Research on Topology and Security Behaviour Intelligent Analysis by Hybrid P2P Traffic

Yike Wang*, Yuhang Feng, Longshan Shang, Haonan Zhang, Boshen Zhang
Dalian University of Technology, Dalian, Liaoning, 116024, China

*Corresponding author: yikewang@dlut.edu.cn

Abstract. At present, most of the popular P2P topology sharing systems are based on unstructured P2P topology. This topology uses flooding method to spread query information, which has good stability, but the efficiency is very low. According to the changing characteristics of the communication attribute state of the network behavior which occupies a large network bandwidth and communicates more frequently, an attribute migration state-oriented network communication behavior analysis method is proposed. The research on the identification of P2P traffic is of great significance for the management of P2P network. In view of the shortcomings of the current P2P traffic identification methods, such as large error and unstable identification results, in order to improve the identification effect of P2P traffic, a P2P traffic identification method based on neural network is proposed. This paper proposes a network file sharing system Kapa, which is based on a hybrid hierarchical P2P topology, which combines the advantages of unstructured and structured P2P topology and has strong practicability. The recognition effect and reliability of this method for P2P applications are verified.

Keywords: P2P, Unstructured, Attribute migration, Neural network, Reliability verification

1. Introduction

Topology structure determines the performance of P2P file sharing system. According to the relationship of topology structure, P2P network topology can be divided into three basic forms: centralized topology, fully distributed unstructured topology and fully distributed structured topology, also known as DHT network.

For network behavior analysis have been compared and effective technical method, can be in a different environment for different object is used to identify the behavior analysis [1], but for large network operations unit, there are the following limitations: 1) is limited to the network at any time generated a huge number of network management log data mining methods can't be used in the online whole [2];2) Due to the complexity brought by the random and disordered behavior of network users, the existing technical methods are often unable to respond effectively and the detection effect is poor [3];3) At present, although a large number of derived attributes can be used for statistical analysis of traffic behavior, most of these analyses are oriented to the state changes among attributes, while neglecting the quantitative evaluation of the attribute gap state that reflects the user in the network space level [4].
The network file sharing system based on P2P has become the most popular software on the Internet at present. Compared with the traditional file sharing system based on C/S (Client/Server), the P2P file sharing system has the advantages of decentralization, good scalability, strong robustness, high cost performance and load balance, so it has a broad application prospect.

In order to obtain the ideal effect of P2P traffic identification, peer-to-peer (P2P) network traffic identification method based on neural network is proposed, and compared with classical P2P traffic identification method on the same platform simulation comparison experiment, the results show that this method can describe the change of the P2P network traffic characteristics of P2P network traffic identification accuracy is higher, it is very obvious advantages.

Based on the characteristics of P2P network traffic identification technology mainly through TDD traffic has proportion, packet size distribution, etc, and then combined with machine learning algorithms, such as support vector machine (SVM), bayes algorithm and decision tree [5], peer-to-peer (P2P) network traffic identification model was constructed through training, but in the practical application of these machine learning algorithms have some deficiencies, such as support vector machine training process is complex, peer-to-peer (P2P) network traffic identification is long; Bayesian algorithm and decision tree P2P network traffic identification accuracy is low [6].

2. P2P network

The traditional network usually adopts the working mode of client/server, in which the server is the center of the whole network. The client sends the task to the server, and the server processes the task, so the server completes most of the operation, which has a heavy burden and difficult to improve the work efficiency. However, the P2P network weakens the server.

The fully distributed unstructured P2P network adopted the organization method of random graph, and the Flooding method was adopted to transfer the querying information among the nodes. P2P applications based on this structure have a good fault-tolerant ability in the face of the dynamic changes of the network. The most typical application is the network file sharing system Gnutella. Gnutella is a variant of Napspter, and its topology is different from Napspter.

The biggest difference between them is that Gnutella is a more pure P2P system. Gnutella is a fully distributed P2P communication protocol that only finds nodes. In the Gnutella network, where there is no central indexing server, each client is truly peer, free to share any file, and it can be transferred in a friendly manner; in terms of document retrieval, the Flooding algorithm based on completely random map was adopted, and the codes of any client were made public.

