第 38 卷 第 8 期
Vol.38 No.8

计 算 机 工 程
Computer Engineering

2012 年 4 月
April 2012

• 网络与通信 •

文章编号：1000—3428(2012)08—0073—03 文献标识码：A 中图分类号：TP393

校园网络流量自相似性分析与研究

张 浩[a]，吴 敏[b]

(东华大学 a. 计算机科学与技术学院；b. 信息办，上海 201620)

摘 要：对于校园网等小规模的局域网，通过计算网络流量自相似值的方法无法有效检测网络异常流量。针对该问题，在分析校园网络流量特点的基础上，将网络流量分解成趋势项和随机成分等其他项，使用经验模式分解消除网络流量中的趋势项，使得网络流量序列的自相似值能直接反映随机成分状态。实验结果表明，该方法能提高异常流量检测的准确性。

关键词：自相似性；希尔伯特黄变换方法；经验模式分解；网络流量；Hurst 值

Analysis and Research on Self-similarity of Campus Network Traffic

ZHANG Hao[a], WU Min[b]

(a. Institute of Computer Science and Technology; b. Information Office, Donghua University, Shanghai 201620, China)

【Abstract】For the campus network and other small-scale local area network, the abnormal traffic can not be detected rightly by calculating self-similarity value directly. Based on analyzing the characteristics of the campus network traffic, it decomposes traffic into trend and random components and other items, uses Empirical Mode Decomposition(EMD) to eliminate traffic trend items, and makes the self-similarity value direct response to the state of random items. Experimental results show that this method can improve the accuracy of abnormal traffic detection.

【Key words】self-similarity; Hilbert-Huang Transform(HHT) method; Empirical Mode Decomposition(EMD); network traffic; Hurst value

DOI: 10.3969/j.issn.1000-3428.2012.08.024

1 概述

文献[1]指出网络业务符合自相似特性，大量的研究表明正常的网络流量在大时间尺度上具有自相似性，网络攻击或者滥用造成的流量突发会对这种自相似特征产生明显的影响。因此，通过对网络流量的自相似值的计算，可以发现网络可能存在的异常现象。

文献[2]使用自相似理论分离分布式拒绝服务(Distributed Denial of Service, DDoS)流量和正常流量，提出一种有效检测 DDoS 攻击流量的混沌模型。文献[3]基于不同情形中自相似函数的小波变换系数的关系，提出一种能定位异常信息的异常检测方法。文献[4-7]证明基于网络流量的自相似特性能够在网络异常检测上有很好的应用意义。这些研究都是基于网络结构比较复杂、网络应用范围比较广的情况，对于校园网络这种局域网，有它特定的周期和规律，白天流量明显大于夜晚流量，流量分布不随机，在不同时间尺度上的自相似性不明显。如果直接使用常规的异常检测方法，则会产生一定的误差。网络流量作为一种非常复杂的非线性时间序列，其中存在表示流量序列长期行为规律的趋势项。本文对原始流量中的趋势项进行消除，通过计算流量中随机成分[8]等其他成分的 Hurst 值发现网络中的异常。本文通过对原始流量信息进行经验模态分解(Empirical Mode Decomposition, EMD)算法处理，消除趋势项。

2 自相似

2.1 定义

自相似性或者称为分形型是复杂系统中常见的一种现象。直观上看，具有自相似特性的系统会在某种测度的不同尺度下表现出相似的特性，这种测度可以是空间、时间或者任何其他合理的定义。

自 Mandelbrot B 最初提出分形的概念以来，分形理论已被逐渐应用到许多科学领域。在计算机网络领域，之前一直是沿用应用在传统电话网的泊松分布模型，但到了 20 世纪 90 年代，由于 WWW 等网络应用的出现及网络用户的激增，Leland W E 等人发现数据通信量可以用自相似过程进行描述，对网络自相似的研究逐渐应用到异常检测、流量预测等领域中。

2.2 自相似值估计方法

网络流量业务量的自相似性由一个 Hurst 参数(H)刻画，H 取值在(0.5,1)之间时，过程具有自相似性，并且 Hurst 参数值越大，表示自相似程度越高。估计 Hurst 系数的方法很多，可以分为时域和频域 2 类，常用的时域方法包括方差时间法(VT)、绝对值法(Abs)、留数法(Res)和 R/S 法。频域方法有周期图法、Whittle 法和小波法[9]。鉴于 R/S 方法的实现简单和稳定性，本文的自相似值使用 R/S 方法计算。

