第 35 卷 第 3 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.3 Computer Engineering February 2009
·安全技术· 文章编号：1000—3428(2009)03—0173—03 文献标识码：A 中图分类号：TP393.08

P2P 环境下的蠕虫检测算法

王秀英[1,2]，邵志清[1]，刘百祥[1]
(1. 华东理工大学信息科学与工程学院，上海 200237；2. 上海新侨职业技术学院计算机信息系，上海 200237)

摘 要： P2P 下载与网络蠕虫具有相似的搜索机制，导致网络蠕虫难以被检测并定位。该文提出一种融合危险理论和 ID3 分类算法的检测算法 D-ID3。利用熵理论分析 P2P 应用、蠕虫、正常主机的属性特征，得到轴属性。利用 ID3 分类算法得到可以区分蠕虫、P2P 和正常流量的分类规则。实验结果表明，该算法能成功检测出网络蠕虫，其误警率较低。

关键词： 网络蠕虫；ID3 算法；危险理论；P2P 流量；熵

Worm Detection Algorithm Under P2P Circumstances

WANG Xiu-ying[1,2], SHAO Zhi-qing[1], LIU Bai-xiang[1]
(1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237;
2. Department of Computer and Information, Shanghai Xinqiao Vocational and Technical College, Shanghai 200237)

【Abstract】 It is difficult to detect Internet worm in LAN, because of the similarity of probing mechanism between the P2P and internet worm. This paper proposes a detection algorithm names D-ID3 that is based on danger theory and ID3 classification algorithm. The entropy theory is used to analyz P2P application, worm, natural mainframe, and extract axis attributes. ID3 and danger theory are applied to get classification rules that can differentiate worm, P2P and natural traffic. Experimental results show that this algorithm can detect worm and P2P successfully with a low false alarm rate.

【Key words】 Internet worm; ID3 algorithm; danger theory; P2P traffic; entropy

1 概述

近年来，网络蠕虫对计算机系统和网络安全的威胁日益增加。以 CodeRed, Blaster, Slammer 等为代表的主动探测蠕虫和以 Melissa, LoveLetter, MyDoom 等为代表的 E-mail 蠕虫流行时间长、覆盖面积广，对信息系统造成了巨大危害。

目前对网络蠕虫的研究主要集中在 3 方面：模型，检测和防治。文献[1]分析 Code-Red 蠕虫的传播机制并提出蠕虫传播模型。很多网络蠕虫检测和防治方法基于对网络流量异常行为(如网络中突然增加的失败连接数[2]，网络蠕虫连接超出地址范围的目的地址[3])的分析。但 P2P 技术在校园网内的广泛应用给蠕虫的检测和定位带来了困难。由于多数 P2P 应用在搜索机制上采用洪泛法，会产生大量无效连接，因此难以区分其行为与蠕虫扫描。鉴于此，本文提出一种适用于复杂网络环境的 D-ID3 检测算法。此算法通过分析数据包中各属性熵值的分布，得到轴属性[4]。

2 D-ID3 算法

蠕虫扫描具有以下 2 个特征：(1)会产生大量无效 IP 地址和无效连接请求；(2)能达到每秒几十次~几百次的扫描速率。因为 P2P 流量具有类似特点，所以要检测混有 P2P 的局域网上蠕虫的传播，就要分析数据包的会话信息。数据包属性共 6 个：源 IP 地址 *srcAddr*，目的 IP 地址 *dstAddr*，源端口 *srcPort*，目的端口 *dstPort*，上层协议 *Protocol*，数据报长度 *len*。

2.1 属性选择

属性选择能在不失去数据原有价值的基础上选择最小的属性子集，并去除不相关的属性和冗余属性。通过观察 P2P、蠕虫和正常数据流的上述 6 个属性，可以发现它们的属性取值特征是不同的。本文通过熵理论确定蠕虫检测中所使用数据集的轴属性[4]。

熵被用来描述数据集合的不确定性，文献[5]给出了熵的定义，即

$$H(X)=-\sum_{i=1}^{n}p_i \mathrm{lb}(p_i) \quad (1)$$

其中，X 为会话的属性，$X\in\{srcAddr, srcPort, dstAddr, dstPort, Protocol, len\}$；$H(X)$为属性 X 的熵；p_i 是 X 中第 i 个取值的概率。

定义 分别考察 P2P、网络蠕虫、正常流量数据集各属性的熵值，如果某 2 个数据集的某个属性的熵值相近，而与另一个数据集的该属性的熵值差异较大，则将该属性定义为轴属性。

