Journal of Theoretical and Applied Information Technology
30th September 2012. Vol. 43 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE RESEARCH ON LPA ALGORITHM AND ITS IMPROVEMENT BASED ON PARTIAL INFORMATION

[1] SHENG XIN
[1] School of Finance, Shandong Polytechnic University, Jinan, 250353 China

ABSTRACT

With the growing expansion of data size, in-depth study on the social network clustering algorithm gets prominence. A wide range of researches, from the spectral clustering based on modularity, the hierarchical overlapping clustering algorithm to clustering algorithm based on local information, are mainly based on local neighbor information. The current network node cluster typically has a lower time complexity, thus more suitable for large-scale network data sets. In this paper a detailed analysis of the LPA algorithm based on local information is done. The algorithm is not only able to effectively tap the cluster structure of social networks, but also the concept is more simple and easy to be understood. Further research and analysis show that the LPA algorithm in the network cluster structure of the division can be further enhanced, so this article is committed to this direction. The LPA algorithm is proposed to improve the LPA-the SNA algorithm.

Keywords: *LPA Algorithm, Partial Information, Clustering Algorithm*

1. LPA ALGORITHM BASED ON LOCAL INFORMATION

1.1 The Basic Idea Of LPA Algorithm

For most of the social network， clustering algorithm requires a priori knowledge as guidance or computing time costs, Raghavan, et al. Probe cluster structure of the label propagation algorithm, its basic idea is: initially, all nodes designate a representative of their respective cluster structure of the digital label (Numerical Label) algorithm to enter the next cycle, the nodes according to the label of its neighbor nodes; most of the neighbors have a label as its label. Digital label of the node with the iteration of the conduct and continue to change, so the connection is relatively dense nodes gradually to achieve the same tag number. Iterations terminate when the numeric labels of all nodes no longer change the algorithm terminates, the node with the same label form a cluster structure, the entire network performance characteristics are of the cluster structure.

1.2 LPA Algorithm Process Description Of LPA Algorithm

In 2007, Raghavan, et al proposed a label propagation algorithm based on local information search to exploit the cluster structure of the network, network clustering. LPA algorithm uses only the network structure as a guide to clustering in large-scale network, without requiring a priori knowledge, such as the size and number of cluster structure, the computational cost and objective function of optimization which doesn't need to be pre-defined. It is simple in concept, low complexity, easy to operate to achieve efficiency.

LPA algorithm assumes neighbor nodes v for each adjacent $v_1, v_2, ..., v_k$ node has a cluster label to which it belongs, the node v according to the label of its adjacent nodes to determine their own label, the node v to the cluster of most of its neighbors belong structures.

The algorithm is initialized with a unique label for each node, and then spreads in the network tab. Connecting the node with the spread of the label more closely, the label quickly reaches a consensus, shown in Figure 1. Repeat the implementation of the communication process, in each iteration; each node selected most of the neighbors shared label number to update its own label. The update process is divided into synchronous and asynchronous two ways. Sync node $t-1$ in the first iterations, according to its neighbor in the second iteration label to be updated, so $C_v(t) = f(C_{v_1}(t-1), ..., C_{v_k}(t-1))$, where is $C_v(t)$ the label of the node v in the t first iterations. However, if the subgraph in the network is two (Bi-partite), or nearly two chart, it will easy

Journal of Theoretical and Applied Information Technology
30th September 2012. Vol. 43 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to cause a label oscillation (Oscillations of Labels), shown in Figure 2, due to the node labels in accordance with the first iterations. This leads to the node labels constantly change in a and b between the networks and cannot be effectively divided. Such cases especially in the star graph. To overcome this problem, usually using asynchronous updates

$C_v(t) = f(C_{v_{i1}}(t),...,C_{v_{im}}(t),C_{v_{i(m+1)}}(t-1),...,C_{v_k}(t-1))$

, which is $v_{i1},...,v_{im}$ in the current iteration has been updated v to the neighbor, $v_{i(m+1)},...,v_{ik}$ but does not update v the neighbor. In each iteration, the nodes N in the network are updated according to the random arrangement of the order. The algorithm initially, since each node is assigned a different label, the total N tab symbols. With the increase in the number of iterations, the number of tags is gradually reduced; the final number of tags is with the same number of the formation of the cluster structure.

