Improved BSP Clustering Algorithm for Social Network Analysis

Sanjiv Sharma and R. K. Gupta
Madhav Institute of Technology & Science, Gwalior (INDIA)
E-mail : er.sanjiv@gmail.com

Abstract

Social network analysis is a new research field in data mining. The clustering in social network analysis is different from traditional clustering. It requires grouping objects into classes based on their links as well as their attributes. While traditional clustering algorithms group objects only based on objects' similarity, and it can't be applied to social network analysis. So on the basis of BSP (business system planning) clustering algorithm, a social network clustering analysis algorithm is proposed. The proposed algorithm, different from traditional BSP clustering algorithms, can group objects in a social network into different classes based on their links and identify relation among classes dynamically & require less amount of memory .

Key Words: SNA, BSP, Link list, clustering algorithm, Data mining.

1. 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 analysis of 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 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 [2], classification [8], link prediction [7]. Other achievements include PageRank [9] and Hub-Authority [4] in web search engine.

Present paper, clustering analysis of social network is studied. In the second section, a social network clustering algorithm is proposed based on BSP clustering algorithm. 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, an example of social network clustering algorithm is presented.

2. Social Network Analysis Based on BSP Clustering

There has been extensive research work on clustering in data mining. Traditional clustering algorithms [3] 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, thus find algorithms that can meet this challenge.

The BSP (business system planning) clustering algorithm [11] is proposed by IBM. It 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 sub-systems, 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.

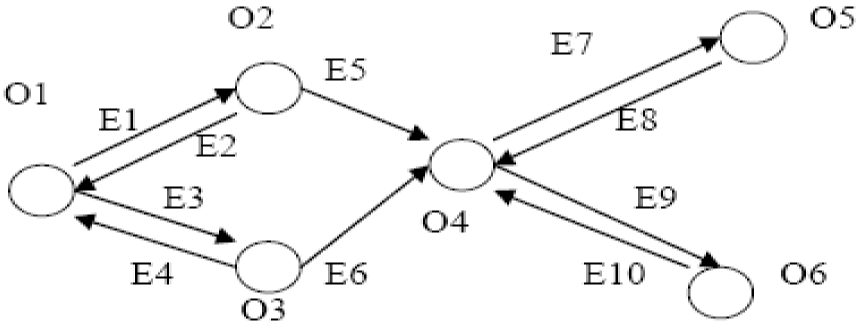


Figure 1: A sample of social network.

In figure 1, Let *Oi* be an object in social network (*i* = 1...*m*), let E_j which means directed link between two objects, be a directed edge of the graph (*j* = 1...*n*).

After definition of objects and directed edges, also define reachable relation between two objects. There are two kinds of reachable relation among objects, shown as following:

1) One-step reachable relation: if there has directed link from *Oi* to O_j through one and only one directed edge, then O*i* to *Oj* is a one-step reachable relation. For instance in figure 1 there has 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 has 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. For instance in figure 1

has a directed link from O_1 to O_4 through directed edges E_1 and E_5, then O1 to O4 is a 2-steps reachable relation.

After these definitions, use BSP clustering algorithm to analyses a social network. The analysis processes are as following steps:

Generate edge creation matrix and edge pointed matrix

First according to the objects and edges in the graph, define two matrixes *Lc* and *Lp.*

Let *Lc* be a $m\times n$ matrix which means the creation of edges. In the matrix, $Lc\ (i, j)$ =1 denotes object *Oi* connects with the tail of edge *Ej* , which means that object *Oi* creates the directed edge *Ej* . $L\ (i, j)\ c$ =0 denotes *Oi* doesn't connect with the tail of edge *Ej* , which means *Ej* isn't created by object *Oi* .

Calculate one-step reachable matrix between objects

After the definition of *Lc* and *Lp*, calculate one-step reachable matrix between objects through the following equation.

$$G=Lc*L_p^T = g_{i,j}=\underset{k=1}{\overset{n}{V}}(l_c(i,k)\wedge l_p^T(k,j))$$

$$,i=1\ldots\ldots m, \left[j=1\ldots\ldots m \right] \qquad (1)$$

^ is Boolean product, V is Boolean sum.

$G(i, j)$ =1 means *Oi* to *Oj* is a one-step reachable relation, $G(i, j)$ = 0 means there hasn't a one-step reachable relation from *Oi* to *Oj* . Through G, calculate all one-step reachable relation between objects.

