摘要

随着大规模电子商务行业中数据传输量的增加，管理如此大量的请求成为一个关键问题。通过在分布式服务器系统上应用任务调度来处理请求是一种有效的方法。但是，当服务器节点在短时间内处理极大量的请求时，过高负载量是不可避免的。没有有效的高负载监视方法，很难高效的对后端服务器集群进行监控和管理。据我们所知，监控高负载异常节点的现有方法既不灵活也不直观，并且不能检测服务器节点异常，例如定位客户端发送的不合理请求。在本文中，我们提出了一种基于真实数据集的可视化分析的作业调度监控方法，该方法允许监控人员了解该区域中运行节点的状态，并通过各种视图组件观察导致高负载服务器的可疑请求。这种思路为服务器端任务调度集群监测提供了一种全新的方法，并且经过测试显示是可行且有效的。

关键词：数据可视化 任务调度 异常检测

Abstract

With the increasing amount of data transmission in large-scale e-commerce industry, managing such an enormous amount of requests simultaneously becomes a key problem. It is an effective method to handle the requests by applying task scheduling on the distributed server system. However, high load on the servers is inevitable in scheduling when requests of a server node process extreme large amount of tasks in a short period of time. It is difficult to maintain load balance without an effective high-load surveillance approach. To best of our knowledge, existing methods of monitoring high-load abnormal nodes are neither flexible or intuitive and are not capable of detecting server node anomalies such as positioning unreasonable requests sent by the clients. In this paper, we propose a job scheduling monitoring method based on visual analysis from a real dataset, which allows the monitoring personnel to know the status of running nodes in the area and observe the suspected requests causing a high load of servers via various view components we inject into the system.

Keywords: Task scheduling, visual analysis, abnormal detection

第一章 绪 论

1.1 研究工作的背景与意义

现如今网络电子商务项目正在蓬勃发展，人们在享受着足不出户购物和信息浏览，但是往往当使用者的数目变得足够多时，安全隐患往往随之到来。尽管各大互联网企业采取了一些应对措施，比如将用户节点置于类比虚拟机的容器中，或者使用算法对用户请求进行调度处理，使指令被分散在不同的服务器上进行处理，但如果控制这台用户节点的服务器因为某些原因出现超负荷运转，甚至宕机的话，目前整个整个相关领域是没有很好的监视系统来全程监视负责处理用户请求的服务器的运行状态的。比如全球先进的超级计算机研究所——美国德克萨斯州高级计算中心 (TACC) 现在存在检测正在运行服务器的监视器，但是功能和用户操作流程十分不人性化：他们必须通过具有丰富经验、工作年限很长的后台管理者，通过报错信息得到运行异常的服务器编号，再去监测系统手动输入，通过查看后端返回的一系列运行参数，通过经验判断这台服务器的出错原因从而采取补救措施。类似这样的做法，缺点十分明显，比如在整个操作的过程中，操作流程十分僵硬，充斥着很多人为感知和人为判断，使得操作效率的低下；或者在某一时刻，如果因为在某一地区的用户普遍都出现了系统故障导致在短时间内出现了大面积的服务器运行异常，在如此庞大的工作量之下，工作人员的补救效率在不先进的系统下是很迟钝的，这种中间环节的纰漏就会导致解决过程的滞后，从而引发连带问题。

1.2 本文的主要贡献与创新

本论文的贡献主要是，通过数据可视化手段，实现了对服务器工作状态的全过程实时监测，克服了以前人为操作的种种不利，在功能上也进行了扩充，可以实现的具体功能和创新点如下：

- 及时而直观地了解到时间节点下，正常和异常服务器的运行状态参数 (CPU 利用率、memory 利用率、disk 利用率等) 和服务器本身的各项硬件指标 (CPU 数目、memory 大小等)，通过时间和服务器节点实时切换；
- 通过服务器所带的任务 (task)、作业 (job) 和实例 (instance) 信息，分析出导致服务器运行异常的原因是由什么任务，或者哪一台用户的异常操作而引起的，在可视化视图中分析异常原因；
- 直观地观察到服务器任务 (task)、作业 (job) 和实例 (instance) 之间的包含关系和数目大小，及时得到主要占用服务器资源的作业信息，做出相应调整；

