DOI:10.19789/j.1004-9398.2019.05.005

基于准实时流量数据报出与信息熵技术的网络流量异常监测研究*

邓小清 王伦浪

(四川文理学院智能制造学院,四川 达州 635000)

摘要: 网络流量异常监测主要通过数据仓库、OLAP、数据挖掘、WEB、ETL等相关技术,来提高网络监测数据处理和存储的效率,同时还可以通过信息熵技术保证精确性、简洁性,能够确保精确性的前提下更为简洁高效. 针对这些技术,文中通过合理的架构,使得信息熵技术在功能性方面达到了决策更加有效合理的效果,同时还验证了信息熵技术框架的灵活性和可复用性.

关键词: 网络流量异常监测; OLAP; 信息熵

中图分类号: TP393

0 引 言

随着个人网络设备尤其是移动网络设备的快速普及和各种移动互联网应用的迅速发展,移动互联网已成为人们日常生活中离不开的信息传播平台. 然而,伴随着网络流量的增大以及各种类型流量的不断出现,网络流量中形形色色的异常流量也随之而来,严重影响了网络通信的质量和用户主机的安全使用. 攻击者会使用木马或者僵尸网络等技术控制相当大规模的僵尸计算机来攻击用户主机、投递垃圾邮件,更有甚者进行分布式拒绝服务攻击(distributed denial of service, DDoS),严重威胁用户主机甚至整个网络的安全[1]. 最近几年,随着硬件设备的提升和网络带宽的增加,网络攻击也在向着快速化、自动化、隐蔽化和复杂化发展,传统的基于特征分析的网络流量监测和攻击识别技术已不能达到现实的需要[2]. 目前,在网络攻击检测领域,网络异常检测方面,越来越多的人开始重视针对DDoS攻击的检测技术,DDoS攻击作为现今网络攻击最为主要的手段,其相应的检测技术已被资深网管和分析员所研究和利用[3].

1 相关理论概述

1.1 数据采集及传输

数据传输是指数据源与数据宿两者之间经由一条或多条线路进行数据传送,也可以理解为是一种把数据从某处送往另一处的操作[4]. 可以预计到,由于流量监测中心会在短时间内接收到来自网络设备子信息熵技术的大量数据传输,这就要求信息熵技术有能处理大量并发请求的能力,必要时需要能对接收到的数据进行异步处理[5].

数据采集由监测建筑中的各计量装置、数据采集器和数据采集软件组成,主要把各计量装置采集的流量数据采集到数据库中. 分项流量数据的采集频率为15 min/次~1 h/次之间,数据采集频率可根据具体需要灵活设置. 本信息熵技术采用异步传输和同步传输两种传输方式,以宽带传输的形式进行传输,以达到实现多路复用,增加信道容量,并且提高传输速率的目的[6].

1.2 数据处理和存储

数据的分析主要通过联机分析处理(on-line analytical processing, OLAP)和数据挖掘(data mining, DM)来实现[7].

1.3 数据仓库

数据仓库就是一个数据的集合,是面向主题的、相对稳定的、集成的、并且会随时间改变的一个数据集合,主要用于支持管理决策. 集成性是数据仓库所有特性中最重要的部分[8]. 面向应用的操作型数据库往往相互之间是独立并且异构的. 由于监测平台中所采集到的数据来自于不同的子信息熵技术,数量庞大,对数据的处理和存储有极高的要求,鉴于数据仓库的特性,本信息熵技术通过对使用单位需求的分析,建立企业数据仓库的逻辑模型和物理模型,并规划好信息熵技术的应用架构,将使

收稿日期: 2018-09-25

* 四川省教育厅一般项目(16ZB0360)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

用单位各类数据按照分析主题进行组织和归类[9].

