International Journal of Grid Distribution Computing
Vol. 8, No.5, (2015), pp.97-110
http://dx.doi.org/10.14257/ijgdc.2015.8.5.09

Design of Complex Network Distributed Computing Information Mining Method

Yiran Wang and Guang Zheng

School of Computer Science and Technology, Zhoukou Normal University, Henan Zhoukou, 466001, China
School of Computer Science and Technology, Zhoukou Normal University, Henan Zhoukou, 466001, China
wangyiran76@163.com

Abstract

The information that caused by the complex network is massive, but because of a large amount of information, so the use of traditional data analysis has been unable to meet the search and mining complex network data, so the distributed computing mode, i.e., cloud computing has become the main method of calculation method of complex network, according to the calculation of cloud computing, it needs to consider computing topological partition method and the computational performance. This paper presents a data mining model matrix, according to this model can integrate different information, optimization of data mining, so as to improve the efficiency of complex network distributed computing.

Keywords*: Complex Network, Data Search, Data Mining, Computational Efficiency*

1. Introduction

Complex network has become a hot research issues, and many of the network can be attributed to the complex network problems, with the rapid development of Internet, the application of a wide variety, analysis and use of complex information network has become a current in the field of data mining important. There are a large number of published related research results. The analysis of complex large-scale information network to the field of data mining has brought a lot of opportunities, it provides data and a large number of application scenarios for data mining and analysis, and it provides a new stage for data scientists. But at the same time, the analysis and mining of large-scale information network is also faced with many challenges. First, the information network data size, such as web data before 800816 included 3 has reached hundreds of billion, the number of users of social networking software has exceeded 1000000000, while the general biological information network is greatly in excess of this magnitude. This makes the traditional data mining technology and algorithm cannot use in such a large-scale data analysis [1]. In the face of large-scale information network, any more than the linear complexity analysis of the algorithm is feasible. Secondly, the analysis of large-scale information network has brought forward new challenges to computing power. Analysis of single high performance computer is not suitable for the large-scale data; distributed computing has become a new trend of data mining.

The social, natural and information systems are often composed of large in the role of different types of individuals, we can model these complex systems using information network. Node and edge information in the network has to express the relationship between individuals and different attribute complex with certain properties. For example Webpage Internet links and complex relationship is a typical information network. Webpage with tags and attributes are different, links

ISSN: 2005-4262 IJGDC

can be attached on the behavior of user information. The social networking site friend relationship, user attributes and does widely shared multimedia information also constitute a complex information network. In addition to the scientific collaboration network, calls and semantic net and so on, analysis of information network and using method of various fields from graph theory, machine learning, pattern classification [2], data mining and statistical inferences to uncover patterns and rules hidden behind the web, so that people on the complexity of the information network is a comprehensive and in-depth understanding.

At present the analysis of large-scale information network has the following main contents: 1) finding important nodes in large-scale complex information network. For example, find valuable applications to users in the massive Webpage in. Certain areas of the most influential person found in social networks. 2) Mining community structure in large scale information network. Community structure exists widely in different information network. For example, a social network of an organization, including a theme of "having a certain function and dyeing biological information network in genome. 3) The development and evolution of information network. A lot of information network is a dynamic change, plays an important role in the analysis of complex information network of these dynamics. For example, change the evolution and research groups in the scientific collaboration network enables us to understand the evolution of scientific research to a future development and discovery of new areas of research and mining. 4) Information of network in the link prediction problem. Method for analysis of complex scale information network mainly contains the following two kinds of analysis and representation: analysis model based on graph theory and analysis model based on matrix [3]. Analysis paradigm based on graph theory is to make the graph structure abstraction of information network, the information of complex network entities as nodes in the diagram, the mutual relationship between entities, with edges between nodes to represent. The analysis paradigm often adopts the traditional chart algorithm to analyze complex information networks, such as the shortest path map algorithm, graph theory segmentation algorithm of equilibrium. Matrix analysis method based on complex network is to calculate the information was analyzed by means of matrix. For example, using matrix decomposition characteristics to predict the link, using the spectral analysis method to divide the community, the current model of massively parallel computing, such as MapReduce, MPI etc., and the analysis is not applicable to large-scale and complex information network. While the design of these systems was not to the information network analysis, analyses of the algorithm developed in these distributed systems often are not efficient use of computing resources group. At the same time, the system programming model does not apply to information network analysis and algorithm design. Although MapReduce is suitable for the large-scale data calculation [4-5], but sometimes not very well in some graph algorithms, such as iterative number more, graph algorithms require frequent access to the topology information of other nodes. MPI programming interface is too complex, on the design algorithm is not easy and flexible.