- 通过热力图刷新迅速掌握某服务器节点在过去一段时间的工作状态，帮助工作人员寻找规律，辅助判断异常节点的出错动机；
- 了解到在整个有记录的时间线上所有服务器节点的运行状态和某一时刻下的全平台的服务器异常数目，定位异常高峰，联动其他组件有针对性地确定需要观察的时刻。

总而言之，该方法的创新点就是在传统的监测系统的基础上，通过增加可视化元素，使操作过程变得简洁流畅；在此思路的基础上，增加了各种功能，来帮助监测人员合理地定位异常，通过可视手段分析异常原因，定位故障源，从而有效地实时控制服务器的运行状态，为庞大的商业系统的正常运行保驾护航。

1.3 本论文的结构安排

本文的章节结构安排如下：

第一章：绪论部分，简述研究背景和现阶段研究进展，分析需求从而得出需求；

第二章：详细阐述本文的相关研究、研究意义、研究手段、简述数据可视化方法和研究数据集等研究基础；

第三章：分析系统设计思路，按模块化阐述系统的各部分功能，通过操作视图来具体演示系统的工作流程从而实例化分析，并且加入了测试用例来检验我们想法的正确性和系统的可行性；

第四章：全文的总结和归纳，分析系统仍然存在的缺陷，以及对未来工作的展望;

第五章：参考文献

第二章　项目意义及背景介绍

2.1　项目意义

对电子商务任务调度模式下的异常监测的可视化实现，对此类控制平台提出了一种新的思路和方法，也是数据可视化在交叉领域的全新进展。

2.1.1　电商集群的规模化

2018 年 7 月 10 日，2018 中国互联网大会发布了新一版的《中国互联网发展报告 (2018)》。报告中指出，2017 年，中国电子商务交易服务营收规模为 5027 亿元，首次突破 5000 亿大关。2017 年第三方互联网支付也达到 143.26 万亿，网络购物市场交易规模达 5.33 万亿元，而网络零售的市场交易规模为 7.18 万亿。中国网上支付用户规模达 5.31 亿人，其中手机支付用户就达 5.27 亿人，较 2016 年底增加 5783 万人，年增长率为 12.3%，规模增长迅速。正如网民在现实生活中体会到的，电子商务交易已经成为我们购买日常用品的首选手段，在中国庞大的网民基数和中国网络技术飞速发展的双重影响下，这一数字在以后还会以大增幅增长。根据商务部统计，2020 年预计中国网络零售市场规模为 9.6 万亿，是 2012 年的 10 倍之多。电子商务在深度影响着国民生活的同时，也掌握着货架经济的命脉，倘若电商集群由于某种原因崩溃的话，给整个国家带来的影响即将是灾难性的。

中国电子商务的龙头企业——阿里巴巴网络技术有限公司 (以下简称阿里巴巴)，在国内的电子交易平台里扮演着举足轻重的地位。根据 2018 年阿里巴巴公布的财年财报显示：2018 全年，阿里巴巴营收 2502.66 亿元人民币（约 398.98 亿美元），同比增长 58%，核心电商业务收入 2140.20 亿元人民币，同比增长 60%，均创下 IPO（Initial Public Offerings，指股份公司首次向社会公众公开招股的发行方式，简称 IPO）以来年度最高增幅，2018 财年净利润为 832.14 亿元人民币（约 132.66 亿美元）。如此庞大的交易额和成交额使专家和研究人员不自主地将目光放在其身上。2018 年阿里巴巴在 Github 上公布了一组数据，该数据刻画了阿里巴巴在 8 天地范围内 4000 台服务器的运行状态的数据，而这组数据也将成为本项目的研究数据集，在本章第三部分，将会着重对该数据进行详细介绍。

2.1.2　故障问题及手段

电子商务平台的实现其实并不是一个不能达到的要求，但是任何系统架构在达到一定规模之后，大大小小的问题往往会接踵而至，而这也是一个企业生存的关键。2018 年 8 月 1 日，阿里巴巴旗下的淘宝网交易平台，淘宝服务器出现大范围的

