Data Flow Control Algorithm based on Feature Matching in Mobile P2P Network

ZHANG Yuan
Department of Electronics and Information Engineering, Anhui Institute of Architecture, Anhui Hefei, 230601, China
Email: zyztjm@126.com

SHEN Qing-wei
Department of Electronics and Information Engineering, Anhui Institute of Architecture, Anhui Hefei, 230601, China
Email: qwshen@ai.ai.edu.cn

Abstract—This paper proposes a characteristic matching based mobile P2P network data flow control algorithm, to detect the illegal data flow in the mobile P2P network, finally achieve the purpose of control the data flow. The algorithm is proposed based on the SVM feature matching algorithm, firstly on the basis of the traditional feature selection, the paper put forward 3 d feature selection method, so as to improve the precision of detection, secondly uses RBF kernel as the kernel function of SVM feature matching algorithm, and finally divides the test process into iterative detection and smooth detection. Simulation experiments show that characteristic matching based on mobile P2P network data flow control algorithm has a higher recognition rate and better scalability.

Index Terms—MobileP2P; Feature matching; The SVM; Data flow control

I. INTRODUCTION

With the development of the mobile network technology and the gradually mature of P2P network technology, the mobile P2P network combining with mobile computing and P2P technology is emerged [1]. Mobile P2P Network (Mobile Peer-to-Peer Network) is also called Mobile peer-to-peer Network, it is based on the Mobile wireless Network, the nodes in the Network is usually a Mobile phone, PDA and other Mobile devices, these devices have a certain particularity, such as small storage capacity, low processing power and limited Network bandwidth and limited battery energy, etc [2]. Therefore, compared with the traditional P2P network, mobile P2P network has more constraints and restrictions.

On P2P in “research and realization of work together in mobile network”, Yang Debing (2008) proposed the problem of P2P’s real-time performance is not good and delay response for the important information, so he realizes the simple cooperative work on the basis of choosing Kadhamia algorithm (Kad) as the research object. SuXingHua (2007) has studies P2P technique in the “P2P video download system based design and implementation” and proposed a P2P video download model based on Kadhamia protocol, then put forward the design scheme of the model and design for implementation, this P2P based approach is superior compared with the traditional C/S architecture server. In order to guarantee the cloud service security, in “two kinds of P2P protocol based data self destruction technology”, Zhang Dezhi, He En (2012) used two P2P protocols of Chord and Kadhamia to designe individual application data based self destruction protocol, using the owned disturbance characteristics of P2P protocol to carry out the automatically destroy of individual network data, using OMNEST simulation software to carry out the performance analysis of the data destruction of these two protocols, the simulation results show that the Kadhamia protocol data has a better destruction performance.

With the rapid development of mobile P2P technology, in order to avoid being easily detected, mobile P2P applications have adopted the technology of dynamic port and encryption protocol, mobile P2P traffic detection and control technology has been faced with serious challenges [4]. Due to port jump, and the development of hiding technology of load encryption, the method of using application port to test mobile P2P traffic has been basically eliminated [5]. Dynamic port, for example, makes the traditional traffic detection method based on port is powerless, as a result, mobile P2P traffic detection is turning to the transport layer characteristics based method and the method based on data mining [6]. And transfer layer flow characteristics based P2P traffic detection method is a kind of effective method at present. This method is based on the analysis of transport layer packet (TCP and UDP packets), and combined with the traffic feature of P2P system, to detect whether a network flow belongs to P2P traffic [7]. This kind of flow characteristics based method has the advantages of accurate, efficient, and able to detect the encrypted data flow compared with previous port test and bag net load testing method.

The proposed Mobile P2P traffic detection scheme is using this detection and control method, the unknown and encrypted mobile P2P applications can be effectively detected in flow feature matching phase, so as to overcome the inherent defects of deep packet inspection technology in this aspect, and overcome the defect of weak classification ability based on transport layer’s flow characteristics detection technology.
II. MOBILE P2P NETWORK

A. Definition and Character of Mobile P2P Network

Mobile P2P network is also called mobile peer-to-peer network, it is the overlay network superposition in the mobile wireless network’s network layer, belongs to the session layer, it can use a variety of underlying access technology of bandwidth, the purpose of the network is achieving the data and resource sharing [8] between mobile nodes based on direct data exchange.