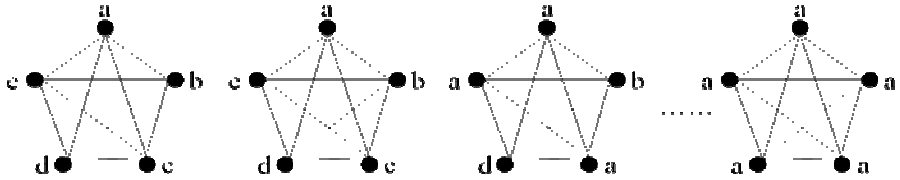

Figure 1 The Label Propagation Diagram

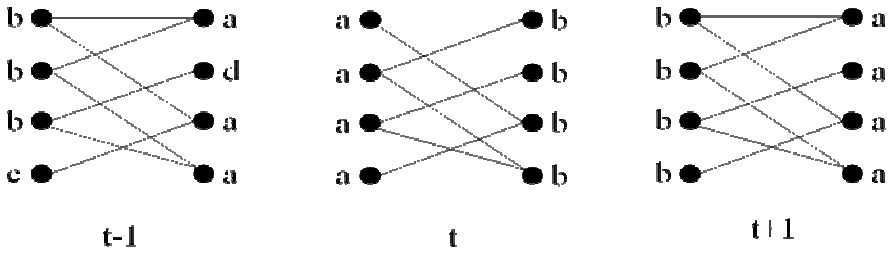

Figure 2 Two Figure Network Label Oscillation Instance

The label of each node in the network with the increase of the number of iterations changes, this algorithm is an ideal iterative process. However, there is some node, its neighbor, the highest number of number of tags may not be the only, in such cases, label of the most frequently composed of a candidate set, and from randomly selected. Implementation of the algorithm, until each node in the network is only one of its neighbors, is the highest number of tags. So, the network last divided into a number of not connected cluster structure. Let the label $C_1,...,C_p$ in the current network, $d_v^{C_j}$ said neighbor nodes v with the label for the number of nodes C_j, the algorithm termination condition: If v you have a label C_m, then $\forall j$, $d_v^{Cm} \geq d_v^{C_j}$. The end of the iterative process, the same node label is assigned to a cluster structure.

The LPA algorithm is described as follows:

The first step: the initialization, the network of each node is assigned a unique numerical label. For each node $C_v(0) = v$.

Step 2: Let the iteration counter $t = 1$.

The third step: the network nodes arranged in random order, and sort the results stored X in the vector.

The fourth step: In accordance with the vectors X stored in the order, for each node $v \in X$, according to the asynchronous formula : $C_v(t) = f(C_{v_{i1}}(t),...,C_{v_{im}}(t), C_{v_{i(m+1)}}(t-1),...,C_{v_k}(t-1))$ in order to update the label symbol. f Then back to the highest number of labels in the current neighbor. The highest number of labels is not unique, and then randomly selected one from the candidate set.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 5: If each node has a label, its neighbor, the highest number of labels, then the algorithm stops. Otherwise, order $t = t + 1$, go to the third step.

2. IMPROVEMENT OF LPA ALGORITHM BASED ON NODE ATTRIBUTES SIMILARITY

2.1 Description Of The Problem

Therefore, the LPA algorithm ignores the attribute information of the node itself, and consider only connected to local information, and the LPA algorithm has a larger random, which would have led to the algorithm may not be optimal; the network is divided into even the node of the error divided. The following image is an example to visually illustrate the problems of the algorithm.

A small interaction network is shown in Figure 3, where each node represents a person, everyone belongs to the school to use the letters next to the node. According to the label propagation algorithm, it is clear that networks can be divided into two clusters structure A University $\{v_1, v_2, v_3, v_4, v_5, v_6\}$ B University $\{v_7, v_8, v_9, v_{10}\}$. Careful analysis of the v_6 node, the label propagation process, according to the connection point of view of the connection node and its neighbor, the candidate set, two, $\{v_4, v_5\}$ and $\{v_9, v_{10}\}$ the corresponding label for the University of A and B, University, randomly choosing one, such as : A University, but according to its actual properties of view, it is B University, so the node error division, affecting the quality of the network clustering. The next round of the iterative process, it is divided into the correct cluster structure, but it is to pay the cost of time.