下面简要介绍 R/S 方法的计算原理：

对于时间序列 X、部分和 $Y(n)$ 及样本方差 $S^2(n)$，R/S 方法统计量为：

$$\frac{R}{S}(n)=\frac{1}{S(n)}\left[\max_{0\leqslant t\leqslant n}\left(Y(t)-\frac{t}{n}Y(n)\right)-\min_{0\leqslant t\leqslant n}\left(Y(t)-\frac{t}{n}Y(n)\right)\right] \quad (1)$$

对于分形高斯噪声(FGN)有：

作者简介：张 浩(1987－)，男，硕士研究生，主研方向：网络安全；吴 敏，副研究员

收稿日期：2011-06-10 **E-mail：**zhhao23@foxmail.com

$$E\left[\frac{R}{S}(n)\right] \propto C_H n^H \tag{2}$$

3 网络流量自相似性分析

本文以某高校校园网流量为对象，分析网络流量的自相似性。网络拓扑结构见图 1，为了解整个校园网的网络状态，收集网络拓扑结构中所有层级中的网络流量，然后进行分析。总共收集 4 个节点的流量信息，节点 1 为整个校区的网络出口处，是整个校区的流量总和，具有最高汇聚度；节点 2 为汇聚层中一个子网络结构的总流量，该子网络节点的总流量是 3 个子网络中流量最小的；节点 3 为汇聚层中一个子网络结构的总流量，该子网络节点的流量是 3 个子网络中流量最大的；节点 4 为最下层路由器连接的交换机的流量总和。

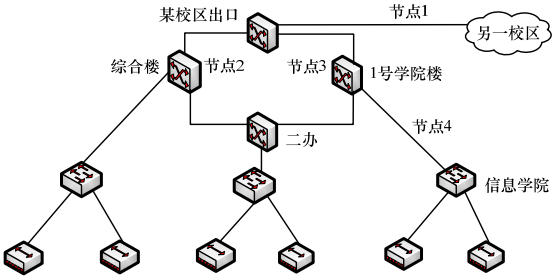

图 1 某高校网络拓扑结构

网络的自相似性的分析只需要一个基于时间序列的流量统计信息，不需要网络流量中的数据包，使用网络管理中普遍应用的简单网络管理协议(Simple Network Management Protocol, SNMP)协议采集交换机中对应端口的的 4 个节点的流量信息。采集日期为 2011 年 4 月 11 日-4 月 17 日，周期为 10 s，数据通过 RRDTOOL 存储，文件存储格式为 RRD，RRDTOOL[10]是一个开源的轮询数据存储和展现工具，其在数据存储和数据展示上有很强的能力。

对 4 个流量数据进行大时间尺度的自相似度计算，因为采集周期为 10 s，如果计算整个星期的自相似值，数据量比较大，因此采用增大时间间隔的方法，以 5 min 为单位，计算自相似值。使用 R/S 方法对 4 个节点的 4 组数据进行 Hurst 值计算，计算结果见表 1。

表 1 不同网络节点自相似值的计算结果

节点号	流量方向	Hurst 值
1	In	0.558 5
1	Out	0.647 0
2	In	0.570 1
2	Out	0.690 3
3	In	0.635 2
3	Out	0.687 3
4	In	0.661 3
4	Out	0.679 1

由表 1 的结果可以看出，4 个节点的自相似值介于 0.5~0.7 之间，所以，节点流量都符合自相似性，说明在一个星期的尺度上，校园网的网络状况比较正常。对于每个节点，上行流量的自相似值大于下行流量的自相似值。说明上行流量的网络状况偏离正常值更远，上行流量的异常性更强。基于校园网络的网络应用情况，推测也许因为校园网络中的上行流量主要还是以 P2P 为主，其他的正常网络应用都属于正常网络动作，不会对网络状况产生很大的影响。

表 1 说明了网络中不同拓扑结构中的节点在大时间尺度上都符合自相似性，如果要了解每个节点的详细网络状况，必须对更小的时间间隔的流量数据进行分析。节点 1 是整个网络的流量出口，其汇聚度最高，它的流量状态能够基本反映整个网络的状态，所以，本文通过提取节点 1 中 RRD 文件中保存的流量信息，以 10 s 为单位，5 min 为一个时间段计算一天中不同时间点的 Hurst 值。使用 R/S 方法计算 Hurst 值，结果如表 2 所示。