Mobile P2P network is based on the mobile wireless network, the network nodes is usually the mobile devices, such as mobile phone, PDA, etc. Compared with the traditional P2P network, mobile P2P network mainly has the following characteristics [9]:

1. The highly dynamic nature. The traditional P2P network is based on fixed wired network; the nodes will not move randomly in the network, the topology of the entire network is stable. Mobile P2P nodes has a strong mobility, their location in the network change constantly, the highly dynamic nature lead to the entire network topology changes drastically, which could lead to a poor network search performance, also causes the inefficiency of data transmission and reduce resource search success rate.

2. It becomes more random and frequent of node join and leave. In traditional P2P network, nodes’ join and leave is mainly due to the factors such as start-up, shutdown [10]. Mobile P2P network’s transmission medium mainly is the electromagnetic waves, and electromagnetic waves have a limited scope and susceptible to interference and shielding, so in addition to the factors such as node boot and shutdown, the characteristics of the transmission medium will also lead to node’s join and leave, this has increased the randomness and the possibility of network node’s join and leave [11].

3. Each node has limited communication range.

4. The connection reliability. Compared with fixed wired networks, mobile wireless network connection reliability is poorer, mainly due to the wireless transmission medium used by the wireless network; it also will increase the failure rate of the network nodes.

Network structure of mobile P2P network

due to the central server in centralized network, making the network's fault tolerance and expansibility relatively poor, and including failure of single point; the flood type way of news spread of total distributed unstructured network increased the network load; network information maintenance of fully distributed network topology is also increased the network load; Half distributed network is a compromise for centralized network and the distributed network, which effectively control the spread of redundant information in the network, also has good extensibility.

Reference [12] proposed a half distributed and super node based mobile P2P system, as shown in figure 1. in which the super node (super node) is responsible for caching, content management and the communication with cable P2P network; Edge node (edge node) is responsible for providing a mirror image of mobile node, content virtual storage and data sharing; Mobile node’s function is unitary, mainly be responsible for the management of the dynamic data. the entire network is composed by the domain centralized by super node, domain, carry out the centralized control in the domain, realize distributed management inter-domain; At the same time, the mobile node in the domain can directly communicate in peer-to-peer manner without through edge nodes.

![Figure 1. Mobile P2P network structure](image)

B. Existing Problems

1. Relatively poor channel environment and mobility

In traditional P2P network, when searching for resources, information must go back in accordance with the original search path, but due to the instability of the physical connection of mobile devices, it is easy to cause the information unable to return [13].

A huge difference between cellular network and fixed network is the communication environment and mobility [14]. wireless links environment of cellular network is relatively much bad than fixed network, it will often happens the multipath fading of the wireless signal and channel congestion, which caused the connectivity between P2P nodes unguaranteed, to a certain extent, it affected the structure of the P2P overlay network, and thus it affects P2P overlay network’s structure, and it affects the inquire or the routing mechanism of P2P network, etc. [15]. Nodes’ frequent movement under mobile environment is also a big challenge for existing P2P business, which would cause a series of technical problems.

2. The self-defects of mobile devices

Another problem that can not be ignored is the self-ability of the mobile device. Due to the mobile device's own resources, which limit the calculation processing power, storage capacity, energy supply of mobile terminal. Although, in recent years, mobile computing and storage capacity has been greatly promoted, but with the P2P network equipment such as PC, laptop, there is a big gap, and that makes the software functions for mobile devices limited a lot, most successful P2P applications are not suitable for the existing situation, such as large file transfer and sharing.

3. Bandwidth consumption

In a fixed data online, P2P are generally considered "broadband killer". According to the relevant authority’s statistical report, it showed that in the fixed IP network,
more than 60% of the business flow is produced by all kinds of P2P business.

