网络行为研究环境的设计与实现*)

Network Behavior Researching Environment Design and Implementation

吴晓江　帅典勋　刘东林　邓志东

(华东理工大学计算机科学与工程系　上海200237)

(清华大学智能技术和系统国家重点实验室　北京100084)

Abstract We design and implement a researching environment for network behavior. This paper mainly presents our considerations on how to combine the network monitor with the network behavior analysis in an organic and interacting way, so as to establish an effective, dynamic and stable modeling environment for network behavior. This also laid a foundation for the further research on network behavior.

Keywords Network behavior, Network monitor, Network flux, Fractal dimension

1. 引言

随着网络应用的拓广和规模的扩大，网络正逐渐成为一个复杂的开放系统。在越来越多因素的作用下，网络不再是简单、可预测的线性系统，它演变为了一个具有复杂行为的非线性系统，如具有社会行为、新陈代谢行为以及非线性动力学行为等。对复杂网络行为的有限了解使我们还没有办法掌握网络行为背后的内在规律。这给网络的管理控制带来了困难，进而使许多网络问题无法得到全面的解决。

近年，不少研究对网络行为有了新的发现和认识，如对网络流量自相似性的发现[1]，网络行为中社会行为的表现[2]、TCP/IP 机制的混沌特性[3]，路由器 BUFFER 排队的自组织现象[4]等等，很多方法和理论如代数模型方法、混沌孤立子竞争方法、遗传进化理论也逐渐被利用到网络行为的研究中。随着网络行为研究的不断深入，需要建立一个合适的研究环境，把这些方法与理论有效地结合到网络行为的研究中。

目前，网络行为的研究环境，一方面是通过网管设备对网络系统的监测来收集反映网络状态的数据；另一方面是通过建立网络的数学模型来进行理论研究。然而目前的缺陷是大量先进网管技术所获得的数据样本没有进一步被很好地分析、利用而导致大量的浪费，而另一端对网络的理论模型研究有时又脱离了真实的网络环境从而使得出的结论缺少较强的说服力。网络环境和理论模型没有得到有机地结合阻碍了网络行为的进一步研究。

本文设计并建立了一个网络行为研究环境，它将网络数据的监测采集功能和对网络数据进行的理论分析建模功能融入了一个整体的环境，在网络数据的采集和网络行为的分析之间建立了互动的关系，使之能充分有效地利用数据和新的理论方法。这一环境为网络行为的各种理论分析创造了统一的研究平台，并且为网络行为的长期性研究奠定了基础。

2. 概要介绍

2.1 总体设计

网络行为建模环境包括了数据监测采集和数据理论分析两大功能，以及联系这两大功能模块之间的互动机制(如图1)。

数据监测采集工作包括选择所研究的网络环境，如何从网络中获得需要的原始信息，如何存储样本数据，数据以什么样的方式提供给后面的理论分析等。数据理论分析包括如何将得到的数据信息进一步转化为可供分析的数据，根据不同的分析目的选择自己需要的理论形成实际的算法，得出分析或建模的结果、比较分析结果等工作。

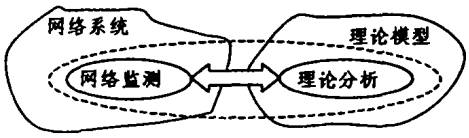

图1　两大功能模块的关系

模块间的互动机制实现了数据的传递，样本的组织转换和信息的反馈，使网络系统和理论模型这两个功能模块在环境中并不孤立，而是有机地结合，互相协调地工作，进而使网络行为的研究能在合适的数据结构和有效的理论算法上开展、同时它也将起到了降低模块间耦合度的作用。

2.2 特点与功能

网络行为研究环境中，通过两个模块间的互动机制，实现了两者的有机结合。既要通过理论分析的结论来帮助确定或调整网络系统的监测方向，同时也要通过研究监测到的样本数据来确定分析的算法，甚至确定理论模型。这样的结合既为所监测到的网络数据提供了进一步的分析手段，又为理论和方法的应用提供了一个搭建在真实网络系统上的平台。这样的结合可以方便地更换不同理论进行研究，在改变网络监测工具、方式，或改变理论方法时不需要跟着改动环境的其他部分，为长期研究网络行为提供一个稳定的环境。

在这样的环境中，我们可以将网络监测模块作为观测器对真实的网络系统进行观察，发现一些表面的网络行为如突发性流量特征、自相似性现象和表面的相变现象。另外，可以通过利用相关理论进行深入的实验分析，进一步发现网络行为的更深层次的表现如网络流量在时间和空间上所表现出来的有序性，网络的社会行为，网络系统中的混沌运动和自组织现象等等。此外，可以根据分析模块的要求指导网络管理监测

