

新颖的网络域名用户关键角色识别方法

许小丰^{1,2}, 杨力³, 王巍^{1,2}

(1. 通信信息控制和安全技术重点实验室, 浙江 嘉兴 314033;
2. 中国电子科技集团公司第三十六研究所, 浙江 嘉兴 314033;
3. 西安电子科技大学网络与信息安全学院, 陕西 西安 710071)

摘要: 提出了一种新颖的网络角色识别算法。该算法首先利用 ISODATA 方法对网络事件用户行为进行聚类分析; 在此基础上, 改进混合式蚁群 (HBACA) 算法, 对网络事件中各用户积累行为进行持续分析, 从而识别该事件执行过程中各用户所起的作用。仿真结果表明, 该算法可以分清网络用户层次关系, 获得事件中起主导作用的网络用户。

关键词: 域名用户; 聚类分析; 行为积累; 蚁群算法

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00128

Novel role analysis method for network domain users

XU Xiao-feng^{1,2}, YANG Li³, WANG Wei^{1,2}

(1. National Laboratory of Information Control and Security Technology for Commutation System, Jiaxing 314033, China;
2. Electronics Technology Group Corporation No. 36 Institute, Jiaxing 314033, China;
3. School of Network and Information Security, Xi'an University of Electronic Science and Technology, Xi'an 710071, China)

Abstract: A novel role analysis method for network domain users was proposed. Firstly, the algorithm used ISO-DATA algorithm to cluster network events for user behavior analysis, on the basis of conditions, the method search relationships between varieties of grouping sets in the network instance, and identify each user role in the performing process of network instance. Finally, simulation results show that the effectiveness of the proposed algorithm can distinguish the network user hierarchy and obtain the dominant roles about network users.

Key words: domain user, clustering analysis, accumulation of behavior, ant-colony algorithm

1 引言

随着网络基础设施建设步伐的加快和网络应用的日益广泛以及用户数量的不断增多, 网络的连通性、复杂性和不可控性也急剧增加^[1]。利用网络进行不法活动存在一定隐蔽性, 因此网络中的犯罪行为越来越多地呈现出来。

目前, 在研究网络域名用户在网络行为中角

色问题的领域, 如分析恐怖主义、网络犯罪相关技术过程中, 文献[2]采用将相关人员 (角色) 可视化的方法进行显示, 形成角色之间的信息链条。文献[3]通过收集事件成员之间的相互关系, 从网络层面对恐怖分子的信息与社交进行抽取, 得到相关的证据和关联。文献[4]利用层次贝叶斯对立方法, 成功构建了犯罪组织之间的相互关系, 并设计开发出软件系统 NETST; 通过该系

收稿日期: 2016-11-01; 修回日期: 2017-01-12。通信作者: 许小丰, 554171641@qq.com

基金项目: 国家自然科学基金资助项目 (No.61671360)

Foundation Item: The National Natural Science Foundation of China (No.61671360)

统,可以准确地发现成员之间的结构关系,即角色分配。

国内在网络角色方面的研究处于起步阶段,文献[5~6]利用基因遗传算法对网络角色进行聚类分析,得到网络犯罪行为的各类角色,但是角色重要性排名方面有所欠缺。文献[7]利用网络开放性邮件挖掘的手段,开发出一套角色与犯罪分析系统 EMT。

2 用户角色识别算法

用户角色识别算法主要分为两步:1) 采用类似于 ISODATA 算法的聚类分析方法^[8,9],对用户行为进行聚类分析,通过设置用户各类行为的预期标准(即用户各类行为的一个预设值),利用 ISODATA 方法对无限相近的类似行为用户进行聚类;2) 在各种活动聚类分析完成的前提下,利用改进的混合式蚁群算法对单位时间内网络数据流进行分析,得到网络事件执行过程中各用户角色。

2.1 用户聚类分析

本节对聚类分析方法进行描述,包括基本思想、控制参数、初始条件、实现步骤等^[9],然后将网络中用户行为的事件类型、事件频率、事件信度、事件代价等根据具体情况进行综合,代入聚类方法中进行计算,实现用户行为类型分析。

具体描述如下。

2.1.1 ISODATA 基本思想

ISODATA (iterative self-organizing data analysis techniques algorithm)^[9]是迭代、自组织数据的分析策略与方法,主要适用于动态、行为多变客体的聚类分析,其基本思想如图1所示。

