



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Improved BSP Clustering Algorithm for Social Network Analysis

¹T.B.Saranya Preetha, ²G.T.Prabavathi

¹Research Scholar, ²Assistant Professor in Computer science

Department of Computer Science

Gobi Arts and Science College

Gobichettipalayam, Tamil Nadu 638453, India

Abstract: A social network is a social structure of individuals, who are linked (directly or indirectly to each other) depending on a common relation of interest, e.g. friendship, trust, etc. Social network analysis is the study of social networks to recognize the structure and behavior of social actor. Social network analysis has gained importance due to its usage in different applications - from product marketing to search engines and organizational dynamics. The conventional clustering approaches group objects based only on objects' similarity which does not suite for social network data. Social network objects should be grouped into classes depending on their links as well as their attributes. In this paper, a clustering algorithm based on BSP (Business System Planning) clustering with Principal Component Analysis (PCA) technique is proposed. This algorithm produces significant improvement in clusters, as it groups objects in a social network into different classes based on their links and identify relation among classes.

Keywords: Social Network Analysis, BSP Clustering, PCA

I. INTRODUCTION

Social Networks are graph structures whose nodes or vertices represent people or other entities embedded in a social context, and whose edges represent interaction or collaboration between these entities [10]. Social networks are highly dynamic, evolving relationships among people or other entities. This dynamic property of social networks makes studying these graphs a challenging task. A lot of research has been done recently to study different properties of these networks. Such complex analysis of large, heterogeneous, multi-relational social networks has led to an interesting field of study known as Social Network Analysis (SNA).

Social network analysis, which can be applied to analyze the structure and the property of personal relationship, web page links, and the spread of messages, is a research field in sociology. Recently social network analysis has attracted increasing attention in the data mining [1] research community. From the viewpoint of data mining, a social network is a heterogeneous and multi-relational dataset represented by graph [3]. Research on social network analysis in the data mining community includes following areas: clustering analysis [1], classification [8], link prediction [7], Page Rank [9] and Hub-Authority [4] in web search engine.

This paper deals with the social network cluster analysis. In the second section, a social network clustering based on Business System Planning (BSP) clustering algorithm is given. The algorithm can group objects in a social network into different classes based on their links, and it can also identify the relations among classes. In the third section, importance of Principle Component Analysis technique is dealt. Fourth section describes the methodology of BSP clustering. Section five shows how BSP clustering is improved by applying the PCA reduction technique.

II. BSP Clustering in Social Networks

There has been extensive research work on clustering in data mining. Traditional clustering algorithms [1] divide objects into classes based on their similarity. Objects in a class are similar to each other and are very dissimilar from objects in different classes. Social network clustering analysis, which is different from traditional clustering problem, divides objects into classes based on their links as well as their attributes. The biggest challenge of social network clustering analysis is how to divide objects into classes based on objects' links, and to find algorithms that can meet this challenge.

The BSP clustering algorithm [11] is proposed by IBM. It is designed to define information architecture for the firm in business system planning. This algorithm analyses business process and their data classes, cluster business process into subsystems, and define the relationship of these sub-systems.

Basically BSP clustering algorithm uses objects (business processes) and links among objects (data classes) to make clustering analysis. Similarly social network also includes objects and links among these objects. In view of the same pre-condition, the BSP clustering algorithm can be used in social network clustering analysis. According to graph theory, social network is a directed graph composed by objects and their relationship. Figure 1 shows a sample of social network, the circle in the figure represents an object; the line with arrow is an edge

of the graph, and it represents directed link between two objects, so a social network is a directed graph. In figure 1, Let O_i be an object in social network ($i = 1 \dots m$), let E_j which means directed link between two objects, be a directed edge of the graph ($j = 1 \dots n$). After definition of objects and directed edges, reachable relation between two objects is also defined.

