Algorithm of Distributed Mining Community Structure Used in Sports Psychology Based on Local Information

Cui Sidong
Institute of Physical Education
Yunnan Normal University
Chenggong
China 650500

Journal of Digital Information Management

ABSTRACT: *The paper is studying and implementing a mining algorithm of community structure based on local information. When compared with the algorithm based on global information, our algorithm has a better performance on efficiency which only looses just a little accuracy. Firstly, we analyze the basic feature of community structure of complex network, then, lead to the character of community structure of complex network. Secondly, we introduce some previous work on complex networks and analyze some algorithms of community structure. Finally, we raise a new function to decide a certain node belongs to which community which based on the local modality raised by Clauset. We use the new function to improve the Bagrow algorithm and get a better result on accuracy and also get a new algorithm named Bagrow - β. After that we define two heuristic methods. According to the heuristic methods we created a new algorithm which called BLC algorithm. The new algorithm improves the speed of Bagrow-â algorithm while keeping the accuracy. Then we implement the algorithm with the model of MapReduce, and run the program on the Hadoop platform.*

Categories and Subject Descriptors:
H.2.8 **[Database Applications]:** Data Mining; I.1.2 Algorithms

General Terms:
Data Mining, Data Algorithms, Community Data Analysis

Keywords: Distributed Mining, Community Structure, Sports Psychology, Local Information

Received: 5 January 2013; Revised 1 March 2013, Accepted 14 March 2013

1. Introduction

Complex network is composed of many interconnected parts, and each part can be called an agent. A complex network of every agent demonstrated by the overall behavior will be different from the behavior of the individual [1].

Complex network is difficult to be understood through the following aspects: (1) structural complexity. (2) The network is constantly changing, as in the Internet, page and the connections between them at any time may establish or disappear. (3) Connected between the diversity, different connections between nodes can have different weights, different directions, different meaning. Nervous system synapses are strong or weak, synaptic state can be either inhibitory or excitatory. (4) Dynamic complexity, the node is in a nonlinear dynamical system, such as in gene network, a node in a different time and may have a variety of state. (5) A plurality of nodes, the nodes in the network may have different functions, such as biochemical networks in the control of cell division of various enzymes. (6) Complicated with complexity, many concurrencies may affect each other in [1].

2. Six Degrees of Separation

Small world also known as six degrees of separation theory, the so-called small world refers to the actual network is much smaller than regular network average distance between two nodes and than random networks are much larger than the average clustering coefficient. The United States western power network, star network of cooperation shows the small world network features of [1]. Small world properties of dynamical systems model showed a stronger signal propagation speed, computing, synchronization capability. For example, infectious virus in the small world network spread more rapidly and more easily [2].

Measured from two aspects of structural features, as shown in Figure 1, the path length and clustering coefficient

$C(p)$ $L(p)$ the $L(p)$ measure of degree of separation between nodes, and $C(p)$ measured between adjacent nodes of the community. When $p = 0$ $L \propto n/k$, then $C = 3/4$. When $p = 1$ $L \propto \ln n/k$, and $C = k/n \to 0$. Here the need for $n >> k >> \ln(n) >> 1$, where $k >> \ln(n)$ assurance random graph is connected. Conclusion the conclusion is $p = 0$, graph clustering is high, big world (big world) this is because the L will increase with the growth of the n linear growth. When $p = 1$, class graph is low, small world, because with the growth of the only exhibits a logarithmic growth. This may make people have such doubt: value of the larger always and value larger linked.The next figure 1 reveals the intuition and the opposite is true. As can be seen from the graph has a value of extensive region, in this region has always maintained a very small values of L_{random} and, $C(p) >> C_{random}$ and that a regional presence is an illustration of some of the small world network is exist. Based on the natural existence of network study, many networks go to this area, so that the small world network is widespread.