2.1.2 ISODATA 聚类实现

在具体实现过程中,设定算法最初的聚类数目为 d (网络用户数),初始状态的聚类中心个数可以设置为 $n_i (i=1,2,3,\dots,c)$,其中, c 为网络用户最终角色类型数目),步骤如下。

1) 设置控制参数

① K 表示最终希望得到的类型数目。即网络用户行为中,最终希望分成多少个类别,根据这些类别,再利用改进的混合蚁群方法进行各类角色的重要性排名。

② θ_N 表示聚类过程中,至少应该达到的样本数目。即一类用户角色至少需要多少个类似网络用户行为的支撑。

③ θ_s 表示设定的标准偏差。如果在分类过程中类与类之间的偏差大于该值,则网络用户角色需要分裂(一分为二),当成2个角色对网络事件进行考虑。

④ θ_c 表示聚类中心(网络角色)之间的最小

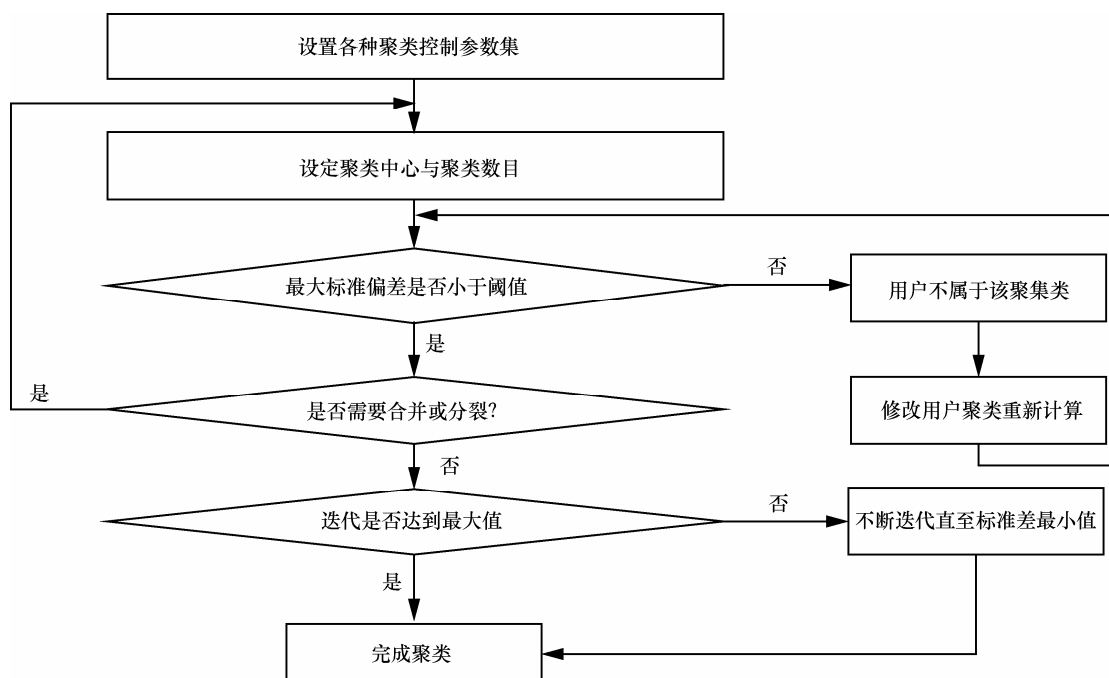


图1 ISODATA 聚类方法基本思想

差距,如果在聚类过程中小于该值,则2种网络角色可认为是一类角色,即角色间的合并。

⑤ N 表示聚类过程中允许迭代的最大次数,即网络角色识别的收敛性。

2) 若 $\|X - n_i\| < \|X - n_j\|, (j=1,2,\dots,c; i \neq j)$, 则 $X \in \omega_i$ 。将已知的所有样本(网络用户的具体行为,如单击某网页、邮箱、持续时间、具体IP地址等)分别对应放入到各个类别中。