Calculate multi-steps reachable matrix between objects

Besides one-step reachable relation, there are multi-steps reachable relations between objects too. Also need calculate multi-steps reachable matrices (2-steps, 3-steps, …,*m*-1-steps).

According to graph theory and the BSP clustering algorithm, calculate multi-steps reachable matrix *G*2, G3, G4,…., *Gm*−1 . Following equations show the calculation of multi-steps reachable matrix:

$$G^2=G*G = \quad g^2_{i,j}=\underset{k=1}{\overset{m}{V}}(g(i,k)\wedge g(k,j))$$

$$,i=1\ldots\ldots m, \left[j=1\ldots\ldots m \right] \qquad (2)$$

$$G^3 = G^2*G$$
$$G^4 = G^3*G$$
$$\ldots\ldots$$
$$G^{m-1} = G^{m-2}*G$$

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 ,...,*Gm*−1 .

Calculate reachable matrix

Because only consider whether reachable relations exist between two objects, but do not care these relations are one-step or multi-steps, so calculate reachable matrix R based on G,$G2$,$G3$,$G4$,...,$Gm-1$. The calculation of R is shown as following equation:

$$R=I \vee G \vee G2 ... \vee Gm-1 \tag{3}$$

V is Boolean sum, I is unit matrix.

$R(i, j) = 1$ means reachable relation exists from Oi to Oj , but the reachable relations existing in matrix R is not mutual, for instance $R(i, j) = 1$ means reachable relation exists from Oi to Oj ,but it doesn't means reachable relation exists from Oj to Oi . Mutual reachable relations between two objects are important in a social network, so

calculate mutual reachable matrix based on R .

Calculate mutual reachable matrix and generate clusters

The mutual reachable matrix can be calculated through following calculate equation.

$$Q=R^\wedge RT \tag{4}$$

^means Boolean product

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

Thus according to mutual reachable matrix Q, divide a social network into classes based on strong submatrices in Q or adjusted Q. While strong sub-matrix is defined as follows.

Strong sub-matrix: if all elements in a sub-matrix of Q are 1, this sub matrix is strong sub-matrix.

Identify relationships among classes

After clustering of social network, also need 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, so directed links exist between classes. Through$G$, identify all relations among classes.

After pervious 6 steps, divide a social network into classes. Social network clustering analysis algorithm can be given:

```
Input:
$L_c$ : Edge creation Matrix
$L_p$ : Edge pointed matrix
Begin
$G = L_c * L_u^T$
for k=3 to m do
$Gk-1 = Gk-2 * G$
$R = I \vee G \vee G2 ... \vee Gm-1$
$Q = R \wedge R^T$
$Q_k -> C$
$(C_k, Q)$->Relation $(C_k)$
End
```

$Q->C_k$ means generating clusters through mutual reachable matrix$Q$, and (C_k ,Q)->Relation(C_k) means identifying relationships among clusters base on clusters and one-step reachable matrix$G$.

3. Improvement over BSP Clustering Algorithm

In previous paper based on BSP clustering algorithm, an algorithm of social network clustering analysis is proposed. It divides 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, can't load into main memory. But because these matrices are very sparse, so design an efficient data structure to overcome this shortcoming.

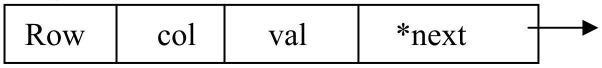
Present paper propose modification of existing BSP algorithm using Link list data structure. Using this data structure can overcome shortcoming (which have been mention above) .Following procedure is require for converting this sparse metrix in to link list:

A matrix is a two-dimensional data object made of m rows and n columns, therefore having *m*, *n* values. When *m*=*n*, call it 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 a 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 (i,j,value) can be represented by using a node having four fields as shown in the following:

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



1. In order to add two sparse matrices represented using the sorted linked lists as shown in the preceding program, 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.

