面向国防安全的网络大数据分析与应用系统

许洪波，陈　波
中国科学院计算技术研究所 北京 100190

摘要
在调研国内外大数据分析与应用研究现状的基础上，针对国防安全领域现有业务体系中存在的数据碎片化、不规范、难共享等突出问题，提出面向国防安全的网络大数据分析与应用方案，将国防安全现实需求与大数据技术有机结合，既能够发挥大数据技术在多源异构数据融合、深层次安全信息挖掘、打破信息孤岛实现广泛共享等方面的优势，又能够适应现有的业务体系，快速产生实际效果。最后，对面向国防安全的网络大数据挖掘和分析相关技术进行了系统性介绍。

关键词
大数据；国防安全；大数据分析；多源异构数据融合

doi: 10.11959/j.issn.2096-0271.2015038

Network Big Data Analysis and Application Systems for National Defense Security

Xu Hongbo, Chen Bo
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract
Based on the state of the art of big data research, a national security-oriented network big data analysis and application system was proposed, against existing problems of national security systems, such as data fragmentation, nonstandard, difficult to share, and so on. In this system, the current national security requirements and big data technologies were organic combined. It could not only play the advantages of big data technology in multi-source heterogeneous data fusion, deeply mining security information, and breaking information island, but also share the advantages of the existing business architecture and quickly producing the actual effect. Finally, a systematic introduction to the national security-oriented network big data mining and analysis technologies was given.

Key words
big data, national defense security, big data analysis, multi-source heterogeneous data fusion

1 引言

目前，大数据已经发展成为科技界和企业界甚至世界各国政府关注的热点。Nature和Science等杂志相继出版专刊来专门探讨大数据带来的挑战和机遇[1]。在这样的背景下，网络空间的数据主权将成为继海、陆、空、天4个空间之后大国博弈的另一个空间。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为国家之间和企业之间新的争夺焦点[2]。美国认为大数据是“未来的新石油”，其2012年3月发布的《大数据研究和发展计划》不仅是一个推动美国在高技术领域继续领先的战略计划，更是一个保护美国国家安全、推动社会经济发展的计划[3,4]；2012年5月，英国政府注资建立了世界上第一个大数据研究所；同年，日本也出台计划重点关注大数据领域的研究。以美国为代表的发达国家在国家顶层推动下，正在通过大数据向更高的现代化水平的综合国力迈进。我国“十八大”报告中明确提出，网络空间与深海、深空是我国核心利益的关键领域。大数据领域的落后，意味着失守产业战略制高点，更意味着国家安全将在网络空间出现漏洞。

可以预见，未来国家之间的经济、政治和军事竞争将是大数据引领的竞争。网络大数据是其中重要的一环，通过对网络大数据进行定量分析和定性分析相结合的综合分析研判，能够进一步发现安全线索，掌握舆论倾向，追踪敏感及热点事件，预测发展趋势，对可能的危机情况进行预报预警，维护国家安全和社会稳定，提高国家竞争力。

2 国内外研究发展现状

2.1 国外相关领域发展情况

关于大数据的研究已经引起了包括美国在内的许多国家政府的极大关注。2012年3月22日，奥巴马政府宣布投资2亿美元启动“大数据研究和发展计划”，该计划旨在提高和改进人们从海量和复杂的数据中获取知识的能力，进而加速美国在科学与工程领域发明的步伐，增强国家安全。这是继1993年美国宣布“信息高速公路”计划后的又一次重大科技发展部署。美国政府认为大数据是“未来的新石油”，将“大数据研究”上升为国家意志，必将给未来的科技与经济发展带来深远影响。根据该计划，美国的国家科学基金会（NSF）、国立卫生研究院（NIH）、国防部（DoD）、能源部（DoE）、国防高级研究计划局（DARPA）、地质勘探局（USGS）6个联邦部门和机构，共同提高收集、存储、保留、管理、分析和共享海量数据所需的核心技术，扩大大数据技术开发和应用所需人才的供给。该计划还强调，大数据技术事关美国国家安全、科学和研究的步伐，并引发教育和学习的变革。例如，DARPA的大数据研究项目：多尺度异常检测项目旨在解决大规模数据集的异常检测和特征化；网络内部威胁计划旨在通过分析传感器和其他来源的信息，进行网络威胁和非常规战争行为的自动识别；Machine Reading项目旨在实现人工智能的应用和发展学习系统，对自然文本进行知识插入。美国能源部的大数据研究项目包括：机器学习、数据流的实时分析、非线性随机的数据缩减技术和可扩展的统计分析技