1.4 OLAP 和 DM 技术

1.4.1 OLAP 技术

数据库之父 Codd 提出了联机分析处理的概念,并且他同时提出的相关 12 条准则将 OLAP 与 OLTP(on-line transaction processing)联机事务处理区分开来. 联机分析处理技术的显著优点在于它可以将数据分析结果可视化,分析数据十分灵活,而且数据操作很直观,从而使庞大的数据分析工作得到简化,变得轻松和高效. 简而言之,联机分析处理技术是一种多维数据分析工具的集合[10].

1.4.2 DM 技术

数据挖掘技术是经过了人们对数据库技术的长期研究开发之后的产物. DM 技术是数据库技术的更高级阶段,不仅可以做到数据库能做的查询遍历,还能找到数据与数据之间的关联,通过这些强大的功能可以促进数据的传递. 在网络流量监测技术中,所采集到的数据是大量、不完全的,数据挖掘技术就是可以从中能提取到隐含、事先不知道的内容,然而是潜在有用的信息的一个过程. 换言之,DM 技术就是更深层次的一种数据分析方法[11].

数据仓库、联机分析处理和数据挖掘是网络流量异常监测技术的三大骨干技术. 在解决问题的同时,数据仓库也利用 OLAP 中间件实现了数据的整合. OLAP 中间件使数据保持了各数据集市之间的一致性,并能在短时间之内完成数据的快速输出. 而联机分析处理技术和数据挖掘技术都属于进行数据分析的过程,只是程度深浅不同而已. OLAP 技术是一种程度较浅的数据分析过程,是一种交互的分析过程. DM 技术则是深层次的分析过程,还是一种自动的分析过程,正因为 DM 技术的这两个特性,管理人员由此可以获得更细致化以及更深层次的信息.

2 技术分析

2.1 总体需求概述

需求分析是实施的第一步,在其他活动开展之前必须明确地定义使用单位对检测平台的期望和需求,包括需要分析的主题,各主题可能查看的角度(维度). 需要发现使用单位哪些方面的规律,用户的需求必须明确. 网络流量异常监测技术主要是针对节能监测平台应用于领导的决策层而设计,侧重于数据统计、分析和直观的展现功能. 在这一平台上,通过建立专属的网络流量数据仓库,提炼出价值数据,经过统计、分析,以信息发布的方式,多角度直观地将流量情况展现给管理部门(这些对比分析以动态曲线图、饼图、报表等多种形式展现). 为现场操作人员、领导提供了不同的网络流量管理平台,实现了对网络流量消耗信息熵技术的各级管理需求. 满足建筑流量监测的基本功能,实现流量监测数据实时上传、接收,并兼具人工采集数据录入、数据丢失报警、修补及纠错功能. 同时,信息熵技术应具备操作简单、易用、直观、易维护的特点,方便操作人员的使用. 实施过程中需要确定信息熵技术的核心需求,包括当前要解决的主要问题和未来几年信息熵技术扩展的需要.

2.2 业务流程描述

监测流程大致可分为三个层次: 通过安装在网络设备中的监测终端来采集数据,由各个子信息熵技术上传至上级中心监测技术. 再由中心监测平台对数据进行处理和分析. 最后,将分析结果呈现给决策者,供决策者进行决策,如图 1 所示. 在平台运行过程中,建立定额管理制度是流量管理的一个重要部分.