根据定义选择出的轴属性可以降低算法复杂度，去除无用属性，从而提高检测率。

2.2 ID3 算法

通过选用分类方法区分蠕虫、P2P 和正常流量。分类的目的是构造一个分类函数或分类模型(分类器)，该模型能把数据库中的数据项映射到某个给定类别。即给定数据库 $D=\{t_1,t_2,\cdots,t_n\}$，元组 $t_i\subseteq D$，类的集合 $C=\{C_1,C_2,\cdots,C_m\}$，分类问题定义为从数据库到类集合的映射 $f: D\rightarrow C$，即数据库中的元

基金项目： 上海高校选拔培养优秀青年教师科研专项基金资助项目
作者简介： 王秀英(1971－)，女，讲师、博士研究生，主研方向：网络异常行为检测；邵志清，教授、博士、博士生导师；刘百祥，副教授、在职博士研究生
收稿日期： 2008-10-21 **E-mail：** xywang71@163.com

组 t_i 分配到某个类 C_j 中，则有

$$C_j = \{t_i \mid f(t_i) = C_j,\ 1 \leqslant i \leqslant n, \text{且} t_i \in D\}$$

根据定义的属性选择结果，分类时采用如下 2 个步骤：

(1)模型训练阶段。段根据给定的训练集，找到合适的映像函数 f: $f(t) \rightarrow C$ 的表示模型。

(2)使用步骤(1)完成的函数模型预测数据类别，或利用该函数模型扫描数据集中的每类数据，从而形成分类规则。

决策树方法是一种较通用的分类函数逼近方法，它基于贪心算法，采用自顶向下的递归方式构造决策树。Quinlan 于 1979 年提出著名的 ID3 方法，ID3 算法采用基于信息熵定义的信息增益度量来选择内节点的测试属性。

信息增益表示给定的属性 X 对于分类 Y 的不确定性的减少，用 $I(Y; X)$表示。分类 Y 的不确定性用熵值表示，即 $H(Y)$。分类 Y 在给定属性 X 时，其条件熵表示为 $H(Y|X)$。属性 X 的信息增益(互信息)描述如下：

当属性 X 和分类 Y 的取值都为离散值，即 Y 的取值范围为$\{y_1, y_2, \cdots, y_k\}$时，$X$ 的取值范围为$\{x_1, x_2, \cdots, x_l\}$，则 Y 的熵表示为

$$H(Y) = -\sum_{i=1}^{i=k} P(Y = y_i)\mathrm{lb}(P(Y = y_i))$$

Y 关于 X 的条件熵表示为

$$H(Y \mid X) = -\sum_{j=1}^{j=l} P(X = x_j)H(Y \mid X = x_j)$$

信息增益 $I(Y; X)$是指由于知道属性 X 的值后，使得熵的不确定性较少。$I(Y;\ X)$越大，说明选择测试属性 X 对分类提供的信息越多。Quinlan 的 ID3 算法在每个节点上选择信息增益 $I(Y; X)$最大的属性作为测试属性。

实验中选用的属性值是连续的，而 ID3 算法要求属性值是离散的，为了计算信息增益，需要对连续属性进行二值化处理。选择合适的阈值θ，将属性 X 二值化，则信息熵的计算公式转换为

$$I(Y; X) = \arg\max X_\theta I(Y, X_\theta)$$

通过信息增益计算可以得到分类规则。

2.3 危险理论

利用 ID3 算法可以达到将数据集初步分类的目的，但还需要进一步考虑如何降低分类误检率。

Burnet 提出的 SNS(Self/Non-Self)理论建立了传统的免疫系统学，该系统的核心功能是区分“自我”与“非我”。免疫细胞通过“阴性选择”的检查过程，使免疫细胞只对“非我”成分的抗原做出免疫应答，对“自我”成分形成免疫耐受，不产生免疫应答。因此，“自我-非我”识别是人工免疫系统的关键。上述理论不能解释一些现象，例如肠胃每天都接触大量细菌，虽然这些细菌都不会被规类为“自我”，但不会引起免疫响应。危险模式理论的主要创导者 Matzinger 在文献[6]中认为，免疫系统不能区分 SNS，而只能区分危险信号。

危险模式和 SNS 模式的根本区别是免疫应答的触发信号不同。SNS 模式认为免疫系统只区分 SNS。对异己抗原产生的外源性信号产生应答，危险模式认为免疫系统只区分危险信号，对损伤细胞产生的内源性信号，即危险信号产生应答。

由于网络蠕虫和 P2P 流量占用大量网络带宽，可能引起网络阻塞，而影响其他用户对网络的正常使用，因此将一定时间窗口内超过一定阈值的网络流量定义为危险信号。只有产生危险信号时，系统才会启动检测模块。从而实现降低误检率、提高检测效率的目的。