故障，全国多地网友在微博反馈称自己的淘宝崩溃无法查看订单，淘宝 App、PC 版网页均出现“网络竟然崩溃了”的提示，即使切换网络和重启手机也无效，在长达数小时的等待之后，该漏洞得到了修复。具体的原因，阿里巴巴并没有给出明确的回复，之后这件事也就不了了之。然而这已经不是阿里巴巴遇到的第一次服务器崩溃事件，每隔数月就会时不时的发生类似的服务器故障。由此可见，即使是如此宏大的电子商务企业也会因为后端服务器的宕机事件，造成企业形象的不良影响的同时也造成了利润的亏损。换句话说，如果能快速发现故障，找出导致该进程异常的原因，再利用分支等手段同时进行维护，结果将会截然不同。而本系统的设计初衷就是以阻止此类问题的发生而提出的。

2.1.3　设计优势

本系统设计将基于可视分析，在对服务器异常进行图形化展示的同时，还拥有异常定位、时序性监测等特点，对出现异常信号的服务器节点进行实时监测，并直观的展示造成该异常的任务，从而达到快速、准确的定位故障的目的。

2.2　项目背景及相关理论

2.2.1　研究现状

面对异常检测问题 [1]，Xiaowei Qin 等人提出了一种面向对象的检测框架 [2]，它具有两步聚类，称为沙漏聚类。两个参数，关键质量指标和因果参数，通过结合自组织映射（SOM）和 k-medoids 的混合算法，将它们聚类成不同的类型。Pei Yang 等人 [3]。建议使用生成的拮抗网络 (GAN) 来检测异常。Daojing He 等人 [4], 介绍使用软件定义网络 (SDN) 检测流量异常的优势。A.R.Jakhale[8] 使用数据挖掘 [5][6][7] 技术，利用滑动窗口模型和聚类技术检查网络流的异常数据包。Alireza Tajary[9] 等, 提出了一种吞吐量感知的瞬态故障检测方法，它利用了多核服务器处理器的特性。

为了识别大型，动态和异构数据中的异常 [10]，Nan Cao 等，介绍一种视觉互动 [11][12][13] 系统和框架，称为 Voila。该系统主要实现在线监测和与用户的互动。Y.B. Luo 等人 [14], 提出了一种基于流量限制可穿透能见度图（FL-LPVG）的异常检测方法。该方法基于网络流序构建复杂网络，挖掘相关图的结构行为模式，提取网络流特征序列，利用 LPG 将统计特征序列转换为关联图，通过数据挖掘和信息检测异常流量基于熵的理论技术。其优点是该方法大大简化了异常检测过程，有效降低了高维数据的维数。但是为了提高这个系统的效率，我们必须从大量数据中完全挖掘行为特征。因此，它肯定会带来如何处理大数据以及如何提取有效信

息的挑战。

Josef Kittler 等人 [18], 在解决异常检测问题时引入域异常的概念 [15][16]。异常有许多方面，每个域都是异常的许多方面之一。在此基础上，他们参考贝叶斯概率推理设备，并提出统一的异常检测框架 [17]，以识别和区分每个领域的异常。该设备通过定义各种异常属性来提供域异常事件的分类。该框架的创新特点是它暴露了异常的多方面性质，并且可以识别可能导致异常事件的各种原因，以及相应的检测机制。

2.2.2　数据可视化

本项目的特征是基于数据可视化的电商平台集群管理。数据可视化是由于视觉给人类带来的感知是最直观最有效率的一种获取信息的手段。目前国内的数据可视化行业也在蓬勃发展，涌现除了不少有能力的人才和惊艳的科技成果，下面将用一小部分篇幅来介绍一下这门领域的相关特点及其主要用途。

从生物学的角度来考虑，人类所有器官能接收到信息的 80% 都来自于视觉，在大数据时代下，对信息的表达则就显得尤为关键，而数据可视化也就承担起了这一重要任务，充当着数据和人类之间的关键载体。顾名思义，数据可视化是关于数据视觉表现形式的科学技术研究。其中，这种数据的视觉表现形式被定义为，一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。它是一个处于不断演变之中的概念，其边界在不断地扩大。主要指的是技术上较为高级的技术方法，而这些技术方法允许利用图形、图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性以及动画的显示，对数据加以可视化解释。

2.2.3　电子商务分布式任务调度

2.2.4　数据集分析

2.3　关键问题

2.3.1　数据处理

2.3.2　集群表示及定位故障源

2.3.3　故障域算法

2.3.4　时序性及空间性

参考文献

[1] J. Al Dallal, P. Sorenson, “System testing for object-oriented frameworks using hook technology”, Proceedings 17th IEEE International Conference on Automated Software Engineering, 2002.