After pervious 5 steps, divide a social network into classes. Social network clustering analysis algorithm can be given:

```
Input:
 Lu : Edge creation Lists
 Lp : Edge pointed Lists
Begin
G = $L_c$ * Swap ($L_u$ ) //perform swapping row column of $L_u$
for  k=3 to m do
Gk −1 =Gk −2 *G
Lp : Edge pointed Lists
Begin
G = $L_c$ * Swap ($L_u$ ) //perform swapping row column of $L_u$
for  k=3 to m do
Gk −1 =Gk −2 *G
R = I V G V G2 ... V Gm−1
Q = R ^ Swap(R) //perform swapping row column of R
 $Q_k$− > C
($C_k$,Q)->Relation ($C_k$)
End
```

Working Example

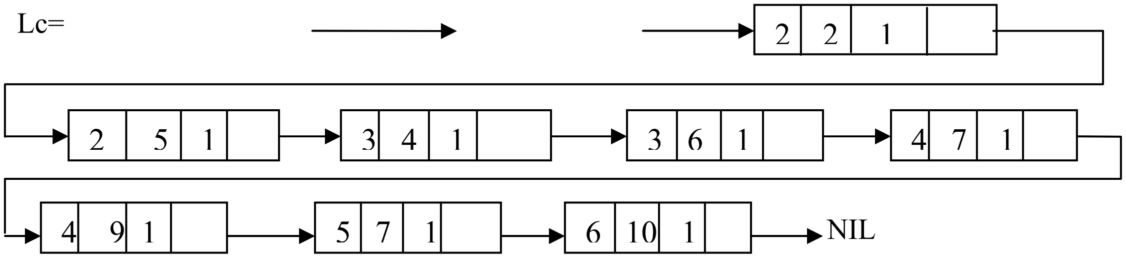
Now an example is given to show process of the cluster analysis of social network. Suppose a social network as figure 1 shows. According to the figure, can give the edge creation matrix *Lc* and edge po inted matrix *Lp* as following.

In the form of link list

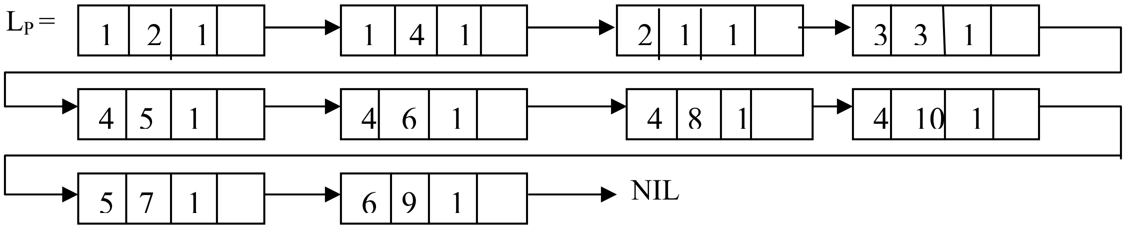
1	1	1	

1	3	1	

Lc=



LP =

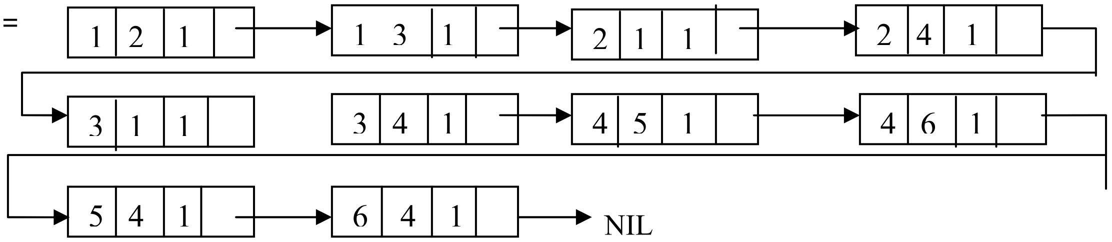


According to the social network clustering algorithm, *Lc* and *Lp* clustering the social network show as following steps:

Calculate one-step reachable matrix between objects

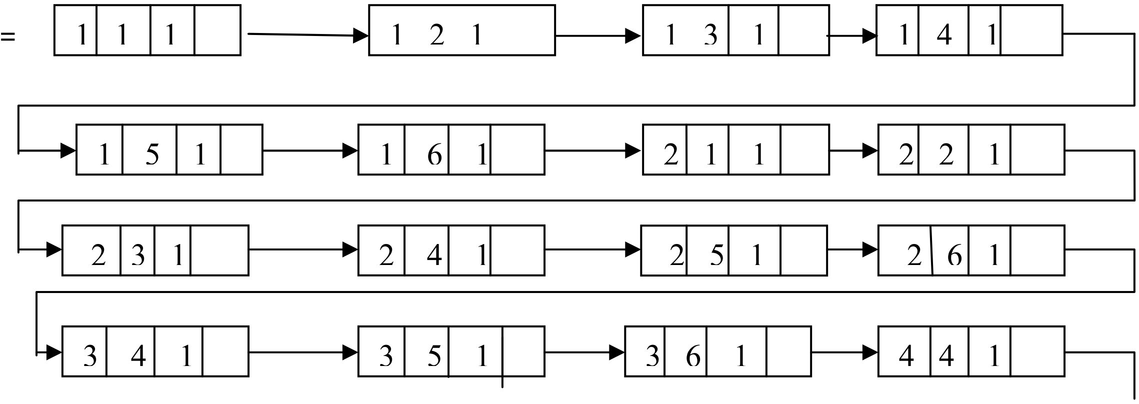
$$G=\begin{pmatrix}1&0&1&0&0&0&0&0&0&0\\0&1&0&0&1&0&0&0&0&0\\0&0&0&1&0&1&0&0&0&0\\0&0&0&0&0&0&1&0&1&0\\0&0&0&0&0&0&0&1&0&0\\0&0&0&0&0&0&0&0&0&1\end{pmatrix}\bullet\begin{pmatrix}0&1&0&1&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&0&0\\0&0&1&0&0&0&0&0&0&0\\0&0&0&0&1&1&0&1&0&1\\0&0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&0&1&0\end{pmatrix}^T=\begin{pmatrix}0&1&1&0&0&0\\1&0&0&1&0&0\\1&0&0&1&0&0\\0&0&0&0&1&1\\0&0&0&1&0&0\\0&0&0&1&0&0\end{pmatrix}$$