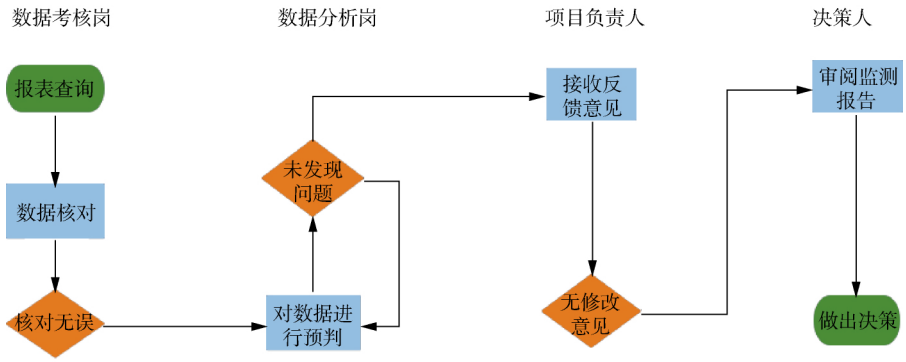


图 1 监测平台运行流程图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.3 数据分层处理

整个信息熵技术整体采用五层架构体系设计，包括: 数据采集层，数据传输层，数据处理层，应用展示层和数据上传层 5 层.

2.3.1 数据采集层

数据采集层的主要功能: 数据采集器对下面各种智能计量仪表、控制设备的数据进行采集，并以 TCP/IP 的方式向上传输到数据中心.

数据采集采取定时自动模式，可根据事先设定的采集间隔自动进行数据采集和存储(信息熵技术支持秒级数据采集) . 数据采集采取了多种信息熵技术安全性措施，如上层网络状态的侦测，下层仪表设备的故障判断与定位等.

2.3.2 数据传输层

数据上传采用 VPN(virtual private network) 通道以加密数据报文方式上传，定时主动往中心服务器传输流量数据. 信息熵技术默认是每天上传一次，当采集器端网络出现临时故障时，采集器自身提供不少于 30 d 的临时数据存储空间，并在网络再次恢复时主动把未上传的数据传输到中心服务器，同时信息熵技术支持数据中心的补调上传指定的数据，以保证流量数据的完整性.

2.3.3 数据处理层

在流量管理平台，本地的数据采集器中缓存最新的 30 d 的历史流量数据，并每天将数据都上传至网络流量管理中心.

“数据存储区”中存储的是按照数据抽取频率，把原始采集数据与用能单位基础信息、地理基础信息、流量报警信息结合处理后，形成网络流量计量和远程监测技术实时监测、明细查询、报警提示所需要的原始明细数据. 这样各类数据的大批量加工既可不影响实时的数据采集，而只是在数据库上进行处理，同时也为各类明细查询显示提供给出数据源. 其物理结构与原始采集数据库有所不同，增加了明细查询所需的内容.

“后台分析数据库”中存储的是按照数据抽取频率，把数据存储区的基础数据按照统计、分析、比较功能的要求进行不同时间段的分类、汇总和折算. 这样提高整个信息熵技术的运行效率，最大化缩短信息熵技术的响应时间. 这样各类数据的大批量加工不会影响到网络流量计量和远程监测技术的查询响应速度，只是在数据库内进行处理，处理后的结果可以让信息熵技术根据用户的需求直接进行调用，大大减少信息熵技术页面功能的计算工作量，提高整个信息熵技术的运行效率，最大化缩短信息熵技术的响应时间.

2.3.4 应用展示层

对经过数据处理后的分类分项流量数据进行分析汇总和整合，通过静态表格和动态图表方式将流量数据展示出来，为节能运行、节能改造、信息服务和制定政策提供信息服务.

展示层采用 B/S 软件体系结构，有权限的用户可以直接利用 Internet 浏览器方式访问工业流量监测分中心服务器，查看数据报表和图表等信息，操作方便，免安装、维护.