Mobile network bandwidth is increasing in recent years, our country’s third generation mobile communication technology; td-scdma has entered trial commercial stage, but still has a gap compared with wired network. P2P network’s consumption of network bandwidth has always been the main obstacle to its development. For mobile operators, the business traffic problem generated by P2P business will be a huge challenge to all mobile operators.

(4) Security problem
The security problems in Internet P2P also exists in the mobile P2P, there are a lot of security problems need to be solved such as DRM (digital rights management) problem, P2P network spam filtering, P2P network virus isolation, P2P network structure.

III. SVM FEATURE MATCHING
A. SVM
Support Vector Machine, referred to as SVM, was proposed by ms Cortes and Vapnik in 1995, it is the most representative small sample learning machine of statistical learning theory which is based on structure risk minimization [16]. It shows a lot of advantages in solving small sample, nonlinear and high dimensional pattern recognition. The SVM has been well applied in areas such as data mining, pattern recognition, function fitting and nonlinear system control [17].

SVM method is established based on VC dimension theory of statistic learning theory and structural risk minimum principle, searching for the optimal compromise between the complexity of the model (i.e., on a particular learning accuracy of training sample) and learning ability (i.e., any samples without error identification) according to the limited sample information, in order to get the best generalization ability (or generalization ability) [18]. The VC dimension is a measure of function which can be simply understood as the complexity of the problem, the higher of VC dimension the more complicated of the problems. The so-called structural risk minimizing refers to as the minimum sum of empirical risk and confidence risk when it is solving problems, this makes SVM very suitable for solving the problem of text classification.

Based on pattern recognition theory, the low dimensional and space linear inseparable model mapped to high-dimensional space through nonlinear may realize linearly separable. Kernel function’s (kernel ticks) basic function is to accept two lower dimensional space vectors, it can calculate the inner product variable values in high dimensional space after some transformation, therefore, kernel function technology is successfully applied in support vector machine, thus it create a variety of methods based on kernel function. Kernel function is defined as: $K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j))$. According to Mercer verification principle: in normed linear space $L_2$, the kernel function $K(x, y)$ can spread form as follows:

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x) \varphi_i(y), \lambda_i > 0$$

's necessary and sufficient condition is: for any non-zero function $f(x)$ that make $\int f^2(x)dx < \infty$, the condition $\int \int K(x, y)f(x)f(y)dxdy = 0$ can be established. This condition is also called the Mercer.kernel function that satisfy the Mercer conditions must be represented as an inner product in the feature space, thus it can be seen as a reasonable kernel function. Kernel function method can be combined with different algorithm to forme the different kinds of methods based on kernel function, and can choose a suitable kernel function algorithm according to different application.

The common function is

1. (Polynomial function)

$$K(x, x') = ((x \cdot x') + c)^d$$

2. (Gaussian radial function (RBF))

$$K(x, x') = \exp(-\gamma ||x-x'||^2), \gamma > 0$$

3. (Sigmoid kernel function)

$$K(x, x') = \tanh(k(x,x') + v)$$

where $k > 0, v < 0$.

SVM's core idea is to utilize a nuclear function which satisfies the Mercer condition to replace a nonlinear mapping, making the sample points in the input space can be mapped to a high-dimensional feature space, and make the linear be divided in the space, and then constructs an optimal hyperplane to approximate ideal classification results. Optimal hyperplane structure method is equivalent to "linearly separable support vector classification machine", the classifier transform the problem of using largest interval method to solve the optimal classification optimization into its dual problem, and solving the relatively simple dual problem to solve the original classification problem. Then the slack variables and penalty factor are introduced to solve the problem of nonlinear classification, a certain classification errors are allowed interval (soft), and finally got the the standard C - SVM of nonlinear soft interval. The clever is simplifying a complicated
optimization problem solving into the inner product operation of original sample.

B. Feature Selection

Because the support vector machine uses kernel function to handle all the input data, we can know by the relation $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ between kernel function $K$ and the mapping function $\Phi$, as long as it conforms to this definition, there is no limit to the form of $x_i$ and $x_j$, which can not necessarily be a vector. Currently used common input forms are vector, sequence, tree, set, the graph, etc. this paper uses the most common form - vector as input of support vector machine. Each dimension data of the vector is referred as feature , feature selection is to transform the low data into vector form accepted by support vector machine. Select features must be able to reflect the difference between P2P and the P2P traffic, so as to achieve the purpose of improving the detection accuracy.