*)本课题得到973国家重点基础研究规划项目(G1999032707)、国家自然科学基金资助项目(60073008和69773037)和清华大学智能技术和系统国家重点实验室基金资助和支持。吴晓江　硕士研究生，研究方向：网络、人工智能。帅典勋　教授、博士生导师。

工具制造特定的网络环境，来主动地研究不同情况甚至极端情况下的网络行为。最后通过对行为现象进行深入的分析对比，可以了解网络行为的变化规律，从而进一步揭示网络行为的内在规律。

3. 详细设计和实现

图2是根据上述思想所实现的网络行为研究环境的体系结构。各个模块通过数据信息的流动和模块间的信息反馈有机有序地结合，协调工作。同时各模块应用各种工具实现各自独立的功能。其中采集、分类、样本库模块属于网络数据采集功能，预处理、分析、对比模块属于理论分析功能。

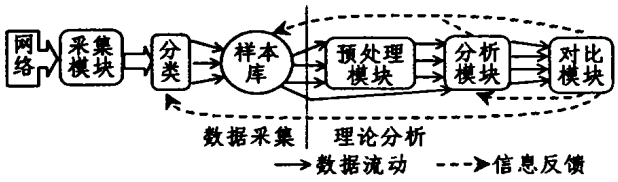

图2 网络行为研究环境的体系结构

采集模块的主要任务是从实际网络系统中收集样本数据。这一任务主要是由一个或多个网络管理监测工具承担。不同的研究内容对样本数据采集方式和采集内容有不同的要求。但一般一个功能强大、方便易用的管理监测软件就可胜任多种网络采集管理工作。例如，目前我们使用的是 Network General 公司的 Netxray 网管软件。它可以安装在基于 NT/98环境下的服务器或客户机上，对主机间的通信进行完全的监视，捕捉几乎任何类型的报文。Netxray 的主要功能有：

1)对各种报文的头字节进行详尽的解析，通过统计分析后，从不同角度得到多种反映网络特征的信息；

2)提供了各种网络流量指标不间断的历史数据记录；

3)可以主动发送自定义数据包，按要求制造各种类型的网络流量；

4)对报文的采集允许设置多种过滤条件，实现针对性的监测分析；

5)设置不同的时间间隔来调整采样的周期；

6)以多种图表方式动态实时显示网络上数据流动的特征。

7)可以根据设定好的流量模式向网络环境发送数据流。

Nexray 提供的大量针对流量的功能与数据，为网络行为特别是网络流量行为研究提供了方便。但它也有不足之处，如不能进行分布式数据的采集。根据不同需要或实际情况可以使用不同网络管理监测设备。

分类模块中，将采集模块所提供的原始数据按照一定的分类原则分类。对原始数据进行分类主要是为了从不同角度进行网络行为的研究。不同类别的原始数据可以体现不同情况下的网络特征。以下列举由 Nexray 工具所提供的原始数据的各种分类情况。

按流量指标分类主要是针对采集得到网络数据流量的各类指标，如 Packets/s(每秒数据包个数)，Octets/s(每秒传送的字节数)，Runt/s(每秒的过短包计数)，Collisions/s(每秒冲突的包的个数)，Utilization/s(每秒网络的利用率)等等，它们形成了各自的数据类别；

按照反映网络正反常两方面分为正常指标和异常指标两类数据；

按照采集的时段分类，可得到不同时段的数据，同时它们还可以归结为某一指标如流量的高峰时段和常规时段样本数据；

按照对比原始样本数据而发现特征来分类，如分成突发流量的样本和平稳流量的样本。

通过对报文解析可以得到报文原地址和目的地址、报文类型、报文内容等信息，这些信息可以根据不同的需要进一步被确定分类原则；

最后，还可以根据在分析对比模块中所发现的规律模式的反馈，来对原始数据进行分类。这表示分类过程中的分类原则在网络行为的研究过程中不是一开始就固定的，而是可以不断地按照分析要求生成新的分类原则。

样本库负责存放监测采集到的样本数据。被分类的数据保存在相应类别的文件中，纳入样本库内，并建立相应的检索机制和样本说明。样本库内样本数据文件的有序整理使得后面的处理变得灵活、方便。可以根据分析和对比的需要调出样本库内不同类别的原始样本。同时样本库使采集的数据不易被丢失或遗忘，且能在不同的实验项目中得到重用。一个完善的样本库可以为不同的理论分析提供统一的数据界面，因而方便了不同研究的转换。

