International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-3 Issue-6, November 2013

A Novel Algorithm for Automatically Detecting Number of Clusters for Mining Communities in Heterogeneous Social Networks

Renuga Devi. R, Hemalatha. M

Abstract— Social media have attracted millions of user's attention in recent years. In a distributed social network a community mining is one of the major research areas. Mining of network communities is a major problem now a day. This problem should be avoided. Several methods were proposed, but most of the methods of community mining consider the homogeneous network. But in distributed network there are multiple networks are interconnected with each other which are known as heterogeneous networks. Each network represents a specific kind of relationship. Same time each relationship plays an important place in a distinct situation. Mining of such an important community in a distributed environment is a difficult task. To overcome the above mentioned problem, this paper presents a novel Convergence aware Dirichlet Process Mixture Model (CADPM) for automatically mining the network communities in heterogeneous networks. The earlier Dirichlet Process (DP) mixture model is unsuitable in some situation. The number of clusters for community mining is unknown in prior. So the CADPM is proposed to handle the large number of data-cases.

Index Terms— Community, Dirichlet Process, Heterogeneous Network, Hidden Communities, Social Network.

I. INTRODUCTION

Most of the social networks are modeled as a graph for analysis. The nodes in a network represent individual person, or an organization, or a computer, or any resources. The edges between the nodes show a relationship between the human. Several algorithms have been proposed to analyze the social networks. But existing methods take for granted that there is a single kind of social network is available, which is known as homogenous network. But in real world social networks various types of relationships between the nodes is existing. We can model each relationship as a different graph or network. Having multiple relationships in a network is known as multi relational network or heterogeneous network. Each relation plays different important roles in various situations. Main problems in social network analysis are discovering a particular group of individuals who is sharing the some common information in a network. Finding the importance nodes in a heterogeneous network is also an important task in social network analysis. To overcome these problems, first we need to find relation who plays a significant role in a community. But that kind of relations may not be present explicitly. We have to find such hidden community's relationship in a network before performing community mining operation. For example in a real world network community consists of several relations. Few people study at the same place, few people share the common knowledge, few works in the same place, few will meet same place, etc. In mathematics we can represent the above scenario as graph format. The nodes indicate human and the edges between the nodes show the relationship strength. Various relationships are there, so the network should contain heterogeneous relationship. It is possible to model this network as a number of homogeneous network or graphs at some situation. Every graph indicates a particular kind of relation. The relationship between the nodes cannot contain equal strength. In some situation few relations play an important role, in some other situation other relations may play an important role. Here the problem is finding a relation which is most related to the user queries. There some hidden relations may exist which is more suitable to user requirements. In general community mining problem has several comparable properties to the graph cut problem. Community mining and social network analyzing problem are considered as graph mining. Community mining problem also called as sub graph identification [1].

Now a day an evolutionary clustering plays an important role in data mining applications. Evolutionary clustering methods used with the data set, where it is increasing over the particular time, when the number of data has a large number of clusters. And also useful if the group of data created from a particular time to different time and added data may connect to the group and already available data may depart. Same time newly created clusters will appear on the network and already available clusters have disappeared from the network. Alternatively, the evolutionary clustering problem has numerous solutions in different applications [2].

The Dirichlet Process Mixture model also known as DPM model. From the evolutionary data DPM model automatically takes the quantity of the clusters. Also the cluster combination amount of information at various time intervals is used to reproduce an even cluster revolutionize over the particular time. The Dirichlet Process Mixture model is a statistical model. It was developed to capture the sharing reservations in the space of prospect measure in the statistics literature. Dirichlet Process Mixture model extends the DP if a random measure is no longer in a single distribution. Mathematically clustering problem really fits into a mixed mode and makes use of DPM model naturally. Another important process is, the Dirichlet Process Model allows an unlimited amount of mixture components, detaching the illumination on solving the clustering representation assortment problem [3].

