

基于云计算和分布式技术的流量分析模型

盖 璇

(东北石油大学,黑龙江 大庆 163318)

摘 要:传统海量网络流量分析模型采用串行分析方式,在运行中存在时间开销大的问题,为此提出基于云计算和分布式处理技术的海量网络流量分析模型。首先结合云计算与分布式处理技术的运行方式,搭建模型结构,在该结构下实时采集网络中的流量数据,并作为模型的输入值。通过对初始数据的存储、分类以及异常检测等处理,分别得出海量网络流量的分析结果,综合多个方面的分析结果得出模型的输出项。实验对比结果表明,通过云计算和分布式处理技术的应用有效地降低少量用户在线环境和大量用户在线环境时间开销,在分析速度上具有明显优势。

关键词:云计算;分布式处理技术;网络流量;海量网络数据;流量控制

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2022)0114-06

Traffic Analysis Model Based on Cloud Computing and Distributed Technology

GAI Xuan

(Northeast Petroleum University, Daqing 163318, China)

Abstract: The traditional massive network traffic analysis model adopts the serial analysis method, which has the problem of large time cost in operation. Therefore, a massive network traffic analysis model based on cloud computing and distributed processing technology is proposed. Firstly, combining the operation mode of cloud computing and distributed processing technology, a model structure is built, under which the traffic data in the network is collected in real time and used as the input value of the model. Through the initial data storage, classification and exception detection, the analysis results of massive network traffic are obtained respectively, and the output items of the model are obtained by synthesizing the analysis results of many aspects. The experimental results show that the application of cloud computing and distributed processing technology can effectively reduce the time cost of a small number of users' online environment and a large number of users' online environment, and has obvious advantages in the analysis speed.

Key words: cloud computing; distributed processing technology; network flow; massive network data; flow schedule

0 引 言

网络流量是能够直接反映出网络性能的好坏的重要指标,也能够反映出网络运行的安全性,相关人员通过收集并分析网络中的流量数据,即可确定当前网络的运行状态,从而对网络进行调整与维护^[1]。

当前对网络流量进行控制一般采用网络拓扑结构的模型分析,分析模型中存在的流量监测值,对网络结构中节点调度分配进行优化。目前主流的海量网络流量分析方法包括基于 PTN 流量平台的城域网流量分析方法^[2]和基于小波分析的网络流量异常检测方法^[3],在完成在进行海量数据处理时,大部分采用云计算技术,如基于 Hadoop 平台的海量数据分析和处理方法^[4]。以上的主流海量网络流量分析处理方法存在

分析精度低、时间开销大的问题。

为了解决上述问题,引入云计算与分布式处理技术,利用云计算技术满足网络中不同用户的不同需求^[5]。

通过搭建海量网络流量分析模型,利用云计算技术来完成网络流量的分类,利用网络节点对流量进行控制优化。能够在传统网络流量分析模型的基础上,对网络流量分配进行动态规划,降低时间消耗成本。

1 海量网络流量分析模型设计

网络流量监测分析一直是网络优化、网络安全等领域的重要研究内容。通过监测网络流量分析出网络中出现的一些异常(如木马攻击)。随着网络应用高

收稿日期:2022-03-10

基金项目:引导性基金(2017YDL-07)

作者简介:盖 璇(1990-),女,硕士研究生,讲师,研究方向为大数据。

速,网络中异常数据常常通过流量而表现出来,传统的流量检测分析一般采用了两种方式:基于服务主机的流量检测和基于网络的流量检测,其中服务主机的流量检测是通过分析网络中每个主机的包数据、网络数据访问日志,分析出网络中流量异常,基于网络流量检测一般通过多种通信协议来实现流量检测,由于 Internet 协议较多,基于协议的流量分析不适当当前企业大数据流量应用场景中,传统模式下对海量网络流量进行分析和控制时,分析模型中采用串行分析方式,在运行中存在时间开销大的问题一般采用采集流量数据,使流量检测效率较低。