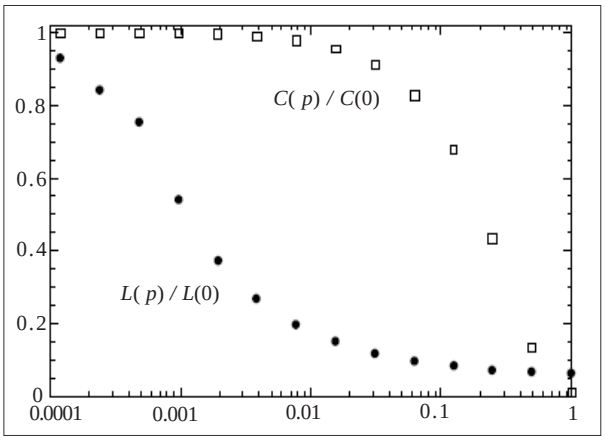

Figure 1. Relationship of path length and clustering coefficient

The so-called scale-free property refers to the nodes in the network degree distribution with power-law features, as shown in figure 2. In fact, the network has shown no scaling is because of the following two reasons: (1) usually by giving a network to add a new node to achieve network expansion. (2) General easy to add a new neighbor node to the existing network has a lot of neighbor node.

Based on the two characteristics of LáSZL óRé Ka presents a network growth model to illustrate the nodes in the network and scale-free distribution [3]. The model of the process is such that, in the initial model with m_0 nodes, each step of adding a new node, the new node adds a m ($m < m_0$) edges, connection model in existing m node. The assumption of a new node is connected to the node in the model probabilities depending on the node of i degrees k_i then, $\Pi(k_i) = k_i / \Sigma_j k_j$ after t steps, model with $t + m_0$ nodes and mt edges. In this way the whole network will enter a scale-free state, at one of the nodes in the network with k edges probability exponential $Y = 2.9 \pm 0.1$ power-law distribution.

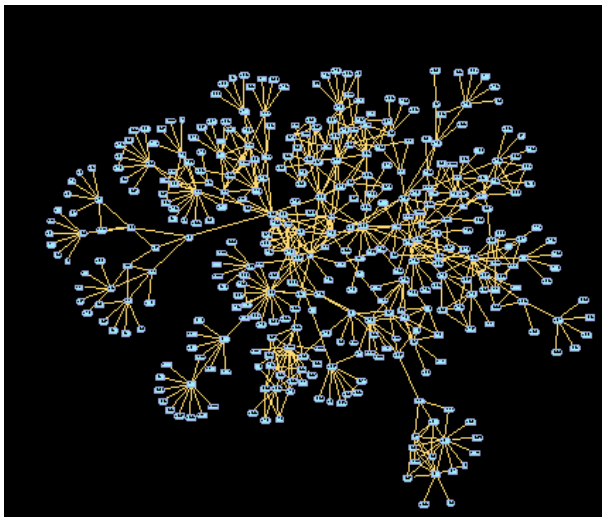

Figure 2. Network diagram consistent with a power-law distribution

Robustness is a network of nodes in the network changes the overall network topology preserving invariant ability. In practical applications, it can be achieved by removal of the network a number of nodes after the network connectivity strength to measure the network robustness. In scale-free networks (connected with power-law distribution), randomly selected number of nodes so that its failure, then the network still maintain connectivity may be large; however if the characteristics of network after selecting the particular node to attack, then the network paralysis of the possibility of increasing. This is because the scale-free network has some notable function node. In the ER random graphs (connected with Poisson distribution), a scale-free network is just the opposite, this is because the connection is uniformly distributed, no apparent key node in [4] network.

3. The Fundamental Concepts of Community in complex network

Community structure of complex network topological structure is the main one of the attributes; mining community structure is to reveal the relationship between the nodes and complex network of functionally important means for [5].

3.1 Definitions of Community

So far, it has not made a widely recognized community structure definition, different researchers give the formal definition from different point of view. The most common definition: network of vertex if show some of the community, will have such characteristics as communities within the node connection between dense and sparse [6] community connections.