Manuscript published on 30 November 2013.
***Correspondence Author(s)
Renuga Devi. R, Department of Computer Science, Karpagam University, Coimbatore, India.
Hemalatha. M, Department of Computer Science, Karpagam University, Coimbatore, India.

The remaining part of the paper is prepared as follows. In part 2 consists of literature survey about heterogeneous network and Dirichlet Process Mixture model and its uses I n community mining problem, Next in part 3, our proposed Convergence Aware Dirichlet Algorithm for heterogeneous community mining problems. Performance evaluation of proposed algorithm is discussed in part 4. In part 5 we provide a few suggestions and conclusions.

II. EXISTING METHODS

Most of the existing work in social network analysis and statistics was not much efficient. Because the lack of collecting data and most of the existing methods focused only on very small kind of networks. Now a day the usage of the Internet has grown with the several social networking sites. And large numbers of data are available and several research works started to analyze the data [1]. Social network Community mining and detection, clustering in social networks has been considered for a reasonably long time. The existing research methods are trying to divide the large network into several independent parts and group similar nodes into the same clusters.

First community mining and detection problem started with the homogeneous social networks. Existing methods, such as modularity based methods [4, 5], spectral clustering algorithms [6, 7, 8], and based on probabilistic models [9, 10, 11, 12], and presently to bipartite networks [13, 14], and recently on heterogeneous networks [15, 16]. Recently, the heterogeneous network with star network schema was proposed for real world networks. It was different from the existing works and other community detection and mining methods. It mainly focused on the model of the dynamic development of the net cluster based multi relational communities. In general at every time new nodes may join the dynamic network, some nodes may leave from the network. Finding evolutionary clusters of such dynamic network sequences will help the people to understand the development of communities in the network [16].

Apart from the traditional methods few other methods have been proposed on homogeneous networks, which were extended from the basic clustering methods. Recently, in [17] studied the development of heterogeneous social networks. On the other hand, the network communities are distinct as a single kind of objects, and cluster count for every community need to be prefixed and precise by users. These methods will not optimize clustering performance. Studies in community detection with automatic cluster numbers is needed for heterogeneous networks.

Dirichlet method provides the simplest way to feature priors to the cluster range of mixture models, and this is often terribly useful to make a decision the cluster range mechanically. Recently, some works have extended the Dirichlet method into considering time data. Another DP-based extension planned to model organic process clump [18, 19]. The variations of these models of these strategies include, the proposed method offers a particular answer to net-cluster evolution in heterogeneous networks and author outlined a completely unique generative model for net-cluster evolution. This may model the evolution of constant cluster in several timestamps, whereas several existing works need constant clusters (atom distributions) don't amendment among totally different timestamps, final model doesn't claim a worldwide reasoning of the model, however greedy reasoning at every time stamp, that is additional sensible for timely change the evolution [20].

In the year of 1999, researchers Kumar et al., Used the method called bipartite graph to search out the center of the community, then they expanded the center to urge the specified community [21]. Flake et al., applied the maximum-flow and minimum cut framework of the community mining [22]. The authority-and-hub plan [23] was conjointly utilized in the community mining [24, 25, and 26] and has many extensions [27]. The concept of frequent item set in association rule mining has conjointly been utilized in community mining [28].

Schwartz and Wood well-mined social relationships from email logs [29]. The Referral Web project was projected to mine a social network from a large kind of net information, and use it to assist people notice consultants WHO might answer their queries [30]. Adamic and Jewish calendar month tried to get the social interactions between folks from the data on their homepages [31]. Agrawal et al., analyzed the social behavior of the folks on the newsgroups [32]. Moreover, the online it may be truly viewed as an outsized social network. The famous link analysis algorithms, like Google's PageRank [33] and Kleigberg's HITS rule [23], may be seen as social network analysis on the online. Because of some statistical considerations the existing methods are not much suitable for mining hidden communities in the large scale heterogeneous networks. We proposed a CADPM algorithm to handle these issues in heterogeneous networks.