此次海量网络流量分析模型的设计是在传统分析模型的基础上引入了云计算和分布式处理技术,设计的分析模型由多个子模块组成,通过分布并行方式实现对各个子模块调用。

1.1 海量网络流量分析模型的搭建

由于分布式网络流量分析模型由若干本地云、任务协调节点、Web 服务器和关系数据库等组成。其中本地云用来存储和处理海量网络流量数据,任务协调节点作为各个本地云的用户,向云端发出执行流量分析任务的请求,并汇集分析结果^[6]。结合分布式技术和网络结构,确定海量网络流量分析模型的基本结构,如图 1 所示。

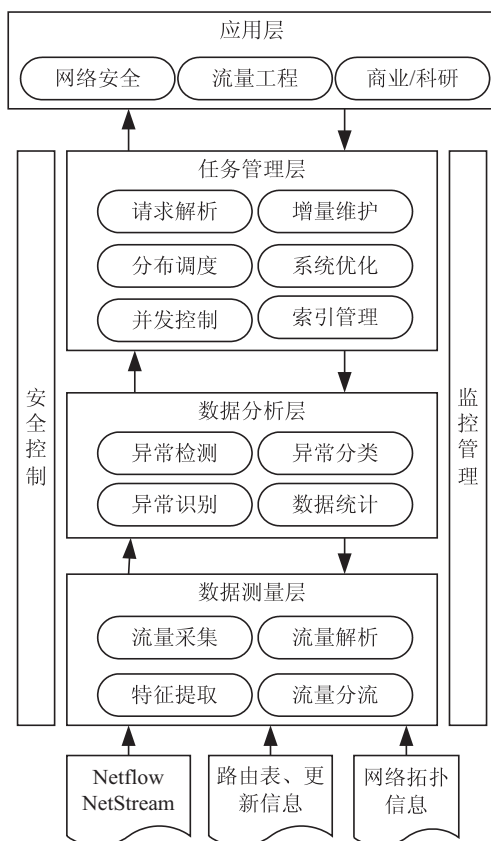


图 1 海量网络流量分布式云处理分析结构
在大规模网络中通信节点数量放大,使用传统的

分析方式会占用较大的处理空间,造成存储开销与运算时间开销^[7]。因此利用分布式处理技术对模型内部的运行进行并行化改造,根据海量网络流量的存储特性,将流量矩阵按列划分,对输入的海量网络流量矩阵 $X_{N \times M}$ 中的元素 x_{k_1, k_2} 分解成如下序列:

$$\begin{cases} C_{m_1, m_2} = \sum_{k_1, k_2} h_{k_1, k_2} x_{(k_1+2m_1), (k_2+2m_2)} \\ D_{m_1, m_2}^1 = \sum_{k_1, k_2} g_{k_1, k_2}^1 x_{(k_1+2m_1), (k_2+2m_2)} \\ D_{m_1, m_2}^2 = \sum_{k_1, k_2} g_{k_1, k_2}^2 x_{(k_1+2m_1), (k_2+2m_2)} \\ D_{m_1, m_2}^3 = \sum_{k_1, k_2} g_{k_1, k_2}^3 x_{(k_1+2m_1), (k_2+2m_2)} \end{cases} \quad (1)$$

式中, h 表示的是尺度系数, g^i 则代表小波系数的有限长度值。计算输出序列在点 (m_1, m_2) 的值只需要输入序列在点 $(2m_1, 2m_2)$ 附近点的值,则可以表明计算在一定程度上只依赖局部性的数据。在模型的运行过程中假设有 P 个网络流量处理分析任务,可以将初始数据 C_{j, m_1, m_2} 按列块分配给对应的 P 个模块节点,按照各个模块的运行方式得出对应的分析结果^[8]。在各个模块运行的过程中,控制各个子模块的并行运行方式和进度,实现对模型的并行化处理和改造。在资源分配时,一般采用轮询原则,设置资源监测点,如果某节点模块流量减少到某值时,监控管理中心负责分流分配此节点任务,分配任务后,监控中心刷新海量网络流量矩阵。