Later Radicchi and others in the research community structure mining algorithm is presented in the process of new communities, respectively. It defines the strength of two kinds of communities. The community is defined as: the subgraph of any one of its child nodes with vertex map edge is connected more than and subgraph nodes

with edges [7]. Its formula is expressed as:

$$k_i^{in}(V) > k_i^{out}(V), \qquad \forall i \in V \qquad (1)$$

Weak neighborhood is defined as: all subgraphs within the interconnect edge number and larger than the subgraph internal vertex and the external vertices of even the sides and [8]. Its formula is expressed as:

$$\sum_{i \in V} k_i^{in}(V) > \sum_{i \in V} k_i^{out}(V) \qquad (2)$$

You can see the strong community is a very strict definition, but there are stronger than the community more stringent definition of community: *LS* set. *LS* set is composed of a set of vertices, for the collection of an arbitrary subset of the subset, and complement the edges of internal nodes and set the edges should be bigger than the external nodes and the edges of [9] collection.

3.2 Levels of community structure

In a large network, network may exhibit a certain level of. For example, Facebook a large number of active users and Google index of one billion documents can notice that they have some form of hierarchy: several small communities comprised a larger community, while the larger communities are combined into a higher level of the community [10]. You can use a tree to represent the hierarchical, Newman et al. research community structure algorithm repeatedly using tree algorithm to carry out auxiliary. As shown in Figure 3, the figure shows a 12 node of the tree, they very well reflected that the network has a hierarchy. The bottom of each node can be thought of as a single individual or as a community, to said community combination, finally merged into a single community. The tree diagram is sociology and biology researchers widely used.

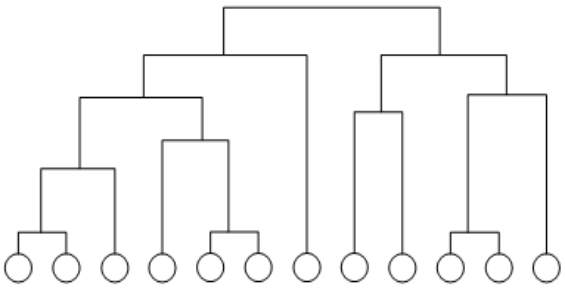

Figure 3. Tree diagram of heuristic clustering algorithm

3.3 The overlapping community structure.

Real networks of community structure after the presence of another important feature is the overlap of [11], the so-called overlap is refers to the network of the existence of such a node; they may belong to two or more community. The *w*9, *w*10 shown in Figure 4 and *w*11, *w*12 are two community overlap. In this paper a exist, such as in a social network, one of the students attended primary school, middle school, University respectively form a community, that the student will belong to the community. At this time the community would show overlapping. Mining community of overlapping significance is, community overlapping portions of the node as in multiple communities, so it has high information capacity.

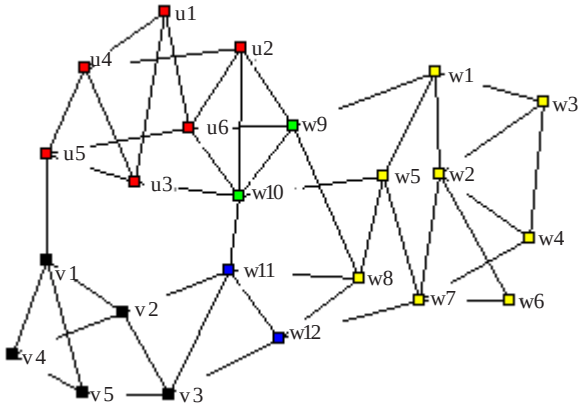

Figure 4. Community overlapping

4. Community Structure Mining Algorithm

4.1 GN Algorithm *w*5

GN algorithm is a classification algorithm, through the continuous removal of community between sides to realize network automatic classification. Girvan and Newman uses the edge set of coefficients to mark each edge effect, or is used to connect different community node, or is connected with a community's internal node. Can define an edge the edge set of coefficients for the whole network in all including the edges in the shortest path number. In addition can also use other values to measure the edge the edge set of coefficients, this paper is to illustrate the algorithm of GN, so choose the shortest path definition. Community between side edges set coefficient is large, and the inner edge of the edge set of coefficients of the community is relatively small. Application of GN algorithm network types: without considering the differences between the vertices, edges, no weight to. GN algorithm is the fundamental step:

(1) computing network each edge of the edge set coefficients

(2) to remove the edge set the maximum coefficient of edge

(3) to calculate each edge in the network edge set of coefficients

(4) repeated 2, 3 two steps until all edges are removed in a network

The problem existing in GN algorithm is complex degree is high, can reach $O(nm^2)$ where n is the number of vertices, m is the number of edges. (1) when do not know in advance and eventually into the community the number, then the GN algorithm "*is the number of iterations should*" cannot be identified. (2) because the GN algorithm cannot from the aspects of quantity to define network community structure, so, in spite of the topology of the network, but does not directly from the structure on which the community was judged, which cannot determine whether these community structure is or is not a real network community structure.

Algorithm of edge set of coefficients calculation way of using the network traversal methods — the breadth-first search, the specific steps are as follows:

(1) For the initial node distance of s $d_s = 0$ and $w_s = 1$

(2) Each of the s neighbor i, $d_j = d_s + 1 = 1$ and $w_i = w_s = 1$.

(3) i neighbor j does as follows:

A) If the node j has not been processed, then $d_j = d_i + 1$ and $w_j = w_i$

B) If the node j has been processed and $d_j = d_i + 1$, then the node weight increases w_i $w_j = w_j + 1$.

C) If the node j has been processed and $d_i < d_i + 1$, not processed

(4) Third step from the beginning to repeat the process, until no node has not been processed

In actual calculation, you can use a queue or FIFO cache to storage has been processed node, this method is the most efficient.

Then calculate all edge set of coefficients, the calculated node representative of the values from the source node to node s i minimal path. The following statistics is got through a side of the minimum number of paths. If node i and j is connected to the node j and distance from the source node distance is farther from the node, then the node to the node of the minimum path contribution is w_i / w_j, then the edges of the edge set of coefficient values increased by w_i / w_j. In order to calculate all starting from minimum path on the edge set factor contribution should be treated as follows:

(1) for each leaf node t, the absence of a node is made from SS to the node through the t

(2) each of the t neighbor i, from t to i is arranged on the edge of a fractional w_i / w_t

(3) the distance from the source node s the farthest edge computing, until s. From node to node j on the i side, wherein the node j s node i distance is farther, on the edge of the assignment all adjacent edge weights are then multiplied by w_i / w_j

(4) repeat step third until it reaches the node s

The n network node as the source node for a above calculation, each calculated each edge values are summed to get the final edge set of coefficients, the time complexity is $O(mn)$. The final overall time complexity is $O(m^2 n)$, this is because of the need to remove all edges.

4.2 Kernighan-Lin Algorithm

Kernighan-Lin Algorithm is a local optimal heuristic algorithm, the algorithm tries to be a network of S is divided into two size consistent subgraph A and B, and makes the connection between at least two subgraph. Matrix representation of C is between any two vertices weight. The algorithm begins network S were randomly divided into two parts A and B , then goes through a series of switching minimization of AA and BB between the external cost T.

Where $A^* = A - X + Y, B^* = B - Y + X$. The definition of A^* and B^* for external links for at least two minutes results, then for arbitrary initial partition of A and $B(|A| = |B|), X \subset A, Y \subset B$ and make the exchange X and Y can get A^* and B^*. So the question is how to select the X and Y, to illustrate the algorithm, to be defined as follows

$$\forall a \in A, E_a = \sum_{v \in B} C_{av}, I_a = \sum_{x \in A} C_{ax} \tag{3}$$

Where E_a is the external links, I_a as the internal links.

Definition of

$$\forall b \in B, E_b = \sum_{v \in A} C_{bv}, I_b = \sum_{x \in B} C_{bx} \tag{4}$$

Get $\forall z \in S, D_Z = E_Z - I_Z$. For any $a \in A$, $b \in B$, a, b induced by A and B exchange links between changes in $D_a + D_b - 2C_{ab}$. The implementation of the algorithm for:the nodes in S, calculate the value of D, $a_i \in A$, $b_i \in B$, such that $g_1 = D_{a_i} + D_{b_i} - 2c_{a_i b_i}$ maximum the elements of $A - \{a_i\} B - \{b_i\}$, D value, to calculate, the formula for

$$\begin{cases} D_x' = D_x + 2c_{xai} - 2c_{xbi} & x \in A - \{a_i\} \\ D_y' = D_y + 2c_{yaj} - 2c_{yaj} & x \in B - \{b_j\} \end{cases} \tag{5}$$

repeat step second, until all the nodes have been processed, node is switched after not doing the second exchange.