Based on this, the paper gives the following 3 d feature selection method:

$$\text{vector(flow)} = \langle f(src, dst), g(src, spt), h(src, dpt) \rangle$$

In this feature selection method, using function $f$, $g$, $h$ calculate the former 3 dimension, the first dimension is to reflect the difference of P2P and non-P2P flow in the address, the second and the third is to reflect the port difference discussed by previous text. The function $f$ defined as follows:

$$f(src, dst) = \text{dif(src, DST) / same(src, dst)}$$

This function has two independent variables-- resource IP src and target IP dst , $\text{dif(src, DST)}$ calculate the IP number with different goals connected with IP, $\text{same(src, dst)}$ calculate the linking number established by target IP and dst IP src . Such as resource IP p0 has connections with 4different goal of IP p1, p2, p3 and p4, the connections are all 1 with the previous three IP, and it established 2 connections with the fourth IP, thus the $\text{dif(p0, DST)}$ value is 4,

$$\text{same(p0, p1)} = \text{same(p0, p2)} = \text{same(p0, p3)} = 1 , \text{same(p0, p4)} = 2$$

so the $f$ value is 4 of previous three IP to resource IP p0, the last one is $4/2 = 2$. For P2P flow, its $\text{dif}$ value is always big, but $\text{same}$ value is relatively small, the expected $f$ value will be bigger, for non-P2P on the contra. We can define the function of resource port and target port according to the same method, it shows as follows:

$$g(src, spt) = \text{dif(src, SPT) / same(src, spt)}$$

$$h(src, dpt) = \text{dif(src, DPT) / same(src, dpt)}$$

We use the figure 2 and figure 3 to illustrate the 3 d data does play a role to distinguish the effect of P2P and non-P2P. Figure 2 is a two dimensional data consisting of previous two dimensional images, two kinds of the P2P application web and FTP data concentrated near the origin, and the previous two dimension P2P data is generally greater than 10, but the boundaries between the P2P and the non-P2P data is not clear. Figure 3 is the 3 d data image, we can see clearly that 3 d data have been able to better distinguish P2P and non-P2P.

B. Kernel Function Selection

In addition to carefully selecting feature, another important aspect of improving the detection accuracy is to select the appropriate kernel function. Unsuitable kernel function can not map the input space data to high dimensional linear separable feature space.

In general, RBF kernel is the ideal first choice, for the following reasons:

1. The RBF kernel map the sample into high-dimensional feature space with non-linear, so it has more advantages in dealing nonlinear data processing than the linear nuclear;
2. The linear kernel is a special case of the RBF kernel, confirmed: linear kernel uses parameters $C$ which has the same value performance in some $(C, \gamma)$ with RBF kernel.
3. The Gaussian kernel is similar on the performance with RBF in some parameters;
4. Polynomial kernel needs to adjust three parameters, RBF only need to adjust one, and so model is less complex;
5. RBF has less calculation difficulty, because its output limit is between 0 and 1, but the output of the polynomial kernel may tend to infinite, Gaussian kernel is invalid on some parameters.

Based on the above reasons, we adopt RBF kernel as the kernel function of SVM.
IV. SIMULATION RESULTS

A. SVM Feature Matching based Iterative Detection

Test process as shown in figure 4, in the first training, we use small sample data of P2P and non P2P to train support vector machine, adjust the parameters of SVM to the best state. After the completion of training, output a model, this model is in fact a decision function, the SVM forecasting tool classify the test data refer to this model. If the false alarm rate is very high, it indicates a big possibility of non-P2P identified as the P2P, if the missing report rate is high; it indicates a large proportion of P2P identified as non-P2P. For these two cases, it needs to further improve the detection accuracy. So feedback the wrong fractional non-P2P data to P2P training samples, feedback the wrong fractional P2P data to P2P samples, and then repeat the previous process, until we get a satisfactory rate of false alarm and non-response rates, this is for the end of training.