III. THE PROPOSED CONVERGENCE AWARE DIRICHLET PROCESS MIXTURE MODEL (CADPM)

The existing Mixture model is mainly used for clustering purposes. The earlier model assumes that an observation Oi is created from a fixed number of different statistical methods, which is K. $(clusters)\{\phi_k\}_{k=1}^{K}$ with different component weights π_k. The model works by increasing the log values for all the observations, each the part weights and also the parameters for every cluster are collected, and a soft agglomeration will be achieved consequently.

However, it's typically troublesome for individuals to specify the correct cluster variety K within the mixture model. Dirichlet method Mixture Model may be a distinctive method to solve the matter, wherever the cluster variability is taken into account as denumerable infinite, and also the distribution of part weights follows a Dirichlet method (an extension of Dirichlet Distribution of infinite space) with a base distribution G0.

The authors defined the Dirichlet Process Mixture model as

$(o_i \mid \phi_i \sim f(\phi_i))$ **Error! Bookmark not defined.**

$(o_i \mid G \sim G)$

Where ϕ_i is the parameter of the cluster associated with o_i and it follows the distribution of G. The distribution G is $\alpha\ G_o\ \alpha$ the concentration parameter.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In this existing model, the cluster number K is given, the parameters for all the clusters are tired from the similar prior distribution G0, and the component weights are drained from a Dirichlet Distribution as the previous.

In this proposed CADPM model we slightly differentiate the variational model for q that allows families over T to be nested. L goes to infinite but we tie the parameters of all models after a specific level T.

Especially we enforce the situation that for all the components with index i>T the variatiotnal 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. $q_{vi}(vi,\phi_i^v)=pv(vi\mid\alpha)$ and $q_{vi}(vi,\phi_i^\eta)=p\eta(\eta i\mid\lambda)$.

In our model we define the free energy F as the limit $F=\lim_{L\to\infty}F_L$ Where F_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.

Variational model for variation distribution q can be proposed to allow the family of network to be nested with truncate value. Than Impose condition that the variations distributions for the stick-lengths and the variational distributions for the components should be equal with index value which should be greater than Truncation value. $q_{vi}(vi,\phi_i^v)=pv(vi\mid\alpha)$ and $q_{vi}(vi,\phi_i^\eta)=p\eta(\eta i\mid\lambda)$

Calculate the free energy with respect to variatiotnal distribution

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

F =

$$+\sum_{i=1}^{T}E_q\left[\log\frac{q_{z_n}(z_n)}{pz(z_n\mid v)px(x_i\mid\eta_{z_n})}\right]$$

We can define an implicit truncation level or variational mixture. Then have to assign nonzero responsibility to components beyond the level of truncate value. By optimizing the truncation value of network by increasing Truncation value to decrease free energy

$$q_{z_n}(z_n=i)=\frac{\exp(S_n,i)}{\sum_{j=1}^{\infty}\exp(S_n,i)}$$

By minimizing free energy we can get the exact value

$$F=\left\{\sum_{i=1}^{T}E_{q_v}\left[\log\frac{q_{vi}(vi\phi_i^v)}{pv(vi\mid\alpha)}\right]+E_{q_{\eta i}}\left[\log\frac{q_{\eta i}(\eta i\phi_i^\eta)}{p\eta(\eta i\mid\alpha)}\right]\right\}$$

$$-\sum_{n=1}^{N}\log\sum_{1=1}^{\infty}\exp(S_n,i)$$

Then can calculate the density function

$$p(x\mid X,\theta)\approx\left|\sum_{i=1}^{T}E_{q_v}\left[\pi_{i(v)}\right]E_{q_{\eta i}}\left[\log px(x_n\mid\eta_i)\right]\right|+$$

$$\left[1-\sum_{i=1}^{T}E_{p_v}\left[\pi_{i(v)}\right]E_{p_{\eta i}}\left[\log px(x_n\mid\eta)\right]\right]$$

In table 1, we have given the proposed CADPM algorithm procedure.