The above, the paper considers the node attribute information is that the introduction of the concept of node attribute similarity proposed label propagation algorithm based on node attributes similarity is committed to improve the effect of the network by reducing the time overhead of the algorithm.

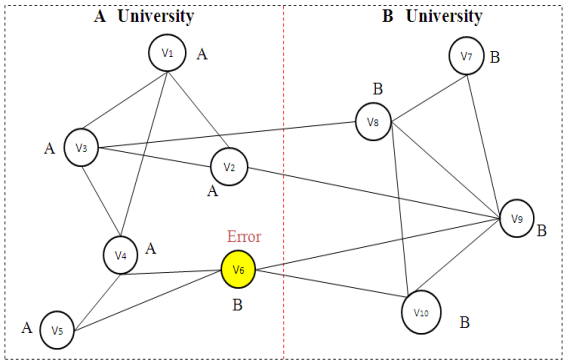

Figure 3 Interpersonal Network Is Divided Instance in

2.2 Node Attributes Similarity

First give some basic concepts and definitions of compute node attributes similarity required before given a specific description of the LPA-the SNA algorithm. First, Table 1 summarizes the symbols mentioned above, some said, and defines a notation of this section to use.

Table 1 Symbols Defined In Table

Symbol	Definitions
$I = \{1, 2, ..., N\}$	Collection of entity objects
$V = \{v_1, v_2, ...v_N\}$	A collection of nodes in the graph model
$S = \{s_1, s_2, ..., s_N\}$	Set of node attributes
$A_{N \times N}$	Between the nodes of the adjacency matrix
$X = \{x_1, x_2, ..., x_N\}$	Label collection
$N_i (i \in I)$	Nodes adjacent systems
N_{ir}	Nodes v_i according to the label frequency w is divided adjacent subsystems, N_{ir} as one of the subsystems

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The social network model described by section 2.1.1, has been in the social network graph model, which indicates a collection of nodes in the network, denoted by $\{v_1, v_2, ..., v_N\}$, E indicating the connection between the nodes. For purposes of research, this paper proposes the following definition:

Definition 1 the collection of entity object $I = \{1, 2, ..., N\}$ set is called a collection of objects, corresponding to the various entities in the social network.

Definition 2 node attribute set to $S = \{s_1, s_2, ..., s_N\}$ set is called the node set of attributes, which s_i indicates the node v_i attribute data.

Definition 3 Adjacency matrixes denoted between nodes $A_{N \times N}$, which $w_{ij}(w_{ij} \geq 0)$ nodes v_i and v_j connection weights $w_{ij} > 0$, expressed v_i and v_j are interconnected, and the size of the weights is equal to w_{ij}.

Definition 4 tag collection $X = \{x_1, x_2, ..., x_N\}$ of collection as the collection of labels, labeling the spread of the iterative process, each node v_i must correspond X to an element in the collection; each node must belong to a label represented by the data set.

Definition 5 adjacent node v_i adjacent to the system is defined as: When a node v_i adjacent node, that v_i is greater than zero, the adjacent system is $N_i = \{j; w_{ij} > 0, i \neq j\}$, or v_i when is 0. $N_i = \varnothing$ The label propagation process, the number of occurrences in the adjacent label is sometimes more than one, then, depending on the label adjacent to the system divided into several subsystems $N_{i1}, ..., N_{ir}, ...N_{iw}$. Each subsystem is with X_{ir} a label to label $X_{ir} \in X$.

2.3 LPA-SNA Algorithm Process Description And Implementation

On the basis of the original of LPA-SNA algorithm, when most of the neighbors of the node to be updated belongs to the cluster structure is more than one, not only uses adjacent node subsystem LPA-the SNA algorithm to calculate each adjacency subsystem node the average value of the property, but also need to be calculated update attribute similarity of the nodes and the neighboring subsystem, and select to make the highest similarity $MaxSimi(S_i, S_{N_{ir}})$ subsystem's label as the label of the current node.