术，其中生物和环境研究计划的目标是大气辐射测量等气候研究设施，系统生物学知识库项目是对微生物、植物等生物群落功能的数据驱动的预测。美国国家人文科学捐赠基金会（NEH）项目包括分析大数据的变化对人文社会科学的影响，如数字化的书籍和报纸数据库、从网络搜索结果数据、传感器和手机记录交易数据。美国国家科学基金会大数据项目的重点也是突破关键技术，包括：从大量、多样、分散、异构的数据集中提取有用信息的核心技术；开发一种以统一的理论框架为原则的统计方法和可伸缩的网络模型算法，以区别适合随机性网络的方法[3,4]。

欧盟方面也有类似的举措。过去几年欧盟已对科学数据基础设施投资1亿多欧元，并将数据信息化基础设施作为Horizon 2020计划的优先领域之一。而2012年1月截止的预算为5千万欧元的FP7 Call 8专门征集针对大数据的研究项目，还是以基础设施为先导[3,4]。

在大数据基础设施和应用方面，美国已经走在世界前列。美国国家安全局投资20亿美元建设的“犹他数据中心（Utah data center）”占地超过9 000 m^2，负责对海量网络情报数据进行过滤、处理和融合。该中心的存储能力是yottabyte（10^{24} bit），足以存储未来100年整个人类的电子信息。2013年5月，美国国家安全局宣布将建设规模超过犹他数据中心6倍的新型“高性计算中心（the high performance computing center-2）”，预计2016年投入使用。

下面对美国相关的项目和机构进行介绍。

2.1.1 PRISM（棱镜）计划

NSA从2007年起开始尝试利用大数据开展反恐工作，追踪和发现潜在的恐怖活动。其中典型的是PRISM计划，其正式名称为“US-984XN”。PRISM项目能够对实时通信和历史数据进行深度监听，被监控的信息可以说是一切事物，包括电子邮件、即时消息、视频、照片、存储数据、语音聊天、文件传输、视频会议、登录时间和社交网络资料的细节。通过棱镜项目，国家安全局甚至可以实时监控一个人正在进行搜索的网络内容。许可的监听对象包括任何在美国以外地区使用参与项目公司服务的客户，或者是任何与国外人士通信的美国公民。几乎所有的美国网络大公司都已加入该计划：微软公司在2007年9月首个参与该项目，雅虎公司于2008年3月、谷歌公司于2009年1月、Facebook公司于2009年6月、Paltalk公司于2009年11月、YouTube公司于2010年9月、Skype公司于2011年2月、AOL公司于2011年3月、苹果公司于2012年10月相继参与其中。此外，Dropbox公司也被指控“即将加入”该项目。《卫报》获得的热力图显示，2013年2-3月，美国国家安全局在短短30天内，就从全世界互联网上收集到970亿条数据，其中近30亿条来自美国。PRISM项目背后的关键技术是Accumulo，该系统基于Apache Hadoop系统框架设计，类似于谷歌公司的BigTable存储系统。Accumulo擅长分析庞大的数据，从而生成众多的图表，发现和强化这些数据间的连接。系统可以管理数月甚至数年的资讯，轻易发现怀疑恐怖分子的通话网路以及涉及的参与者。NSA通过已掌握的恐怖分子的活动数据与嫌疑者比较，从而决定是否需要进一步行动。

2.1.2 Recorded Future系统

Recorded Future是美国马萨诸塞州一家创业企业，号称世界上第一个利用时序分析引擎（temporal analytics engine）

预测未来的工具，该系统可以通过扫描并分析成千上万个网站、博客、Twitter账户的信息来找到目前和未来人们、组织、活动和事件之间的关联性，可以给出任何事件的在线发展趋势。其预测基础包括语义分析、统计数据、时间推理等，简单地说就是基于过往的历史数据，利用搜索引擎对关键词进行分析，最终以图表、数据和文字的形式展现一个预测的结果。目前，该系统已经吸引了很多重量级客户，包括美国国防部。利用客户提供的可靠性很强的数据，Recorded Future通过自己的搜索计算方法进行预测。其预测的关键词包括：who、when、where等关于人物、时间、地点的基础数据，然后通过图表和数据展现出来。据有关案例考证，Recorded Future有能力辨认出事件和早期趋势，2010年3月21日，以色列总统佩雷斯指控黎巴嫩真主党有飞毛腿导弹，Recorded Future搜索了黎巴嫩真主党领袖纳斯鲁拉以前的言论，发现一个月前就有确凿的证据证明佩雷斯的指控是没错的。