Table 1: The proposed CADPM Algorithm

Step1: Find the log-likelihood of all the observations to calculate similarity for clustering
Step 2:Initialize object Oi
Step 3:Construct mixture model

$$O_i\sim\sum_{k=1}^{K}\pi_{k0}P(O_i\mid Z_{i=k})$$

Step 4: Define the DPM model

$$(o_i\mid\phi_i\sim f(\phi_i))$$
$$(o_i\mid G\sim G)$$
$$(G\sim DP(G_O,\infty))$$

Step 4:Finite infinite number of clusters with cluster number k

$$O_i\mid Z_i\{\phi_k\}_{K=1}^{K}\sim f(\theta_{zi})(o_i\mid G\sim G)$$
$$Z_i\mid\pi\sim Discreate(\pi_1,\pi_2,.....\pi_k)$$
$$\phi_k\sim G_O$$
$$\pi\sim Dirichlet(\alpha\mid K,.....\alpha\mid K-1$$
$$Dirichlet(v,q,z)$$

Step 4:End

IV. RESULTS AND DISCUSSIONS

The proposed CADPM algorithm is evaluated using the Stanford data set. The Stanford data set consists of the web page details of the Stanford University. That is stanford.edu. In this Stanford network, the nodes represent the web pages, and the edges represent the hyperlinks between the nodes. The entire network has 281903 numbers of nodes and 2312497 numbers of edges. The average co-efficient of the network is 0.5976. The proposed algorithm's clustering efficiency is compared with the existing Dirichlet Process Mixture model. For the evaluation three parameters are used. Namely Accuracy value, Precision rate and Recall rate. The values are represented in table 2. For the testing process 5000 nodes are used.

Table 1: Performance Comparison of the proposed CADPM algorithm

Parameters Used	Dirichlet Process Model	Convergence Dirichlet Process Mixture Model
Accuracy	90.75	95.04
Precision	0.91	0.95
Recall	0.91	0.95

The graphical representations of the above comparison values are given below. The following graph shows the clustering comparison of the existing Dirichlet process model and the proposed CADPM algorithm.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

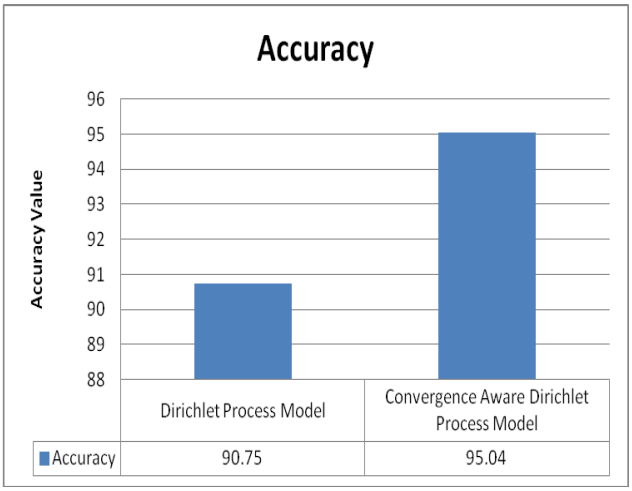

Figure 1: Accuracy Value Comparison

The above Figure 1 indicates that the comparison of the accuracy value of the existing Dirichlet Process model and the proposed CADPM algorithm. In the graph, the accuracy values are represented in % value at Y-axis, and the algorithm is represented in X-axis. The proposed system accuracy value is higher than the existing Dirichlet process model's accuracy value. The proposed CADPM algorithm gives 95.04% accuracy and the existing method gives 90.75% of accuracy. From this results analysis, it is clear that the proposed algorithm is producing accurate clustering results compared to the existing method. The proposed system is 4.29% better than the existing system.

