学校编码：10384　　分类号_____密级_____
学号：24320111152281　　UDC_____

厦门大学

硕 士 学 位 论 文

大数据环境下实时流量异常检测算法的研究

Research on Real-time Abnormal Detection of Network Traffic under The Big Data Environment

黄志敏

指导教师姓名：吴清锋 副教授

专 业 名 称：软 件 工 程
论文提交日期：2015 年 6 月
论文答辩日期：2015 年 7 月
学位授予日期：2015 年 9 月

指 导 教 师：_____

答辩委员会主席：_____

2015 年 7 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

随着互联网的快速发展，数据网规模不断扩大，基于数据网的各种应用业务也越来越广泛，对各大运营商以及业务数据量大的企业，在运维管理方面不得不投入了大量的人力与物力。在运维管理中，故障实时预警是最重要的一部分，快速并准确地预警故障，可以以最快的速度发现业务环境中的问题，从而及时地避免故障带来的严重损失。在大数据环境下，研究实时故障预警技术的理论和实现机制，在现实运维管理中有很大的实际意义。

其中流量实时预警是最重要的一个环节，通过流量的实时预警，可以很大程度反馈业务的故障。因此，本文的研究重点主要是针对流量序列的异常检测展开。

本文的研究内容主要包含两个部分：流量异常检测的准确性和在大数据环境下，流量异常预警的实时性。

1、在流量异常检测的准确性方面，提出了 N-ARMA 的流量异常检测算法，该算法基于 ARMA 的时间序列的预测模型，对模型参数进行提取与模型优化处理，并使用序列预测偏差拟合正态分布，构建异常可置区间。最后通过对比小波算法和排列熵算法，实验表明 N-ARMA 的检测算法在准确性最优。

2、在大数据环境下实时预警方面，本文提出了一个基于 Storm 的实时分布式的计算框架，结合了 Kafka 分布式可靠传输消息队列，并使用 nginx 与 jetty 作为前端负载均衡处理。最后通过对比传统的集中式服务框架，实验表明，本文提出的框架具有更小的性能开销，同时异常检测时延更短。

关键词：大数据；流量；实时预警

Abstract

With the rapid development of Internet, the data network scale expands unceasingly, on the basis of data network business is becoming more and more widely applied, the operators and enterprises, business data in operational management had to spend a lot of manpower and material resources. In operations management, real-time fault early warning is one of the most important part of early warning and fault rapidly and accurately, the questions in the business environment can be found at the fastest speed, to avoid failure to bring serious loss in a timely manner. Under the environment of big data, the real-time fault early warning theory and implementation mechanism, in the real operational management has a great practical significance.

The real-time traffic warning is one of the most important, through traffic real-time warning, can greatly big feedback failure of the business. Therefore, the research emphasis of this article is mainly aimed at flow sequence of anomaly detection.

The research content mainly includes two parts: traffic anomaly detection in large data environment, and the accuracy of real-time of traffic anomaly warning.

1, In terms of the accuracy of traffic anomaly detection, N-ARMA model of traffic anomaly detection algorithm is proposed, the algorithm based on prediction model of time series ARMA model, the model parameters extraction and optimization model, and use the sequence prediction deviation of normal distribution fitting, build anomalies can set range. Finally by comparing the wavelet algorithm and permutation entropy algorithm, the experiment showed that N-ARMA model of optimal detection algorithm in accuracy.

2, Under the environment of big data in real time warning, Propose a real-time distributed computing framework based on the Storm, is a combination of Kafka distributed message queue, reliable transmission and use nginx and jetty as a front-end load balance. Finally through comparing the traditional centralized service framework, the experimental results show that the proposed framework has a smaller performance

overhead, anomaly detection delay shorter at the same time.

