第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·软件技术与数据库· 文章编号：1000—3428(2007)02—0040—04 文献标识码：A 中图分类号：TP311.13

基于小波技术的网络时序数据挖掘

郭四稳[1]，何 维[2]，王 鹏[3]
(1. 广州大学计算机教育软件研究所，广州 510006；2. 中国农业银行四川分行，成都 610017；3. 信息产业部第三十九研究所，合肥 230000)

摘 要：网络安全日志数据库是一种历史数据，对它的分析具有十分重要的实际价值，作为一种时序数据库，针对它的信息挖掘已研究出许多方法。该文提出了一种新的对此类时序数据库的信息挖掘方法，利用小波变换多分辨率分析的方法对信号化后网络安全日志数据库中的数据在不同的时间尺度上进行分析和信息挖掘，从中提取出单位时间内网络受到攻击次数的时间周期规律，并对这种方法的分析特性进行了阐述，而且利用小波阈值重建的方法对原始信号数据进行去噪处理，收到了良好的效果。

关键词：小波变换；信号化；网络安全日志数据库；数据挖掘

Network Time Serial Data Mining Based on Wavelet Technique

GUO Siwen [1], HE Wei [2], WANG Peng [3]
(1. Institute of Computer Education Software, Guangzhou University, Guangzhou 510006; 2. Sichuan Branch, Agriculture Bank of China, Chengdu 610017; 3. The 39th Institute, Ministry of Information Industry, Hefei 230000)

【Abstract】 Network security log file database is a kind of historical database. It is very important to research it. A lot of data mining methods to research it are found. A new method is provided to analyze and mine this kind of time serial database. Multi-resolution analysis of wavelet transform method is used to analyze and mine the data of network security log file database on different time scale. The period law of the attack number every hour is found by this method. Finally the wavelet threshold method is used to de-noise from the data. It is proved that this method can get good results.

【Key words】 Wavelet transform; Signalizing; Network security log file database; Data mining

1 概述

随着网络技术的飞速发展，网络安全日志数据库里的数据量也飞速增长，并且随着信息粒度的变细，数据量将进一步增加，因此有必要对这些数据进行分析挖掘，寻找出我们感兴趣的模式。这类数据库一般属于时序数据库，而传统的分析挖掘方法一般是基于统计学的算法把时间序列数据作为随机变量 $X_i(i=1,2,\dots,n)$ 来处理的。如果将数据值看作信号幅度，时序数据库中的每一列均可看作是时域中的一个一维信号 $f(t)$，这样就可以采用信号分析领域中的方法来对日志数据库中数据进行分析挖掘。

信号分析领域小波(wavelet)变换方法已经得到了越来越广泛的应用[1~3]。小波变换是由法国数学家Morlet[3]于 1980 年在进行地震数据分析工作时提出的，而小波研究的热潮始于 1986 年，随后，S.Mallat[4,5]于 1989 年提出了多分辨率分析的概念，统一了在此之前Stromberg[6]、Meyer、Lemarie、Battle[7]等提出的各种具体的小波构造方法，给出了构造正交小波基的一般方法和FFT相对应的快速算法 ——Mallat算法。小波变换克服了傅立叶分析的局限，在Heisenberg测不准原理的约束下在时域和频域上同时具有良好的局部化性质，因而已有效地应用于如信噪分离、编码解码、检测边缘、压缩数据、识别模式以及将非线性问题线性化、非平稳过程平稳化等问题，正是在这种意义下，小波变换被誉为数学显微镜。对于Gabor变换一旦选定了窗函数，时频分辨率就已固定，只是在相空间进行平移。虽然受Heisenberg测不准原理的约束小波变换时域和频域分辨率的乘积值在相空间是不变的，但由于其多分辨率的特性能在不同尺度上得到越来越精确的函数。

本文采用小波分析的方法对基于网络安全日志的时序数据库信息进行分析挖掘，充分利用了小波的优良特性，为今后对网络日志数据信息的分析挖掘提供了新的思路和方法。

2 小波分析算法

小波变换是用一簇小波函数系去逼近一个信号，小波充当了傅利叶变换中正弦和余弦函数的角色，而小波函数系是通过一个基本小波函数经过伸缩和平移后构成的。

$$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right)\qquad a,b\in R,a\neq 0 \tag{1}$$

