网络监控日志可视化分析

栾鑫月，王 美，聂俊岚

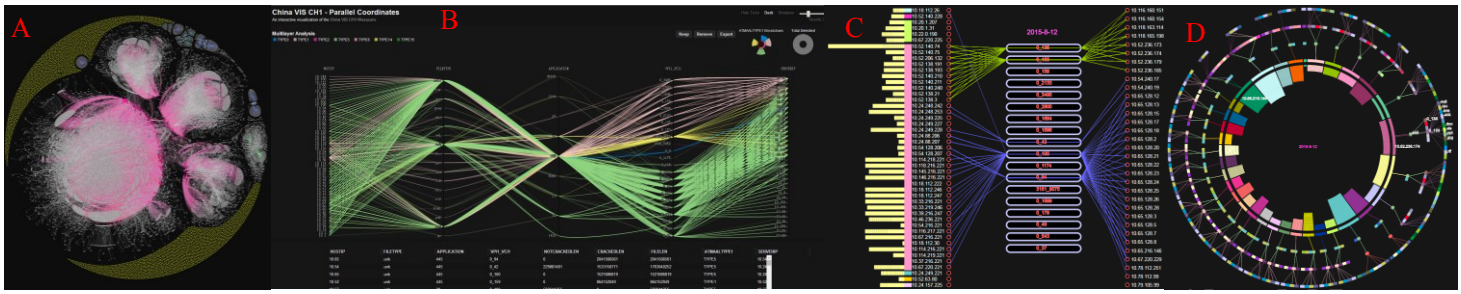

图 1.A：网络结构力导向图，B：结合玫瑰图的平行坐标系，C：HSDC 图，D: 环形坐标系

摘要—基于 BigBusiness 公司两个月网络监控日志数据，对内部网络的通信进行分析。使用主机的度、主机通信次数、主机的各端口使用率对服务器进行初步筛选，适当的调整后，将其在图 1A 中标出，验证了选出服务器的正确性。用结合了玫瑰图的平行坐标系、HSDC（Host-Server-Daily-Communication）图和环形坐标系来展示数据的多层次属性之间的关系，以更好地探索用户的行为模式。HSDC 图和环形坐标系还分别实现对多用户多层次和单用户多层次连接传输的实时模拟。

关键字—服务器识别；可视分析；网络异常；行为模式；环形坐标系

1 引言

ChinaVis 2016提供了由网络监控系统在BigBusiness公司的骨干通信链路上抓取的数据包信息，它记录数据包在链路层、网络层和应用层的相关信息[1]。通过分析该监控系统试运行期间两个月的网络监控日志，识别该公司内部网络中的客户端与服务器，实现对数据连接的多层次可视分析，探索其可能反映的用户行为模式。

2 概述

首先对数据进行清洗，根据主机间的通信画出力导向图，然后使用基于主机的度、主机通信次数、主机的各端口使用率的公式进行服务器的初步筛选，适当的调整后，将其标于力导向图中，如图1A所示，验证选出服务器的正确性。在可视化设计阶段主要使用热度图、堆叠图、折线图和力导向图来发现网络中的异常。树地图适于展现经同一虚拟管道进行通信的源IP和目的IP的分布规律和连接模式；甘特图、河流图可以更简明地展示虚拟管道随时间的推移而出现的各种模式。用结合了玫瑰图的平行坐标系、HSDC图和圆形坐标系展示数据的多层次属性之间的关系，以更好地探索用户的行为模式。HSDC图和环形坐标系还分别实现对多用户多层次和单用户多层次连接传输的实时模拟。

3 区分客户端与服务器

3.1 客户端与服务器的识别方法

由于通常情况下服务器会与较多的主机通信，且根据其提供服务的不同，某个端口的使用率会远高于其它端口；我们用下面的公式来识别内网中的服务器：

$$Z_i = ((r_i - r) > 0) \wedge ((d_i - d) > 0) \wedge ((c_i(p_j)) \Big/ \sum_{j=m}^{k} c_i(p_j) - w) > 0)$$