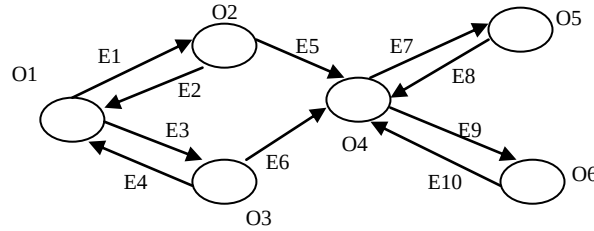


Figure 1: A Sample of Social Network

There are two kinds of reachable relation among objects, shown as following:

1) One-step reachable relation: if there is directed link from O_i to O_j through one and only one directed edge, then O_i to O_j is a one-step reachable relation. For instance in Figure 1 there is a directed link from O_1 to O_2 through the directed edge E_1 , O_1 to O_2 is one-step reachable relation.

2) Multi-steps reachable relation: if there is directed link from O_i to O_j through two or more directed edges, then O_i to O_j is a multi-steps reachable relation. There is a directed link from O_1 to O_4 through directed edges E_1 and E_5 , then O_1 to O_4 is a 2-steps reachable relation. After these relations, BSP analyses a social network by generating edge pointed matrix and calculates one-step and multi-step reachable matrix between objects.

III. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) [13] in multivariate statistics is widely adopted as an effective unsupervised dimension reduction method and is extended in many different directions. The main justification of dimension reduction is that PCA uses singular value decomposition (SVD) which gives the best low rank approximation to original data. However, this essentially noise reduction perspective alone is inadequate to explain the effectiveness of PCA. In this paper, we provide a new perspective of PCA based on its close relationship with the BSP clustering algorithm. We show that the principal components are actually relaxed cluster membership indicators.

The central idea of Principal Component Analysis (PCA) is to reduce the dimensionality of the data set while retaining as much as possible the variation in the data set. Principal components (PC's) are linear transformations of the original set of variables. PC's are uncorrelated and ordered so that the first few PC's contain most of the variations in the original data set [Jolliffe, 1986].

PCA is a classic tool for analyzing large scale multivariate data. It seeks linear combinations of the data variables (often called factors or principal components) that capture a maximum amount of variance. Numerically, PCA only amounts to computing a few leading eigenvectors of the data's covariance matrix, so it can be applied to very large scale data sets. One of the key shortcomings of PCA however is that these factors are linear combinations of *all* variables; that is, all factor coefficients (or loadings) are non-zero. This means that while PCA facilitates model interpretation and visualization by concentrating the information in a few key factors, the factors themselves are still constructed using *all* observed variables. In many applications of PCA, the coordinate axes have a direct physical interpretation; in finance or biology for example, each axis might correspond to a specific financial asset or gene. In such cases, having only a few nonzero coefficients in the principal components would greatly improve the relevance and interpretability of the factors. In sparse PCA, there is a trade-off between the two goals of *expressive power* (explaining most of the variance or information in the data) and *interpretability* (making sure that the factors involve only a few coordinate axes or variables). When PCA is used as a clustering tool, sparse actors identify the clusters with the action of only a few variables.

The original n data points in m -dimensional space are contained in the data matrix $(x_1, \dots, x_n) = X$. In general data is not centered around the origin. The centered data matrix $Y = (y_1, \dots, y_n)$, where $y_i = x_i - \bar{x}$ and $\bar{x} = \sum_i \frac{x_i}{n}$. The covariance matrix is given by

$$S = \sum_i (x_i - \bar{x})(x_i - \bar{x})^T = YY^T$$

The principal eigenvectors u_k of YY^T are the principal directions of the data Y . The principal eigenvectors v_k of the Gram matrix YY^T are the principal components; entries of each v_k are the projected values of data points on the principal direction u_k . v_k and u_k are related via: $v_k = Y^T \frac{u_k}{\lambda_k^{1/2}}$ where λ_k is the eigenvalue of the covariance

matrix YY^T . Thus after applying PCA to the data, the principal components are identified from the data to be clustered. Thus only the important and the vital data are going to be clustered. In this proposed approach, the social network data is analyzed based on the BSP and PCA.

