基于流数据挖掘的网络流量异常检测及分析研究

魏桂英, 姜亚星

(北京科技大学 经济管理学院, 中国 北京, 100083)

[摘 要]网络流量异常检测及分析是网络及安全管理领域的重要研究内容。本文探讨了网络流量异常的种类、网络流量异常检测的方法, 分析了基于传统检测方法在网络流量异常检测应用中存在的问题。并重点对基于流数据模型的网络流量异常检测进行了研究, 综述了已有流数据挖掘研究方法在网络流量异常检测中的研究进展。最后, 本文对现有研究工作存在的问题及未来的研究方向进行了探讨。

[关键词]网络异常; 异常检测; 流数据; 流数据挖掘

doi: 10.3969/j.issn.1673-0194.2009.15.012

[中图分类号] TP274; TP393.06 [文献标识码] A [文章编号] 1673-0194(2009)15-0039-04

1 引 言

随着 Internet的快速发展与日益普及, 越来越多的信息通过网络来传输和存储, 网络安全越来越重要。网络流量异常检测及分析是网络及安全管理领域的重要研究内容。网络流量突发异常是指网络业务流量突然出现的不正常的重大变化。及时发现网络流量的突发异常变化对于快速定位异常、采取后续相应措施具有重要意义。

目前网络规模和速度的不断增加, 流量突发异常检测算法需要实时准确地分析处理海量的网络业务量数据, 具有很大的挑战性。流数据模型的提出[1], 使得采用流数据模型来描述网络通信量, 解决现有网络流量异常检测模型存在的不足成为可能, 基于流数据挖掘的网络流量异常检测及分析得到了广泛的研究。

1.1网络流量异常分类

网络流量异常是指对网络正常使用造成不良影响的网络流量模式, 常见的网络流量异常有如下几类:

(1)网络扫描。网络扫描是一种常见的网络异常流量, 它表现为在单位时间内, 同一个源 IP访问大量不同的目标 IP或同一目标 IP的不同端口, 目标 IP通常是连续的。

(2) DDoS攻击。拒绝服务攻击通常以消耗服务器端资源, 迫使服务停止响应为目标。它表现为大量不同的源 IP对同一目标 IP发送数据包。单位时间内数据包的量大, 数据包长度长, 占用大量的带宽资源。

(3)网络蠕虫病毒。蠕虫病毒利用操作系统的漏洞主动传播, 并且可以在局域网或者广域网内以多种方式传播。这种网络蠕虫病毒的攻击方式, 除了造成大量的网络流量外, 也会消耗大量的系统资源。而且这类异常通过局部链路上的流量测量数据很难检测, 往往需要对全网的流特征进行分析或采用全网的流量统计分析方法进行检测。

(4)由网络故障和性能问题造成的异常。典型的网络性能异常是文件服务器故障、网络内存分页错误、广播风暴和瞬间拥塞等引发的网络流量行为的异常。另外, 恶意下载、对网络资源的不当使用, 会造成流量异常, 导致网络带宽浪费。

1.2网络流量异常检测方法

近年来, 有很多针对网络流量异常检测的研究工作, 概括起来, 针对网络流量异常检测的方法主要有以下几种[2]: 基于特征 /行为的研究方法、基于统计的异常检测、基于机器学习的方法和基于数据挖掘的方法等。

基于特征 /行为的检测研究通过在网络流量数据中查找与异常特征相匹配的模式来检测异常。因此需要分类描述网络异常的流量的特征及行为特征、构造蠕虫分类和 DDoS攻击行为等, 其缺点是无法检测出未知的攻击类型, 而且需要对规则特征库不断进行更新。

基于统计的研究不需要事先知道异常的特征, 使用时间序列的流量数据, 采用统计分析技术检测异常。机器学习的方法更强调如何基于更新的信息和以前的结果来提高系统的性能, 异常检测中常用的机器学习技术包括基于系统调用的序列分析、贝叶斯网路、主成分分析法、马尔可夫模型等。数据挖掘技术可以用来从大量审计数据中挖掘出正常或入侵性质的行为模式[3]。后 3种异常检测的方法是对正常的系统网络行为进行建模, 通过与正常模型的比较来进行异常检测, 因此能有效地发现已知和未知的攻击。