在海量网络流量分析模型中除了分布式处理技术之外,还使用了云计算手段,为了保证云计算技术在模型子模块中的正常运行,需要在模型结构中加入云处理层,该层的基本结构如图 2 所示。在本层收集流量,按照协议检测分布式处理层,对数据进行协议解码,识别出流量中的用户数据项,若用户有特殊需求,可存储到特征库中,通过控制接口实现数据的处理、过滤。经过流量数据处理后输出最后结果到应用层。

根据云模型定义^[9],假设 U 是一个用精确数值表示的定量论域, S 和 T 分别是 U 空间上相关联的定性概念,当 S 中的元素 s 对 T 的隶属度确定满足公式(2)时,表示 U 是一个存在稳定取向的随机数。

$$CT(S) \in [0, 1] \quad (2)$$

那么此时,概念 T 从论域 U 到区间 $[0, 1]$ 的映射在属于空间的分布即为云端数据,相应的映射关系可以表示为:

$$CT(S): U \rightarrow [0, 1], s \rightarrow CT(S) \quad (3)$$

在网络环境中实时产生的海量流量数据存储到硬件设备中时,模型自动通过上述映射方式形成对应的云端数据,方便分布式处理。网络数据进行传输时,在处理层收集到数据,分别按照节点处理任务,对不同类

型的数据进行分发存储,采用 HDFS 数据存储,分别对 Text、Binary 等不同类型进行处理,处理时根据用户特征库中的特征,对数据过滤处理,经处理层处理后,把网络数据进行输出。

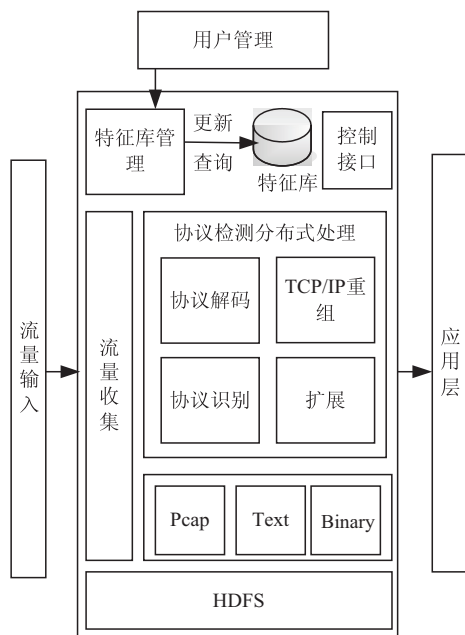


图 2 云处理层详细设计图

1.2 海量网络流量数据的实时采集和处理

实时网络流量数据也就是流量分析模型的输入值,网络流量数据的采集分为两个步骤,分别为数据的采集和数据的传输。此次网络流量数据的实时采集应用了 NetFlow 流量统计协议,在采集过程中一条数据流是由一个源主机和一个目的主机间的单方向传输的网络数据包组成,其中源主机和目的主机由各自的 IP 地址和端口号标识^[10]。利用分析 IP 数据包的属性,快速区分网络中传输的不同类型业务的数据流,记录其传送的方向和到达节点,统计起始和终止时间,服务类型包含的数据包数量和字节数等流量信息,得出的记录结果即为网络流量数据的实时采集结果。NetFlow 流量采集在网络测量节点上安装采集器,并对路由发送的 NetFlow 数据包进行收集,并将其发送至分析服务器进行进一步处理。

在实际的网络流量数据的采集过程中,数据的质量决定了数据分析质量的高低,海量的杂乱数据无法进行有效的数据分析,因此需要在存储相关网络流量数据之前,对数据进行预处理。此次网络流量数据预处理包括数据清洗、数据集成、数据转换和数据归约四个步骤,其中数据清洗的目的是消除数据中的噪声和不完整数据,通过光滑、聚类 and 去重三种清洗方式得出海量网络流量数据的初始清洗结果^[11]。删除、均值补充、近邻补充等方式均为不完整数据的处理方式,利用其属性值计算 x 和 y 的邻近测度,将与具有空值的数