Therefore, a fall exclusively used for computing system diagram calculation began. According to the large-scale graph mining encountered in parallel in the problem, Google proposes a more suitable for mining distributed computing system. The system provides a use easy FFL households graph computing interface and has good tolerance. But the system of mining algorithms cannot efficiently solve for some complex diagram, the node oriented computing model in some graph algorithm and can't get good performance. Because in this system, map data are randomly assigned on each machine, so some complex graph algorithm is very easy to cause the human scale data movement, resulting in network congestion.

Data migration is a reasonable distribution of data can effectively reduce the distributed computing system, avoiding network congestion and improve the calculation efficiency. The graph partitioning algorithm can carry on the reasonable segmentation on large graph data, effectively reduces the communication overhead in distributed computing. Now hit the graph partitioning algorithm, such as spectrum method and multilayer segmentation algorithm complexity is high, cannot on large graph data segmentation. And the existing graph partitioning algorithm, the algorithm can in a parallel manner to carry out large-scale chart data aggregation and segmentation, suitable for large-scale chart data segmentation. To obey the power-law distribution map data, graph partitioning algorithm of almost all cannot achieve good segmentation results. In view of this situation, this paper presents a partitioning strategy a new map data, the degree of nodes with larger segmentation to different machines, effectively reduces the data communication overhead and achieves load balance calculation.

2. Related Work

Intrusion detection is defined as: identification for the computer or cyber source malicious intention and behavior, and the process that to respond to this. Intrusion detection system is a system to complete the above functions. Intrusion detection system can detect unauthorized object (person or program) against the system intrusion attempt or behavior, at the same time the illegal operation monitoring the authorized object of system resources [6]. Intrusion detection system generally consists of three parts: information collection and pretreatment, data analysis and response system. Intrusion behavior mainly refers to the unauthorized use of system resources, may cause harm to the system data loss and damage and the system of denial of service. For the intrusion detection of network attacks can be divided into the following 4 categories: (1) through the examination of a single IP data of Baotou Department (including TCP and UDP) can be found in the attack (2) by examining a single data Department of Baotou, but also to check the data segment information to find the attack, (3) the frequency to top off find the attack detection, (4) using the slice of attacks, such attacks using fragment reassembly algorithm in the intrusion examination system loopholes to circumvent the fin check and attack. To check this kind of attack, it must advance as recombinant try.

A design of intrusion detection system requirements include: (1) the real-time requirements: if trying to attack or attack as soon as possible to be found, it is possible to find out the location of the attackers, and to prevent further attacks, and it is likely to damage control to a minimum, the system also can record all the activities under the attacker the attack process, and can be used as evidence of playback. Real time intrusion detection can reduce the administrator through to audit the system log to find the intruder or intrusion behavior cues when many inconvenient [7]. (2) Scalability requirements: because there are thousands of different species of known and unknown attack, attack behavior characteristics are also each are not identical. So we must establish a mechanism to separate the architecture of intrusion detection system and the use of strategies. To an established intrusion detection system must be able to guarantee that the new type of attack occurs, can through certain mechanisms (such as updating attack signature database mechanism) without changes in the situation of intrusion detection system itself, so that the system can detect novel attacks. The function of the whole design of intrusion detection system, also want to use architecture can be extended, so that the system can adapt to the extension of the structure itself possible future requirements. (3) Suitability requirements: intrusion detection system can be used in different environment. For example: in the high speed and large capacity computer network

environment, increase the number of computer system in the environment, change the types of computer system, intrusion detection system should still be able to carry out normal work[8]. Suitability requirements also include adaptability, intrusion detection system itself on the host platform that is: the ability to work across platforms, adapt to the different situation of the host platform of soft, hardware configuration. (4) Safety requirements: intrusion detection system must have a robust (robust), cannot bring security problems and new security risks to the host computer system and the computer environment. (5) The validity claims: intrusion detection system must be effective, that is to say, for misstatements and omissions attacks can be controlled in a certain range.