The time complexity of the algorithm is $O(n^2 \log n)$ algorithm lies in, must be randomly assigned two sub communities, according to the different initial distribution, calculation process and algorithm of consumption also may not be the same.

4.3 Newman Fast Algorithm

Fast Newman algorithm is a local clustering algorithm, the initial of each node as a community, by calculating the community the closeness of the community merged, finally forming one or more community. Algorithm using the module of division is good or bad, each ΔQ maximum merge operation. With two communities merged community modularity is set to $Q' = \sum_k (e_{kk} - a_k^2)$ obviously be the combined total number of nodes and edges is the merger of two communities and, according to definition:

$$e_{kk} = e_{ii} + e_{jj} + e_{ij} + e_{ji} \tag{6}$$

and

$$a_k = \sum_l e_{lk} = \sum_l e_{li} + \sum_l e_{lj} = a_i + a_j \tag{7}$$

Algorithm specific calculation procedure is as follows:

(1) The first clear each node as a separate community

(2) Computing community consolidation induced module of incremental $\Delta Q = 2\,(e_{ij} - a_i a_j)$ select ΔQ increases the maximum or minimum combinations were associated with reduced

(3) Repeat step second until the remaining in a community

Algorithm implementation process in the formation of a similar tree structure, in different gradation of the graph can get different results; generally choose the maximum Q value division. Algorithm in the worst case complexity is $Q(n\,(m+n))$, for the sparse matrix can be simplified as $O\,(n^2)$.

4.4 CNM Algorithm

CNM algorithm is Clauset, Newman and Moore et al proposed an improved algorithm, its time complexity is $O(md\log n)$, where m is the number of edges vertices, n, d tree depth. Its essence is a kind of according to the module of greedy algorithm, it is improved by applying a new data structure.

Based on the modularity function definitions, such as formula (2-13) below, can put the network as a multiple graph, where a vertex represents a community. Multiple graph adjacency matrix with elements of $A'_{ij} = 2me_{ij}$, i and j combined community means using them and the substitution matrix line i j column elements, in Newman fast algorithm of this replacement process is in the matrix explicitly called out, but if the matrix is a sparse matrix would be using a special data structure to improve efficiency. Calculation of ΔQ_{ij} and find a community of i, j was time consuming. In the algorithm CNM, no longer maintained community adjacency matrix and then calculating the ΔQ_{ij}, but directly to maintain and update a ΔQ_{ij} on the values of the matrix. Algorithm of advantage lies in that he uses the data structure, as shown below.

(1) ΔQ_{ij} on a sparse matrix, wherein i and j community the existence of at least one side. The matrix of each line is saved as a balance two forks tree (lookup operation cost $\log n$ time) and a maximum heap (constant time to find the maximum value)

(2) One of the biggest pile H, each row contains the maximum value of matrix ΔQ_{ij}

(3) A primitive vector array, including a_i

According to the above description, algorithm is still up in the belief that every node is a community, it can be an initial matrix,

$$\Delta Q_{ij} = \begin{cases} 1/2m - k_i k_j/(2m)^2 & \text{if } i, j \textit{ are connected} \\ 0 & \textit{otherwise} \end{cases} \quad (8)$$

And for each community in i $a_i = k_i / 2m$ The specific algorithm is as follows:

(1) According to the above formula ΔQ_{ij} and a_i

(2) To find the largest ΔQ_{ij}, combined with the corresponding community, update matrix ΔQ_{ij}, H and a_i, increased value Q for the $Q = Q + \Delta Q_{ij}$