2.4 D-ID3 算法

基于危险理论的 ID3 算法，即 D-ID3 算法流程如图 1 所示。图 1(a)描述了训练阶段，研究的数据对象为训练数据集，具体过程如下：(1)根据网络异常用户(包括 P2P 下载、网络蠕虫)每秒发送的数据包量远大于正常用户所发送数据包量的特性得到危险信号；(2)根据 ID3 算法得出分类规则。图 1(b)描述了测试阶段，研究的数据为测试数据集。在一定时间窗口内，如果网络流量超过危险信号的阈值，则根据图 1(a)得到的分类规则对其进行分类，否则归类为正常网络行为。

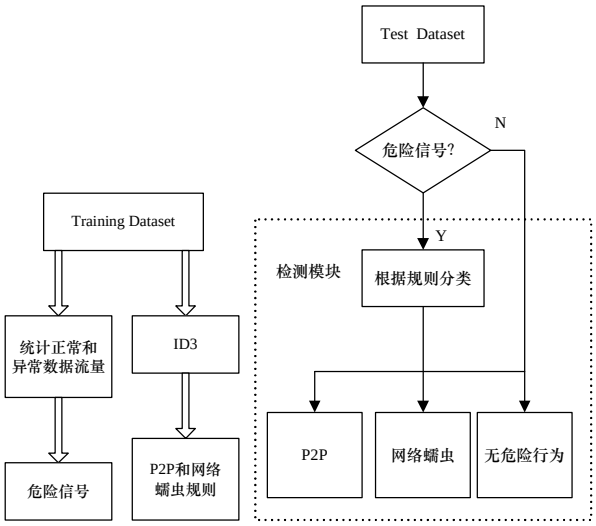


图 1　D-ID3 算法流程

3 实验

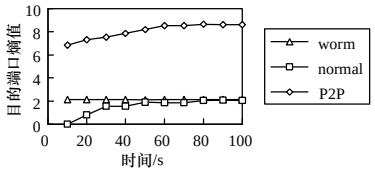
3.1 实验数据描述

实验数据从带有防火墙的某校园网上采集。抽取的蠕虫样本为一个感染了 Sasser 蠕虫的主机流量和一个感染 Blaster 蠕虫的主机流量；抽取的 P2P 样本为一个用 BT 下载的主机，一个为使用 PPStream 的主机；抽取 4 个正常使用网络的主机作为训练集数据。测试集的数据来自校园网的实时流量。P2P 下载和网络蠕虫进行主机探测时主要利用 ICMP 的 Ping 包、TCP SYN, FIN, RST 和 ACK 包。网络中 ICMP 的流量较少，因此，实验中只考虑带有 TCP SYN 标志和 UDP 的数据包，且数据包的源 IP 地址为内网 IP。测试集的数据为 100 s 内的实际数据流，经处理后作为测试集。

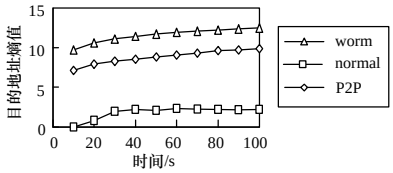
3.2 ID3 算法

3.2.1 属性选择

根据式(1)考察训练集数据 6 个属性的熵值，即 *srcAddr*, *dstAddr*, *srcPort*, *dstPort*, *Protocol* 和 *len*，考察在 10 s, 20 s,…, 100 s 内属性的熵值，实验结果见图 2。由图 2(a)可以看出，网络蠕虫和正常流量的目的端口的熵值远小于 P2P 流量目的端口的熵值，这是因为 P2P 下载需要访问大量目的端口(尤其是大于 1023 的端口)，正常的数据流量访问的目的端口大多为常用端口如 53, 80, 21, 25 等，蠕虫感染的主机访问的目的端口除了 53 端口外，还有蠕虫特定的目的端口(不同蠕虫访问的特定目的端口不同，如 Blaster 探测 135, 139, 445, 593 端口)。由图 2(b)可以看出，网络蠕虫和 P2P 流量的目的地址的熵值大于正常流量目的地址的熵值，这符合蠕虫和 P2P 下载的扫描特性。根据定义，最终选择的轴属性为 *dstAddr* 和 *dstPort*。



(a)3 种流量目的端口的熵值对比



(b)3 种流量目的地址的熵值对比

图 2　3 种流量样本的熵值对比

3.2.2　基于 ID3 算法的分类

用 3.2.1 节得到的轴属性对网络流量进行分类。测试集数据采用网络中 100 s 内的数据流量。分别以 2 s, 5 s 和 10 s 作为时间窗口，根据 ID3 算法进行分类，分类结果如图 3 所示。图 3 表明，当选择的时间窗口为 5 s 以上时，网络蠕虫的检测率可达 100%，P2P 流量的检测随时间窗口的增大而检测逐步上升。为了得到较好的实时检测效果，选择 5 s 的检测窗口，以得到较高的检测率，并较好地满足检测的实时性要求。