B. Smooth Detection based on SVM Feature Matching

When we use the figure 5, 6, 7 to compare the previous three dimensions of the P2P (web traffic), BT, PPlive, we found that as the same F2P data, although BT and PPlive show the different characteristics from the non-P2P on these 3 dimensions, such as on the first and the third dimension, the value of BT and PPlive is always larger than the P2P, but the relations is blurred between them on the second dimension.

In order to avoid affecting accuracy and detection efficiency due to the data differences between P2P applications, we introduce the smooth functions

$$q(x) = \begin{cases} x, & 1 < x < t_1 \\ t_1 - \frac{t_2 - t_1}{e^{x-1}}, & x > t_1, t_2 > t_1 \end{cases}$$

In this function, $x$ is a particular d before smooth, the output results is the value of this d after smooth. Due to $\text{dif} \geq \text{same}$, the value of function $f$, $g$, $h$ are all constant greater or equal to 1. It can be seen by the smooth function form that the value after smooth changes between $l$ and $t_1$, if the value before smooth is between $l$ and $t_1$, then the value invariant after smooth, if the value is greater than $t_1$, then decrease according to the index. Determine the proper $t_1$ and $t_2$ for each dimension data, thus it can achieve the purpose of smoothing differences between all kinds of P2P applications.
applications. Figure 8, 9, 10 is the good smooth results obtained by experiment, the 3 dimensional data $t_1$ and $t_2$ is respectively 10 and 20.

After smoothing, differences between two types of P2P data decreases obviously, and the difference with non-P2P data keep unchanged.

Smooth test inspection process as shown in figure 11, handing the training set to SVM to train after smoothing process, due to the smoothing reduces the differences between P2P data, it is expected to use faster training speed to get a accurate model. Test data set, of course, can be predicted after smoothing.

Experimental data set described in table 1, a total of 6 sets of data. The first set of data is "training data set", contains 28.2% BT flow, the other is non-P2P traffic, including web, FTP, DNS, mail, MSN and QQ. The second, third and fourth group includes 3 different P2P and some of the non-P2P traffic. The fifth flow collected from 10 computers, including the P2P traffic, and four kinds of P2P traffic. The last group is non pure P2P traffic; the purpose is strictly test the false alarm rate of the method.

Here we use "training data set" as the training sample. In smooth operation, the 3 d data $t_1$ values are all taking 10, $t_2$ values are all taking 20. Searching $(\lambda, \gamma)$’s optimal parameter $(2.0, 0.0078125)$ through the grid – search.

Table 2 compares the different experiment result with or without the smooth processing. It can be seen from the table that for the data without the smooth process, the false-alarm traffic is at least 292 connections of test set, then we can prove that the proposed scheme is an effective method of low false alarm and low omission.

V. CONCLUSIONS
Current users use the Internet mostly for the use of network television, P2P download, and other entertainment activities, the great use of such applications can lead to serious network bandwidth consumption, thus affect the network service quality. While P2P applications use strategies of dynamic port and encryption application layer data to escape conventional flow detection mechanisms, making troubles on traffic matching algorithm, the paper proposed mobile P2P network data flow control algorithm. Experimental simulation results show that the proposed SVM feature matching algorithm based mobile P2P network data flow control algorithm has a higher recognition rate and better scalability.

ACKNOWLEDGEMENTS

This work is supported by The Education Department of Anhui Province Natural Science Project (KJ2010B034), the Education Department of Anhui Province Natural Science Project (KJ2011A059).

REFERENCES


ZHANG Yuan is a lecturer in the School of Electronic and Information Engineering, Anhui Institute of Architecture. She received master’s degree in Computer science from Anhui University in 2003, She is a PHD student in the Department of Computer Science and Technology at Anhui University. Her research interests are in the area of computer Network, quotient space, subspac clustering algorithm.

SHEN Qing-wei is a professor in the School of Electronic and Information Engineering, Anhui Institute of Architecture. She received master's degree in Computer science from Hefei University of Technology in 2002 and was entitled Anhui Province outstanding young backbone teacher. Her research interests are in the area of new generation of network protocols, computer system architecture.