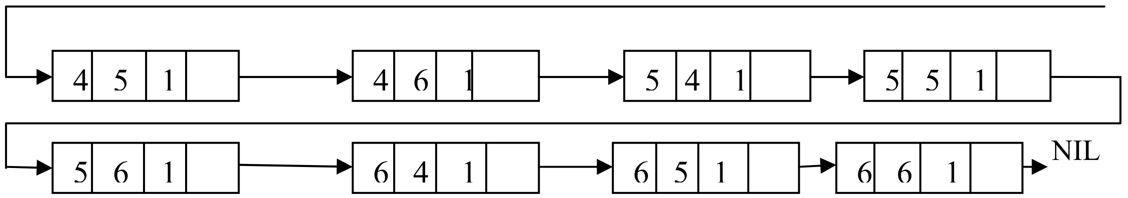
G =



Calculate reachable matrix based on one-step and multi-steps reachable matrix

R=

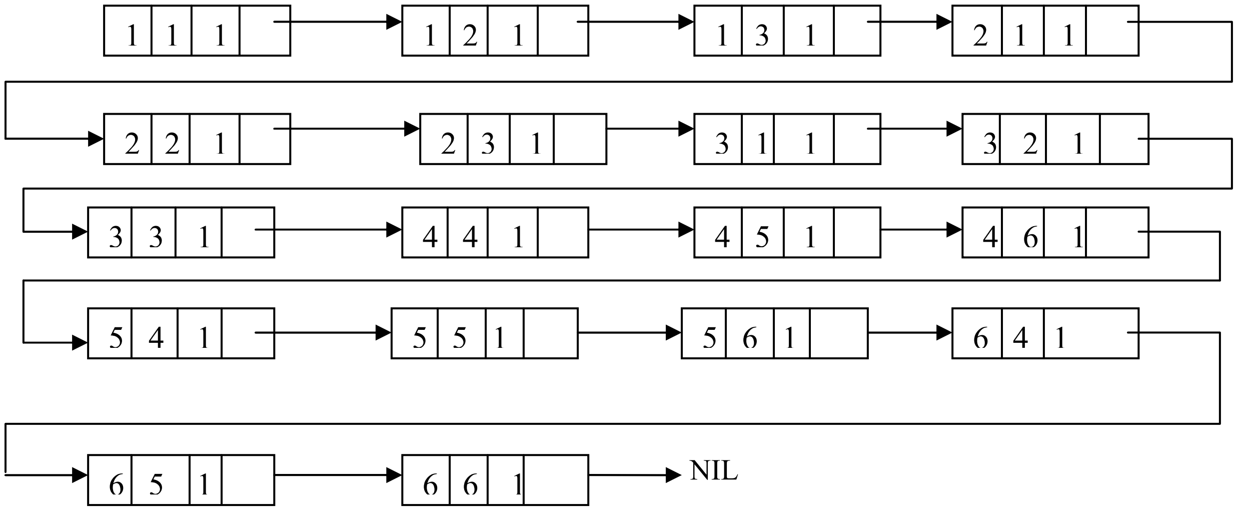




Calculate mutual reachable matrix and generate clusters

$$Q = R \wedge R^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \wedge \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Q=



According the mutual reachable list *Q*, it includes two strong sub list. So divide figure 1 to two classes, the first class C_1 includes object O_1 ,O_2 ,O_3 , and the second class C_2 includes O_4 ,O_5 ,O_6 .

Identifying relationships among classes

According to one-step reachable matrix G , there have one-step reachable relations between two classes ($O_2 -> O_4$ and $O_3 -> O_4$), so identify relations between two clusters C_1 and C_2 , as figure 2 shows.

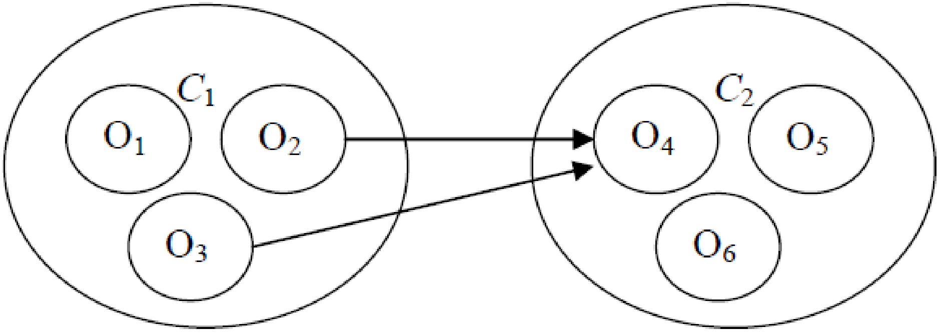


Figure 2: Identify relationships between two clusters.

C points to C_2 i*n* figure 2, but C_2 not points to C_1, so identify relations between two classes.

4. Conclusion

In this paper based on BSP clustering algorithm, an algorithm of social network clustering analysis is proposed. It divides 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. Also in our algorithm the edges between objects have same weight; however in real world such edges may have different weights. Meanwhile the property of each cluster has not been analyzed. These will be solved in our future research.

References

[1] Bhattacharya I, Getoor L.(2004). Iterative Record Linkage for Cleaning and Integration. Proceeding SIGMOD 2004 workshop on research issues on data mining and knowledge discovery, Paris, France,11-18.

[2] Gao X, Wu S, Yu B. (2002). Management Information System. Beijing: Economy and Management Press (in Chinese).

[3] Han J, Kamber M. (2006). Data Mining: Concepts and Techniques 2nd edition. San Francisco: The Morgan Kaufmann Publishers.

[4] Kleinberg J. (1999). Authoritative sources in a hyperlinked environment*. Journal of the ACM,* 5,604–632.

[5] Krebs V. (2002). Mapping networks of terrorist cells. *Connections,*24,43-52.

[6] Kubica J, Moore A, Schneider J. (2003). Tractable Group Detection on Large Link Data Sets. Proceeding 3rd IEEE international conference on data mining, Melbourne, FL,573-576.

[7] Liben-Nowell D, Kleinberg J. (2003). The Link prediction problem for social networks. Proceeding 2003 international conference on information and knowledge management, New Orleans, LA,556-559.

[8] Lu Q, Getoor L. Link-based classification. (2003). Proceeding 2003 international conference on machine learning, Washington DC, 496-503.

[9] Page L, Brin S, Motwani R, Winograd T. (1998). The PageRank citation ranking: Bring order to the web. Technical report, Stanford University.

[10] The Link Prediction Problem for Social Networks (2003) David Liben-Nowell, Jon Kleinberg, 556 – 559.

[11] 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.

[12] Algorithms by thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Published by Prentice-Hall India (1999).