通过上述公式计算每台主机的Z值，Z值大于0，该主机为服务器，否则不是。上式中：r_i、d_i、$c_i(p_j)$分别代表主机i的通信次数、度、使用端口 p_j通信的次数； r、d和w分别表示主机通信次数、度和端口使用率的阀值。本题中r=53*2=106次,d取通信次数大于106的主机的度的期望值，w=0.5。

为保证服务器对整个网络提供服务的全面性，针对不同的应用，分别在力图中将找到的服务器标出，对结果进行筛选、补充。最后在力图表示的内网中将找到的服务器用粉色结点标出，将内网的客户端用白色结点标出进行验证，结果如图1中A图所示，图中红色的结点位于主干网上，且能把内网的各个主机连接起来，验证了所选出的服务器的正确性。

4 可视化设计

4.1 力导向图

用节点表示主机，用节点之间的连线表示二者的通信关系，绘制出网络中主机通信的结构图。观察其结构，发现独立于主干网络的子网，进一步分析发现，此子网存在时间较短，认为这些子网是异常的结构。图1中A图绿色和蓝色结点所在的子网即为这种异常结构。

4.2 堆栈图

堆栈图可以发现随着时间的推移，各个组成部分数量的变化及总量的变化。图 2 展示不同的端口每天的使用次数。

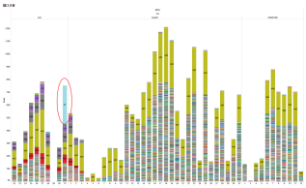

图 2：端口访问量异常

- *栾鑫月，燕山大学， E-mail: 464176845@qq.com.*
- *聂俊岚，燕山大学 教授, E-mail: niejll3@163.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.

4.3 折线图与河流图相结合

由于折线图可以清晰地展现变化量较大的数据，而河流图又可以展现数据的组成成份。我们使用如图3所示的折线图和河流图来发现网络中的流量异常，左侧“流量分析—总”视图展示了内网中的总流量随时间的变化，使用此视图，我们可以确定异常点（左侧紫框内），在右侧的源和目的IP段进一步分析，锁定到发生异常的确切的主机（右侧紫框内）。此方法也可用来发现内网通信次数异常行为。

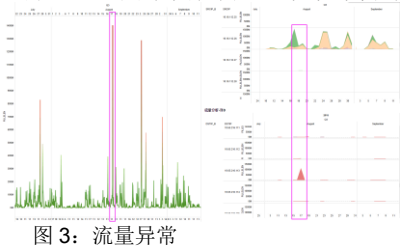

图 3：流量异常

4.4 双向柱形图

用图4所示的双向柱形图来展示虚拟管道的链路层数据传输模式及文件传输模式，上方的柱形图表示虚拟管道总的传输流量情况，不同颜色的柱形条表示不同的链路层数据类型。上方各个柱形条颜色唯一，可知各虚拟管道传输的链路层数据的类型是唯一的。下方的柱形图表示每个虚拟管道传输的传输层不同类型的文件产生的流量，柱形条上不同颜色的部分表示不同的文件类型。由图可知，各虚拟管道传输文件类型不唯一。该图不仅可以详细展示两种模式，还可以清晰对比两种传输模式。

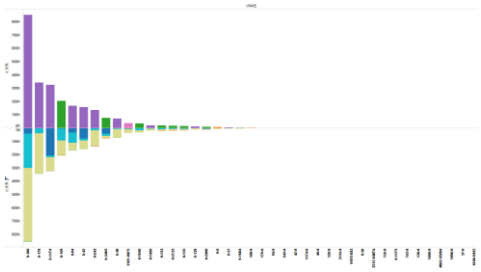

图 4：虚拟管道传输数据的链路层类型及传输层文件类型

4.5 河流图

虚拟管道通信质量变化主要表现在单位时间内当虚拟管道的使用次数或传送数据的流量变化时，通过虚拟管道的数据包的被损坏率是否随之增加。使用河流图可以清晰地展示这一变化，绘制各虚拟管道通信质量的河流图，如图5所示，发现，部分虚拟管道传输的数据包被损坏的比重不会因使用次数剧增而增加，而部分虚拟管道传输的数据包被损坏的比重则会因使用次数剧增而增加。