表 2 节点 1 在不同时间段的 Hurst 值变化

时间段	Hurst 值	时间段	Hurst 值
00:00-00:05	0.563 1	3:00-3:05	1.058 2
00:05-00:10	0.811 1	3:05-3:10	0.790 9
00:10-00:15	0.677 0	3:10-3:15	0.942 2
00:15-00:20	0.931 4	3:15-3:20	1.029 9
00:20-00:25	0.926 2	3:20-3:25	0.429 4
00:25-00:30	0.862 4	3:25-3:30	0.776 2
00:45-00:50	0.846 1	14:00-14:05	0.779 7
00:50-00:55	0.898 7	14:05-14:10	0.712 7
00:55-01:00	0.753 9	14:10-14:15	0.607 1
01:00-01:05	1.010 5	14:15-14:20	0.839 9
01:05-01:10	0.705 4	14:20-14:25	0.362 0
01:10-01:15	0.608 3	14:25-14:30	0.543 3

由表 2 的数值可知，晚上 Hurst 均值为 0.812 3，而白天均值为 0.640 8。晚上用网低峰时，Hurst 值普遍偏高，从网络使用习惯可知，一般晚上的流量都是由下载软件产生，当晚上下载流量与整个网络的流量的比率增大时，也同时增大了下载流量特性对整个网络的影响比例，因此引起 Hurst 参数的增大。网络流量作为一种非常复杂的非线性的时间序列，必定存在表示流量序列长期行为规律的趋势项，表示为网络流量在一定范围的平均值。对于校园网这种小规模的局域网，网络受单点影响较大，网络流量随时有增大和变小的趋势，但这种趋势并不是一种异常，本文对原始流量中的趋势项进行消除，通过计算流量中随机成分[8]等其他成分的 Hurst 值发现网络中的异常。

基于以上推测，本文在使用 R/S 方法计算 Hurst 时，引入希尔伯特黄变换(Hilbert-Huang Transform, HHT)方法。HHT 变换主要是对非稳态和非线性信号的一种自适应时频分析方法。通过 HHT 变换中的 EMD 可消除网络流量中趋势信息，从而提高计算 Hurst 值的准确度。

4 异常流量检测

4.1 HHT 思想

HHT[11]是一种新的信号分析方法，适用于分析非线性、非平稳信号。HHT 变化有 2 个步骤。首先，将要处理的数据分解为若干个本质模态函数(Intrinsic Mode Functions, IMF)，分解流程称为经验模态分解(Empirical Mode Decomposition, EMD)。然后将 IMF 做希尔伯特转换(Hilbert Transform, HT)，获得希尔伯特范围。本文中主要用到第 1 个步骤，做经验模态分解。它是 HHT 的核心算法，能够分解出网络信号中的趋势项，以消除其对自相似值的影响。用 $X(t)$ 表示网络流量的时间序列，通过 EMD 处理，最后把 $X(t)$ 用式(3)表示：

$$X(t) = \sum_{j=1}^{n} c_j + r_n \tag{3}$$

网络流量时间序列最后能被分解为 n 个 IMF 和一个趋势函数 r_n。r_n 是原始序列中包含的长期趋势序列，影响计算 Hurst 值时的精确度。

在基于以上算法的预处理后，原始的时间序列依旧具有自相似的特征，而 r_n 可以从原始序列中清除，以消除网络信号中的趋势项给 Hurst 计算带来的影响。

4.2 HHT 处理结果分析

为了与没有经过 HHT 处理的 Hurst 计算结果比较，本文

使用图 1 中节点 1 的流量信息。对原始流量信息使用 EMD 变换，分离出趋势项 r_n，然后对分离后时间序列使用 R/S 方法计算 Hurst，计算结果与原计算结果数据对比见表 3。

表 3 节点 1 流量信息经分离趋势项后 Hurst 值的变化

时间段	改进 Hurst 值	时间段	改进 Hurst 值
00:00-00:05	0.667 0	3:00-3:05	0.867 3
00:05-00:10	0.846 7	3:05-3:10	0.537 4
00:10-00:15	0.545 1	3:10-3:15	0.837 8
00:15-00:20	0.853 5	3:15-3:20	0.637 4
00:20-00:25	0.768 8	3:20-3:25	0.370 2
00:25-00:30	0.809 1	3:25-3:30	0.804 5
00:45-00:50	0.654 8	14:00-14:05	0.942 9
00:50-00:55	0.584 6	14:05-14:10	0.835 2
00:55-01:00	0.725 9	14:10-14:15	0.951 6
01:00-01:05	1.110 8	14:15-14:20	0.780 5
01:05-01:10	0.924 7	14:20-14:25	0.401 8
01:10-01:15	0.584 3	14:25-14:30	0.588 2