At present, the technology of intrusion detection has two categories: anomaly detection and pattern matching detection technology. Anomaly intrusion detection technology refers to acts or resource users use to judge whether the intrusion, and does not depend on specific behavior whether appear to detect; and the pattern matching detection technology is through the judgment and reasoning of some specific behavior, so as to detect the intrusion [9]. Anomaly detection and tries to find some unknown intrusion behaviors, and pattern matching detection is the identification of some of the known intrusion behavior. A real-time network security tools first it must be safe, that is to say not because of its introduction to the system to bring the security problem of other. It is a system with a reasonable structure to ensure the effectiveness and real-time detection, at the same time, we should fully consider the requirements of practical applications [10], and so it can determine the design principles and strategies are distributed intrusion detection system:

(1) Using the distributed monitoring, centralized management mode, namely through a plurality of monitor to monitor the distribution of site management on the web;

(2) Using the idea that modular component, it makes the system has good scalability to detect new intrusion behaviors emerging;

(3) Minimizing the effects on the system and network performance and resource occupation;

(4) Through the rule knowledge base to intrusion detection, this can adapt to different computer environment from a single computer system to hundreds of different computer systems

(5) Intrusion events for false negatives and false positives can be controlled in a reasonable range.

It is mainly composed of five parts of a distributed network intrusion detection system based on Network: engine, storage system, analysis system, response system, and console.

Storage system is the role of storage network engine produced from raw data and analysis results and other important data analysis system. The original data storage to provide conclusive evidence in legal sanctions against intruders, the shared database storage system is different between the components; it is different parts of the system to provide their data of interest. Therefore, the storage system should provide flexible data maintenance, query and processing services. Analysis of the system's role is to come from the network engine data packets for analysis and processing. Analysis system including the module, rule knowledge base, the protocol analysis module, data analysis module and secure communication of five parts pretreatment [11]. Analysis of the system is the core of intrusion detection system based on protocol analysis, pattern matching detection technology is the guarantee of high efficiency system. The characteristics of highly standardized protocol analysis technology based on network protocol to detect the attack behavior quickly, this detection technique of small amount of calculation, even in high load on the network, but also can detect a variety of attacks, but not loss. The response

system is a system to adopt corresponding measures of intrusion behavior has confirmed the. Response measures include: (1) warning measures, such as: to administrator email, send warning messages to the console; (2) protective measures, such as: cut off the invaders TCP connection, modify the router access strategy[12]. The console is the intrusion detection system and user interface. The user can through the various monitor console configuration system, but also through the console to understand the operation of the monitor.

3. Establishing a Model to Describe

Network engine function will be the network interface is set to promiscuous mode to monitor the network interface, and to reach the network packet interception down, for usage analysis system. Network engine will read all Internet traffic, including all protocol port, all subnet host all over 100 data, but in practical application, there need not be concerned with the data of a plurality of users, called garbage data, garbage data has an important proportion in all traffic, it has seriously affected the work efficiency of the system to improve ,therefore, efficient information filtering mechanism is an important part of information monitoring, which allows the user to specify the subnet host specific and specific protocols such as HTTP, FTP, SMTP port of the filter, only will be submitted to the user sensitive data concerned to the upper layer, so as to improve the efficiency of the system.

Analysis system is analyzing and processing network data from network lead climbing. System analysis is the core of the whole network based distributed intrusion detection system, and the analysis of detection methods using system is the key to the whole system performance, we use the improved method, also is the traditional pattern matching, pattern matching detection technique based on protocol analysis. Rule knowledge base is the feature library known network attack behavior, if there is no rule knowledge base, and then the system will not recognize any aggressive behavior. How to use simple, easy to use and effective language to describe an attack behavior characteristics [13], it is a key problem of the system, in judging a main network data packet is a when the attack is to judge the port number and the data section of the contents of the packet, IP address, protocol type and the TCP flag is auxiliary feature code. So whether it should be at the beginning of the analysis of original network data packet, directly matching port and the data contents of the section? For the detection of some network packets, judge the port and the data segment content directly, the detection efficiency is relatively high. But because the intrusion system is aimed at all packet are detected one by one to determine whether exhibit aggressive behavior, so should follow the common features of first detection of all attacks and then the individual characteristics of the principle, for example: if the first detection of 1P address [14], once found that does not belong to the detection range, immediately detects the next data package and not continue to test the data packets in the other field. This ensures that the efficiency of the whole detection system, but also improves the real-time alarm.