图 5：虚拟管道单位时间通信次数与数据包损坏率的关系

4.6 HDFC(Host-Daily-Flow-Coummunication)图

图6中，右侧上、下两个热度图展示主机每天的流量、通信次数，供用户查看总体情况。点击热度图上某一点时，出现左侧的图，图中的玫瑰图表示此主机在当天24小时主动和被动通信的次数。饼图表示该主机与各个服务器通信的流量的多少。下侧两个弧形图展示与其通信的所有主机，弧的宽度代表流量的大小。红色和蓝色的弧线分别连接的是目的IP和源IP。该图可以清晰地对比不同时间或不同用户间的行为。

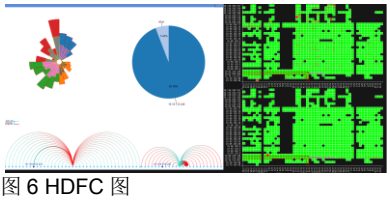

图 6 HDFC 图

4.7 结合玫瑰图的平行坐标系

如图1B所示，平行坐标系从左到右的五个轴分别表示客户端所在网段、文件类型、应用类型、虚拟管道、以及服务器网段，线的颜色表示链路层数据类型；玫瑰图表示每种链路层数据的比重和传输成功与失败的比重；环图表示选中的记录数占总记录数的百分比。

此模块既可以查看数据的整体关联，又可通过交互分析任意两种属性的关联。玫瑰图和环图的使用，最大程度地实现了对数据的多维分析。

4.8 HSDC 图

如图1C所示，右边的IP表示客户端；中间有管道标识的为虚拟管道；虚拟管道上方为当前时间标注；左边的IP表示服务器；左侧的矩形条带表示不同的应用类型；横置的柱形图表示各个服务器的相应应用截至当前时间的流量累加；客户端与虚拟管道、服务器与虚拟管道间连线的颜色表示数据链路层的数据类型。

HSDC图可实时模拟多用户多层次传输，直观地展示了数据在应用层、链路层和网络层的对应关系。

4.9 环形坐标系

如图1D所示，第一层（以最外层为第一层向里依次递增）为每个用户传输文件的类型；第二层为每个用户传输数据使用的虚拟管道；第三层为每个用户使用的不同应用；第四层为用户IP地址；第五层为截至当前时间用户传输的流量累计；圆心为当前时间标注。

环形坐标系可实时模拟单用户多层次连接，提取每个用户的数据单独建立平行坐标系，环形布局多用户的平行坐标系，既有助于单用户行为模式分析，也有利于对多用户行为的整体感知。

5 讨论

本文从网络数据属性出发，结合实际问题，设计了合适的可视方案，主要特点如下：

在服务器判定过程中，本文提出的识别服务器的方法综合考虑了服务器应具有的各种特征，有效地实现了对服务器的初步筛选；

在异常查找中，按照从整体到局部再到个体的顺序设计可视方案，达到了快速有效找到产生异常的根源的效果；

在虚拟管道模式分析中，根据数据的不同特性设计不同的可视方案，最大程度地展现了数据的多维特性，有效地对数据进行了多角度分析；

为满足多层次数据可视分析，以客户端、服务器、虚拟管道为载体，从实时监控和结论分析的角度设计了HSDC图、环形坐标系以及交互式的玫瑰图和平行坐标，既详尽展示了数据的时序性，也不丢失数据的总体特性。

综上，此系统达到了实用性，有效性，新颖性的要求。

6 总结

本文提出了判定服务器的方法，根据多层次网络监控日志特性设计了合理的可视方案实现对网络日志的可视分析。有助于网络管理人员识别服务器、分析服务器和客户端特性、理解网络结构、发现网络中的异常通信模式，从而有效、正确地评估网络的运行状况。

7 参考文献

[1] ChinaVis 2016. http://chinavis.org/2016/