上述传统网络异常检测方法, 通常是建立在对整个数据集进行等同学习的基础上的, 检测结果受历史数据

[收稿日期] 2009-06-18

[作者简介] 魏桂英(1969-), 女, 北京科技大学经济管理学院讲师, 在读博士, 主要研究方向: 数据挖掘、网络安全与管理、信息系统。

的影响较大，难以真实反映当前网络数据的行为特征。而检测网络异常行为是否发生，通常根据最近的网络行为就可以做出判断，并不依赖于整个历史数据集。另外，现有异常检测算法的时间、空间复杂性较高，且受内存等系统资源的限制，难于对持续、快速到达的大规模原始网络数据进行处理，不适合进行在线检测。

2 流数据模型及特点

流数据就是大量连续到达的、潜在无限的数据的有限集合。令 t 表示任一时间戳，a_t 表示在该时间戳到达的数据，流数据可以表示为：$\{a_1, \cdots, a_{t-1}, a_t, a_{t+1}, \cdots\}$。与传统数据相比，流数据有以下特点[4]：

（1）数据高速到达，实时性要求高。

（2）流数据是一种海量的数据，因此不可能对流数据的每一个数据项都进行存储。

（3）由于数据量无限增长，对流数据的扫描次数仅限于一次。

（4）流数据的无限性使得流数据挖掘无法保存原始数据，仅能在内存中保留原始数据的概要信息，并基于这些概要信息生成最终结果。因此，流数据挖掘结果实际上是在一定误差范围内的近似结果。

由于网络数据流符合以上的流数据特点，所以采用流数据模型来描述实际的网络通信量，解决现有网络流量异常检测模型存在的不足是非常合适的。

3 基于流数据挖掘的网络异常检测研究进展

目前基于流数据挖掘的网络异常检测研究工作主要从以下几个方面展开：流数据概要结构设计、流数据变化挖掘、流数据聚类挖掘、频繁项挖掘以及多维流和多流的挖掘。

3.1 基于流数据概要结构设计的研究

流数据的特性决定了流数据算法的核心是设计高效的单遍数据集扫描算法，由于流数据量远大于可用内存，系统无法在内存中保存所有扫描过的数据，因此流数据概要结构设计是流数据挖掘的基础和首要工作。

针对网络流量的异常检测研究中，基于采样、小波变换、哈希函数等概要结构设计得到了广泛研究。文献[5]提出了基于概要结构数据的变化检测算法，设计了一个概要数据结构的变体——k-ary Sketch，减少了算法所需的时间和空间复杂度。文献[6]提出了一种改进的概要数据结构——The Count-Min Sketch（CM概要），该结构可用于网络大规模点击的发现。文献[7, 8]采用了层次化的概要技术，实现了在多维流中发现 HHH（Hierarchical Heavy Hitters，层次大流）的算法；文献[9]提出一组位图算法，解决使用小存储空间实现高速链路上数据流的聚类问题。文献[10]提出了两层小波树摘要数据结构及基于此结构的突发检测算法；并基于该算法设计了一个网络流量突发检测原型系统。文献[11]设计了一种面向大规模网络异常发现的高频概要算法（FSA, Frequent Sketch Algorithm），进行网络突发高频事件检测。

3.2 流数据变化挖掘方法

流数据变化检测及异常发现是流数据挖掘研究领域中的一个重要分支，流数据突发检测属于一种形式特殊的流数据异常变化检测，是指发现流数据中的异常数据聚集。在网络管理中突发检测可应用于对短时间内丢包个数进行监控。