Each rule is logically divided into two parts: the rule head and rule option. The rule head defines the rules of behavior, network data to match the packet protocol, source address, destination address, the network mask, source port and destination port; rule options include to the user if the alarm information and used to determine whether the data packet to attack other packet data for display. The general format of rules: <rules for operating> <agreement > < source host IP>< source port operator > direction of target host IP>< target port > (< rule option 1: value 1>: < rule option 2: value 2>;..., < rule option n: value n>); in parentheses on the left side of the part is the rule head, in parentheses the part is the rule options. In front of colon rule option parts called option keyword. Rule option parts are not any rules are required, it is

used to explicitly define some aggressive behavior and the need to take some action (such as advertising for the packet). Only the composition rules must satisfy in order to perform the corresponding operation, limiting the message is "the logical relationship between elements and"; at the same time, between the rules in the knowledge base is the relationship between a "logical or". The rule head comprises data packet protocol address information, information and when each part of the packet conforms to the rules of the attack of what action to take in information. The first field rules of the head is the behavior of rule, the second field is the protocol, the third field is the address and port information. Rules of conduct that what should be done when it is discovered that meets the requirements of data packet [15]. Including three kinds of operation behavior: alert, log, pass, the rules of the knowledge base of the rule set is divided into two categories: one category is according to the type of service application layer to establish a set of rules. For example: ftp.urels set rules containing FTP services[16], dns.urels set the relevant rules contain DNS service attack, telnet.urels set contains relevant rules at P services attack; another kind is according to the type of attack classification, for example: dos.urles set contains refused to related rules service attack class, backdoor.urels set contains a backdoor attack rules etc.. Through the classification management rules, we can more easily to rule knowledge base update management.

Pretreatment is mainly to complete the restructuring of network data packets. Protocol analysis includes all specific protocol analysis sub module: the analysis of HTTP protocol, FTP protocol, TENLET protocol analysis, analysis of POP3 protocol analysis, SMTP protocol analysis, analysis of IMAP protocol, TFTP protocol analysis and protocol analysis of RPC module. Protocol analysis mainly completed the following work: two TENLET, S P, I protected price 4, Yang P3 protocol is a character oriented protocol, so the data preprocessing, from single protocol to transmit the message, requires protocol processing module to cache, in accordance with the agreement of the command node beam mode (with a carriage return characters that end), to several continuous data packet reform a complete command for attack detection feature matching model, process in the reorganization of data packet, to carries on corresponding processing to the specific client protocol package in some special type characters, including the protocol itself commands, space etc.. In protocol analysis, intrusion behavior of some simple determination, by the analysis of HTTP protocol as an example, after receiving the message of FTP protocol, first determine the message integrity, then calculate the HTTP command operation and the length of the parameter, if found greater than a given length, can be considered HTTP long command buffer overflow attacks. Directly generate alarm signal to the response module. Without having to give a data analysis module, these not only enhance the real-time detection of the intrusion detection system, and reduce the attack feature matching mode resource intrusion detection burden.

The method of data analysis is the core of intrusion detection system, so as far as possible using the fast pattern matching algorithm to improve the performance of the system. Methods all attacks are represented as rules stored in the rule in the knowledge base. Package search intrusion features in network data, requires a valid string search algorithm, when the knowledge base of network data packet and rules in a rule model matching, and then determine which is the network intrusion behavior, for example, if you find a HTTP request to a server on the "/cgi/bin", it could be an attacker is looking for system GCI vulnerabilities.

The original network data capture network engine package submitted to the analysis of the system, by analyzing the system realize the function of data storage, it converts the original data recorded in a temporary database. If the intrusion events, it also make the analysis results write temporary database corresponding table. The storage system to achieve a permanent database management and maintenance, its

timing from the temporary database batch loading data, doing so reduces the real-time requirement of the system. The main reason is to load data into database, we should really take the physical storage of the data written to a database, you must call the submit command, if each data was submitted to a, so for the information processing will provide but will affect the system performance and convenient.