其中 a 为伸缩因子与局部频率相对应，b 为平移因子，ψ 叫基本小波或母小波(Mother Wavelet)。

对于任意函数 $f\in L^2(R)$，关于 $f(t)$ 的连续小波变换定义如下：

$$W_\psi f(a,b)=\langle f(t),\psi_{a,b}(t)\rangle=\frac{1}{\sqrt{|a|}}\int_{-\infty}^{+\infty}f(t)\overline{\psi(\frac{t-b}{a})}dt \tag{2}$$

为方便计算机处理必须将变换进行二进制离散化：

如果

$$(a,b)=(\frac{1}{2^j},\frac{k}{2^j})\quad,\quad \psi_{j,k}=\psi_{a,b}(t)=2^{j/2}\psi(2^jt-k) \tag{3}$$

作者简介：郭四稳(1963－)，男，副研究员、博士，主研方向：网络安全，人工智能，计算机软件技术；何 维、王 鹏，博士、工程师

收稿日期：2006-02-11 **E-mail：**guosiwen@tsinghua.org.cn

那么 $f(t)$ 可以表示成如下的小波序列：

$$f(t)=\sum_{j,k=-\infty}^{+\infty} c_{j,k}\psi_{j,k}(t) \tag{4}$$

其中

$$c_{j,k}=\left\langle f(t),\psi_{j,k}(t)\right\rangle=2^{j/2}\int_{-\infty}^{+\infty} f(t)\overline{\psi(2^j t-k)}dt \tag{5}$$

以上为离散二进小波变换的过程。

本文采用了小波变换的 Mallat 算法来对网络安全数据进行处理。由于 Mallat 算法大大提高了小波变换的效率，因此今后还可将小波变换应用于网络安全的实时信号的分析和监控。Mallat 算法的一维情形如下：在式(4)、式(5)中，$\psi_{j,k}$ 一般不具有初等解析表达式，并且在实际中 $f(t)$ 往往是由数值方法给出的。因此直接求 $c_{j,k}$ 是不方便的。Mallat 首先得到了一个计算式(5)的离散算法，从而避免了上述困难。

空间$L^2(R)$中的多尺度分析是指$L^2(R)$中满足如下条件的一个空间序列 $\{V_j\}\ j\in z$：

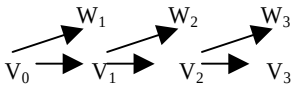
(1)单调性 $\cdots V_{-1}\supset V_0\supset V_1\supset\cdots$；

(2)渐近完全性 $\cap_{k\in z}V_k=\{0\}$ 和 $\overline{\cup_{k\in z}V_k}=L^2(R)$；

(3)伸缩规则性 $f(t)\in V_k\Leftrightarrow f(2t)\in V_{k-1}$；

(4) $\exists\phi\in V_0$ 使 $\{\phi_{0,n}\}$ 是V_0的Riesz基。

这里 $\phi(x)$叫尺度函数(Scale function)，定义W_k为 V_k 在 V_{k-1} 中的正交补空间，例如 $V_{k-1}=V_k+W_k$，并且 $V_k\perp W_k$。因此可按如下方法进行分解：



尺度函数和小波函数满足如下的双尺度方程：

$$\phi(t)=\sqrt{2}\sum_{k=0}^{L-1}h_k\phi(2t-k) \tag{6}$$

$$\psi(t)=\sqrt{2}\sum_{k=0}^{L-1}g_k\phi(2t-k) \tag{7}$$

这里 $\phi(t)$ 和 $\psi(t)$分别为尺度函数和小波母函数，L 为滤波器长度。假如小波是正交的，那么 $g_k=(-1)^k h_{L-k-1}$，离散小波分解和重构就能够分层进行，对每一层的平滑信号进行分解，分成低一层的平滑信号和细节信号，这就是 Mallat 算法。如果把双尺度方程的系数看着滤波器，那么 Mallat 算法实际上就是起到双通道滤波器对的作用，尺度函数和小波函数分别作为低通和高通滤波器。h_k 和 g_k 分别为低通和高通滤波器系数。

