Social Network Analysis Based on BSP Clustering Algorithm

Gong Yu
School of Business Administration China University of Petroleum

Follow this and additional works at: https://scholarworks.lib.csusb.edu/ciima

Recommended Citation
Yu, Gong (2007) "Social Network Analysis Based on BSP Clustering Algorithm," Communications of the IIMA: Vol. 7 : Iss. 4 , Article 5.
Available at: https://scholarworks.lib.csusb.edu/ciima/vol7/iss4/5

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Communications of the IIMA by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
Social Network Analysis Based on BSP Clustering Algorithm

Gong Yu
School of Business Administration
China University of Petroleum

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 clustering algorithms, can group objects in a social network into different classes based on their links and identify relation among classes.

INTRODUCTION

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 (Han & Kamber, 2006).

Research on social network analysis in the data mining community includes following areas: clustering analysis (Bhattacharya & Getoor, 2005; Kubica, Moore and Schneider, 2003), classification (Lu & Getoor, 2003), link prediction (Liben-Nowell & Kleinberg, 2003; Krebs, 2002). Other achievements include PageRank (Page, Brin, Motwani and Winograd, 1998) and Hub-Authority (Kleinberg, 1999) in web search engine.

In this 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, and then the conclusion and the future work direction are given.

SOCIAL NETWORK ANALYSIS BASED ON BSP CLUSTERING

There has been extensive research work on clustering in data mining. Traditional clustering algorithms (Han & Kamber, 2006) 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 we need find algorithms that can meet this challenge.

The BSP (business system planning) clustering algorithm (Gao, Wu and Yu, 2002) 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 $O_i$ 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, we can 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 $O_i$ 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 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 $O_1$ to $O_4$ is a 2-steps reachable relation.

After these definitions, we can 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 matrix $L_c$ and $L_p$.

Let $L_c$ be a $m \times n$ matrix which means the creation 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 $O_i$ doesn’t connect with the tail of edge $E_j$, which means $E_j$ isn’t created by object $O_i$.

For example in figure 1 object $O_1$ connects with the tail of $E_1$, then it means $O_1$ creates $E_1$, so $L_c(1,1) = 1$; $O_1$ doesn’t connect with the tail of edge $E_2$, then it means $E_2$ is not created by $O_1$, so $L_c(1,2) = 0$. 
Let $L_p$ be a $m \times n$ matrix which means the pointed relations of edges. 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 $O_i$.

For example in figure 1 object $O_2$ connects with the head of $E_1$, which means $O_2$ is pointed to by $E_1$, so $L_p(2, 1) = 1$. But $O_2$ doesn’t connect with the head of edge $E_2$, then it means $E_2$ doesn’t point to $O_2$, so $L_p(2, 2) = 0$.

**Calculate one-step reachable matrix between objects**

After the definition of $L_c$ and $L_p$, we can calculate one-step reachable matrix between objects through the following equation.

$$ G = L_c \cdot L_p^T = \left( g_{i,j} = \bigwedge_{k=1}^{n} (L_c(i,k) \land L^T_k(k,j)), i = 1,\ldots,m, j = 1,\ldots,m \right) $$

$\land$ is Boolean product, $\lor$ is Boolean sum.

$G(i, j) = 1$ means $O_i$ to $O_j$ is a one-step reachable relation, $G(i, j) = 0$ means there hasn’t a one-step reachable relation from $O_i$ to $O_j$. Through $G$, we can 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. We also need calculate multi-steps reachable matrices ($2$-steps, $3$-steps, $\ldots$, $m$-1-steps).

According to graph theory and the BSP clustering algorithm, we can calculate multi-steps reachable matrix $G^2, G^3, G^4, \ldots, G^{m-1}$. Following equations show the calculation of multi-steps reachable matrix:

$$ G^2 = G \bullet G = \left( g^2_{i,j} = \bigwedge_{k=1}^{m} (g(i,k) \land g(k,j)), i = 1,\ldots,m, j = 1,\ldots,m \right) $$

$$ G^3 = G^2 \bullet G $$

$$ G^4 = G^3 \bullet G $$

$$ \cdots $$

$$ G^{m-1} = G^{m-2} \bullet G $$

These matrices include 2-steps, 3-steps… $m$-1-steps reachable relations between objects. Now we can know $n$-steps reachable relation between two objects through $G^2, G^3, G^4, \ldots, G^{m-1}$.

**Calculate reachable matrix**

Because we only consider whether reachable relations exist between two objects, but do not care these relations are one-step or multi-steps, so we need calculate reachable matrix $R$ based on $G, G^2, G^3, G^4, \ldots, G^{m-1}$. The calculation of $R$ is shown as following equation:

$$ R = I \lor G \lor G^2 \ldots \lor G^{m-1} $$


\( \vee \) is Boolean sum, \( I \) 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 we need 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.

\[
Q = R \land R^T
\]  

\( \land \) means Boolean product

In the matrix \( Q(i, j) = 1 \) means 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 we can cluster based on \( Q \).

Thus according to mutual reachable matrix \( Q \), we can divide a social network into classes based on strong sub-matrices 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, we 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, we can say directed links exist between classes. Through \( G \) we can identify all relations among classes.

After pervious 6 steps, we can 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**

\[
G = L_c \cdot L_u^T
\]

for k=3 to m do

\[
G^{k-1} = G^{k-2} \cdot G
\]

\[
R = I \lor G \lor G^2 \ldots \lor G^{m-1}
\]

\[
Q = R \land R^T
\]

\( Q \rightarrow C_k \)

( \( C_k \), \( Q \)) -> Relation ( \( C_k \))

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

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, we can give the edge creation matrix $L_c$ and edge pointed matrix $L_p$ as following:

$$L_c = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}, \quad L_p = \begin{bmatrix}
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

According to the social network clustering algorithm, $L_c$ and $L_p$, clustering the social network show as following steps:

Calculate one-step reachable matrix between objects

$$G = \begin{bmatrix}
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}, \quad G^T = \begin{bmatrix}
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix} = \begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}$$

Calculate multi-steps reachable matrix between objects

$$G^2 = G \cdot G = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix}, \quad G^3 = G^2 \cdot G = \begin{bmatrix}
0 & 1 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}$$

$$G^4 = G^3 \cdot G = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix}, \quad G^5 = G^4 \cdot G = \begin{bmatrix}
0 & 1 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}$$
Calculate reachable matrix based on one-step and multi-steps reachable matrix

\[ R = I \vee G \vee G^2 \ldots \vee G^5 = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
\end{pmatrix} \]

Calculate mutual reachable matrix, generate clusters

\[ Q = R \land R^T = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
\end{pmatrix} \land \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
\end{pmatrix} = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 \\
\end{pmatrix} \]

According the mutual reachable matrix \( Q \), it includes two strong sub matrices. So we can 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 \( O_4, O_5, O_6 \).

Identify relationships among classes

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

**Figure 2: Identify relationships between two clusters.**

\( C_1 \) points to \( C_2 \) in figure 2, but \( C_2 \) not points to \( C_1 \), so we can identify relations between two classes.
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.

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 we can design an efficient data structure to overcome this shortcoming.

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