3) 在所有已经存在的数据类别中,若有其基数(用户行为数) $N_i < \theta_N$, 则舍掉 ω_i , 并令聚类数目为 $c = c - 1$ 。

4) 根据关系更新均值向量为

$$\omega_i = \frac{1}{n} \sum_{X \in \omega_i} X (i=1,2,\dots,c)$$

5) 对 ω_i 中的所有样本(网络用户行为)计算其到相应聚类中心 n_i 间平均长度, 则 $\overline{D}_i$ 为

$$\overline{D}_i = \frac{1}{N_i} \sum_{X \in \omega_i} \|X - n_i\| (i=1,2,\dots,c)$$

6) 对整个样本空间中的所有样本(网络行为)距离其相应中心聚类点的平均值 $\overline{D}$ 进行计算, 为

$$\overline{D} = \frac{1}{N} \sum_{i=1}^c N_i \overline{D}_i$$

7) 对聚类类型(即各类网络角色类型)进行分类、合并、迭代等操作, 具体为:

①如果整个算法迭代数达到最大值 N , 设定 $\theta_c = 0$, 然后转到步骤 11);

②如果 $c \leq \frac{K}{2}$, 则转到下一步;

③如果 $c \geq 2K$, 则转到步骤 11)。

8) 对于每一个 ω_i , 其标准偏差的计算为

$$\begin{cases} \sigma_i = \sigma_{i1}, \sigma_{i2}, \dots, \sigma_{in} \\ \sigma_{ij} = \sqrt{\frac{1}{N_i} \sum_{X^k \in \omega_i} (X_j^k - m_{ij})^2} \end{cases} \quad (1)$$

其中, X_j^k 是第 k 个样本中的第 j 个分量 (k 类网络行为中的某个行为), m_{ij} 是第 i 个中心点的第 j 个分量 (角色 i 中的某个分量), σ_{ij} 是第 i 个标准计算偏差的第 j 个分量, n 是样本 X 的维度。

9) 对任意一类的聚类(一类网络用户角色)过程, 计算标准差的最大值 $\sigma_{i\max} (i=1,2,\dots,c)$ 。

10) 当 $\sigma_{i\max} > \theta_s$, 有 $\overline{D}_i > \overline{D}$ 且 $N_i > 2(\theta_N + 1)$ 或 $c \leq \frac{K}{2}$, 则将 ω_i 分类为 2 个聚类中心 (相当于 2 个角色) 进行聚类分析。

11) 对于所有的聚类中心点, 计算互相之间的距离, 为

$$D_{ij} = \|m_i - m_j\| (i=1,2,\dots,c-1; j=i+1, i+2, \dots, c)$$

12) 对步骤 11) 中聚类中心距离小于阈值的类别进行合并, 合并成一个类别 (即合成一个角色), 聚类结束。

2.2 改进混合蚁群算法

本文重点在于改进混合式蚁群算法实现网络事件中用户关键角色的分析, 本算法具体描述如下。

本文主要利用蚁群算法中的侦察蚁和搜索蚁进行分析。因此, 在网络数据流矢量表建立的基础上, 利用各种蚂蚁的不同特性对源用户到目标用户的数据流进行局部及全局分析, 得到节点上用户间关系^[10,11]。

在蚁群算法中, 各类型蚂蚁的职责不同: 1) 负责侦察工作的蚂蚁以每个节点为基点, 进行局部的查询与搜索, 并以其特有的侦察素来记录得到的信息, 以便为下一步进行搜索工作的蚂蚁到达该节点后选择下一个节点提供有用的信息; 2) 搜索蚁根据侦察蚁在各个节点上释放的信息素的多少以及方向进行全局搜索。假设一个节点对应一个用户, 具体做法如下。

2.2.1 侦察蚁

将 b 个节点与 b 只侦察蚁进行一一对应, 每一个侦察蚁以自身为出发点, 选择与自己直接相连的其他 $b-1$ 个节点进行侦察, 将已经知道的先验知识与自身寻找侦察到的相关信息进行结合, 形成侦察信息素, 记为 $Flow[i][j]$, 标记从蚂蚁 i (节点 i) 到蚂蚁 j (节点 j) 所通过的路径。其中, $Flow[i][j] (i, j=0,1,2,\dots,b-1; i \neq j)$ 表示为

$$Flow[i][j] = \begin{cases} \frac{\tilde{Inf}_{ij}}{Inf_{ij}}, & \text{节点 } j \text{ 在节点 } i \text{ 的 3 跳以内} \\ 0, & \text{其他} \end{cases} \quad (2)$$