[2] Xiaowei Qin, Shuang Tang, Xiaohui Chen, Dandan Miao, Guo Wei, “SQoE KQIs anomaly detection in cellular networks: Fast online detection framework with Hourglass clustering”, China Communications, pp.25-37, 2018.

[3] Pei Yang, Weidong Jin, Peng Tang, “Anomaly Detection of Railway Catenary Based on Deep Convolutional Generative Adversarial Networks”, 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2018.

[4] Daojing He, Sammy Chan, Xiejun Ni, Mohsen Guizani, “Software-Defined-Networking-Enabled Traffic Anomaly Detection and Mitigation”, IEEE Internet of Things Journal, 2017.

[5] Vania Bogorny, Shashi Shekhar,“Spatial and Spatio-temporal Data Mining”, 2010 IEEE International Conference on Data Mining, 2010.

[6] Hetal Thakkar, Barzan Mozafari, Carlo Zaniolo, “A Data Stream Mining System”, 2008 IEEE International Conference on Data Mining Workshops, 2008.

[7] G. Williams, R. Baxter, Hongxing He, S. Hawkins, Lifang Gu,“A comparative study of RNN for outlier detection in data mining”, 2002 IEEE International Conference on Data Mining, 2002. Proceedings, 2002.

[8] A. R. Jakhale, “Design of anomaly packet detection framework by data mining algorithm for network flow”, 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), 2017.

[9] Alireza Tajary, Hamid R. Zarandi, “An Efficient Soft Error Detection in Multicore Processors Running Server Applications”, 2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP), 2016.

[10] Chihiro Sakazume, Hiroyuki Kitagawa, Toshiyuki Amagasa,“DIO: Efficient interactive outlier analysis over dynamic datasets”, 2017 Twelfth International Conference on Digital Information Management (ICDIM), 2017.

[11] Nan Cao, Chaoguang Lin, Qiuhan Zhu, Yu-Ru Lin, Xian Teng, Xidao Wen, “Voila: Visual Anomaly Detection and Monitoring with Streaming Spatiotemporal Data”, IEEE Transactions on Visualization and Computer Graphics, pp. 23 -33, 2018.

[12] Cheong Hee Park, “Anomaly Pattern Detection on Data Streams”, 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), 2018.

[13] Li Xinran, “Research on massive data 3D-visualization”, 2012 IEEE International Conference on Computer Science and Automation Engineering, 2012.

[14] Y.B. Luo, B.S. Wang, Y.P. Sun, B.F. Zhang, X.M. Chen, “FL-LPVG: An approach for anomaly detection based on flow-level limited penetrable visibility graph”, 2013 International Conference on Information and Network Security (ICINS 2013), 2013.

[15] Hongbin Xia, Wenbo Xu, “Research on Method of Network Abnormal Detection Based on Hurst Parameter Estimation”, 2008 International Conference on Computer Science and Software Engineering, 2008.

[16] Zhixin Sun, Jin Gong, “Anomaly Traffic Detection Model Based on Dynamic Aggregation”, 2010 Third International Symposium on Electronic Commerce and Security, 2010.

[17] Yong Li, Wei-Yi Liu, “Backward probabilistic logic reasoning algorithm for decision problem with Conditional Event ALgebra on Bayesian networks”, 2008 International Conference on Machine Learning and Cybernetics, 2008.

[18] Josef Kittler, William Christmas, Teófilo de Campos, David Windridge, Fei Yan, John Illingworth, Magda Osman, “Domain Anomaly Detection in Machine Perception: A System Architecture and Taxonomy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.845-859, 2014.