文献[12]根据流分布特征变化确定网络异常。同时，还提出了多路子空间方法（multiway subspace method）在多流特征中提取异常变化。文献[13]也提出了通过流分布异常在网络数据流中发现重要变化的算法，该算法的思想是在高速数据流中发现最重要的 Deltoids，通过 Deltoids 的确定发现网络流量在接口之间以及路由器之间在一段时间内的重要不同。文献[10]针对网络流量数据海量、高速的特点，提出了基于两层小波变换数据结构的突发检测算法。文献[14]提出了基于偏差的异常检测方法，并采用 M—树为存储结构建立正常的用户行为轮廓。

3.3 基于流数据聚类挖掘的方法

基于聚类分析的异常检测在网络异常检测领域已经得到了深入的研究。例如，按数据流中的属性特征将数据流分类聚合，挖掘具有某种异常流量模式（资源使用特征）的聚合流，从而检测及确定网络流量的异常行为；挖掘在特定时间间隔，在某一给定链路上的统治的聚合流，即大流量流（HH, Heavy Hitter）或出现频率高的流。如果某一大流是通过很少的几个端口发送的，这通常代表某一特定的蠕虫的特征；如果异常的大流发送到某一特定 IP 地址，这种情况常指示 Flash 拥挤或 DDoS 攻击。

文献[15]提出了一种多维流量聚类方法，可以从多个不同维度（源地址，目的地址，协议，源端口，目的端口）对流量进行分析，能够基于对实际流量的分析，确定出在特定时间的统治流（大流）或不正常流。文献[16]提出了在大流聚类中检测变化的技术。

文献[17]提出了一种基于数据流聚类的两阶段异常入侵检测方法，首先在线生成网络数据的统计信息，并利用最能反映当前网络行为的统计信息检测入侵行为。该两阶段模型不但能够存储与维护海量原始网络数据的统计信息，而且减少了历史数据对检测结果的影响，提高了入侵检测精度。文献[18]使用聚类离群点算法来分析实时数据与正常数据的偏离程度。

3.4 基于频繁项挖掘的方法

频繁项挖掘主要用于网络流量监控、计费和异常检测中，监测出现频率超过某个设定阈值的流，或者以近似线性的速度识别大流，从而指出某种网络异常的发生。

文献[19]提出了在流数据中近似计算数据项出现次数超过用户给定阈值的频繁项挖掘算法。文献[20]

提出了动态跟踪频繁项——“热门元素”的算法。“热门元素”指频繁出现的元素或出现次数超过某个阈值的元素，在网络数据流中主要指频繁出现的 IP地址。文献[21]通过采样、基于哈希函数运算过滤数据包等方法，计算统治的聚合数据流，目标是使用有限的存储空间支持高速链路上数据流的统计问题。文献[7]提出了在多维流中发现 HHH的算法。文献[16]提出了一维和二维 HHH的在线识别算法，并将该算法应用于在大流中进行变化检测。

3.5基于多维流挖掘和多流挖掘

按流数据挖掘时所依据的网络流量数据的属性可将挖掘分为一维流挖掘和多维流挖掘。一维流挖掘，即按某一特定属性的聚合，进行流挖掘。例如，挖掘从某一 IP地址发出的数据流量超过某个阈值的流。多维流挖掘，即根据多个属性的聚合，进行流挖掘。例如基于源 IP、目的 IP和目的端口三维的大流量流挖掘是指找出所有从某源 IP主机发往某目的 IP主机的某目的端口的流量超过某个阈值的聚合流。一般多维流挖掘算法是基于一维流挖掘实现的。

文献[15]研究了从多个不同维度（源地址，目的地址，协议，源端口，目的端口）对流量进行分析，并确定出在特定时间的统治流（大流）或不正常流的多维流量聚类方法。文献[7]采用多维层次化的概要技术，实现了在多维流中发现 HHH的算法。文献[16]在研究一维聚集流发现算法的基础上，探讨了多维聚集流发现算法。

另外，对于海量网络业务量数据来说，通过单一数据流很难确定大规模网络的整体状态，例如基于目的 IP的高频突发项检测可以发现 DDos攻击，然而对某个网站的大规模突发访问、突发访问和 SYN—Flood攻击的流量模式在 IP上很难区分，必须通过多数据流信息关联进行综合判断。目前基于多流的网络异常检测研究还相对较少。