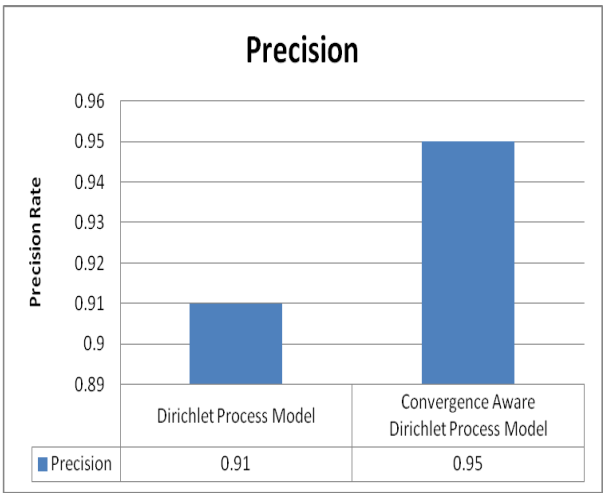

Figure 2: Precision Rate Comparison

The above Figure 2 indicates that the comparison of the precision of the existing Dirichlet Process model and the proposed CADPM algorithm. In the graph, the precision values are represented in % value at Y-axis, and the algorithm is represented in X-axis. The proposed system precision rate is higher than the existing Dirichlet process model precision rate. The proposed CADPM algorithm has 0.95% precision rate and the existing method gives 0.91% of precision rate. From this results analysis, it is clear that the proposed algorithm works better when compared to the existing method. The proposed system has 0.04 higher precision rate than the existing algorithm.

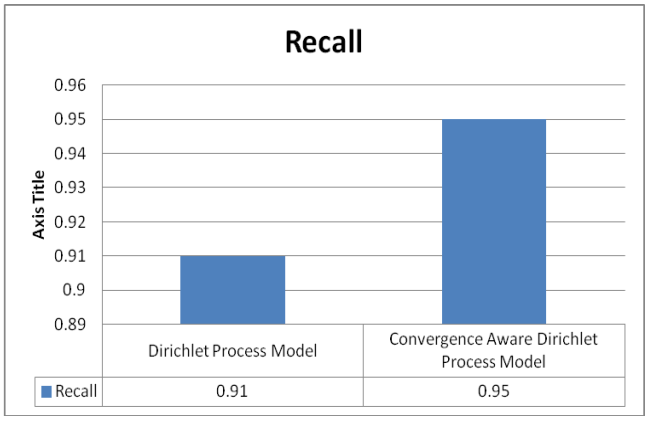

Figure 3: Recall Rate Comparison

The above Figure 3 indicates that the comparison of the recall rate of the existing Dirichlet Process model and the proposed CADPM algorithm. In the graph, the recall values are represented in % value at Y-axis, and the algorithm is represented in X-axis. The proposed system recall rate is higher than the existing Dirichlet process model recall rate. The proposed CADPM algorithm has 0.95 recall rate and the existing method gives 0.91 recall rate. From this results analysis, it is clear that the proposed algorithm works better when compared to the existing method. The proposed system has 0.04 higher recall rate than the existing algorithm.

V. CONCLUSION

This paper presents a novel Convergence aware Dirichlet Process Mixture Model (CADPM) for community mining problem in heterogeneous networks. The existing Dirichlet Process (DP) mixture model was used for the community mining problem. But it has few drawbacks. The number of clusters for clustering process is unknown in prior. Most of the existing methods are not suitable for a large scale mining application. To overcome these problems the CADPM model was proposed, and it gives 4.29% higher clustering accuracy than the existing method. From the experimental analysis, the proposed method works best in large scale networks.