LPA improved the SNA algorithm flowchart, a detailed description of the process is as follows:

Step 1: Initialize the labels of nodes in the network, followed by each node is assigned a unique number label. For the node v order $C_v(0) = v$.

Step 2: Let the iteration counter $t = 1$.

Step 3: Iterative implementation of the following three steps until the network node labels are the highest number of labels in its neighbor.

(1) The arrangement of the nodes in the network in random order, and sort the results stored in the vector X.

(2) In accordance with the vectors X stored in the order, for each X vector $v \in X$, according to the asynchronous formula

$$C_v(t) = f(C_{v_{i1}}(t), ..., C_{v_{im}}(t), C_{v_{i(m+1)}}(t-1), ..., C_{v_k}(t-1))$$

in order to update the node label symbols. f Then back to the highest number of labels in the current neighbor. If the highest number of labels more than one, that is, node there are a number of neighboring subsystem, depending on the node attributes, in accordance with the formula 3-1 or 3-3, the properties of each label corresponds to the subsystem average $S_{N_{ir}}$. And then calculate the nodes v and the similarity of the properties of each subsystem, the highest similarity of the subsystem with the label as the label of the node.

$$t = t + 1 \quad (3)$$

Among them, in the calculation of the average degree of the neighboring subsystem properties, the data sample closes to the subsystem containing is greatly needed, randomly select from a number of data attributes weighted average as the approximation properties of the neighboring subsystems. This is necessary in the network with the size of one million nodes, and can reduce the time consumption of the algorithm. The relative

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

small size of the data sets used in the proportion of data samples randomly selected from the subsystem, this article does not discuss in detail.

After the introduction of node attribute similarity, solving the highest number of labels in the neighbor has been some changes in $MaxFrequencyLabel(ALGraph, index, Label[])$ a function, when the highest number in the neighbor of the judgment node label is not unique, by calculating the various maximum number of labels node candidate set attributes mean to update the current node label. The function parameters and return values with the original LPA algorithm in the function parameters and return values consistent with the implementation process has been great changes, described as follows:

1. $TotalLabelArray[d_{index}] \leftarrow \{0\}$
2. Defined to store a label that contains the node vector $vector < \text{int} > vecLabel$
3. defines the highest number of label not only when the anthology property mean vector $vector < float > vecAverage$
4 Definitions node and subsystem attribute similarity vector $vector < float > vecSimi$
5 $AVER \leftarrow 0.0$
6. If node labeled cross-border
7. returns an error message;
8. If the updated node has no neighbor
9. Return to the last cycle, the label
10. For $i \leftarrow 0$ to the node degrees
11. The number of 11 statistical neighbors for each label
12. $TotalLabelArray[d_{index}]$ The highest number in the If the number of labels is not unique
13. {
14. For $j \leftarrow 0$ to most times the number of tags
15. {
16. Neighboring subsystem to find the highest frequency tags N_{ir}, stored in $vecLabel$ the vector
17. The properties of the mean calculated N_{ir} in accordance with the formula 3-1 or 3-3 $AVER$
18 in accordance with the formula 3-2 or 3-4 to calculate attribute similarity, stored in the vector $vecSimi$
19.}
20. For $k \leftarrow 0$ to attribute similarity vector length $vecSimi.size()$
21. To find $vecSimi$ out that the highest value of the attribute similarity $MaxSimi(S_i, S_{N_{ir}})$
22 .neighboring subsystem corresponding return $MaxSimi(S_i, S_{N_{ir}})$ label
23.}
24. Else
25. Return the highest number of tag number

2.4 LPA-The SNA Algorithm Analysis

LPA algorithm known network topology is as a guide for efficient clustering, in almost linear time. LPA-the SNA algorithm in the use of the network topology at the same time, the introduction of node attribute information, to improve the quality of the network division, to maintain the LPA algorithm with lower time complexity at the same time, reducing the time overhead of the algorithm.