$\tilde{Inf}_{ij}$ 表示以节点 i 为中心到其他 $b-1$ 个节点中最大信息流量。以 $Flow[i][j]$ 的值为基础, 开始

时刻各路径上假设信息量为

$$\tau_{ij}(0) = \begin{cases} CFlow[i][j], Flow[i][j] \neq 0 \\ \frac{C \tilde{Inf}_{ij}}{\hat{Inf}_{ij}}, \text{其他} \end{cases} \quad (3)$$

$\hat{Inf}_{ij}$ 与 $\tilde{Inf}_{ij}$ 的含义相反, 表示路径中所得到的最小信息量; C 设定为开始时刻, 各个蚂蚁之间寻路过程中的信息量。

2.2.2 搜索蚁

假设在 t 时刻, 蚂蚁 $k(k=1,2,\dots,n)$ 从当前所在的节点 i 到下一步希望到达节点 j 的概率 p_{ij}^k , 可表示为

$$p_{ij}^k = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha [\eta_{ij}(t)]^\beta}{\sum_{Flow \notin tabu_k} [\tau_{is}(t)]^\alpha [\eta_{is}(t)]^\beta}, j \notin tabu_k \text{ 且 } Flow[i][j] \neq 0 \\ 0, \text{其他} \end{cases} \quad (4)$$

所有蚂蚁完成一次循环, 各路径上信息量调整为

$$\tau_{ij}(t+1) = \begin{cases} \rho \tau_{ij}(t) + (1-\rho) \Delta \tau_{ij}, Flow[i][j] \neq 0 \\ \rho \tau_{ij}(t), \text{其他} \end{cases} \quad (5)$$

其中, $\Delta \tau_{ij}$ 设定为本轮过程中, 蚂蚁从节点 i 到节点 j 之间一共付出的信息素总和; 信息素的持久性 ρ 与数据流经过的路径及时间间隔有关, 且与时间间隔成指数关系, 故 ρ 为

$$\rho = \begin{cases} \left[\frac{D_Tran_{ij}^k(t) + D_Disp_{ij}^k(t)}{\sum_{k=1}^n (D_Tran_{ij}^k(t) + D_Disp_{ij}^k(t))} \right]^\varepsilon (1-\gamma)^\delta \varphi, \\ 1 < \delta \leq 6 \\ 0, \delta > 6 \end{cases} \quad (6)$$

其中, $D_Tran_{ij}^k(t)$ 、 $D_Disp_{ij}^k(t)$ 分别表示蚂蚁 k 从节点 i 到节点 j 的数据传输、处理时延, γ 为信息调整常量, δ 表示数据离目标节点的跳数, ε 、 φ 分别为时延与传输距离的调整系数, 其值由节点单位时间 t 内所通过数据分组的信息量决定, 即

$$\varepsilon = \frac{T_{ij}^{data}(t)}{R_j^{data}(t) + T_{ij}^{data}(t)} \quad (7)$$

$$\varphi = \frac{\frac{S_{ij}(t)}{S_{ij}(t) + F_{ij}(t)}}{\frac{S_{ij}(t-1)}{S_{ij}(t-1) + F_{ij}(t-1)}} \quad (8)$$

$T_{ij}^{data}(t)$ 为需要传输的数据分组个数, $R_j^{data}(t)$

为节点 i 未收到 ACK 回复需要重传数据分组的个数; $S_{ij}(t)$ 为数据流在节点 i 到节点 j 路径上经过的次数, $F_{ij}(t)$ 为数据流未在节点 i 到节点 j 路径上经过的次数。

另外, $\Delta \tau_{ij}(t) = \sum_{k=1}^n \Delta \tau_{ij}^k(t)$, $\Delta \tau_{ij}^k(t)$ 为蚂蚁 k

本轮从节点 i 到节点 j 之间一共付出的信息素总和, 其值为

$$\Delta \tau_{ij}^k = \begin{cases} \frac{Q \left(\frac{\tilde{Inf}_{ij}}{\hat{Inf}_{ij}} \right)}{L_k}, \text{蚂蚁 } k \text{ 经过 } (i, j) \text{ 且 } Flow[i][j] \neq 0 \\ 0, \text{其他} \end{cases} \quad (9)$$