REFERENCES

1. Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan and Jiawei Han. 2005. Mining Hidden Community in Heterogeneous Social Networks. Proceedings of the 3rd international workshop on Link discovery. Pages: 58 – 65.
2. Tianbing Xu, Zhongfei (Mark) Zhang, Philip S. Yu and Bo Long. 2008. Dirichlet Process Based Evolutionary Clustering. Eighth IEEE International Conference on Data Mining, ICDM '08. Pages: 648 – 657.
3. Jianwen Zhang, Yangqiu Song, Changshui Zhang and Shixia Liu. 2010. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. Pages 1079-1088.
4. M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. Physical Review E, 69(2).
5. M. E. J. 2006 Newman. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74:036104.
6. J. Shi and J. Malik. 1997. Normalized cuts and image segmentation. In CVPR'97, page 731, Washington, DC, USA. IEEE Computer Society.
7. U. von Luxburg. 2006. A tutorial on spectral clustering. Technical report, Max Planck Institute for Biological Cybernetics.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

8. S. White and P. Smyth. 2005. A spectral clustering approach to ¯nding communities in graph. In SDM '05.
9. T. A. B. Snijders. 2002. Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure.
10. P. D. Ho, 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, Pages: 1981-2014,
13. H. Zha, X. He, C. Ding, H. Simon, and M. Gu. 2001. Bipartite graph partitioning and data Clustering. In CIKM '01, pages 25{32, New York, NY, USA, ACM.
14. I. S. Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In KDD '01, pages 269{274, New York, NY, USA, ACM.
15. Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus, 2009. Integrating clustering with ranking for heterogeneous information network analysis. In EDBT '09, Pages:565 - 576, New York, NY, USA, ACM.
16. Y. Sun, Y. Yu, and J. Han. 2009. Ranking-based clustering of heterogeneous information networks with star network schema. In KDD '09, Pages: 797-806, New York, NY, USA, ACM.
17. L. Tang, H. Liu, J. Zhang, and Z. Nazeri. 2008. Community evolution in dynamic multi-mode networks. In KDD '08, Pages 677-685, New York, NY, USA, ACM.
18. T. Xu, Z. M. Zhang, P. S. Yu, and B. Long. 2008. Dirichlet process based evolutionary clustering. In ICDM '08, pages 648-657, Washington, DC, USA, IEEE Computer Society.
19. 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, Pages 658-667, Washington, DC, USA, IEEE Computer Society.
20. Yizhou Sun, Jie Tang, Jiawei Han, Manish Gupta, and Bo Zhao. 2010. Community Evolution Detection in Dynamic Heterogeneous Information Networks. Proceedings of the Eighth Workshop on Mining and Learning with Graphs. Pages 137-146.
21. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. 1999. Trawling the web for emerging cyber communities. In Proceedings of The 8th International World Wide Web Conference.
22. G. W. Flake, S. Lawrence, and C. L. Giles. 2000. Efficient identification of web communities. In Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000).
23. J. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), Pages: 604–622.
24. D. Gibson, J. Kleinberg, and P. Raghavan. 1998. Inferring web communities from link topology. In Proceedings of the 9th ACM Conference on Hypertext and Hypermedia.
25. J. M. Kleinberg. 1999. Hubs, authorities, and communities. ACM Computing Surveys, 31(4).
26. C. Chen and L. Carr. Trailblazing the literature of hypertext: Author co-citation analysis (1989-1998). 1999. In Proceedings of the 10th ACM Conference on Hypertext and hypermedia.
27. D. Cohn and H. Chang. 2000. Learning to probabilistically identify authoritative documents. In Proceedings of the 17th International Conference on Machine Learning.
28. W.-J. Zhou, J.-R. Wen, W.-Y. Ma, and H.-J. Zhang. 2002. A concentric-circle model for community mining. Technical report, Microsoft Research.
29. M. F. Schwartz and D. C. M. Wood. 1993. Discovering shared interests using graph analysis. Communications of the ACM, 36(8), Pages: 78–89.
30. H. Kautz, B. Selman, and M. Shah. 1997. Referral web: Combining social networks and collaborative filtering. Communications of the ACM, 40(3), Pages: 63–65.
31. L. A. Adamic and E. Adar. 2002. Friends and neighbors on the web. Technical report, Xerox Parc.
32. R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. 2003. Mining newsgroups using networks arising from social behavior. In Proceedings of 12th International World Wide Web Conference.
33. L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)