3. Construction of P2P Topology based on Neural Network

3.1. Radial neural network

Neural network is one of the most commonly used artificial neural networks, which is usually composed of three layers, namely, the input layer, the hidden layer and the output layer. Each layer has a different number of nodes according to the actual problem. If the number of nodes is N, P and M, then the topology of radial basis neural network is shown in Figure 2.
3.2. P2P Construction based on Neural Network

For the input vector $X = [X_1, X_2, \ldots, X_m]^T$, the unit output of the $i$th node in the hidden layer can be obtained by transforming the radial basis function $\phi(.)$ to the hidden layer, as shown in Equation (1).

$$h_i = \phi(||X_i - c_i||)$$

$$h_i = \exp\left(-\frac{||X_i - c_i||}{2\sigma_i^2}\right)_{1 \leq i \leq p}$$  

The output of the hidden layer is taken as the input of the output layer, and the output of the $K$th node of the output layer is obtained by connecting the weight $W_{ik}$, as shown in Equation (3).

$$y_k = \sum_{i=1}^{p} W_{ik} \exp\left(-\frac{||X_i - c_i||}{2\sigma_i^2}\right)_{1 \leq k \leq m}$$

P2P network traffic identification characteristic values are collected and normalized, as shown in Equation (4).

$$f'_i = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}}$$

P2P network traffic data are collected, and P2P network traffic and non-P2P network traffic are marked. The learning sample is established according to the characteristic and labeled P2P network traffic and non-P2P network traffic values. The number of nodes in the input layer of RBF neural network is determined according to the number of eigenvectors of P2P network traffic, the number of nodes in the hidden layer is determined according to certain rules, and the number of nodes in the output layer is determined according to the type of P2P network traffic.

4. Hybrid hierarchical P2P framework

4.1. Network Framework

In KAPA, the upper layer applies the structured topological Chord protocol based on DHT, while the lower layer organizes the nodes into clusters based on the flooding method by a strategy with close physical distance. A group consists of ordinaryNodes and superNodes. OrdinaryNode is a group and is equivalent to a virtual node in Chord. SuperNode is the high-performing ordinaryNode within the group,
which is used to execute the upper-chord protocol and manage the entire group. The overall architecture of Kapa network is shown in Figure 3.

![Figure 3. The network structure](image)

4.2. Input-output algorithm
The new group CLI joining algorithm is described as follows. ① The CLI locates itself on the Chord ring based on its ID and obtains information about the successor group CLJ. ② The CLI initializes its own pointer table and establishes Interval and Start fields. According to the value of start item, set up the corresponding super node table for each successor group, and ask CLJ to find the successor records of each item. ③ Subsequent group CLJ receives the message sent by CLI, and its previous subgroup acts as the previous subgroup of CLI. Subsequent group is informed to update its previous subgroup, so that it can be better used by CLI.

When the CLI exits, if there is any new metadata to be stored in the group, the metadata presentation will be stored in the successor group of the CLI, and the existing local metadata will be copied to the successor group. ② After metadata replication, the successor group is notified to exit by the CLI, and its predecessor group is modified to be the CLI's predecessor group. Then the CLI's predecessor group is notified to exit, and the successor of the previous successor group is also modified to be the CLI's successor group. ③ All metadata maintained by the group is deleted. ④ The other groups in the Chord ring are notified by the CLI to exit and update the corresponding entries in the pointer table and the corresponding supernode table.

4.3. Address update
Update the address gap composition of destination address, update the number of destination address changes with time interval Δt.

For the update address set $B_{update-k}$, where each element $b_{update-l} \in B_{update-k} = B_k - (B_k \cap B_{k-1})$, the number of the set $B_{update-k}$ is $m_{update-k}$, $1 \leq l \leq m_{update-k}$. By calculating the average value of $\text{dis}(a, b_{update-l})$ and the ENTSI-dis of all nodes in the set $B_{update-k}$.

The degree of port activity and update is calculated by the ratio of the number of active ports to $\Delta T$ at a fixed time interval. Within the $\Delta T$ interval, the number of active and updated ports can be obtained from Sport in ENTDI, and Sport needs to calculate the corresponding number in the destination address set $B_k$. 