IV METHODOLOGY

a) Generate edge creation matrix and edge pointed matrix

First according to the objects and edges in the graph, define two matrixes $L_c(m \times n \text{ matrix})$ and $L_p(m \times n \text{ matrix})$. L_c means the creation of edges and L_p denote the pointed relations of edges. In the matrix, $L_c(i, j) = 1$ denotes object O_i connects with the tail of edge E_j , which means that object O_i creates the directed edge E_j . $L_c(i, j) = 0$ denotes object O_i has no connection with the tail of edge E_j , which means E_j isn't created by object O_i . In the matrix, $L_p(i, j) = 1$ denotes object O_i connects with the head of edge E_j , which means object O_i is pointed to by the directed edge E_j . $L_p(i, j) = 0$ denotes O_i doesn't connect with the head of edge E_j , which means E_j doesn't point to O_i . After defining the matrix reachable matrices are calculated.

b) Calculate one-step reachable matrix between objects

After the definition of L_c and L_p , one-step reachable matrix between objects is calculated through the following equation.

$$G = L_c * L_p^T = g_{i,j} = \sum_{k=1}^n V(1_c(i, k) \wedge 1_p^T(k, j))$$

$$i = 1 \dots m, [j = 1 \dots m]$$

$\vee$ is Boolean sum and $\wedge$ is Boolean product.

$G(i, j) = 1$ means O_i to O_j is a one-step reachable relation, $G(i, j) = 0$ means there is no one-step reachable relation from O_i to O_j . All one-step reachable relation between objects is calculated.

c) Calculate multi-steps reachable matrix between objects

Besides one-step reachable relation, there are multi-steps reachable relations between objects. According to graph theory and the BSP clustering algorithm, multi-steps reachable matrix $G_2, G_3, G_4, \dots, G_{m-1}$ is calculated. Following equations show the calculation of multi-steps reachable matrix:

$$G^2 = G * G = g_{i,j}^2 = \sum_{k=1}^m V(g(i, k) \wedge g(k, j))$$

$$i = 1 \dots m, [j = 1 \dots m]$$

$$G^4 = G^3 * G \quad G^3 = G^2 * G$$

$G^{m-1} = G^{m-2} * G$ These matrices include 2-steps, 3-steps... $m-1$ -steps reachable relations between objects. Now n -steps reachable relation between two objects through $G_2 G_3 G_4 \dots G_{m-1}$ is calculated.

d) Calculate reachable matrix

The algorithm considers whether reachable relations exist between two objects, but do not care these relations are one-step or multi-steps, so reachable matrix R based on $G, G_2, G_3, G_4, \dots, G_{m-1}$ is calculated as $R = IVGVG_2 \dots VG_{m-1}$ where $\vee$ is Boolean sum and I is unit matrix. $R(i, j) = 1$ means reachable relation exists from O_i to O_j , but the reachable relations existing in matrix R is not mutual, for instance $R(i, j) = 1$ means reachable relation exists from O_i to O_j , but it doesn't mean reachable relation exists from O_j to O_i . Mutual reachable relations between two objects are important in a social network, so mutual reachable matrix based on R is calculated.

e) Mutual reachable matrix and cluster generation

The mutual reachable matrix can be calculated through $Q = R \wedge R^T$ where $\wedge$ means Boolean product, R^T means Reachable Transpose matrix.

In the matrix $Q(i, j) = 1$ indicates there are mutual reachable relation between O_i and O_j . In a social network if two objects that have mutual reachable relation, they should belong to the same class, thus cluster based on Q is generated.

Thus according to mutual reachable matrix Q , a social network is divided into classes based on strong sub matrices in Q or adjusted Q . If all elements in a sub-matrix of Q are 1, then that matrix is a strong sub matrix.

f) Identify relationships among classes

After clustering social network nodes, there is a need to identify relationship among clusters. This can be done through generated clusters and one-step reachable matrix G . If there is one-step reachable relation between two objects in different classes, directed links exist between classes. Through G all relations among classes is identified.

After performing these steps, a social network is divided into classes. Social network clustering analysis algorithm can be given as $Q \rightarrow C_k$ means generating clusters through mutual reachable matrix Q , and $(C_k, Q) \rightarrow \text{Relation}(C_k)$ means identifying relationships among clusters based on clusters and one-step reachable matrix G .