据近邻测度大的值定义为最相似的 N 个邻居,并将该属性值代替空值。将多种不同形式的数据统一进行存储管理的过程即为数据集成的过程,该环节包括实体识别、数据冗余和数据值冲突检测三个部分^[12]。数据的转换包括单位换算、数据泛化、规范化和属性构造四种情况,其中数据的规范化处理是对不同量纲的属性赋予相等的权重,对于数据的规范化方法包括属性的归一化和属性标准化,归一化处理表达式为:

$$rd = \frac{r}{r_{\max}} \quad (4)$$

式中, r_{\max} 表示属性的最大值,得出归一化结果 rd 的取值范围为 $[0,1]$ 。而属性标准化的处理表达式如下:

$$rd = \frac{(rd - \mu)}{\sigma} \quad (5)$$

式中, μ 是属性值 rd 的均值, σ 为标准差。将预处理完成的海量网络流量数据存储于源主机中,并通过云计算技术得出对应的云端数据^[13-14]。

1.3 利用云计算对数据的分类

按照网络数据的采集位置可以将网络流量数据分为移动网络流量、固定网络流量等类型,结合网络的历史分析数据,可以将数据分为正常数据和异常数据两种类型,在实际的数据分类过程中需要先统计测量采集的数据量,再进行数据的分类处理^[15-16]。数据的统计主要就是通过累加算法得出不同网络平台在某一段时间区间内产生的流量数据量^[17-18]。定义每个数据样本 W_i 为 n 维向量,即:

$$W_i = (w_{i1}, w_{i2}, \dots, w_{in})^T \quad (6)$$

利用各个网络流量数据之间的欧氏距离计算数据之间的相似性,那么数据样本 W_i 和 W_j 之间的欧氏距离可以表示为:

$$d(W_i, W_j) = \left(\sum_{k=1}^m |w_{ik} - w_{jk}|^p \right)^{\frac{1}{p}} \quad (7)$$

其中, p 为数据样本的共同属性数量。根据得出的相似性测度计算结果对数据进行分类,设置分类阈值为 η ,若将网络流量数据代入到公式(7)中得出的计算结果小于 η ,则两个数据之间的相似性较低,需要进行下一组的计算。若 $d(W_i, W_j)$ 大于 η ,将两组数据划分到相同的数据组别中^[19-20]。

1.4 利用网络节点对流量的控制优化

由于云平台下,多个网络节点的流量随机分配,针对多因素多目标的优化可从以下方面进行提高。

(1) 对非主要因素进行评价。在多种因素进行优化时,由于影响优化的因素较多,致使在优化处理中,计算效率较低,常因非主要因素的优化,影响了整个目标优化效果。因此需要对一些非主要因素进行评价,以“权重”的形式对其进行赋值,如在某些优化目标的

处理上,对“权重”较低的因素进行忽略,可提高优化计算效率和提高优化效果^[21-22]。

(2)对优化目标减化处理。在最终的优化目标时,由于目标的多维性,致使优化计算效率较低,且影响优化效果,因此在进行动态规划时,需要对优化的目标进行减化处理,以重要的目标、次要目标来分别实现目标的优化,以此得到最优化解。

(3)融合多种优化策略。当前目标优化时,对于多个目标,采用的是同一种优化策略,但在实际应用中,需要把这些目标进行分解,分别采取不同的优化算法、策略,提高每个目标的优化效果^[23-24]。

针对云平台网络节点不同时间段内的现有条件、堵塞率进行规划,这样动态规划过的网络节点流量优化分配方法,具有明显的优点:

(1)能够全面掌握每个不同云平台网络节点的现有特征。其与网络节点类型的条件展示相关,如要求对网络节点类型进行查询时,考虑到网络节点的最大流量和流量限制。

(2)能够体现每天不同的时段流量。如周一上午10点-12点,某办公区域的网络节点为较高流量需求,而在办公时段内,员工宿舍区域的网络节点基本没有流量需求,可在办公时间关闭网络节点。

(3)能够对某节点进行预警。针对网络节点类型的不同,设置不同的预警权重,来区分每个网络节点的流量状况,以平衡各个网络节点的分流。

(4)能够高效利用堵塞率。通过服务系统引入每个网络节点的堵塞率,把堵塞率当作一个因子考虑到对考虑网络节点的堵塞可发生的权重,如果某个网络节点堵塞率较高,说明其在备选上的权重较小。动态规划考虑的目标包括网络节点的使用率、网络节点的条件、堵塞率等。