Even if system has written good will also appear a false alarm. When the analyst will require the system to provide all of the data, including the data of Baotou and content, manual analysis, but we cannot long-term preservation of high fidelity of such data, we need to perform one or more reduction in database. An effective way to reduce the data is to bring the database into two main storage devices: the original data in the database and long-term record database. Original data the original data stored in the database: the database has high fidelity data over a period of time, the data includes the source PI address, destination PI address, the original data, spoofing characteristics, characteristics of new attacks, last seen time etc.. The establishment of multiple index and optimize the database should be, so as to be capable of the most effective search. Another important issue of original data in the database is stored much data, the original data to keep long. Is generally believed that the original data should be kept for 3 days to a week, we always as long as possible to keep the original data, the original data manual analysis will be limited after reduction. But the original data takes up a lot of memory, while increasing the original data will extend the database retrieval time, reduce the speed of data analysis. The system provides two ways to save the original data set: according to the time and according to the storage capacity. According to the time set to analysis as the basis of the original data set should save time, is to ensure that the retrieval performance is the premise to configure the system and according to the storage capacity. Long term records database: long term record database to reduce the format of the data record detection conditions of a long period of time. Long term records database of main support to produce reports rather than interactive query, but also help may have missed some of the events were detected. Reduced data format generally only include time, source PI address, source port, the target's address, destination port, protocol identifier.

4. Path Searching Based On Ant Colony Algorithm

The most short circuit search algorithm, compared with other in large-scale network is the most short circuit search; ant colony algorithm has the advantage that the model is simple and fast computing speed [17]. But there are also two problems: first, ant colony is the movement of the individuals in random, although through information exchange can towards the optimal path of evolution, but a large number of chaotic search in the path of a shorter path is difficult, especially at the beginning of the iteration, due to the information in the network element uniform distribution, the ants in the absence of any induction search path is very long, and the path to the poor quality, not only reduce the efficiency of the algorithm and does not favor the algorithm convergence; Second, the pheromone update rule is the core of the ant colony algorithm, determine the advantages and disadvantages of the final solution[18]. An update is too high will accelerate the convergence leads to fall into the local optimal solution, while low update will reduce the efficiency of the algorithm; the algorithm converges to a short period of time. In large-scale network, each path searching starting point and the end point of different, fixed the pheromone value cannot meet the optimal solution convergence of different query case, aiming at the above problems, this paper puts forward the corresponding improvement scheme. In order to avoid the path of ants to choose too "remote", enacted a new rule of ants marching, ensure that each ant can find from beginning to

end a shorter path. When an ant every arrival of a new node, first to determine whether the current path table trajectory has a shorter reach the current node, delete the ant the current track the middle of the table node exists, then the node next selection.

Attribute value plus 1, path selection in other ants, can be combined with the deletion of numerical on route choice probability weighted to a certain extent, to avoid the induction of ants "remote" section. Weighted ants from node i to node j route choice probability calculation method as shown in formula (1):

$$P_{ij}^{k}=\begin{cases}\dfrac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}0.88^{c_{ij}}}{\sum_{s}\tau_{is}^{\alpha}\eta_{is}^{\beta}0.88^{c_{is}}}, j\in \text{all nodes}\\ 0, \quad \text{otherwise}\end{cases} \tag{1}$$

Where τ_{ij}^{α} is scored probability of each nodes, η_{ij}^{β} represents the probability of expectation between two nodes. According to the different topological structure, it can according to the need, obtaining different α and β needs node length of different topological structures.

In addition, it defined distance each ant's longest running, when the travel distance is too long to stop its route choice behavior and the starting point of the re selection. According to the effect of the design before and after the improvement of search experiment, experiment set population consists of 30 ants, the number of iterations for the 200 time, the shortest distance path search. Therefore, to improve the algorithm of the initial population to complete the search time is improved before 29.9%, and it improved the search path length than the improved dropped 78.7%.

Ant colony algorithm can be viewed as a probability distribution model for the parametric solution space based on the parameters of the solution space, parameterized probabilistic model is the pheromone updating rule, so the pheromone of ant colony algorithm is the core, directly affect the optimal convergence speed and solution quality, all the ants in the completion of a process from the starting point to search destination path, usually to a speed of information updating, the updated M. Dorigo method have been given three different models to calculate the increment of information element, they are the ant density and ant system, ant cycle system, such as the type of calculation method (2):

$$\Delta\tau_{ij}^{k}=\begin{cases}M & \text{routing}\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

The system can calculate the number of formula (3)

$$\Delta\tau_{ij}^{k}=\begin{cases}M/d_{ij} & \text{routing}\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

Peripheral system can be calculated as formula (4)

$$\Delta\tau_{ij}^{k}=\begin{cases}M/L_{ij} & \text{routing}\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