(3) Repeat step second, until cannot merge operation

The data structure enables the algorithm in second step time can quickly achieve the update. First of all, pay attention to only need to adjust the ΔQ of a small portion of the data, if the merger of i and j community, will be merged with the community expressed as j, therefore only need to update the j rows and j columns, and then remove the i rows and i columns. Update rules are as follows, if the community KK and community i and community j are adjacent, then $\Delta Q_{ik} = \Delta Q_{ik} + \Delta Q_{ik}$, and if the community k connected to the and not connected to the j, then $\Delta Q_{ik} = \Delta Q_{ik} - 2a_j a_k$, or $\Delta Q_{ik} = \Delta Q_{ik} - 2a_i a_k$ or . The three formula described in algorithm of the whole process, Q only has one vertex, once each for the largest ΔQ_{ij} becomes negative, then every time they ΔQ_{ij} will also make the Q smaller.

4.5 Radicchi Algorithm

Radicchi et al first proposed on the strength of the community structure community definition, and they therefore improved GN algorithm, concrete steps can be described as:

(1) Choice of a community (community, weak community, of course, you can also define their own)

(2) Computes all the edges of the edge set of coefficients, and removed the edge set the maximum coefficient of edge

(3) If the second step without a graph partitioning, go to step second

(4) If the second step will be a map divided into two parts, test all of at least two subimages with the community, if added to the tree diagram

(5) For all subgraphs from the second iteration until no edges being in the network.

In order to overcome the shortcomings of GN algorithm time complexity high shortcoming, Radicchi proposed a new classification algorithm, only need to consider the local data, and GN algorithm to compute the edge the edge set of coefficients, must consider the overall network topology, the new algorithm obviously than the GN algorithm has more advantages. Classification algorithm is a necessary factor is to build a metric, the metric can point out an edge connecting the two vertexes whether to belong to the different communities. The algorithm uses a named set of edges coefficient as measure, which is defined as a given edge to be part of the triangle number divided by the maximum potential, containing the edges of the triangle number. For a given two adjacent nodes of i, j , edge set coefficient is defined as:

$$C_{i,j}^{(3)} = \frac{z_{i,j}^{(3)}}{\min[(k_i - 1), (k_j - 1)]} \quad (9)$$

One is referring to the $z_{i,j}^{(3)}$ contains a side of the triangle number, and $\min[(k_i - 1), (k_j - 1)]$ refers to the most likely includes $e_{i,j}$ triangle number. If one side is connected to the node in different communities, the edge will contain fewer triangles do not even include any triangle, then the $C_{i,j}^{(3)}$ will be relatively small. But in contrast to that, if one side is located in a community, the edge of the edge set coefficient $C_{i,j}^{(3)}$ will be greater. This metric can be measured with an edge of belonging to a community or the connection of different community. If an edge of the triangle for the number of 0, then $C_{i,j}^{(3)} = 0$, in order to prevent the occurrence of such uncertainty can be improved as follows:

$$\tilde{C}_{i,j}^{(3)} = \frac{z_{i,j}^{(3)} + 1}{\min[(k_i - 1), (k_j - 1)]} \quad (10)$$

Consider the triangle polygon is extended to:

$$\tilde{C}_{i,j}^{(g)} = \frac{z_{i,j}^{(g)}}{s_{i,j}^{(g)}} \quad (11)$$

The definition of $z_{i,j}^{(g)}$ is similar to $z_{i,j}^{(3)}$, that contains the $e_{i,j}$ GG side edge number, and $s_{i,j}^{(g)}$ that contains the maximum edge.

It can be defined that improved partition algorithm, and GN algorithm steps are basically the same, after removing the edge when selecting edge set minimum coefficient boundary is the $\tilde{C}_{i,j}^{(g)}$ minimum edge. Specific steps decribed as:

(1) Choice of a community (community, weak community, of course, can also define their own)

(2) Computes all the edges of the edge set of coefficients, and will have a minimum set of edges is coefficient of edge removed, to the next

(3) If the second step without a graph partitioning, go to step second

(4) If the second step will be a map divided into two parts, test all of at least two subimages with the community, if added to the tree diagram

(5) For all subgraphs from the second iteration until no edges being in the network.