3 时序数据库的信号化

在时序数据库中每一条记录都对应着某一时刻或一段时间内对象各参数的情况。如果只看某一个字段对应列的值，将这些值按一定方法量化后作为信号的幅度值 a_i，与其对应的时间值 t_i 形成一组离散化信号 $f(t)=\{t_i,a_i\}$，数据库中的每一个字段均对应这样一组时域信号。时序数据库在经过信号化以后每一个字段对应的列就作为一个完整的信号呈现在我们面前，这时就可以采用小波分析等信号分析方法挖掘并提取出隐含在信号中的特征和模式，这是一种新的针对时序数据库的数据挖掘方法。

网络安全日志数据库对维护网络的安全、发现网络漏洞、监控黑客入侵等方面具有相当大的参考作用。但随着网络日志数据库的不断庞大，往往无法及时被处理，许多网络中心的日志数据被闲置。因此分析挖掘日志数据库中有用信息的工作势在必行。网络安全日志数据库记录了网络各参数随时间的变化情况，是一类时序数据库，可以将其信号化后用以上所述的方法进行处理。

本文所采用的数据取自国内某著名高校网络中心入侵检测系统(IDS)的日志数据库中的数据。我们只取了一个特性进行分析，即 IDS 系统检测到的该网络中心在单位时间里受到攻击的次数，采样间隔为 1h。检测时间为 2002-10-10 17:00:00~2002-10-18 20:00:00。图 1 为信号化后该字段在时域上的信号。其中横轴为时间 t，纵轴为检测到的入侵次数 n。

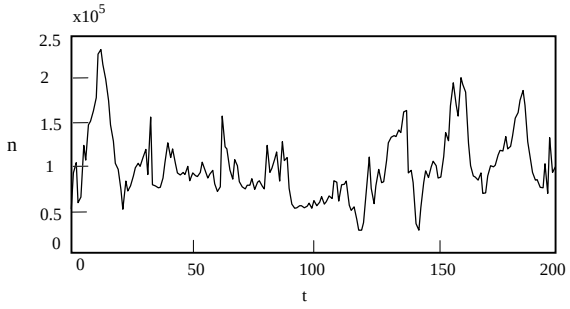


图 1 入侵次数随时间的变化

4 小波分析处理的结果

4.1 多尺度小波分析对网络攻击周期特征的提取

我们首先采用 Haar 小波对图 1 中的一维信号进行三级小波分解。Haar 基是同时具有对称和反对称紧支集的正交小波基。基本 Haar 小波函数定义如下：

$$h(x)=\begin{cases}1 & x\in[0,0.5]\\ -1 & x\in(0.5,1]\end{cases} \tag{8}$$

图 2、图 3 分别为原信号第一级小波分解的近似分量和细节分量，图 4、图 5 分别为原信号第二级小波分解的近似分量和细节分量，图 6、图 7 分别为原信号第三级解的近似分量和细节分量。其中纵轴为小波系数 c。从图中可以看出，每一级小波分解都将上层的近似分量信号分解成下一层的近似分量信号和细节分量，且取样点数减少一半，这是因为在小波变换中，对不同频率成分在时域上的取样步长是调节性的，高频部分步长小，低频部分步长大(subsample)，也就是说高频部分时间分辨率要比低频部分高，这是符合实际要求的，也是小波分析“数学显微镜”特点的体现。