Where M represents the amount of information needed for each path, d_{ij} represents the distance between two nodes, L_{ij} is the time of transmission between two nodes. The above three systems, different path search operations on the information amount, is a fixed value, but according to the search path of different size, each pheromone updating will be very different, the path and three kinds of calculation methods for all search to put a certain amount of pheromone, not only

the large amount of calculation, but also wastes the pheromone release, dilute the preferential rules, resulting in slow convergence speed. Some researchers believe that the current cycle can find the optimal solution or the global optimal solution trajectories are updated in accordance with the three update strategy, but easy to make the pheromone over concentrated in a shortest path algorithm and lead to premature. So the research of how to make the updated information pheromone diversification, make the path selection presents diversification, in order to make the algorithm convergence faster than the best path.

Using ant colony algorithm to solve the dynamic routing problem, the general choice of path travel time as the main evaluation of the shortest path of the standards, in accordance with the size of the travel time information update to induction algorithm to the optimal solution convergence, another advantage of using ant colony algorithm is, in solving the dynamic shortest path, cannot consider the time characteristics of the road network, the search process and route travel time calculation points and, in the search is complete, can be calculated to have the option of travel time on some path, this can avoid some unnecessary calculation.

According to the improved ant colony algorithm and dynamic travel time average velocity based method, this paper obtained the realization steps of calculating dynamic shortest path algorithm using the following:

Step 1: global parameters initialization, including road network data structure (including the mean velocity section number, node number, according to the selected network query time corresponding periods), the road network in each section of the global pheromone initial value;

Step 2: make C=C+1 initialization of N ants, the trace table of each ant is setting to null value, setting the start point and end point;

Step 3: for each ant according to formula (2-4) path selection rules, for each ant, in search of new path, first check whether can optimize its running track, optimization will cut sections of the deletion of attribute value plus 1, then the node according to the path selection method to the next, the deletion of attribute if you choose sections of the value is greater than 0, then it will cut the link attribute value minus 1, will to the trace table in the new path to search into the temporary information, and more sections of the element values.

Steps 4: repeating 2 to step 3, until the iterative coefficient C is equal to NC;

Step 5: the local optimal solution of choice at each iteration in the global optimal solution can be obtained, the shortest path for the global shortest path, and the output of its trajectory, travel distance and travel time.

The existing ant colony algorithm study and application of most only for smaller problems, research and application of ant colony algorithm for large scale problems is not much. And the practical application of the problem is often large-scale problems and even very large scale problems, aiming at these problems, the choice of parallel strategy to design the ant algorithm to solve the right, will greatly reduce the computation time is shortened, which makes the algorithm more timeliness.

For some complex model, the solution space huge problem solving dynamic optimization of complex systems, artificial intelligence algorithm has become the most promising solution method. But the artificial intelligent algorithm also requires a large amount of computing resources and storage space, in solving dynamic optimization or large-scale optimization is restricted to a certain degree. Computing capacity of a single processor is the main bottleneck, with the development of computer technology, computational systems for solving dynamic optimization problems of complex large system is becoming the research hotspot of distributed network connection based parallel. The key of parallel solution to the solution of the problem according to the data and the task is divided into several parts. The

application of ant colony search algorithm for large-scale shortest path, from the data perspective on Algorithm for partitioning is difficult, because the spatial and temporal characteristics of the road network is very obvious, the distribution of stored in different computers, it will cause the ants search in different computer nodes to swim away, resulting in a large amount of relatively poor communication, efficiency [19]. But from the task division of view, ant colony algorithm has the natural advantage of ant colony algorithm in essence is a parallel search technology, due to the ant colony algorithm each time to send IV ants to reach the terminal to find a path from the starting point, the process is very suitable for parallel.

5. Simulation Results

Router computing framework is designed by master and slave, namely the existence of a control node and multiple computing nodes. Work to be responsible for the coordination between nodes, each node in the registered to the management node will be assigned a unique ID. The management node internal maintains a list of nodes to before the event, the list includes each node ID and address information, and which nodes are assigned to the part of the whole graph. The number of data structure size and routing nodes to save this information management in the node related, has nothing to do with the graph of the number of nodes and edges. Therefore, although only one management node, also enough to coordinate calculation of working on a very large graph, a computational node in memory worker maintenance assigned to the basis of router state, the nodes in each super step, open a thread for each router. Each thread will loop through the router load node, function and regulation of each node and passed to the function node to the current value, the iterator of a received message and a side of the iterator. Without considering the case of error, simulation can be divided into the following five steps:

1 the management node process from the metadata HDFS reads graph. Metadata describes a weighted graph of the original data of the top. Master uses the first in a weighted graph on the top floor. Chapter of the segmentation algorithm is the basic segmentation map data, one or more Router and allocation map to compute node operation.