由表 3 可以看出，在经过 EMD 变换后的流量信息依旧具有自相似性，趋势项的消除并没有对网络的自相似性产生影响。原始 Hurst 中异常点有 5 个(01:00-01:05、3:00-3:05、3:15-3:20、3:20-3:25、14:20-14:25)，使用改进的方法计算 Hurst，异常点为 3 个(01:00-01:05、3:20-3:25、14:20-14:25)。

为比较算法的准确性，对 01:00-01:05、3:15-3:20 和 3:20-3:25 3 个时间点的流量数据进行比较，这 3 个时间点分别代表不同的结果类型。图 2~图 4 中的曲线 1 表示原始流量信息，曲线 2 表示该时间段的流量趋势值，由图 2 与图 4 可知，网络流量存在陡增与陡降的流量异常点。图 3 的时间段原始流量信息有慢慢减小的趋势，流量不存在陡增或陡降的异常点，但使用常规算法会得出流量异常的结果，由此可知流量整体的变化对自相似值产生了影响，图 3 中的平滑曲线为该时间段的流量趋势值，消除趋势值后，流量信息的自相似值为 0.637 4，在正常范围内。因此，对流量趋势进行消除，能得到更准确的结果。

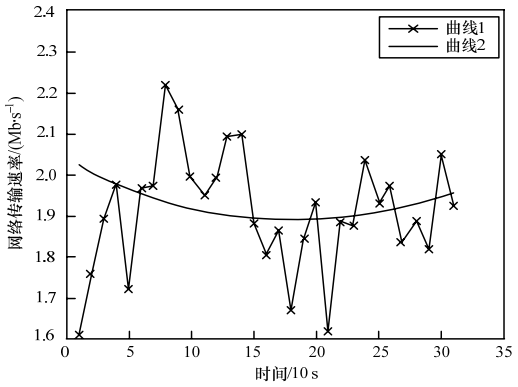

图 2 01:00-01:05 流量值和趋势值

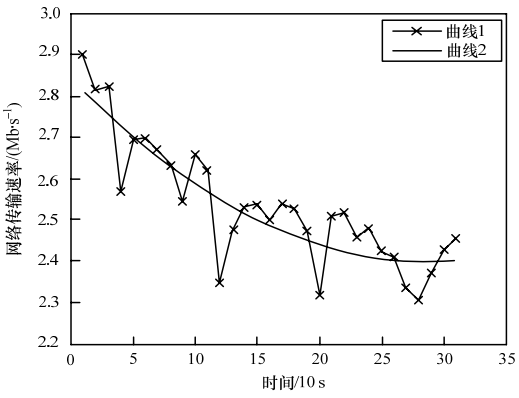

图 3 3:15-3:20 流量值和趋势值

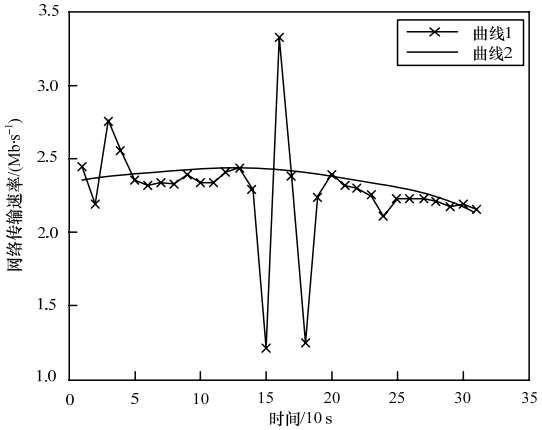

图 4 3:20-3:25 流量值和趋势值

图 5 中的曲线 2 与曲线 3 同图 3 中曲线 1、曲线 2 表示相同，曲线 1 为消除趋势值后的序列。可以看出，消除趋势值后，网络变化特征依旧存在，没有造成网络特征的丢失。图 2 与图 4 由于整体流量变化范围不大，趋势值比较平滑，2 种算法结果一致，判定该时间段为异常。由结果可知，消除小规模网络流量信息的趋势项对于网络流量的自相似值计算更准确，在一定程度上减小误差，提高异常检验准确度。