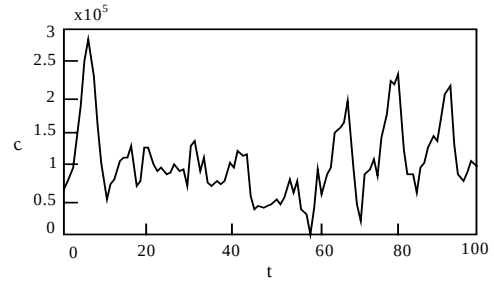


图 2 一级小波分解的近似分量

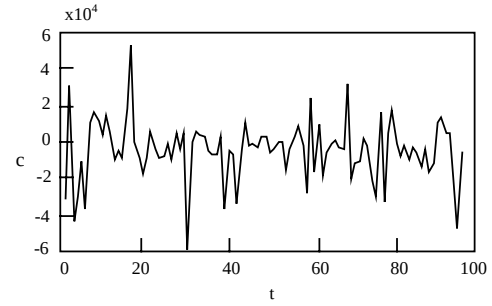


图 3 一级小波分解的细节分量

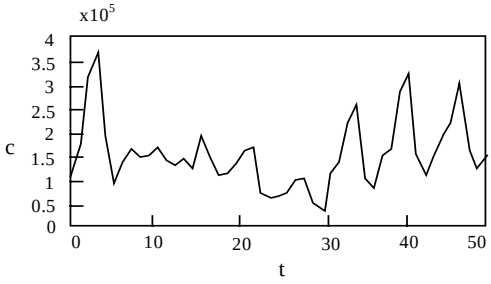


图 4 二级小波分解的近似分量

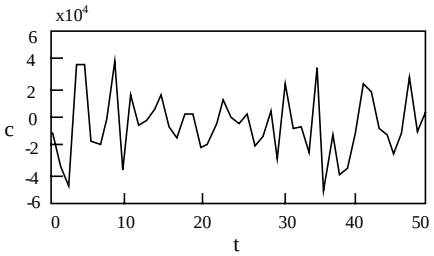


图 5 二级小波分解的细节分量

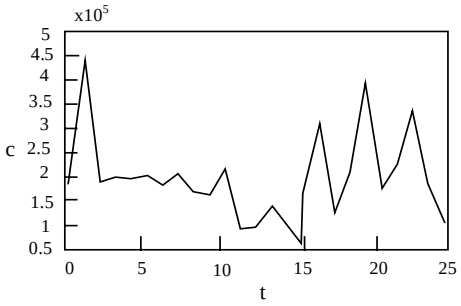


图 6 三级小波分解的近似分量

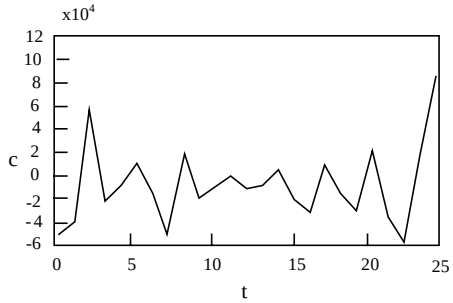


图 7 三级小波分解的细节分量

从原信号在经过三级 Haar 小波分解后的近似分量中可以清晰而准确地分辨出以“天”为周期的信号变化情况。在图 6 上可以明显地看到信号出现 8 个等间距的峰值，其意义对应于单位时间内网络被攻击次数的周期性特征，说明网络被攻击次数的变化存在以天为尺度的周期性特点。对比原信号可以得到攻击次数的峰值一般出现在每天凌晨 6 点左右，而低谷值出现在每天的中午。我们对原信号采用不同的小波进行分解均得到了相似的结果。根据攻击次数的时间周期模式，可以采取相应的防范策略，分析攻击的特征。小波变换的方法为我们提取这一攻击的周期模式提供了有力的工具。