根据每只搜索蚁得到的最优解, 即从蚂蚁出发到最终节点信息素最多的路线, 结合数据流特性, 可得到节点 i 的直接上游节点 (即某用户的直接上层领导), 从而得到各用户在周期 t 内对某网络事件所起的作用。伪流程描述如下。

While (the time (T) of system and nodes are working){

Put m ants into m nodes

if (the distance of node i to $j \leq 3$) {

$$Flow[i][j] = \frac{\tilde{Inf}_{ij}}{\hat{Inf}_{ij}};$$

if (the distance of node i to $j = 1$) {

if (The algorithm is beginning) {
if ($Flow[i][j] \neq 0$)

$$\tau_{ij}(0) = CFlow[i][j]; p_{ij}^k = \frac{[\tau_{ij}(t)]^\alpha [\eta_{ij}(t)]^\beta}{\sum_{Flow \notin tabu_k} [\tau_{is}(t)]^\alpha [\eta_{is}(t)]^\beta};$$

$$\text{else } \tau_{ij}(0) = \frac{C \tilde{Inf}_{ij}}{\hat{Inf}_{ij}}; \quad p_{ij}^k = 0;$$

}//end if

}//end if

```

else{

$$\varepsilon = \frac{L_{ij}^{\text{data}}(t)}{C_j^{\text{data}}(t)};$$


$$\varphi = \frac{\frac{S_{ij}(t)}{S_{ij}(t) + F_{ij}(t)}}{S_{ij}(t-1)};$$


$$\varphi = \frac{\frac{S_{ij}(t)}{S_{ij}(t) + F_{ij}(t)}}{S_{ij}(t-1) + F_{ij}(t-1)};$$

if ( $1 < \delta \leq 6$ )

$$\rho = \left[ \frac{D\_Tran_{ij}^k(t) + D\_Disp_{ij}^k(t)}{\sum_{k=1}^n (D\_Tran_{ij}^k(t) + D\_Disp_{ij}^k(t))} \right]^\varepsilon (1-\gamma)^\delta \varphi;$$

else

$$\rho = 0;$$

if ( $Flow[i][j] \neq 0$ )

$$\Delta \tau_{ij}^k = \frac{Q \left( \frac{C \tilde{Inf}_{ij}}{Inf_{ij}} \right)}{L_k};$$


$$\tau_{ij}(t+1) = \rho \tau_{ij}(t) + (1-\rho) \Delta \tau_{ij};$$

else

$$\Delta \tau_{ij}^k = 0; \tau_{ij}(t+1) = \rho \tau_{ij}(t);$$

} //end else
} //end while

```

3 仿真分析

在仿真实验中，每个蚂蚁代表网络事件中的一个用户节点，假定用户节点（蚂蚁）在三维空间上随机分布，信息素调整常数 $\gamma=1$ ；转移概率、信息素挥发系数等在程序运行过程中自动调整。实验结果如图2和图3所示。

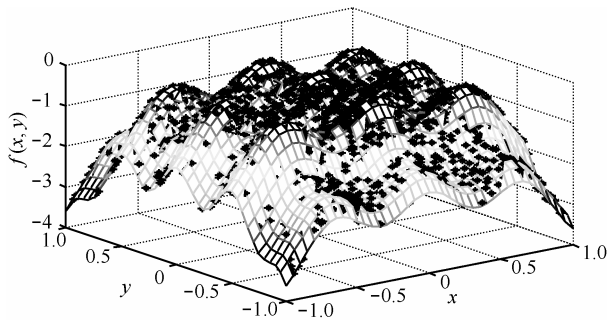


图2 用户节点初始分布

经过本文算法的改进，从图3可以明显看出最重要用户的位置坐标，且各个波峰用户的重要程度非常清晰，网络事件的关键节点极易提取。这是因为本文算法对信息素的持久性、节点转移概率等关键参数进行了动态化处理，并将网络容量、传输时延、调整系数等因素加入其中，更加能够反映出节点的重要程度。