数据图表是反映各项采集数据和统计数据的数值、趋势和分布情况的直观图形和对应表格，可分为数据透视表、饼图、柱状图、线图、仪表盘或动画等，格式灵活，可交互操作. 代码如下:

```
using System. Reflection;
using System. Runtime. CompilerServices;
using System. Runtime. InteropServices;

// 有关程序集的常规信息通过下列属性集
// 控制. 更改这些属性值可修改
// 与程序集关联的信息.
[assembly: AssemblyTitle( "DAL") ]
[assembly: AssemblyDescription( "") ]
[assembly: AssemblyConfiguration( "") ]
[assembly: AssemblyCompany( "SkyUN. Org") ]
[assembly: AssemblyProduct( "DAL") ]
[assembly: AssemblyCopyright( "版权所有 ( C) SkyUN. Org 2013") ]
[assembly: AssemblyTrademark( "") ]
[assembly: AssemblyCulture( "") ]

// 将 ComVisible 设置为 false 使此程序集中的类型
// 对 COM 组件不可见. 如果需要从 COM 访问此程序集中的类型，
// 则将该类型上的 ComVisible 属性设置为 true.
[assembly: ComVisible( false) ]

// 如果此项目向 COM 公开，则下列 GUID 用于类型库的 ID
```


(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
[assembly: Guid("6c98509f-9bb1 -4d41-aceb-cde0239e89f9")]

// 程序集的版本信息由下面四个值组成:
//
//      主版本
//      次版本
//      内部版本号
//      修订号
//
// 可以指定所有这些值,也可以使用“修订号”和“内部版本号”的默认值,
// 方法是按如下所示使用“* ”:
[assembly: AssemblyVersion("1.0.0.0")]
[assembly: AssemblyFileVersion("1.0.0.0")]
```

BLL 的 Properties 如下:

```
using System.Reflection;
using System.Runtime.CompilerServices;
using System.Runtime.InteropServices;

// 有关程序集的常规信息通过下列属性集
// 控制.更改这些属性值可修改
// 与程序集关联的信息.
[assembly: AssemblyTitle("BLL")]
[assembly: AssemblyDescription("")]
[assembly: AssemblyConfiguration("")]
[assembly: AssemblyCompany("SkyUN.Org")]
[assembly: AssemblyProduct("BLL")]
[assembly: AssemblyCopyright("版权所有 (C) SkyUN.Org 2013")]
[assembly: AssemblyTrademark("")]
[assembly: AssemblyCulture("")]

// 将 ComVisible 设置为 false 使此程序集中的类型
// 对 COM 组件不可见.如果需要从 COM 访问此程序集中的类型,
// 则将该类型上的 ComVisible 属性设置为 true.
[assembly: ComVisible(false)]

// 如果此项目向 COM 公开,则下列 GUID 用于类型库的 ID
[assembly: Guid("9281ac7b-abb2 -4bfa-8ee1-fbb2cb7fd9a7")]