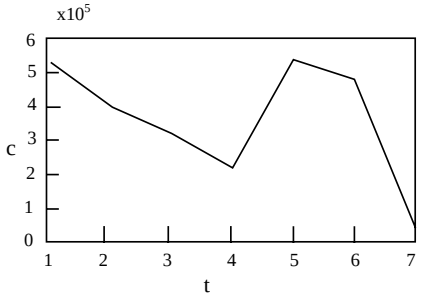


图 8 五级小波分解的近似分量

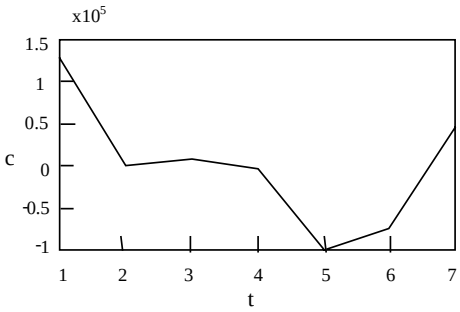


图 9 五级小波分解的细节分量

从不同分解层的图中可以看出各层小波分解的近似分量随分解层次的增加高频成分被逐渐分离出去，如图 3、图 5、图 7、图 9，即被分离出去的高频成分。由于我们采用了二进小波离散化方法，根据采样定理，每增加一级分解，在时域上的采样间隔将增加 1 倍，因此滤波后小波近似分量与细节分量的采样点数将减小 1 倍，越往下分解越能从大的时间尺度去观察信号的变化情况。图 8、图 9 是第五级小波分解的近似分量和细节分量，其中图 8 中的近似分量相对第三级小波分解是从更大的时间尺度去观察信号的变化。可以看出它描绘了网络单位时间被攻击次数以“周”为时间尺度的信号变化情况。事实上，采用小波分析的方法可以对信号进行任意精度的分析，但由于原始数据存在一个采样精度的问题，因此小波分析的最大分析精度不可能超过原始数据的采样精度，原始数据的采样是 1 小时 1 次。其最小分析精度为整个信号的时间跨度。在第三级分解时可以清晰地分辨出以“天”为尺度的信号变化情况，在第五级小波分解时就可以分辨出以“周”为尺度的信号变化情况，这就是所谓的多分辨率分析的方法。根据这一方法，如果时序数据的时间跨度更大一些，比如是“年”，那么便可以通过更多层的分解得到信号分别在“天”“周”“月”“年”等不同时间尺度下的变化模式，为分析网络日志数据库提供更丰富的信息，这是小波多尺度分析方法优于其它信号分析方法的特点之一。

4.2 小波阈值重建法对网络攻击数据去噪声

在原信号中一般都包含有噪声的成分，为便于正确地分析网络数据可以将噪声信号去除。但由于噪声信号和有用信号的细节部分均分布在高频区，传统的如傅里叶变换去噪的方法无法将重叠区很大的有用信号和噪声信号区分开。利用小波分析的方法，可以构造出一种既能降低信号噪声又能保持信号细节的方法。由于有用信号其小波系数幅值大、数目

少，而对于噪声信号其小波系数幅值小，数目较多，因此可以设定一个阈值，使大于这个阈值的小波系数认为是有用信号的系数，而小于这阈值的小波系数认为是由噪声成分供献的，可以去掉。在小波函数满足容许条件式(9)的情况下：

$$\int_{-\infty}^{+\infty}\frac{\left|\hat{\psi}(\omega)\right|^2}{\omega}<\infty \tag{9}$$

$\hat{\psi}(\omega)$ 为小波函数的傅里叶变换，可以用经过阈值处理后的小波分解结果重建原信号，从而得到滤除噪声后的原信号，以便进一步对其进行分析。

图 10 是经过小波去噪后的原信号。可以看出，通过小波去噪处理后原信号中的噪声信号已得到了有效的抑制，对原信号的进一步处理就可以在去噪后的信号上进行，这样将会得到更符合实际的结果。

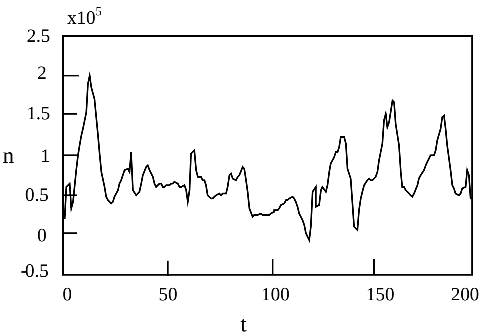


图 10 采用小波阈值去噪的重建信号

5 结论

本文针对网络安全日志数据库这一类时序数据库，采用小波多分辨率分析这种信号分析工具对其进行分析挖掘和处理，从不同尺度成功地提取出了网络被攻击次数的时间周期特征，并将小波去噪的方法初步应用于时序数据库的分析。这证明将时序数据库信号化后，利用信号处理领域的分析方法对其进行分析，挖掘出有用的模式这一思路是可行的。下一步工作准备采用更多的网络安全时序数据库的属性值，用这种方法进行分析，再对多属性特征的分析结果进行信息融合，挖掘出更多、更有意义的模式。