When deleting an edge, first of all need to judge whether the whole graph is divided into two independent sub graph, and then update the deleted edge neighbor node of the $\tilde{C}_{i,j}^{(g)}$ value, assuming that the whole network edge number m, then the first operation need to consume the time $O(m)$, and the second step operation need to consume time is less than. The above two steps needed for all of the edge performing again so it has time complexity is $am + bm^2$, denoted by $O(m^2)$.

5. Community structure partition result measure

Girvan and Newman put forward the concept of modularity is used to measure the whole network partitioning result [12]. It is based on such an intuition, namely random networks generally do not show obvious community structure.

Modularity function used to measure connectivity within the same community the number of sides accounted for the overall ratio of the network by subtracting a value (the value of the formula unchanged, but the network following adjustment: keep the community the same structure, and the connection between nodes randomly rewiring, keeping constant node degree) [13]. In general if the network community the number of sides worse than random networks, then $Q = 0$, maximum Q value of 1, generally Q value between 0.3 to 0.7, value very rare cases of [14].

Since the concept of modularity is put forward, it has become a measure of the community is divided on the module of an important indicator of the level, here are severalfunction expression:

$$Q = \frac{1}{2m_{ij}} \sum [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \quad (12)$$

$$Q = \sum_i (e_{ii} - a_i^2) = Tr(e) - \| e^2 \| \quad (13)$$

$$Q = \sum_{v=1}^{n} \left[\frac{l_v}{L} - \left[\frac{d_v}{2L} \right]^2 \right] \quad (14)$$

Greedy algorithm used to optimize the Q function for the purpose of the algorithm is very effective, but the use of Q function means that it is based on the overall network topology method, need to know the network's overall structure. The function value bigger shows the division effect is better, but the function has some defects: the function maximum value determined by the community, may be more satisfying modular community definition of small community combined with [15-17]. At the same time, some other technology has been proposed to address community structure mining problem of [18] [19], but these need the global knowledge network.

6. Experimental results and analysis

We set the algorithm running on each node has 50 Mapper, 2 Reducer at work. Such selection is based on the3 section introductions and experiments using the machine configuration.

We see in Figure 5, along with the input data is growing, community structure mining algorithm running time is on the increase. Due to the large amounts of data copy, preprocess to occupy the largest proportion at the initial stage. After the pretreatment algorithm BLC started running, and the whole system running smoothly. And BLC algorithm operation and community duplicate checking can be done at the same time.

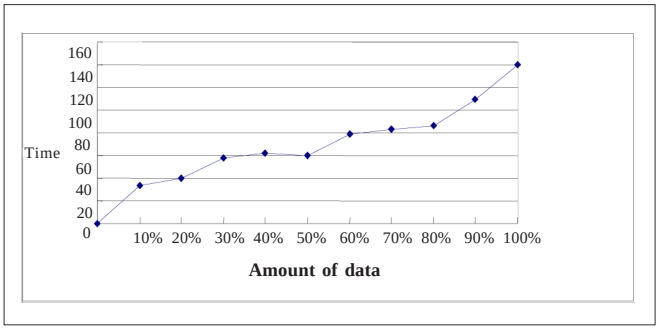

Figure 5. Relationship of the amount of data and the processing time

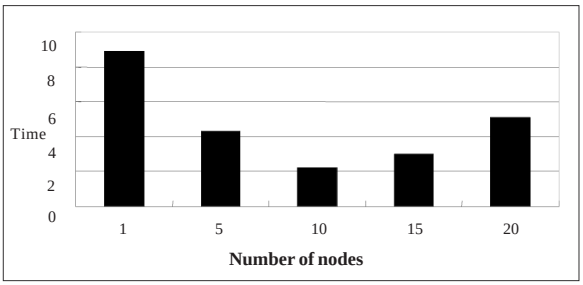

Figure 6. Relationship of the number of nodes and the running time

Algorithm is a key factor for the initial selection of the number of nodes, is the largest amount of parallel, the next figure 6 reveals the initial node number and the algorithm running time relationship. We see in the node number to 10, the performance of the algorithm to achieve the best, this is due to the limited number of clusters, our algorithm consists of 3 MapReduce task, one of the mostly used the data pre-processing, in the following two MapReduce task requires iterative execution, and so cannot achieve maximum parallel, this is due to the inherent nature of community structure mining of decision. And because every second MapReduce task completed, third MapReduce tasks, and we figure 6 of the initial number of nodes will directly affect the second MapReduce task Map number. In second MapReduce task completed, followed by the third MapReduce task will occupy a large consumption, this explains why at 15 initial node algorithm running efficiency will be slightly lower than the 10 initial node properties.