具体算法描述如下:

输入:网络节点流量的使用统计情况、网络节点条件信息数据、网络节点类型统计数据。

输出:网络节点流量优化分流信息。

步骤:

(1)初始化网络节点流量优化分流表为 B ,对每个时间段的网络节点置为空;

(2)设优化考虑的目标因素。记为因素特征表集合为 C ,其中 C 表述为: $C = \{ C_1, C_2, \dots, C_n \}$,其中 C_1, C_2, \dots, C_n 为优化的目标;

(3)从 C 中循环取出网络节点目标因素 C_n ,按照时间分配进行统计,对于有利因素则记为 $+n$,非有利因素则记为 $-n$,其中 n 取值大小为系统的优化时的权重,如使用因此为3,条件相符度为2等;

(4)对时间使用性因素进行筛选。如网络节点的

使用率,统计出在该时间内的使用情况,如流量低于某个值使用则置为-1,单次累加,设定因素的上限与下限;

(5)分析网络节点的数据堵塞情况。根据网络节点的堵塞率,若在时间内堵塞率大于某值,则因素值-1;若堵塞率小于某值,则因素值+1;

(6)对于符合性因素,计算网络节点的因素与历史时段上是否相符。如某网络节点的流量高于某个值类型,则置为2,一般性符合则置为1;

(7)优化网络节点。经过多目标因素的计算,为每网络节点得出因素得分分布表 D ,其结构与 B 相同。在这些因素置为不同权重时,如果某 C_i 值低于中心设置的标准值,则对应的 B_i 进行剔除,或以负无穷值来代替,表示 B_i 不能采用。如果在选择策略中,某节点的 C_i 超出某个值时需要优先选择此路径,则无需考虑其他因素;

(8)生成规划分流信息表。对因素得分表 D 中的每个时段 D_i 进行排序,如果某个 D_i 值为最大,则第 i 个网络节点流量优化分配至 B_i ,最终生成网络节点的规划分流信息表 B ;

(9)优化选择某些特殊因素值要求的节点。如某特殊因素值达到了预置标准,则优先考虑节点,对 D_i 值进行排序分析。按照规划分流信息表 B 实现节点流量的分配。

通过以上设计,把整个网络节点的划分最优目标划分为一个个节点,通过监控中心,对历史时间内的所有的节点流量进行记录,记录每个节点的在划定时间段内的流量值。按不同时间周期进行统计,一般采用一天为一个全周期,时间间隔为10分钟,计算每个时间段的堵塞率,记录在统计表中,在进行网络流量分流时,按节点编号检索出当前时段的堵塞率来划定流量分配值,通过动态规划流量达到对其进行控制的目的。

2 模型测试实验分析

以测试设计的基于云计算和分布式处理技术的海量网络流量分析模型的应用性能作为实验目的,设计模型测试实验。在传输数据流量一定的条件下,通过时间开销的多少来说明海量网络流量控制算法的优劣是切实可行的。此次实验测试的主要指标是海量网络流量分析模型的优化控制后的运行速度,因此分析模型输出分析结果的时间即为模型测试实验的测试指标。实验中除了设计的基于云计算和分布式处理技术的海量网络流量分析模型之外,为了形成实验对比这两个实验对比模型,分别为文献[2]方法分析模型和文献[3]方法分析模型。

实验方法:分别在相同的硬件资源、网络环境下进

表 1 少量用户在线环境的时间开销对比结果

开始时间 /s	数据大小 /GB	文献[2]提出的网络 流量分析模型时间 开销 t_1 /s	文献[3]提出的网络 流量分析模型时间 开销 t_2 /s	设计的网络流量分 析模型时间开销 t_3 /s
21:00:00	0.3	1 156	489	201
21:01:00	1.0	2 682	1 655	945
21:02:00	20.0	27 356	19 925	10 321
21:03:00	40.0	49 241	30 251	12 575
21:04:00	100.0	103 406	89 623	18 923
21:05:00	120.0	139 274	109 855	20 391
21:06:00	140.0	168 239	123 741	22 754
21:07:00	160.0	189 023	149 823	24 963
21:08:00	200.0	225 280	182 024	28 523