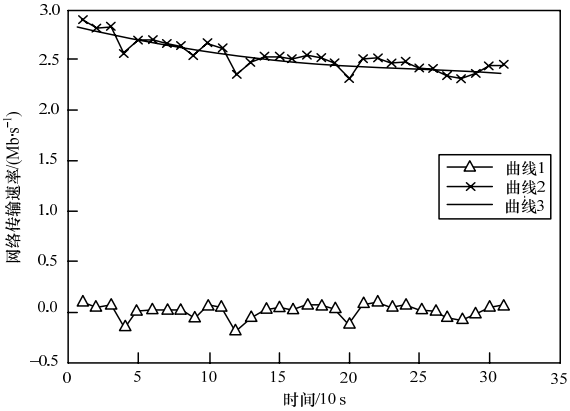

图 5 3:15-3:20 消除趋势值前后流量对比

5 结束语

对于校园网络等小规模的局域网，直接进行自相似值计算不能很好地检测异常流量。本文在分析校园网络流量特征的基础上，采用 R/S 方法对校园网络流量进行自相似分析，结合网络拓扑结构，对每个层级中的网络节点自相似值进行计算，并对不同时间尺度下网络节点的自相似性进行详细分析，发现校园网络在大时间尺度上符合自相似性，白天的网络自相似普遍小于晚上。R/S 方法具有计算简单和稳定的特性，但是对于直接计算的校园网等小规模网络流量时有其不完善的地方，因此，先对收集的原始网络流量信息做预处理，通过 HHT 变化，使用 EMD 流程消除网络时间序列流量中的趋势项给计算带来的影响，在不改变原始序列自相似特性的基础上提高自相似计算的准确性，从而提高对校园网络状况预测的准确性。

参考文献

[1] Leland W E, Taqqu M S, Willinger W, et al. On the Self-similar Nature of Ethernet Traffic[J]. IEEE/ACM Trans. on Networking, 1994, 2(1): 1-15.

[2] Chonka A, Singh J, Zhou Wanlei. Chaos Theory Based Detection Against Network Mimicking DDoS Attacks[J]. IEEE Communications Letters, 2009, 13(9): 717-719. (下转第 78 页)

该数据集由每隔 30 s 采样所得采集的温度、湿度、光强和节点电压等 4 种监测值组成(http://berkeley.intelresearch.net/labdata)。用该数据集的子集进行模拟实验，并和 IDW 算法进行比较。实验结果如图 1~图 4 所示。以监测数据集所属监测区域的具体坐标为输入，不规则网格划分算法得到的结果如图 1 所示。图 1 中的网格大小不一，呈不规则分布，没有出现一个网格里有大量节点的情况，符合预期。从图 2~图 4 可以看出，Kriging 和 IDW 的插值结果与原始数据的趋势一致，这说明数据插值的方法是有效的。对于数值变化相对平稳的温度和湿度数据，2 种插值方法的结果基本相同，相比之下，Kriging 有略小的插值误差；而对于变化幅度大的亮度数据，Kriging 插值的误差明显小于 IDW 插值方法。