2 computing nodes according to a management node distributing of relay nodes, to read the original data from hdfs. Read data is completed, to send master READY signal.

3 management nodes receives the READY signal all routers, master sends the START signal to inform all the nodes to iterative calculation.

4 computing nodes and management nodes receive signal after work to do, the node polling on the basis of open a thread for each router. The Compute O function calls to each node router that includes an active message, receiving from the last Super step to messages. Messages were sent asynchronously, it is in order to make the calculation.

The use of single source node routing algorithm based on path vector (AVDO) algorithm in cluster conducted several experiments. Inspect a random graph in size (in order to study the scalability of the run). The measurement results including user initialization node set, generating test topology map, the running time in memory. Topology of randomly generated Figure 1 is one has 50000000 test nodes, using the shortest path as the path computation time calculation. Among them, the number of working nodes increasing eventually increased to 128 nodes.

With the increase of working nodes, the work operation time decrease.

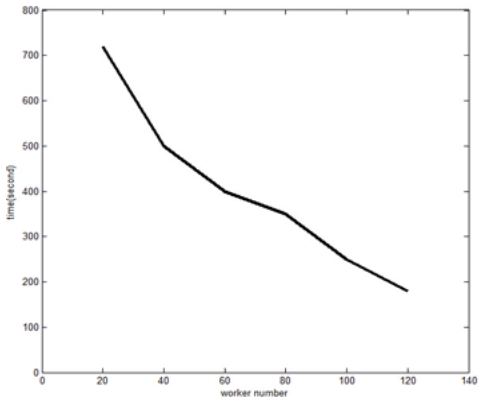

Figure 1

Figure 2 shows a random graph of the number of nodes varies from 10000000 to 100000000 of the shortest path algorithm running time, when the number of working nodes is fixed, a total of 128 working nodes are scheduled on 32 machines. Through the diagram can be seen, extensible router system in the node number increased under the. In this process of change, the operation was called from 37 seconds grows to 287 seconds, the running time is linear with the size of graph data growth, but the growth is relatively slow.

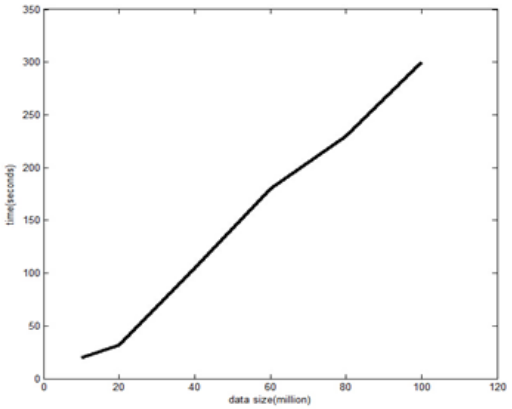

Figure 2 the Relationship between Router Running Time and the Size of the Data

6. Conclusion

This paper presents a complex system of large scale information network analysis framework. The analytical framework focuses on good data distribution, it provides efficient router calculation model and supports the different algorithm design paradigms. The distribution of data, the analysis framework for large node data graph data for segmentation, and achieving load balancing between storage and computation. In the calculation model, the router calculation model based on multiple message dissemination mechanism, it can efficiently carry out analyzing information network of large-scale complex system.

Acknowledgements

This work is financially supported by the National Natural Science Fund, China (No 61103143), basic and frontier project of Science and Technology Department of Henan province, China (No 142300410334), the funding scheme for young backbone teachers of colleges and universities in Henan Province, China.