First, initialize the label for each node, within the time $O(n)$ required. The following iterative process, the LPA-the SNA algorithm takes $O(m)$ time of m which number of the edge of the network. For each node, in order to find the label of the highest frequency of the neighboring points, the first $O(d_v)$ time to find its neighboring points, and then select the label of the most frequently assigned to the node v, the time required in the worst case $O(d_v)$; if the highest frequency label is not unique to calculate the average value of each label corresponds to the node attributes, select the property closest to the label as a label for the current node, the worst case, when all nodes set as a single candidate as the time required. Thus, $O(d_v)$ each iterates the time of each node label is updated s. With the increase in the number of iterations, the number of tags is gradually reduced, and the network gradually shows the characteristics of the cluster structure. In original LPA algorithms, Zachary karate club network, the American University Football League network, the paper co-author network, protein interaction network data sets for verification, and is divided after 5 iterations, 95 percent or more nodes to the correct The cluster structure. In this article the LPA-the SNA algorithm, the introduction of the concept of node attribute similarity, it reduces the algorithm iterations University in the United States Football League network, the paper together focuses on simulation of the network data, find the general

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

iteration four times, 95% of the nodes can be divided into the correct cluster structure significantly reduces the algorithm running time.

When the original LPA algorithm terminates, do not rule out this scenario: Two or more connected cluster structure but the same label, that is, two or more nodes select their common adjacent to a node label, and spread in different directions, leading to the network through the node connected to the cluster structure with the same label. To overcome this situation, the algorithm terminates after running the width of each sub-cluster structure of the network-first search algorithm, to separate the connected cluster structure. This whole process takes time $O(m+n)$. However, the above is a special case, particularly the introduction of node attribute similarity experiments later prove that the LPA-the SNA, it may be negligible.

3. CONCLUSION

This article starts from the algorithmic thinking, the description of the main process, and the LPA algorithm is implemented on the basis of LPA-depth. The randomness of the algorithm can easily lead to the clustering. Firstly, this paper studies the computing node attributes and similarity concepts and definitions, and then describes the SNA algorithm on the LPA-and LPA and LPA-SNA algorithm. Lastly, by introducing the concept of node attributes, the quality of clustering results of the algorithm, can be improved, and the time overhead of the algorithm can be reduced. .

REFERENCES

[1] F. Hormozdiari, R. Salari, V. Bafna and S.C. Sahinalp. “Protein-Protein Interaction Network Evaluation for Identifying Potential Drug Targets”. *Journal of Computational Biology*. Vol.17, No.5, 2010, pp: 669-684

[2] B.H Junker and F. Schreiber, “Analysis of Biological Networks”, *John Wiley & Sons*, March 31，2008, pp: 29-59

[3] B. Viswanath, A. Post, K. P. Gummadi, “Alan Mislove. An Analysis of Social Network-Based Sybil Defenses”. *ACM*. 2010, pp: 363-374

[4] S.Z. Niu, D.L. Wang, S. Feng and Y.Ge, ”An improved spectral clustering algorithm for community discovery”. *Ninth Intl. Conf. On Hybrid Intelligent Systems.*(Shenyang), Aug 12-14, 2009, pp: 262-267

[4] N. Wang and X. Li, *Take the initiative to do the supervision and spectral clustering algorithm based on monitoring information features electronic of 2010*, Vol. 38, No.1, pp:172-176

[5] J.J. Daudin, F. Pichard and S. Robin. “A mixture model for random graphs”, *Statistical computing*, Vol.18, No.2,2008, pp: 173-183

[6] P.W. Ling, “The Needed Optimal Cycle for Prediction Accuracy of Stock Price Behavior for Traditional Industries in Taiwan by Moving Average Method”, IEIT Journal of Adaptive & Dynamic Computing, Vol.2011, No.2, April, 2011, pp:7-13. DOI=10.5813/www.ieit-web.org/IJADC/2011.2.2

[7] J.J Zhou, “The Parallelization Design of Reservoir Numerical Simulator”, *IEIT Journal of Adaptive & Dynamic Computing*, Vol.2011, No.2, April, 2011, pp:33-37. DOI=10.5813/www.ieit-web.org/IJADC/2011.2