预处理模块是样本数据和分析数据的衔接转换部分，在进行分析之前根据具体分析需要和样本类别的不同，模块负责准备不同的预处理程序以对样本库的原始数据文件进行预处理，转化为可供分析的样本数据。这些预处理包括改变数据文件的格式，如过滤数据，选择字段；对文件数据进行标准化处理，减少干扰因素等等。

分析模块是理论分析的确定和实现部分，对处理过的样本数据进行分析得出网络行为的建模或研究结果。网络行为的研究需要利用如协同学、混沌动力学理论、耗散学理论、自组织理论这些系统研究理论中的相应内容，来研究网络行为中的各种复杂非线性现象。如构造时间序列的相空间[5]，计算反映自相似性的分形特性[6]，寻找反映流量特征的宏观参数，计算反映系统混沌状况的李雅普诺夫指数，产生能反映自组织特性的图形等等。分析模块负责对比学习不同理论，确定对应理论模型，实现相应的分析算法。样本库中不同类型的样本文件经过计算分析生成了各自的实验结果，送交对比模块。

对比模块用来对比不同类别之间或同类之间网络参数的特征从而进一步研究网络行为的变化规律，或是将分析模块所得到的模型和实际模型进行对比，从而获得网络行为研究的结论及一些模型的评价结果。根据对比的需要，还可以来进一步指导样本数据的选取，分类和分析算法的确定。

4. 具体应用实例

在网络行为研究环境下，针对研究不同网络流量数据的变化特性，我们计算了不同时间段内的所观测局域网内通过某一节点的数据包个数的分形维值，并且进行分析对比来研究流量行为。

选取一个网络业务量适中，使用时间相对集中的小型局域网为实验对象，通过采集模块(使用 Netxray)，按照一定的采样时间间隔，我们得到了 Octets/s 的历史记录，实验中选择采样间隔为1秒。对样本数据按照不同的时段分类并保存在样本库的对应表中，建立查找表来确定时段与表名的映射关

(下转第63页)

C E}，{B C D}}，则由1)insert 操作产生的 LX_k={{A B C D}，{A C D E}}；由2)修剪过程对 LX_k 作修剪得到 LX_k={{A B C D}}，因为{A C D E}的(3－1)子集{C D E}不在 LX_{k-1} 中，故{A C D E}从 LX_k 中被删除，这样处理的依据是性质1。

过程 prod_subset(LX_{k-1})的作用：去掉满足性质2的导出型关联规则。

上述算法中 search_hash_table，hash_func 的描述略。

3 算法分析

对型如 X={A_1，A_2，…，A_n}的频繁项，对应的规则为{A_{i1}，A_{i2}，…，A_{ip}}=>{$A_{i(p+1)}$，$A_{i(p+2)}$，…，A_{in}}，其中 p=1，2，…，n－1，$A_{ij}\in\{A_1,A_2,\cdots,A_n\}$，j=1，2，…，n，$A_{ij}\neq A_{ik}$(j≠k)。

(1)若所有形如{A_{i1}，A_{i2}，…，A_{ip}}=>{$A_{i(p+1)}$，$A_{i(p+2)}$，…，A_{in}}的规则都是关联规则，则 Apriori 生成方法会产生2^n-2条关联规则，因为前件元素个数 p 为1的关联规则有 C_n^1，为2的关联规则有 C_n^2，依次类推，总的关联规则数=$C_n^1+C_n^2+\cdots+C_n^{n-1}=2^n-2$。而本方法仅产生 n 条关联规则，{$A_{ij}$}=>{$A_{i1}$，$A_{i2}$，…，$A_{i(j-1)}$，$A_{i(j+1)}$，…，$A_{in}$}。

(2)若形如{A_{i1}，A_{i2}，…，$A_{i(j-1)}$，$A_{i(j+1)}$，…，A_{in}}=>{A_{ij}}的规则才是关联规则，则两种方法产生的关联规则数目是一样的，因为此种关联规则不能导出任何其他关联规则。除此之外，本文给出的方法产生的关联规则数目都会比 Apriori 少。