Keywords: Big Data；Traffic；Real-time Alarm

目 录

Contents

第一章 引言

1.1 研究背景

随着互联网的快速发展，数据网规模不断扩大，基于数据网的各种应用业务也越来越广泛。经常网络上的一些故障或者异常都将导致经济上的巨大损失，例如 2001 年 6 月的网络攻击，Code-Red 蠕虫仅在 9 小时内便感染了 250000 机器，直接带来经济损失超过 26 亿美元[1]。在 2003 年 1 月，SQL 的 Slammer 蠕虫爆发，仅在 5 分钟内就导致 12 亿美元的损失[2]。由于蠕虫的侵入，拒绝服务攻击，或者网络的配置不当，严重影响了网络的正常运行。如何有效地管理当前大规模的网络设施，使它更高效，可靠，安全的动作，仍是网络维护领域面临的巨大问题。

对各大运营商以及业务数据量大的企业，在运维管理方面不得不投入了大量的人力，大量的物力。对设备的监控是运维工作环节中的重要组成部分，具体的监测内容包括设备的硬件，系统服务的性能指标，以及设备上的服务的状态。运维通过监测到的数据了解到当前设备和系统的运行状态来评估服务的质量，由于人为的判断存在主观性，同时由于人为的参与必然存在滞后的问题。因此，对服务的实时故障预警是运维管理中必不可少的。其中，流量异常预警是设备监控重要的组成部分。

在传统的运维管理中，比较常见的是通过人为根据设备所跑的业务来手动设定流量的阈值，当流量超出预设阈值范围，便触发告警，该方案依赖于人为的经验判断。同时该方法弹性太小，当设定的阈值过窄时，误报率比较高，严重影响到运维效率，相反设定的阈值过宽，将导致当异常已经出现时，却未达到阈值，经过若干时段才会触发，存在较大延迟，这将会使故障严重影响线上业务，造成巨大的经济损失。为此，必须研究更为可靠并且快速的异常检测方案。

1.2 研究现状

在时间序列的异常检测领域，国内外已经有一些研究成果。小波分析[3, 4]和排列熵[5, 6]是比较常见的时间序列异常检测算法。小波分析的特点在于能提供受

分析序列在时间域、频率域上两个维度的特征，但由于时间序列分析通常是实时性的，小波分析的计算量相对较大，因此限制了其在计算、硬件资源有限的移动设备上的应用。相对来说，排列熵的计算过程较为简洁，排列熵算法自 2002 年问世以来、便受到了学界、工业界的关注。

在网络入侵检测领域，国内外有不少文章发表。Kwitt 和 Hoffman[7]等人提出使用具有鲁棒性的 PCA 模型来检测网络流量异常，Shen[8]等人提出基于聚合网络行为指标的异常检测方法。此外，Karasaridis[9]等人提出了大规范的僵尸网络的检测方法，Jin 等人也提出了在协方差空间的网络入侵检测方法。Sang 和 Li[10]介绍了如何使用 ARMA 模型来预测网络流量，同样地，Cho[11]等人针对广域网流量监控提出了一种聚类的方法，Feldman 等人[12]研究了互联网广域网流量的多重性质，提出了一种基于层级式数据网络的方法，Yurcik 和 Li[13]，以及 PlonKa[14]给我们展示了如何使用网络流量图，Gong[15]提出如何利用网络流量信息来检测蠕虫和其它入侵的方法，Bivens[16]等人提出使用神经网络进行网络入侵检测，也有人提出使用基于小波的方法进行异常检测和以太网流量的预测[17]。

在网管领域，国内外许多组织和企业进行了大量的算法研究和系统开发的工作，许多商品化的网络管理平台被推出。在网络流量预警方面，人们发现传统的 Poisson 模型不适应网络数据流量分析，一些其它流量模型被相继提出。同时，许多相关技术被应用到流量预测中，比如神经网技术[18, 19]、小波方法等。国内也有许多科研机构在这个领域里做了许多工作，如中科院计算所、北京大学等。基于预测模型的管理系统也逐步得到应用，但是，预测精度和广度还有待提高。进一步研究精度高、适应性强、实时性好的流量模型，以及将模型在网管系统中广泛应用，仍是流量预测方面研究工作的重点。