// 程序集的版本信息由下面四个值组成:
//      主版本
//      次版本
//      内部版本号
//      修订号
// 可以指定所有这些值,也可以使用“修订号”和“内部版本号”的默认值,
// 方法是按如下所示使用“* ”:
[assembly: AssemblyVersion("1.0.0.0")]
[assembly: AssemblyFileVersion("1.0.0.0")]
```

2.3.5 数据上传层

数据上传这一模块中,要求实现对老年社区网络设备的基本信息以及精细到每个项目的流量的信息实现同步上传至上级数据中心的功能.其中,数据上传频率根据监测需求灵活设置,数据上传时所采用的数据交换格式是 XML 数据包,通过 NTP/SNTP 协议来达到与上级数据中心的时间一致.并且,在与上级中心的数据上传过程中,会采取数据加密和身份认证等方式来确保安全性.

3 维度报表及 ETL(extract-transform-load)设计

3.1 维度报表设计

以统计表为例,该数据表保存了建筑用电的时间、网络流量等信息,如表 1 和表 2 所示.

3.2 ETL 设计

ETL 数据采集和处理信息熵技术采用 C/S 架构的 WPF 程序和 Windows 服务进行用户交互,应用层采用基于 TCP/IP 协议的方式,由客户端和服务器端通过 XML 格式进行数据传输,并采用 MD5 加密、AES 算法加密,增强数据传输安全性.便捷的安装和配置,无需为客户端采集软件类别不同和数据结构不同而进行定制开发;多客户端同时使用,解决并发的问题;断点续传,解决由于网络产生的丢数的问题.符合国家网络流量管理信息熵技术数据采集处理导则规范.

总体的数据业务流程如图 2 所示.

采用 Windows 服务保护 WPF 桌面应用的方式运行.通过 WPF 桌面应用程序(即 EnergyView 数据采集软件服务端程序)的配置,配置好服务端选项后,即可运行后台服务程序进行数据接收、解析、持

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1 统计节点表

序号	字段	说明	数据类型	允许空	默认值
1	TotalNodeID	统计节点 ID	int	否	NULL
2	TotalNodeName	统计节点名称	nvarchar	是	NULL
3	Type	统计节点类型 ID,关联 Base_TotalNodeType 的 TotalNodeTypeID 字段	int	是	NULL
4	BelongID	所属对象的 id,如: 如果是建筑,那么对应的就是建筑 ID,等等	nvarchar	是	NULL
5	F_EnergyItemCode	分类分项编码 关联 T_DT_EnergyItemDict 的 F_EnergyItemCode 字段	char	是	NULL
6	NodeName	简写名称	nvarchar	是	NULL
7	Flag	数据来源: 1 自动采集点,2 手工录入点	int	是	NULL
8	TotalType	1 为测点汇总,2 为统计节点汇总	int	是	((1))
9	IsAlarm	1 启用统计节点,0 已禁用统计节点	int	是	((1))
10	Remark	描述	nvarchar	是	NULL
11	TotalLevel		int	是	NULL
12	IsSmsAlarm	是否启用报警(1 为启用,0 为不启用)	int	是	((0))

表 2 统计节点维度表

序号	列名	数据类型	长度	小数位	标识	主键	允许空	默认值	说明
1	TotalNodeID	int	4	0	是	是	否	NULL	统计节点 ID
2	TotalNodeName	nvarchar	100	0	否	否	是	NULL	节点名称
3	OrgID	int	4	0	否	否	是	NULL	组织 ID
4	EnergyTypeID	int	4	0	否	否	是	NULL	网络流量类型 ID
5	LoadTypeID	int	4	0	否	否	是	NULL	负载 ID
6	TotalLevel	int	4	0	否	否	是	NULL	等级
7	TotalModel	int	4	0	否	否	是	NULL	统计型号
8	QueryPath	int	4	0	否	否	是	NULL	查询路径