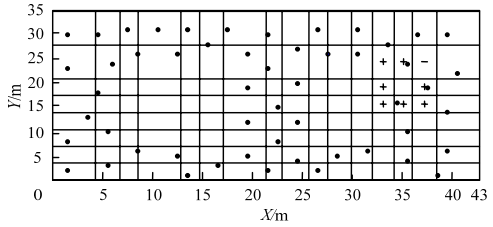

图 1 54 个传感器节点的不规则网格划分

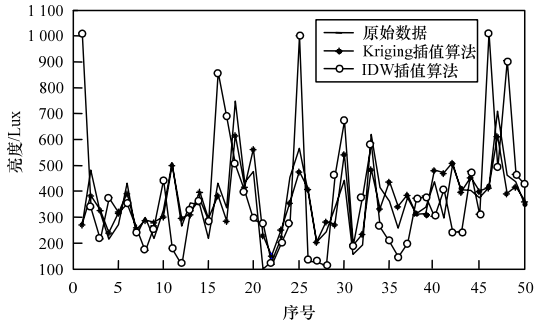

图 2 亮度数据插值结果比较

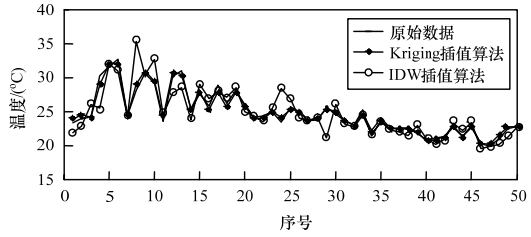

图 3 温度数据插值结果比较

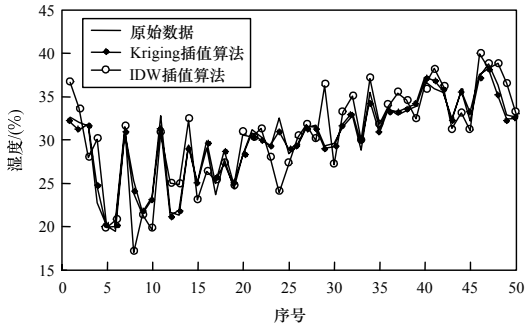

图 4 湿度数据插值结果比较

4 结束语

本文研究传感数据集插值问题，提出一种基于不规则网格划分的快速 Kriging 插值算法。与均匀网格划分方法相比，不规则网格划分能够充分适应传感器节点部署的不规则性。基于不规则网格搜索参估点集，降低了搜索的复杂度，使得 Kriging 算法可以运行于资源受限的无线传感器网络。模拟实验结果表明，与 IDW 相比，本文提出的 Kriging 插值算法有更好的插值精度。

参考文献

[1] 张智勇. 基于GMDH的缺失数据插补方法研究[D]. 成都: 四川大学, 2007.

[2] 周四望, 林亚平, 张建明, 等. 传感器网络中基于环模型的小波数据压缩算法[J]. 软件学报, 2007, 18(3): 679-690.

[3] 付惠娟, 任美睿, 等. 郭龙江无线传感器网络中缺失数据的估计[J]. 计算机工程, 2011, 37(1): 90-92.

[4] Martinez S. Distributed Interpolation Schemes for Field Estimation by Mobile Sensor Networks[J]. IEEE Transactions on Control Systems Technology, 2010, 18(2): 491-500.

[5] 潘立强, 李建中, 骆吉洲. 传感器网络中一种基于时-空相关性的缺失值估计算法[J]. 计算机学报, 2011, 33(1): 1-11.

[6] Richard T, Hare G, David M. Interpolation for Wireless Sensor Network Coverage[C]//Proc. of the 2nd IEEE Workshop on Embedded Networked Sensors. Sydney, Australia: IEEE Press, 2005.

[7] Harrington B, Huang Yang, Yang Jue. Energy-efficient Map Interpolation for Sensor Fields Using Kriging[J]. IEEE Transactions on Mobile Computing, 2009, 8(5): 622-635.

编辑 陆燕菲

(上接第 75 页)

[3] Rawat S, Sastry C S. Network Intrusion Detection Using Wavelet Analysis[C]//Proc. of CIT'04. Hyderabad, India: [s. n.], 2004: 224-232.

[4] Benetazzo L, Giorgi G, Narduzzi C, et al. On the Analysis of Communication and Computer Networks by Traffic Flow Measurements[J]. IEEE Trans. on Instrumentation and Measurement, 2007, 56(4): 1157-1164.

[5] Giorgi G, Narduzzi C. Detection of Anomalous Behaviors in Networks From Traffic Measurements[J]. IEEE Trans. on Instrumentation and Measurement, 2008, 57(12): 2782-2791.

[6] Li Zonglin, Hu Guangming, Zhou Ruqiang, et al. Traffic Model Analysis for Anomaly Detection[C]//Proc. of the International Conference on Information Computing and Automation. [S. l.]: IEEE Press, 2008: 1434-1437.

[7] 吕良福, 张加万, 张 丹. 基于改进小波分析的 DDoS 攻击检测方法[J]. 计算机工程, 2010, 36(6): 29-31, 44.

[8] Cheng Xiaorong, Xie Kun, Wang Dong. Network Traffic Anomaly Detection Based on Self-similarity Using HHT and Wavelet Transform[C]//Proc. of the 15th International Conference on Information Assurance and Security. [S. l.]: IEEE Press, 2009: 710-713.

[9] 张 宾, 杨家海, 吴建平. Internet 流量模型分析与评述[J]. 软件学报, 2011, 22(1): 115-131.

[10] Oetiker T. MRTG[EB/OL]. (2010-05-17). http://www.mrtg.org/rrdtool/.

[11] Huang N E, Shen Zhen, Long R S. A New View of Nonlinear Water Waves——The Hilbert Spectrum[J]. Annual Reviews of Fluid Mechanics, 1999, 31(1): 417-457.

编辑 陆燕菲