V. PROPOSED ALGORITHM

BSP clusters divide a social network into different classes according to objects in the social network and links between objects, and it also can identify relations among clusters. Main disadvantage of this algorithm is that it uses matrices to store edges and reachable relations. In a real social network these matrices will be very huge,

which can't be loaded into main memory. But these matrices are normally sparse, so this feature is used in our work to reduce the size of the matrix.

In this paper we propose a modification of existing BSP algorithm using Link list data structure. Using this data structure we have overcome the shortcomings of BSP. Following procedure is required for converting this sparse matrix in to link list:

A matrix is a two-dimensional data object made of m rows and n columns, therefore having m, n values. When $m=n$, it is a square matrix. The most natural representation is to use two-dimensional array $A[m][n]$ and access the element of i^{th} row and j^{th} column as $A[i][j]$. If large number of elements of the matrix are zero elements, then it is called a sparse matrix. Representing a sparse matrix by using a two-dimensional array leads to the wastage of a substantial amount of space. Therefore, an alternative representation must be used for sparse matrices.

One such representation is to store only non- zero elements along with their row positions and column positions. That means representing every non-zero element by using triples(i, j, value), where i is a row position and j is a column position, and store these triples in a linear list. It is possible to arrange these triples in the increasing order of row indices, and for the same row index in the increasing order of column indices. Each triple can be represented using a node having four fields as shown:

```
Struct snode{Int row,col,val;
              Struct snode *next;}
```

The algorithm to reduce the dimensionality of the matrix is given below:

Input

L_c : Edge creation Lists

L_p : Edge pointed List

Begin

$G = L_c * \text{Swap}(L_u)$

For $k=3$ to m do

$G_{k-1} = G_{k-2} * G$

L_p : Edge pointed Lists

Begin

$G = L_c * \text{Swap}(L_u)$

For $k=3$ to m do

$G_{k-1} = G_{k-2} * G$

$R = IVGVG2 \dots G_{m-1}$

$Q = R^{\wedge} \text{Swap}(R)$

$Q_k \rightarrow C$

$(C_k, Q) \rightarrow \text{Relation}(C_k)$

End

End

1. In order to add two sparse matrices represented using the sorted linked lists, the lists are traversed until the end of one of the lists is reached.
2. In the process of traversal, the row indices stored in the nodes of these lists are compared. If they don't match, a new node is created and inserted into the resultant list by copying the contents of a node with a lower value of row index. The pointer in the list containing a node with a lower value of row index is advanced to make it point to the next node.
3. If the row indices match, column indices for the corresponding row positions are compared. If they don't match, a new node is created and inserted into the resultant list by copying the contents of a node with a lower value of column index. The pointer in the list containing a node with a lower value of column index is advanced to make it point to the next node.
4. If the column indices match, a new node is created and inserted into the resultant list by copying the row and column indices from any of the nodes and the value equal to the sum of the values in the two nodes.
5. After this, the pointers in both the lists are advanced to make them point to the next nodes in the respective lists. This process is repeated in each iteration. After reaching the end of any one of the lists, the iterations come to an end and the remaining nodes in the list whose end has not been reached are copied, as it is in the resultant list.

We performed the above mentioned steps to reduce the dimensionality of the matrix representation, so that, the input to the BSP clustering is reduced in large number and improves the efficiency of the clusters.

VI. EXPERIMENTAL RESULTS

The performance of the proposed social network using the PCA and BSP clustering is evaluated on the Zachary's Karate Club dataset. This data set is a social network of friendships links between 34 members of a karate club at a US university in the 1970. The values '1' in the data set denotes that there is relationship between the two members in the data set. Alternatively, the values '0' denotes that there is no relationship among the members in the dataset. The proposed PCA and BSP clustering is compared with the existing BSP clustering algorithm. From the dataset, 1156 links were extracted. But after applying PCA technique it was reduced to 700 links which was provided as data to the BSP clustering algorithm. The reachable matrix is obtained from the data set. G34 reachable matrix is computed. Then, clusters are obtained. Cluster 1 and cluster 2 is obtained. If, there is relation between the two clusters, then a relation is formed between them. The clustering accuracy of the data sets is obtained. Moreover, the time taken by the proposed approach for clustering the data set is also calculated and is displayed.