Reference

[1] M Rubinov and O. Sporns, "Complex network measures of brain connectivity, uses and interpretations", Neuroimaging, vol. 52, no. 3, **(2010)**, pp. 1059-1069.
[2] K. G. Guruharsha, J. F. Rual, B. Zhai, , "A Protein Complex Network of< i> Drosophila melanogaster", Cell, vol. 147, no. 3, **(2010)**, pp. 690-703.
[3] M. P. Heuvel, R. C. W. Mandl, C. J. Stam, , "Aberrant frontal and temporal complex network structure in schizophrenia a graph theoretical analysis", The Journal of Neuroscience, vol. 30, no. 47, **(2011)**, pp. 15915-15926.
[4] J Wang, H Mo, F Wang, "Exploring the network structure and nodal centrality of China's air transport network, a complex network approach", Journal of Transport Geography, vol. 19, no. 4, **(2011)**, pp. 712-721.
[5] R. V. Donner, J. Heitzig, J. F. Donges, "The geometry of chaotic dynamics—a complex network perspective", The European Physical Journal B-Condensed Matter and Complex Systems, vol. 84, no. 4, **(2011)**, pp. 653-672.
[6] R. V. Donner, J. Heitzig, J. F. Donges, "The geometry of chaotic dynamics—a complex network perspective", The European Physical Journal B-Condensed Matter and Complex Systems, vol. 84, no. 4, **(2011)**, pp. 653-672.
[7] R. V. Donner, J. Heitzig, J. F. Donges, "Ambiguities in recurrence-based complex network representations of time series", Physical Review E, vol. 81, no. 1, **(2010)**, 015101.
[8] P. Yang, X. Li, M. Wu, *et al.*, "Inferring gene-phenotype associations via global protein complex network propagation", PloS one, vol. 6, no. 7, **(2011)**, e21502.
[9] B. Luque, L. Lacasa, F. J. Ballesteros, "Feigenbaum graphs: a complex network perspective of chaos", PLoS One, vol. 6, no. 9, **(2011)**, e22411.
[10] G. A. Pagani and M. Aiello, "The power grid as a complex network: a survey", arXiv preprint arXiv.1105.3338, **(2011)**.
[11] Z. Zhang and X. Zhang, "A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation", Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on. IEEE, vol. 2, **(2010)**, pp. 240-243.
[12] A. M. Knights, R. S. Koss and L. A. Robinson, "Identifying common pressure pathways from a complex network of human activities to support ecosystem-based management", Ecological Applications, vol. 23, no. 4, **(2013)**, pp. 755-765.
[13] Y. Xu, W. Zhou and J. Fang, "Topology identification of the modified complex dynamical network with non-delayed and delayed coupling", Nonlinear Dynamics, vol. 68, no. 1-2, **(2012)**, pp. 195-205.
[14] T. Hossmann, T. Spyropoulos and F. Legendre, "A complex network analysis of human mobility", Computer communications workshops (INFOCOM WKSHPS), 2011 IEEE conference on, IEEE, **(2011)**, pp. 876-881.
[15] T. Hossmann, T. Spyropoulos and F. Legendre, "A complex network analysis of human mobility", Computer communications workshops (INFOCOM WKSHPS), 2011 IEEE conference on. IEEE, **(2011)**, pp. 876-881.
[16] W. C. Yeh, Y. C. Lin, Y. Y. Chung, "A particle swarm optimization approach based on Monte Carlo simulation for solving the complex network reliability problem", Reliability, IEEE Transactions on, vol. 59, no. 1, **(2010)**, pp. 212-221.
[17] Z. Wang, Y. Wang and Y. Liu, "Global synchronization for discrete-time stochastic complex networks with randomly occurred nonlinearities and mixed time delays", Neural Networks, IEEE Transactions on, vol. 21, no. 1, **(2011)**, pp. 11-25.
[18] Z. Wang, Y. Wang and Y. Liu, "Global synchronization for discrete-time stochastic complex networks with randomly occurred nonlinearities and mixed time delays", Neural Networks, IEEE Transactions on, vol. 21, 1, **(2011)**, pp. 11-25.
[19] A. Tordesillas, S. Pucilowski and D. M. Walker, "A complex network analysis of granular fabric evolution in three-dimensions", Melbourne Univ Victoria (Australia), **(2011)**.

Authors

Yiran Wang, he received B. Eng and M. Eng Degree in Computer Science and Technology from Zheng Zhou University, China in 1997 and 2005 respectively. He is currently researching on Internet of Things, Enterprise Informationization.

Guang Zheng, he received B. Eng Degree in Computer Application Technology from Henan University and M. Eng Degree in Computer Application Technology from University of Electronic Science and Technology, China in 1996 and 2009 respectively. He is currently researching on Computer network.