0.95 average of recall. For Gnutella data sets we choose 5000 data, the existing method gives 0.91of recall rate in average, but the proposed method gives 0.96 in average. For Stanford data set, the existing method gives 0.91 in average and the proposed method gives 0.95 recall rates in average. We can say that the proposed CADPM algorithm gives better performance in terms of accuracy, precision and recall rates. Our proposed CADPM algorithm achieves a higher level of performance.

6. Conclusion

We proposed a CADPM algorithm to find the evolutionary communities in heterogeneous networks. The existing Dirichlet Process (DP) mixture models used for the clustering applications, but here the number of clusters is unknown. A CADPM model is used to automatically establish the number of clustering communities in advance. In order to evaluate the proposed algorithm performance we tested our algorithm with three large scale data sets. Proposed method gives 5.43% higher accuracy when compared to the existing method. Through or experimental analysis results, we can say that the proposed method performs well and it is recommended to solving the problem with complex relationships. It gives higher clustering results in heterogeneous networks.

References

[1] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. 2008. Community evolution in dynamic multi-mode networks. KDD '08, New York, NY, USA. ACM. Pages: 677-685.

[2] Yizhou Sun, Jie Tang, Jiawei Han, Manish Gupta and Bo Zhao. 2010. Community Evolution Detection in Dynamic Heterogeneous Information Networks. ACM I978-1-4503-0214-2/10/07.

[3] Dirichlet process. http://en.wikipedia.org/wiki/Dirichlet_process

[4] J. Shi and J. Malik. 1997. Normalized cuts and image segmentation. In CVPR'97, page 731, Washington, DC, USA, 1997. IEEE Computer Society.

[5] U. von Luxburg. 2006. A tutorial on spectral clustering. Technical report, Max Planck Institute for Biological Cybernetics.

[6] S. White and P. Smyth. 2005. A spectral clustering approach to finding communities in graph. In SDM '05.

[7] M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. Physical Review E, 69(2).

[8] M. E. J. Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74:036104.

[9] T. A. B. Snijders. 2002. Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure.

[10] P. D. Hoff, A. E. Raftery, and M. S. Handcock. 2001. Latent space approaches to social network analysis. Journal of the American Statistical Association, 97. Pages:1090-1098.

[11] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. 2007. Model-based clustering for social networks. Journal of the Royal Statistical Society Series A, 170(2). Pages:301-354.

[12] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. 2008. Mixed membership stochastic blockmodels. J.Mach. Learn. Res., 9:1981-2014.

[13] H. Zha, X. He, C. Ding, H. Simon, and M. Gu. 2001. Bipartite graph partitioning and data clustering. In CIKM '01, New York, NY, USA. ACM. Pages; 25-32.

[14] S. Dhillon. 2001.Co-clustering documents and words using bipartite spectral graph partitioning. In KDD '01, New York, NY, USA, ACM. Pages 269-274.

[15] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. 2009. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In EDBT '09, New York, NY, USA. ACM. Pages;565-576.

[16] Y. Sun, Y. Yu, and J. Han. 2009. Ranking-based clustering of heterogeneous information networks with star network schema. In KDD '09, New York, NY, USA, ACM. Pages;797-806.

[17] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. 2007. Evolutionary spectral clustering by incorporating temporal smoothness. In KDD '07, New York, NY, USA, ACM. Pages; 153-162.

[18] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. 2008. Community evolution in dynamic multi-mode networks. In KDD '08, New York, NY, USA. ACM. Pages; 677-685.

[19] M.-S. Kim and J. Han. 2009. A Particle-and-Density Based Evolutionary Clustering Method for Dynamic Networks. International Conference on Very Large Data Bases, Lyon, France.

[20] P. Sarkar and A. W. Moore.2005. Dynamic social network analysis using latent space models. SIGKDD Explor.Newsl., 7(2), pages:31-40.

[21] W. Fu, L. Song, and E. P. Xing. 2009. Dynamic mixed membership blockmodel for evolving networks. In ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, New York, NY, USA, ACM. Pages;329-336.