The proposed PCA and BSP clustering is compared with the BSP clustering algorithm. The performance of the proposed system is evaluated based on the parameters like clustering accuracy and classification time. The accuracy in clusters between BSP and PCA with BSP is shown in Figure.1. It is observed from the graph that the classification accuracy of the proposed approach using PCA and BSP approach is high when compared to the BSP approach. Moreover the time taken by proposed approach using PCA and BSP approach is very less when compared to the existing BSP approach due to the reduction in matrix size.

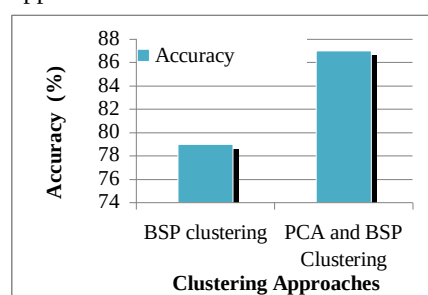


Figure 1. Cluster Comparison

The data set taken for this experiment is small in size. Our future work is to check whether the same efficiency and time reduction exists for a large data set and to combine other reduction techniques with BSP.

VII CONCLUSION

This paper proposes an improved social network clustering technique based on PCA and BSP clustering algorithm. The main idea of this approach is introducing the PCA approach before applying BSP clustering. The main advantage of the PCA technique is that the main and the principal components are identified. Thus after the PCA technique is applied, the important components are given to the BSP clustering. The performance of the proposed approach is evaluated using the Zachary's karate club dataset. The parameters used for the evaluation are clustering accuracy and the time taken for clustering. From the experimental observation, it is clearly observed that the proposed BSP clustering with PCA shows significant performance in terms of accuracy and classification time. Thus the proposed approach is very significant in providing better social network analysis.

REFERENCES

- [1] Gao X, Wu S, Yu B. "Management Information System. Beijing: Economy and Management Press (in Chinese)", 2002.
- [2] Han J, Kamber M. "Data Mining: Concepts and Techniques 2nd edition San Francisco: The Morgan Kaufmann Publishers, 2006.
- [3] W. W. Zachary, "An information flow model for conflict and fission in small groups", Journal of Anthropological Research 33, 452-473 (1977).
- [4] Krebs V. "Mapping networks of terrorist cells", Connections, Vol. 24, pages 43-52, 2002.
- [5] Kubica J, Moore A and Schneider J. "Tractable Group Detection on Large Link Data Sets", Proceeding 3rd IEEE international conference on data mining, Melbourne, FL, pages 573-576, 2003.
- [6] Liben-Nowell D and Kleinberg J. "The Link prediction problem for social networks", Proceeding 2003 international conference on information and knowledge management, New Orleans, LA, pages 556-559, 2003.
- [7] Lu Q and Getoor L. "Link-based classification" Proceeding 2003 international conference on machine learning, Washington DC, pages 496-503, 2003.
- [8] Page L, Brin S, Motwani R and Winograd T. "The PageRank citation ranking: Bring order to the web. Technical report, Stanford University", 1998.
- [9] The Link Prediction Problem for Social Networks (2003) David Liben-Nowell, Jon Kleinberg, 556 – 559.
- [10] Communications of the IIMA (2007) Volume 7 Issue 4 "Social Network Analysis Based on BSP Clustering Algorithm" Gong Yu School of Business Administration China University of Petroleum.
- [11] Algorithms by thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Prentice-Hall India (1999).
- [12] Yeung, K. Y. and Ruzzo, W. L., "An empirical study of Principal Component Analysis for clustering gene expression data", 2001.
- [13] Jiangtao Qiu, Zhangxi Lin, Changjie Tang and Shaojie Qiao, "Discovering Organizational Structure in Dynamic Social Network", Ninth IEEE International Conference on Data Mining, 2009.