4 结束语

基于流数据挖掘技术进行网络流的异常检测及分析是目前研究的重点，并且已经取得了一定的研究成果。未来需要进一步研究的课题有：

（1）高效流数据挖掘算法的研究。在高速网络环境下，为适应大流量网络链路的异常检测的响应时间要求，需要进一步研究高效流数据挖掘算法减小存储开销及降低分析处理算法的复杂性。

（2）探讨新的异常检测方法。随着各种新病毒、攻击等网络安全问题的不断出现，网络异常检测的方法手段也需要不断更新，同时需要扩展研究基于流量的异常检测的内容，例如，基于多维、多流的检测研究等。

（3）异常分析的研究。目前的研究工作多数集中在网络流异常分析的检测，需进一步加强异常确定、定位、异常类型分析及网络异常响应等方法及技术的研究。

（4）研究设计能够有效降低误报率的算法。进一步提高检测的精度，同时消除数据噪声对网络流量异常检测造成的影响。

主要参考文献

[1] Henzinger M, Raghavan P, Rajagopalan S. Computing on Data Streams[M] // External Memory Algorithms, Dimacs Series in Discrete Mathematics and Theoretical Computer Science. Boston: American Mathmatical Society, 1999.

[2] Animesh Patcha, Jung-Min Park. An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends [J]. Computer Networks, 2007, 51(12): 3448-3470.

[3] Lee W, Stolfo S. Data Mining Approaches for Intrusion Detection [C] // Proc 7th USENIX Security Symposium. San Antonio, TX, 1998.

[4] Babcock B, Babu S, Datar M, et al. Models and Issues in Data Stream Systems[C] // In Proc of ACM PODS, 2002.

[5] Krishnamurthy B, et al. Sketch-based Change Detection: Methods, Evaluation, and Applications[C] // ACM/USENIX Internet Measurement Conference, 2003.

[6] Cormode G, Muthukrishnan S. Improved Data Stream Summaries: The Count-min Sketch and its Applications[J]. Journal of Algorithms, 2005, 55(1): 58-75.

[7] Cormode G, Korn F, Muthukrishnan S. Diamond in the Rough: Finding Hierarchical Heavy Hitters in Muti-Dimensional Data[C] // Proc ACM SIGMOD, June 2004.

[8] Cormode G, Korn F, Muthukrishnan S, et al. Finding Hierarchical Heavy Hitters in Data Streams[C] // International Conference on Very Large Databases, 2003: 464-475.

[9] Datar M, Muthukrishnan S. Estimating Rarity Similarity over Data Stream Windows[R]. DIMACS Technical Report, 2002.

[10] 陈婷婷. 基于数据流的网络流量突发异常检测[D]. 哈尔滨：哈尔滨工业大学，2006.

[11] 郑军，胡铭曾，等. 基于数据流方法的大规模网络异常发现[J]. 通信学报，2006, 27(2).

[12] Lakhina A, Crovella M, Diot C. Mining Anomalies Using Traffic Feature Distributions [C] // ACM SIGCOMM, Philadelphia, August 2005.

[13] Cormode G, et al. What's new: Finding Significant Differences in Network Data Streams[J]. IEEE/ACM Transactions on Networking, 2005, 13(6): 1219-1232.

[14] 生若谷. 一种网络异常数据流的检测和控制技术[D]. 哈尔滨：哈尔滨工业大学，2007.

[15] Estan C, Savage S, Varghese G. Automatically Inferring Patterns of Resource Consumption in Network[C] // Proceedings of ACM SIGCOMM, 2003.

[16] Zhang Yin, Singh S. Online Identification of Hierarchical Heavy Hitters: Algorithms, Evaluation, and Applications[C] // IMC'04, October 2004.

[17] 俞研，郭山清，黄皓. 基于数据流的异常入侵检测[J]. 计算机科学，2007, 34(5).