[22] R. M. Neal. 2000. Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2), pages:249-265.

[23] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476), pages:1566-1581.

[24] X. Z. Zoubin, X. Zhu, Z. Ghahramani, and J. Lafferty. 2005.Time-sensitive dirichlet process mixture models. Technical report.

[25] L. Ren, D. B. Dunson, and L. Carin.2008. The dynamic hierarchical dirichlet process. In ICML '08, New York, NY, USA, 2008. ACM. Pages:824-831.

[26] T. Xu, Z. M. Zhang, P. S. Yu, and B. Long. 2008. Dirichlet process based evolutionary clustering. In ICDM '08, Washington, DC, USA, IEEEComputer Society. Pages: 648-657.

[27] T. Xu, Z. M. Zhang, P. S. Yu, and B. Long.2008. Evolutionary clustering by hierarchical Dirichlet process with hidden markov state. In ICDM '08, Washington, DC, USA, IEEE Computer Society. Pages:658-667.

[28] K. Nowicki and T. A. B. Snijders. 2001. Estimation and prediction for stochastic block structures. Journal of the American Statistical Association, 96(455). Pages:1077–1087.

[29] C. Kemp, T. L. Griffiths, and J. B. Tenenbaum. 2004. Discoverying latent classes in relational data. Technical report, Massachusetts Institute of Technology.

[30] E. Airodi, D. Blei, S. Fienberg, and E. P. Xing. 2008. Mixed membership stochastic blockmodels. JMLR.

[31] B. Long, Z. M. Zhang, and P. S. Yu. 2007. A probabilistic framework for relational clustering. In KDD, pp 470–479.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

[33] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha. 2006. Probabilistic models for discovering e-communities. In WWW, Pages 173–182.

[34] X. Wang, N. Mohanty, and A. McCallum. 2006. Group and topic discovery from relations and their attributes. In NIPS, pages 1449–1456.

[35] A. McCallum, X. Wang, and A. Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on enron and academic email. Journal of Artificial Intelligence Research, (0), pages: 249–272.

[36] M. S. H. Peter D. Hoff, Adrian E. Raftery. 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association.

[37] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. 2007. Model-based clustering for social networks.

Journal Of The Royal Statistical Society Series A, 127(2).

[38] A. Globerson, G. Chechik, F. Pereira, and N. Tishby. 2007. Euclidean embedding of co-occurrence data. J. Mach. Learn. Res., 8, pages: 2265–2295.

[39] I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In KDD, Pages: 269–274.

[40] H. Zha, X. He, C. Ding, H. Simon, and M. Gu. 2001. Bipartite graph partitioning and data clustering. In CIKM, New York, NY, USA, ACM. Pages 25–32.

[41] B. Long, Z. M. Zhang, and P. S. Yu. 2005. Co-clustering by block value decomposition. In KDD.

[42] B. Gao, T.-Y. Liu, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. 2005. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. In KDD, pages: 41–50.

[43] I. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W.-Y. Ma. 2003. Recom: reinforcement clustering of multi-type interrelated data objects. In SIGIR.

[44] B. Long, Z. M. Zhang, X. W´u, and P. S. Yu. 2006. Spectral clustering for multi-type relational data. In ICML, pages:585–592.

[45] G. Kossinets and D. J. Watts. 2006. Empirical analysis of an evolving social network. Science, 311(5757).

[46] P. Sarkar and A. W. Moore. 2005. Dynamic social network analysis using latent space models. SIGKDD Explor. Newsl., 7(2), Pages:31–40.

[47] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. 2007. A framework for community identification in dynamic social networks. In KDD.

[48] Dongsheng Duan, Yuhua Li, Ruixuan Li, Zhengding Lu and Aiming Wen. 2012. MEI: Mutual Enhanced Infinite Community-Topic Model for Analyzing Text-augmented SocialNetworks. The Computer Journal.