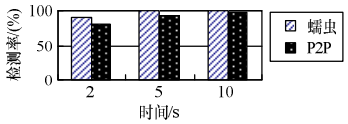


图 3　不同时间窗口下的分类结果

通过选用合适的时间窗口，ID3 算法可以得到较高检测率，但存在较高误检率。造成网络蠕虫误检的来源有 2 个：(1)P2P 视频软件产生的流量;(2)网络上微视频的流量(微视频泛指 30 s~20 min 的视频短片)。

3.2.3　融合危险理论的 ID3 算法

为了降低误检率，需要通过实验得到危险信号的阈值。本文实验选用 3 个感染网络蠕虫的主机、3 个进行 P2P 下载的主机和 78 个正常主机的 TCP SYN 数据包作为训练数据。图 4 考察了不同类型主机每 5 s 发起的连接数。由图 4 可知，感染了网络蠕虫或进行 P2P 下载的主机，其发起的连接数总是高于正常用户的网络连接数。根据危险理论，将超过一定阈值的网络连接行为定义为危险信号(危险信号的阈值确定方法参照 3.2.2 节数据预处理中的连续属性二值化方法)。只有超过危险信号阈值的网络流量才进行如 3.2.3 节所述的分类行为。由表 1 可知，加入危险信号后，在检测率不变的情况下，误检率有所降低。加入危险信号可以提高检测效率，其原因是本文只检测连接数超过阈值的流量，而网络中多数用户的连接数没有超过阈值，因此，提高了系统实时性。

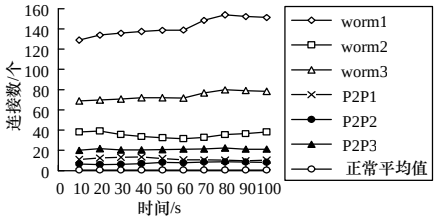


图 4　不同类型主机每 5 s 发起的连接数

表 1　误检率对照　(%)

算法	实验		
	第 1 组	第 2 组	第 3 组
D-ID3	7.69	6.25	7.22
ID3	11.76	12.37	15.41

4　结束语

本文分析网络蠕虫和 P2P 下载的网络流量特征，提出 D-ID3 算法，并将其应用到实际问题中。该算法能正确检测出网络蠕虫和 P2P 流量并降低误检率。本文只分析了蠕虫和 P2P 流量的网络层和运输层特征，下一步工作可以结合应用层的某些特征进行蠕虫检测。

参考文献

[1] Zou Changchun, Gong Weibo, Towsley D. Code Red Worm Propagation Modeling and Analysis[C]//Proceedings of the 9th ACM Conference on Computer and Communications Securit. New York, USA: ACM Press, 2002: 138-147.

[2] Venkataraman S, Song D, Gibbons P, et al. New Streaming Algorithms for Fast Detection of Superspreaders[EB/OL]. (2005-02-04). http://www.isoc.org/isoc/conferences/ndss/05/proceedings/papers/superspreader.pdf.

[3] Bailey M, Cooke E, Jahanian F, et al. The Internet Motion Sensor: A Distributed Blackhole Monitoring System[EB/OL]. (2005-02-04). http://www.csd.uoc.gr/~gvasil/stuff/papers/ims-ndss05.pdf.

[4] Lee W, Stolfo S. Data Mining Approaches for Intrusion Detection[EB/OL]. (1998-10-12). http://www.usenix.org/publications/library/proceedings/sec98/full_papers/lee/lee.pdf.

[5] Gray R M. Entropy and Information Theory[M]. New York, USA: Springer Verlag, 1990.

[6] Matzinger P. The Danger Model in Its Historical Context[J]. Scandinavian Journal of Immunology, 2001, 54(1/2): 4-9.

（上接第 160 页）

参考文献

[1] Parekh A K, Gallager R G. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Network: The Single-node Case[J]. IEEE/ACM Transactions on Networking, 1993, 1(3): 344-357.

[2] Katavenis M, Sidiropoulos S, Courcoubetis C. Weighted Round-robin Cell Multiplexing in a General-prupose ATM Switch Chip[J]. IEEE Journal on Selected Areas in Communication, 1991, 9(8): 1265-1279.

[3] Shreedhar M, Varghese G. Efficient Fair Queuing Using Deficit Round Robin[J]. IEEE/ACM Transactions on Networking, 1996, 4(3): 255-259.