1.3 本文研究内容

如图 1-1 所示，为本文的研究内容关系图及各章节对应关系。本文的研究路线首先介绍了本文提出的理论模型，再到该模型的验证，最后介绍该模型的原型系统的设计。

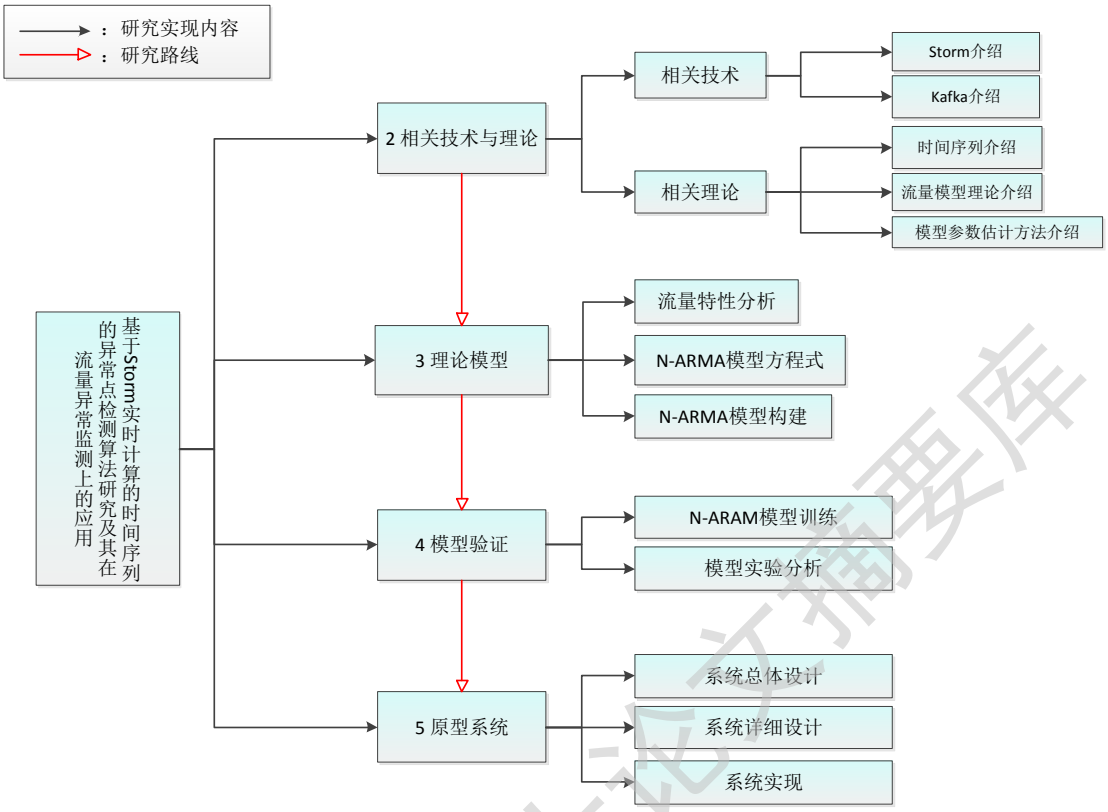

图 1-1 研究内容关系图

本文主要的研究内容主要在两方面，一方面是针对流量异常检测的准确性，一方面是针对大数据环境下，流量异常预警的实时性。具体的论文贡献如下：

1、本文提出了 N-ARMA 流量异常检测模型，通过对比常规的异常检测方法，实验表明该模型具有更高的异常检测准确率和查全率。

2、本文提出了基于 Storm 的分布式实时计算框架，通过对比传统的集中式服务框架，实验表明，本文提出的框架具有更小的性能开销，以及更小的告警时延。

1.4 论文的结构安排

根据对于研究课题所做的工作，将文章组织为下述分章结构：

第一章，绪论。阐述文章选题背景、研究内容以及论文的整体结构。

第二章，相关技术综述和模型理论介绍。技术综述方面包括 Kafka 分布式队列，Storm 实时计算框架。模型理论包括时间序列的简介，以及时间序列的平稳和随机特性的定义，以及介绍相关的时间序列模型，模型参数的估计方法等。

第三章，介绍 N-ARMA 流量异常检测模型。主要介绍了 N-ARMA 流量预测模型的构建方法,预测偏差正态分布模型参数的估计，构建置信区间进行流量异常判定。

第四章，系统建模与实验分析。主要介绍了 N-ARMA 异常判定模型的构建过程，给出构建过程中参数如何提取，模型对比分析，实验结果等内容。

第五章，Storm 分布式实时流量预警框架。主要介绍 Storm 实时分布式平台的构建过程，并对比传统的服务端框架，实验表明系统提出的框架具有更低的告警延时。

第六章，总结与展望。主要是对本文的重点工作进行了总结，包括了流量异常检测模型与 Storm 分布式实时处理的研究工作，同时对下一步的研究进行展望。