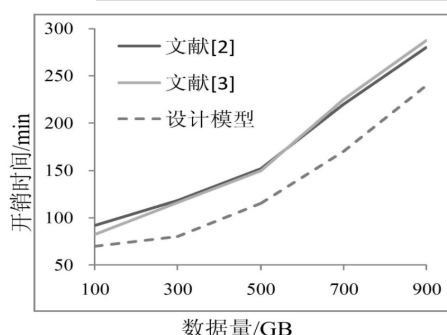


图 5 大量用户在线环境的时间开销对比曲线

从图5中可以明显地看出,设计的基于云计算和分布式处理技术的海量网络流量分析模型的大量用户在线环境时间开销明显低于文献[2]提出的模型和文献[3]中提出的模型。其原因是设计分析模型结合网络的历史分析数据,再进行数据的分类处理,在一定程度上节省大量用户在线环境时间消耗。

因本文模型充分考虑了网络环境下的多节点流量优化处理,根据不同节点的流量需求,实现网络流量的控制。同时在中心控制器中,引入每个网络节点的堵塞率来实现网络,减少不必要的网络开销,因此本文设计的海量网络流量分析模型在时间开销上比文献[2-3]低。

综合两种网络环境下的对比结果,可以看出设计的基于云计算和分布式处理技术的海量网络流量分析模型在运行速度上具有明显优势。

3 结束语

针对不同的网络环境下产生的网络流量数据的类型也会存在差异情况,提出了通过云计算和分布式处理技术的应用,成功实现了对海量网络流量的分析与控制。通过搭建海量网络流量分布式云处理分析结构,详细设计了云处理层,对海量网络流量数据进行实时采集、存储和处理,利用云计算技术来完成网络流量

的分类并行处理,同时利用多因素对网络节点流进行分配,起到控制调节作用,并通过搭建实验环境验证了基于云计算和分布式处理技术的海量网络流量分析模型在运行速度上比其他典型模型具有明显的优势。应用设计的网络流量分析模型进一步分析,还可以统计具体业务的用户偏好度,为运营商增值业务的统计需求提供数据支撑。今后的研究方向可以对云计算和分布式处理参数进一步调优,利用该模型实现对网络流量的深度识别与详细统计。

参考文献:

- [1] 廖先富,刘俊男. 基于 Django 与 HDFS 的分布式三维模型文件数据库构建[J]. 电子技术与软件工程,2018(18):189-191.
- [2] 吴 亮. 基于 PTN 流量平台的城域网流量分析方法[J]. 电信科学,2019(A01):241-248.
- [3] 杜 臻,马立鹏,孙国梓. 一种基于小波分析的网络流量异常检测方法[J]. 计算机科学,2019,46(8):178-182.
- [4] 张趁香. 基于 Hadoop 平台的海量数据分析和处理[J]. 电脑编程技巧与维护,2019,5(1):95-97.
- [5] 杜红军,李 巍,张文杰,等. 基于云计算技术的电力大数据分布式检索系统[J]. 电网与清洁能源,2018,34(9):23-28.
- [6] 朱晓丽,邓惠俊,陈小虎. 基于 Hadoop 云计算平台的数据处理研究[J]. 科技经济市场,2018,4(7):17-18.
- [7] 刘兆禄,赵 英,刘淑梅. 基于 Spark 的网络流量分类方法研究[J]. 通信学报,2018,39(1):30-36.
- [8] 吴 奔,李喜旺,周心圆. 基于流计算的电力调度网络流量监测平台[J]. 计算机系统应用,2018,27(7):59-64.
- [9] 张光卫,李德毅,李 鹏,等. 基于云模型的协同过滤推荐算法[J]. 软件学报,2007(10):2403-2411.
- [10] 卢 楠,杜清河,任品毅. mMTC 网络中基于空口流量的入侵检测[J]. 中兴通讯技术,2018,24(2):34-41.
- [11] 肖 琦,苏开宇. 基于随机森林的僵尸网络流量检测[J].

(下转第 125 页)