Because the BLC algorithm did not do any operation size refinement, operates on Hadoop equivalent to just parallel running several BLC algorithms and each algorithm running results were compared, and therefore gives the final result. Thus the efficiency of the algorithm has not been fundamentally enhanced.

Reference

[1] Strogatz, Steven, H. (2001). Exploring complex networks. nature. 410 (6825) 268-276.

[2] Mucha, Peter, J. (2010). Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. Science. 328 (5980) 876-878.

[3] CNNIC. twenty-eighth statistical report on Internet development in China. China Internet Network Information Center. 2011, 7-11.Newman, M. E. J. (2006). Modularity and community structure in networks. *In*: Proc. Natl. Acad. Sci. USA. 103 (23) 8577-8582.

[4] Sanjay Ghemawat, Howard Gobioff. (2003). The Google File System. *In*: Proceeding SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles,New York,37 (5) 29-43.

[5] Jefirey Dean, Sanjay Ghemawat. (2004). MapReduce, Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, 51(1) 107-113

[6] Kernighan, B. W., Lin. (1970). A efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*. 49, 291-307.

[7] Li Xiaojia. (2008). Community structure in complex networks. Complex systems and complexity science. 5 (3) 19-42. Newma, M. E. J. (2004). Fast algorithm for detecting community structure in networks. Phys Rev E. 69 (6) 1370-1374.

[8] Clauset, A., Newman, M. E. J., Moore, C. (2004). Finding community structure in very large networks. Phys Rev E., 70 (6) 938-943P

[9] Clauset, A. (2005). Finding local community structure in networks. Phys Rev E, 72 (2) 1063-1068.

[10] Luo, F., Wang, J. Z., Promislow, E. (2008). Exploring local community structures in large networks. *Web Intelligence and Agent Systems*, 6 (4) 387-400.

[11] Bagrow, James, P. (2008). Evaluating Local Community Methods in Networks. *Journal of Statistical Mechanics: Theory and Experiment*. (5) 1003-1011.

[12] Nan Du, Bai Wang, Bin Wu. (2008). Overlapping Community Structure Detection in Networks. *In*: Proceeding of the 17th ACM conference on Information and knowledge management, New York. p. 1371-1372

[13] Wan Xuefei. (2010). A heuristic algorithm for detecting overlapping communities. *Computer engineering and applications*,46 (3) 36-38.

[14] Wang Limin. (2010). Based on the maximum node proximity algorithm for detecting local community structure. *Computer engineering*. 36 (1) 25-26, 29.

[15] Zhu Yongzhen. (2011). In the virtual community research on community structure and analysis of computer technology and development. 21 (1) 46-49.

[16] Wu Lingyu. (2010). Considering the object attribute information in the complex network community structure discovery algorithm. Mathematics in practice and theory of 2010, 40 (24) 161-167.

[17] Chen Qiong. (2010). Based on dynamic attributes similarity of nodes in social network community recommendation algorithm. *Computer applications*, 30 (5) 1268-1272.

[18] Li, G. F., Xiong, H. G., Xu, S. Q., Kong, J. Y. (2011). A Hybrid Particle Swarm Algorithm to JSP Problem, *IEIT Journal of Adaptive & Dynamic Computing*, (3) 10-17, Jul. DOI=10.5813/www.ieit-web.org/IJADC/2011.3.3

[19] Zhao, Z. L., Liu, B., Li, W. (2012). Image Clustering Based on Extreme K-means Algorithm, *IEIT Journal of Adaptive & Dynamic Computing*, (1) 12-16, Jan. DOI= 10.5813/www.ieit-web.org/IJADC/2012.1.3