9	MeasureType	int	4	0	否	否	是	NULL	测量类型
10	Quantity	nvarchar	4 000	0	否	否	是	NULL	英文公式
11	CHQuantity	nvarchar	4 000	0	否	否	是	NULL	中文公式
12	NodeName	nvarchar	100	0	否	否	是	NULL	节点名称
13	Flag	int	4	0	否	否	是	NULL	状态
14	TotalNodeCode	nvarchar	50	0	否	否	是	NULL	节点代码

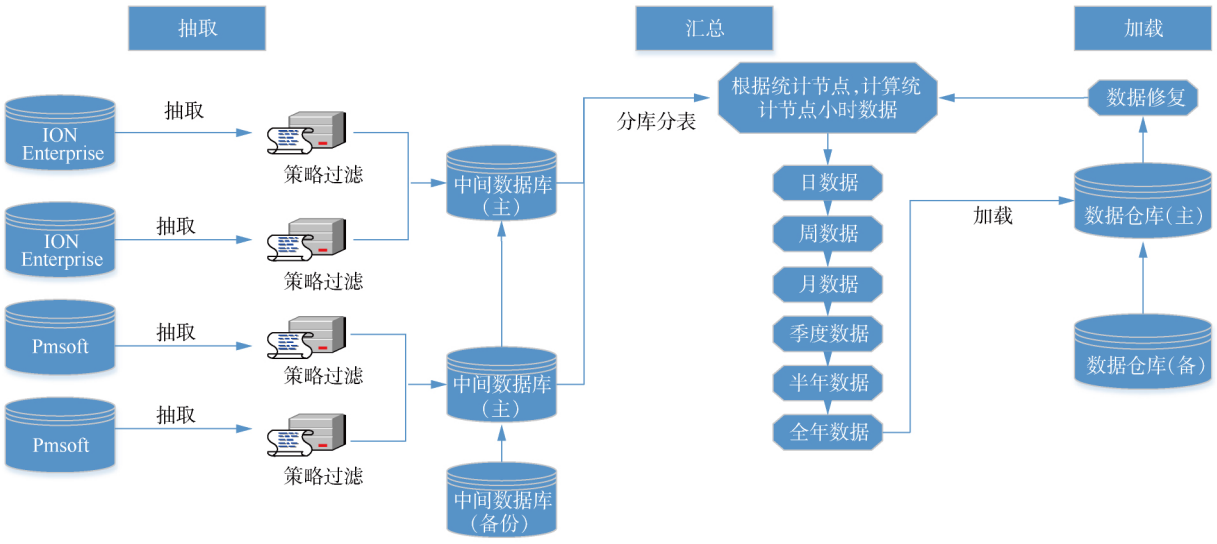


图 2 ETL 数据业务流程图

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

久化存储、流量数据拆分、流量数据处理、分类分项数据汇总等,并与客户端进行保持通讯,监控客户端及仪表状态等. 要点有:

(1) 持久打开端口进行监听,异步处理接收到的数据文件;

(2) 多线程处理,保证多客户端通讯;

(3) 标准化数据接口,实时统一处理各类流量数据;

(4) 持久化数据库服务器存储过程处理,较快提高运行速度.

4 技术测试

4.1 测试方法

从测试阶段上分,软件测试主要分为单元测试、集成测试和信息熵技术测试. 单元测试也叫模块测试,在编程阶段由程序员对自己的编程模块进行自检. 在单元测试期间主要针对以下几个方面进行: 模块接口、局部数据结构、重要的执行通路、出错处理通路和边界条件等. 集成测试也叫组装测试,是对由各模块组装而成的程序进行测试,主要目标是发现模块间的接口和通信问题. 而信息熵技术测试是软件测试中最后的、最完整的测试,是在单元测试、模块测试的基础上进行的,从全局来考察软件信息熵技术的功能和性能要求. 通常信息熵技术测试包括确认测试和验收测试. 确认测试主要依据信息熵技术需求来对软件的功能进行检查,看是否与最初设计目标要求一致. 确认测试的另一项重要内容是软件配置复查,其目的保证软件配置的所有成分都已经齐全,文档与程序完全一致,具有完成软件维护所必须的细节. 在实际操作中,对软件的验收测试可能会维持相当长的一段时间. 本智能监测技术使用 α 测试和 β 测试过程来完成初步验收测试.

4.2 软件测试环境

测试环境分为硬件和软件两种环境. 硬件环境包括了测试必须的服务器,客户端,网络连接设备以及打印机等辅助硬件设备. 软件环境是指被测软件在运行时的操作信息熵技术,数据库及其他应用软件构成的环境. 在本章节中主要分析软件环境. 网络异常监测技术的开发平台选用. Net 开发平台,操作信息熵技术采用 Win Server 2003,因为计算机信息熵技术的核心信息熵技术是操作信息熵技术,所以操作信息熵技术的选取要保证具有极高的可靠性和可用性. 本监测平台的开发工具为 Visual Studio 2008. Visual Studio 2008 能够提供调试,数据库,高级开发工具等强大的功能,让开发者具备了创建更安全可靠的软件的能力.

4.3 测试内容

经过对信息熵技术运行情况的跟踪分析,发现网络异常监测技术运行良好稳定,各项指标符合预期的目标. 下面对信息熵技术性能的测试结果进行展示,包括了主机的性能测试和准确性测试. 其中,对于主机性能的测试,其数据库服务器与应用服务器的运行状态均为良好; 而针对准确性所进行测试,则具体的测试用例如表 3 中所示.

测试结果分析如下: 基于测试结果在利用信息熵技术来进行运行情况的跟踪与分析的基础上,可以认为整个网络中的异常监测体系能够准确监测各项异常,且相关的技术指标与参数也符合预期的目标. 此外,根据准确性的测试用例的具体结果,认为外部的参数文件可以将其数据导入数据库中,能够得到正确的维度分解结果与数据,并最终通过知识库来体现参数的设置结果.

表 3 测试用例

名称	内容
测试环境确认	确认测试时程序版本与上线版本、测试服务器环境和生产环境相关设置、知识库版本和上线版本和其他信息熵技术环境与上线初始环境一致.
案例描述	数据正确性测试.
预期结果	数据库导入参数文件数据,获取维度的分解结果和维度数据,知识库体现参数的设置结果.
实际结果	参数文件的数据成功导入数据库,维度的分解结果和维度的数据获取都正确,参数的设置结果能够在知识库里面正确体现.