[18] 袁福宇. 流数据环境下不确定性入侵检测框架[D]. 长春：吉林大学，2007.

[19] Manku G, Motwani R. Approximate Frequency Counts over Data Streams[C] // Proceedings of 28th International Conference on Very Large Data Bases, 2002, 346-357.

两阶段法求解混装模式下的加热炉调度

梁合兰[1], 李苏剑[1], 邓又好[2]

(1. 北京科技大学 机械工程学院物流研究所, 中国 北京 100083;
2. 首钢迁安钢铁有限公司, 中国 河北 迁安 064404)

[摘 要] 针对目前加热炉调度模型少有考虑混装模式下加热炉调度优化的不足, 建立了连铸—热轧混装一体化模式下的加热炉生产调度优化模型, 并提出了基于贪婪算法和模拟退火算法的两阶段求解方法。生产数据测试表明该模型和算法能有效解决加热炉调度问题。

[关键词] 混装模式; 加热炉; 生产调度优化; 多阶段决策
doi: 10.3969/.j.issn.1673-0194.2009.15.013
[中图分类号] TP399; F273.1; TB112 [文献标识码] A [文章编号] 1673-0194(2009)15-0042-03

1 引 言

加热炉位于钢铁生产中连铸工序和热轧工序之间, 是钢铁生产中的高能耗环节[1]。目前有关加热炉的研究主要从热能的角度侧重于加热炉加热优化控制研究[2], 但对于加热炉的生产调度优化的研究相对较少[3-4]。本文以一体化模式下的加热炉优化调度为研究对象, 建立了连铸—热轧混装一体化模式下的加热炉生产调度优化的数学模型, 并应用基于贪婪算法和模拟退火算法的两阶段方法求解。加热炉生产实际数据仿真结果表明, 该算法能有效求解加热炉调度优化问题。

2 混装模式下的加热炉物流分析

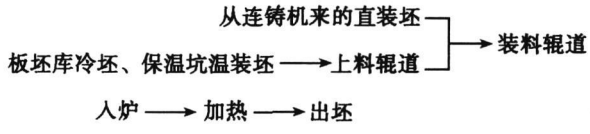

图 1 混装模式下的加热炉调度过程

连铸—热轧混装一体化模式下加热炉调度过程如图 1所示。混装模式下原料既有从连铸机产出的直装坯, 也有库内板坯。其中, 直装坯通过输送辊道直接进入加热炉加热; 板坯库冷坯、保温坑温装坯先进入上料辊

[收稿日期] 2009-06-18

[作者简介] 梁合兰(1981-), 女, 广东江门人, 北京科技大学机械工程学院物流研究所博士研究生, 主要研究方向: 钢铁行业生产调度优化、智能算法等; 李苏剑(1959-), 男, 山西太原人, 北京科技大学机械工程学院物流研究所教授, 博士生导师, 主要研究方向: 物流管理、企业信息系统建模等。

[20] Cormode G, Muthukrishnan S. What's Hot and what's not: Tracking most Frequent Items Dynamically[C] // Proc ACM PODC 2003, July 2003.

[21] Estan C, Varghese G. New Directions in Traffic Measurement and Accounting[C] // Proceedings of ACM SIGCOMM, Pittsburgh, PA, August 2002.

Research on Anomaly Detection and Analysis of Network Traffic Based on Data Stream Mining

WEI Gui-ying, JIANG Ya-xing

(School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, P. R. China)

Abstract: Anomaly detection and analysis based on network traffic are used for network and security management. In this paper, we give a list of 4 kinds of of traffic anomalies and the methods of traffic anomaly detection are presented. Also we give an analysis of the limitations in traditional anomaly detection methods. After that, the research work focus on the traffic anomaly detection methods based on data stream mining. The available research work of data stream mining methods for anomaly detection and analysis are summarized. Finally, we discuss the open problems and challenges in this area.

Key words: Nework Anomaly; Traffic Anomaly Detection; Data Stream; Data Stream Mining