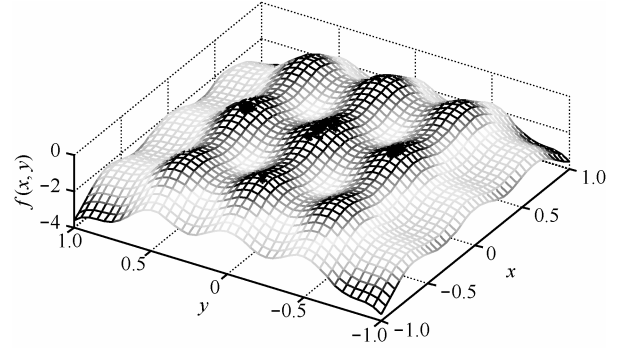


图3 经过本文算法后，用户节点（蚂蚁）的最终分布

4 结束语

网络域名用户关键节点的识别对打击犯罪网络及通信网络拓扑推断具有极其重要的意义，本文提出了基于行为累积的网络关键角色识别方法。主要工作有：1)采用K均值算法对用户行为进行聚类分析；2)在各种活动聚类分析完成的前提下，利用改进的HBACA算法分析单位时间内局部及全局网络数据流向，利用侦察蚁和搜索蚁特征得到网络事件执行过程中各用户角色。实验结果表明，该算法可以提取网络成员个性特征与通信行为之间的内在联系，获取特定网络事件中作为关键角色的节点。

参考文献:

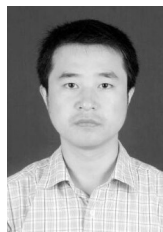
- [1] 乔少杰, 唐长杰, 彭京. 基于个性特征仿真邮件分析系统挖掘犯罪网络核心[J]. 计算机学报, 2008, 31(10): 1795-1803.
QIAO S J, TANG C J, PENG J. The core of the crime network based on the personalized characteristic simulation mail analysis system[J]. Journal of Computer Science, 2008, 31(10): 1795-1803.
- [2] XU J J, CHEN H C. CrimeNet explorer: a framework for criminal network knowledge discovery[J]. ACM Transactions on Information Systems, 2005, 23(2): 202-226.
- [3] KREBS V E. Mapping networks of terrorist cells [J]. Connections, 2002, 24(3): 43-52.
- [4] DOMBROSKI M J, CARLEY K. Estimating a terrorist network's

- structure[J]. Computational and Mathematics Organization Theory, 2002, 8(3): 235-241.
- [5] QIAO S J, TANG C J, PENG J. VCCM Mining: mining virtual community core members based on gene expression programming[C]//The Workshop on Intelligence and Security Informatics. 2006:133-138.
- [6] QIAO S J, TANG C J, YU Z H. Bim mining virtual community structure based on SVM[J]. Computer Science (Supplement A), 2005, 32(7): 208-212.
- [7] STOLFO S J, HERSHKOP S, WANG K, et al. A behavior-based approach to securing email systems[C]//The Second International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security. 2003:57-81.
- [8] 杨博, 刘大有, 金弟, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66.
- YANG B, LIU D Y, JIN D, et al. Complex network clustering method[J]. Journal of Software, 2009, 20(1): 54-66.
- [9] 许国根, 贾瑛. 模式识别与智能计算的 Matlab 实现第 2 版[M]. 北京: 北京航空航天大学出版社, 2014: 22-27.
- XU G G, JIA Y. Pattern recognition and intelligent computing based on Matlab. The second Edition[M]. Beijing: Beijing University of Aeronautics and Astronautics Press. 2014:22-27.
- [10] VAN D E AALST P, WEIJTERS A. Process mining a research agenda[J]. Computer in Industry, 2004, 53(3): 231-244.
- [11] GALBRAITH J R. Organization design: an information processing view[J]. Interfaces, 1974, 4(3):28-36.

作者简介:



许小丰 (1980-), 男, 河南济源人, 博士, 中国电子科技集团公司第三十六研究所高级工程师, 主要研究方向为无线/有线网络安全技术。



杨力 (1977-), 男, 陕西咸阳人, 博士, 西安电子科技大学副教授, 主要研究方向为网络安全、可信计算、网络溯源。



王巍 (1980-), 男, 河北张家口人, 博士, 中国电子科技集团公司第三十六研究所高级工程师, 主要研究方向为无线/有线网络安全技术。