参考文献

1 Daubechies I. The Wavelet Transform, Time-frequency Localization and Signal Analysis[J]. IEEE Trans. on Information Theory, 1990, 36(5): 961-1005.

2 Shapiro J M. Embedded Image Coding Using Zerotrees of Wavelet Coefficients[J]. IEEE Trans. on Signal Processing, 1993, 41(12): 3445-3462.

3 Morlet J, Arens G, Fourgean E, et al. Wave Propagation and Sampling Theory and Complex Waves[J]. Geophysics, 1982, 47(2): 222-236.

4 Mallat S G. Multifrequency Channel Decompositions of Images and Wavelet Models[J]. IEEE Trans. on Speech and Signal Processing, 1989, 37(12): 2091-2110.

5 Mallat S G. A Theory of Multiresolution Signal Decomposition: The Wavelet Transform[J]. IEEE Trans. on PAMI, 1989, 11(7): 674-693.

6 Stromberg J. A Modified Haar System and Higher Order Spline Systems on R^n as Unconditional Bases for Hardy Spaces[C]. Proc. of Conference on harmonic Analysis in Honor of Antoni Zygmund Ⅱ, Wadsworth, Belmont, California, 1981: 475-493.

7 Battle G. Phase Space Localization Theorem for Ondelette. J. Math. Phys. 1989, 30(10): 2195-2196.

(上接第 21 页)

(5)如果 $(Nset_i(N_l)\neq\phi)\cap(Nset_i(N_m)\neq\phi)$，则调用上节给出的最大权匹配算法，并为集合中的节点打上“ 已经处理 ”的标识，否则转入(8)；

(6)根据式(13)求出第i次的φ_i；

(7)$i=i+1$，转入(4)；

(8)将节点上“ 已经处理 ”的标识清除；

(9)转入(2)。

$\vartheta(i)$ 是单调递减的，保证距离N_l,N_m越远的近邻对N_l,N_m相似性的影响越小。通过精确语义相似性分析，最终得到的结果保存在精确语义相似性字典(ESSD)中。

$$ESSD=\eta^{\infty}=\left\{\langle N_l,N_m,\varphi\rangle \middle| N_l\in N(IS_1),N_m\in N(IS_2)\right\} \tag{14}$$

4 结论

在从半结构化的信息源中建立本体的过程中，为了统一分析和处理不同的信息源，本文提出了一个统一的概念模型，将各信息源转换为 SDS-G，该模型不仅完整地表现了各信息源的内涵，在计算语义相关性和语义距离时，还考虑了外延的影响。基于转换后得到的 SDS-G 模型，以节点名、节点属性、节点近邻为比较特征，使用节点名相似性分析方法、0近邻相似性分析方法计算节点的基本语义相似性分析，然后使用i近邻相似性分析方法修正已得到的相似性，提炼模式间更精确的语义相似性。

本文提出的方法还解决了相似性分析中类型冲突的问题，方法是半自动化的，仅在早期需要少量的人类专家的参与。该方法在实验中也取得了较好的效果。

参考文献

1 Gruber T. What is an Ontology?[EB/OL]. 2002-01-28. http://www-ksl.stanford.edu/kst/whst-is-an-ontology.html.

2 Bergamaschi S, Castano S, Vincini M. Semantic Integration of Semistructured and Structured Data Sources[J]. SIGMOD Rec., 1999, 28(1): 54-59.

3 Buneman P, Davidson S, Fernandez M, et al. Adding Structure to Unstructured Data[C]. Proc. of International Conference on Database Theory, 1997: 336-350.