测试过程	将源数据与目标数据核对,核对源数据和目标数据.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

5 结 论

在网络攻击检测领域和网络异常检测领域,针对分布式拒绝服务攻击(DDoS 攻击) 的检测技术,已被资深网管和分析员所研究和利用. 而利用网络流量异常监测来提高网络监测数据处理和存储的效率的同时,还可以利用信息熵技术来提高检测准确度. 本文基于这些技术,通过合理的架构,使得信息熵技术在功能性方面对于决策更加有效合理,同时测试结果也利用信息熵技术来进行运行情况的跟踪与分析,可以更好地为整个网络中的异常监测体系准确监测各项异常服务.

参 考 文 献

[1] 田杨. 基于 NetFlow 的异常流量检测研究与实现 [D]. 长沙: 国防科学技术大学,2009.
[2] 陈卫屏. 网络数据流高速采集系统设计与实现 [D]. 成都: 电子科技大学,2009.
[3] 王海龙. 大规模网络流量异常分析 [D]. 长沙: 国防科学技术大学,2006.
[4] 陈晖. 面向一体化标识网络多因素路由决策机制的研究 [D]. 北京: 北京邮电大学,2015.
[5] 张秀英. 智慧标识网络资源感知机制设计与实现 [D]. 北京: 北京交通大学,2015.
[6] 张龙. 基于云计算的异常流量检测系统的实现 [D]. 天津: 天津理工大学,2013.
[7] 吴小花. 网络异常流量识别技术的研究 [D]. 长春: 长春工业大学,2013.
[8] 张波. 基于小波变换及块匹配技术的网络异常流量检测算法 [D]. 西安: 西安电子科技大学,2013.
[9] 孙琳. 手机安全支付异常流量监测技术的研究 [D]. 北京: 北京交通大学,2014.
[10] 廖建飞. 网络流量的异常检测与业务类型识别方法的研究 [D]. 南京: 南京邮电大学,2013.
[11] 黄茂成. 网络异常流量检测模型研究 [D]. 哈尔滨: 哈尔滨理工大学,2009.

Research on Network Traffic Anomaly Detection Based on Quasi Real Time Traffic Data Reporting and Information Entropy Technology

DENG Xiaoqing WANG Lunlang

(School of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou Sichuan 635000)

Abstract: Network traffic anomaly monitoring mainly through data warehouse, OLAP, data mining, WEB, ETL and other related technologies to improve the efficiency of network monitoring data processing and storage. At the same time, it can also ensure accuracy and simplicity through information entropy technology. It is more concise and efficient to ensure accuracy. According to these technologies, the information entropy technology has achieved more effective and reasonable results in terms of function, and the flexibility and reusability of the information entropy technology framework have also been verified.

Keywords: network traffic anomaly monitoring; OLAP; information entropy

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net