Port activity $D_{\text{port-active-k}} = \frac{n_{\text{active-k}}}{\Delta t}$ where nactive-k is the number of ports whose sport status is active in the ENTDI of the corresponding node element in the set Bk.

5. Empirical results

In order to analyze the recognition effect of P2P network traffic of neural network, the public data set provided by An-DrewMoore is selected as the experimental object. There are 11 data sets in total. Each data set contains P2P network traffic and non-P2P network traffic, and a certain number of P2P network traffic and non-P2P network traffic samples are selected for each data set.

For a P2P network traffic identification method, only high precision P2P network traffic identification results are not enough, but also to analyze its P2P network traffic false alarm rate and missing alarm rate. The false alarm rate and omission rate of P2P network traffic with different methods are shown in Figure 4 and Figure 5.

Figure 4. Comparison of network topologies with different methods

Figure 5. Comparison of error reporting rate of different P2P methods

As can be seen from Figure 4 and Figure 5, the method in this paper has a lower false alarm rate or missing alarm rate of P2P network traffic, and the comparison results show that the overall performance of the method in this paper is better for P2P network traffic identification.
The number of information query hops for Kapa is close to Chord, which shows that Kapa has only a small query latency. When nodes exceed 215, the average number of hops for Kapa is less than Chord. As can be seen from Figure 6, if the node number is greater than 215, the average query success rate of the Kapa network is less than 50%. However, compared with Gnutella network, Kapa still has a relatively high query success rate.

Since the computational performance of P2P network traffic identification is also an important evaluation index, the learning time of P2P network traffic of different methods is counted (seconds, s), and 5 learning experiments are carried out for each method.

In addition, Kapa has great advantages both in terms of query efficiency and load on the network. Moreover, the two-layer structure based on Kapa can accommodate more nodes, so the query efficiency of Kapa will not decline too fast and it has better scalability.

6. Conclusion
P2P network lending platform itself is necessary to understand and grasp the real information of multi-parties, including screening and reviewing the credit level and asset status of customers, so as to ensure the security of transactions. This puts forward higher requirements for network topology performance.
With the continuous expansion of the application scope of P2P network, the application types are becoming more and more complex and diverse, which brings some challenges to the management of P2P network. In order to manage P2P network more finely, combined with the changing characteristics of P2P network traffic, a P2P traffic identification method based on neural network is proposed. Kapa, a distributed network file sharing system based on hybrid hierarchical P2P, introduces the architecture and related algorithms of Kapa network, and gives the simulation results. At present, the Kapa system has been implemented and has been tried out in a certain range. The simulation test is carried out through a specific P2P network traffic data example, and the results show that this method is a P2P network traffic identification method with high precision and high speed. And the stability of P2P network traffic identification result is better than other P2P network traffic identification methods, so it has a very wide application prospect. Combined with the definition of the extended node transmission information, the measurement parameters of the relevant attribute state are obtained.

References
[1] Suhas Kuruba Manjunath Yoga, Kashef Rasha F.. Distributed Clustering Using Multi-Tier Hierarchical Overlay Super-Peer Peer-to-Peer Network Architecture for Efficient Customer Segmentation [J]. Electronic Commerce Research and Applications, 2021 (prepublish).
[2] Chen Yunxue, Song Jinyu. Legal risk and Supervision of P2P Network loan Industry [J]. Legal system and Society, 2021 (07): 49-50.
[3] Nitin Singha, Yatindra Nath Singh. New incentive mechanism to enhance cooperation in wireless P2P networks [J]. Peer-to-Peer Networking and Applications, 2021 (prepublish).
[4] Shiyong Li, Wei Sun, Huan Liu. Optimal resource allocation for multiclass services in peer-to-peer networks via successive approximation [J]. Operational Research, 2021 (prepublish).
[5] Yang Li, Ding Yan, Lin Xiyang. The application of big data in preventing the financial risk of P2P network loan industry chain [J]. Chinese Business Theory, 2021 (02): 86-88.
[6] Wang Huijuan, Liao Li. Research on Credit Authentication Mechanism of P2P Network Lending platform in China--empirical evidence from Renren loan [J]. China's Industrial economy, 2014 (04): 136-147.