(3)本方法产生的关联规则数最多为 $C_n^{[n/2]}$条，如 X={A，B，C，D}，无论支持度如何分布，用本方法产生的关联规则数最多为 C_4^2=6条。因为，若要产生的关联规则多，则必须使形如{A_{i1}，A_{i2}，…，A_{ip}}=>{$A_{i(p+1)}$，$A_{i(p+2)}$，…，A_{in}}的关联规则所导出的关联规则越少，也即前件的元素个数要尽可能多，后件的元素个数尽可能少，极端情况下为(2)所示的形式，但此时总的关联规则数是最少的，因此前件的元素个数不能太多，后件的元素个数不能太少，必须是使 C_n^p(p 为前件的元素个数)达到最大值的 p，从组合规律可知，p=[n/2]时，C_n^p有最大值。该结论的另一作用在于确定 Hash 表的大小。

(4)Hash 表的大小可设为|Ht|=$|L_m|*C_m^{[m/2]}$。Hash 策略可采用如下策略：

addr(A_1，A_2，…，A_p)=(order(A_1)*10^0+ order(A_2)*10^1+…+ order(A_p)*$10^{(p-1)}$) mod |Ht|。

结论 作者在深入研究关联规则特性的基础上，发现了关联规则的冗余特性，本文给出了相关的定义及其若干性质的数学证明，并利用这些性质提出了一个有效地减少关联规则数目的生成算法。该方法生成的关联规则数目虽然减少了，但并没有遗失与之相关的知识，若需要，可以利用这些规则自动生成那些被消除的关联规则。此外，由于产生的关联规则数目显著减少了，因此生成关联规则所耗费的时间也自然地减少了。作者在 Win98环境下利用 vc6.0实现了该算法，运行表明了该算发的正确性和有效性。

参考文献

1 Agrawal R，Imielinski T，Swami A. Mining Association Rules between Sets of Items In Large DataBase. In：Proc. ACM SIGMOD Conf. on Management of Data，Washington，D. C. 1993. 207～216

2 Agrawal R，Srikant R. Fast Algorithms for Mining Association Rules：[IBM Research Report RJ9839]. IBM Almaden Research Center，San Jose，Calif. 1994

3 Agrawal R，Mannila H. Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining. AAAI Press /The MIT Press，1996

4 Jong S P，Philip S Yu. Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. IEEE Transactions on Knowledge and Data Engineering，1997，9(5)

5 欧阳为民，等. 国际上关联规则发现研究综述. 计算机科学，1999，26：41～44

(上接第94页)

系。对每个样本的流量进行标准化处理后，构造流量时间序列相空间。根据分形维的定义实现计算分形维的算法，生成分形维数。之后将数值生成图表进行下一步的处理。数据如表1所示。

将得出的数值递交对比模块进一步地分析得出结论，来决定是否再要计算其它采样间隔时间其它时间段的分形维，如以天为单位，或是否要改变监测采集方式，如监测其它指标。并根据分形维的变化情况来得出流量变化的规律。我们将在环境中利用这些实验得出的数据，根据分形维的变化情况来深入研究流量行为变化规律。

表1 样本分形维数值表

采样时段	采样间隔	样本维数
9:00 -10:00	1s	约5.07
13:00 -14:00	1s	约 4.26
16:00 -17:00	1s	约 7.38
19:00-20:00	1s	约 6.25

结束语 本文根据网络行为研究需要设计并实现了一个网络行为建模环境，有效地融合了网络数据采集监测功能和理论模型分析功能，为网络行为长期性的研究创造了一个有效、互动、稳定的研究环境。我们还在这一环境中，对不同时段所采集的网络流量时间序列进行分析，得到了一些网络流量模型的特征参数分形维数。对网络行为的研究还处于起步阶段，我们希望网络行为研究环境的建立能为实际的研究工作提供方便，同时我们也将不断地应用新的技术和理论来充实和完善现有的研究环境。

参考文献

1 Taqqu L，W. On the Self-similar Nature of Ethernet Traffic(Extended Version). IEEE/ACM Transactions on Networking，1994，2(1)：1～55

2 Huberman B A，Lukose R. Social Dilemmas and Internet Congestion. Science，1997，277：25

3 Veres A. The Chaotic Nature of TCP Congestion Control. IEEE INFOCOM. 2000：1715

4 袁坚，任勇，山秀明. 一种计算机网络的元胞自动机模型及分析. 物理学报，2000，49：398

5 陈式刚. 映象与混沌. 北京：国防工业出版社，1992. 180～181

6 Gao C，Cong S Wu. G. Measurement-based Multifractal Traffic Modeling. In：Proc. of the ISCA 13th Intl. Conf. Las Vegas，Nevada，USA，Aug. 2000：355～360