2.1.3 2049研究所

成立于2008年1月的2049项目研究所，官方的宗旨是指导美国政府决策者至21世纪中期构建一个所谓的“更安全”的亚洲。该组织通过前瞻性的、特定区域的安全研究和政策解决方案，在公共政策领域填补了一项空白。他们采用跨学科方法对社会、经济、管理、军事、环境、技术和政治的发展趋势进行严谨的分析，通过对网络上看似微小、散乱、毫无关系的各类报告和论文信息进行深入挖掘、融合分析，形成权威的战略研究报告。例如，该机构研究考察地区的恐怖主义/极端势力的影响，控制流行性疾病、自然灾害、环境和能源安全问题以及被安全专家越来越多地关注但仍相对较新的领域。利用中国大量的可以在线获取的文档、报道、论文、专利等，研究评估中国快速的经济、社会和军事发展对亚太安全环境的影响，专门针对中国军队的武器装备进行数据分析，指导周边国家和美国如何应对中国的崛起。在2010年的《China’s Nuclear Warhead Storage and Handling System》报告中，详细分析了中国核弹头的存储和处理体系，包括中国人民解放军第二炮兵最重要的核弹存储基地——第22基地的位置、组织结构等以及其他核弹基地的分布位置、移动方式、安全性和可靠性、管理体系等。文末列举了报告来源的大量关于中国军队的在线文档信息（报道、章程、会议报告等）。2013年的报告《The Chinese People’s Liberation Army’s Unmanned Aerial Vehicle Project: Organizational Capacities and Operational Capabilities》则是分析了中国无人机的进展，详细分析了研发机构、设计能力、产品形态、主要人员等信息以及在各个部队中的部署情况。这些信息来自于对大量相关网络报道、论文信息的深入分析、融合分析。

2.2 我国网络大数据研究应用现状

网络大数据主要包括互联网、社交网、通信网等多通道信息，网络信息的高效、全面获取是网络大数据分析、预警的前提和基础。目前的信息获取技术面临的困难主要包括：近70%的网站采用了Javascript及AJAX动态脚本技术以及社交网络内的访问授权限制等导致采集困难；新兴网络媒体具有动态交互性，隐蔽性更强，难以实时掌控；多模态、多通道信息广泛分布，相互交织，需要全面获取、融合分析、交叉验证；在信息分析方面，由于网络信息中充斥着大量垃圾信息，需要大海捞针，从海量信息中甄别有价值的线索[5]；

网络信息传播的演化和大量涌现使得发现与追踪非常困难，需要及时识别和监测热点、突发话题和重大事件的出现与扩散；如何结合网络事件的地理定位、总体态势分析与交叉验证等技术手段来进行综合态势的预测与演练，为辅助决策提供多维度、立体化的分析与预测手段[6]；网络信息的内容理解和判定存在很强的不确定性和特征空间的高维诅咒问题，需要对网络信息进行多立场、多视角的精确研判和分析验证[7]。

针对国防安全需求，我国相关管理部门已经开始了对网络大数据进行分析挖掘的系统研发与应用部署工作[8]。目前存在的主要问题是获取不全面、分析不深入、研判不准确、响应不及时等，不少研发工作仍然处于低水平重复状态，迫切需要从理论和关键技术上取得突破。

3 面向国防安全的网络大数据分析与应用系统

3.1 系统架构

研究面向国防安全的网络大数据分析与应用系统，支持网络大数据的感知汇聚、统一管理，针对重要安全应用的融合分析和深入挖掘以及统一的数据服务，可以实现网络大数据的全面感知和深度融合，支持新的安全业务模式或深化传统安全信息分析效果。系统总体架构设想如图1所示，具体的研究内容如下。

3.1.1 多源异构网络大数据汇聚

接入全球各大网络信息来源，采集互联网网页、视频、音频、新闻媒体、商业数据库、报刊杂志、遥感影像、空间地理信息等网络大数据，按照通用的接口标准进行汇聚融合处理，形成有序信息，为深入综合分析奠定数据基础。

3.1.2 网络大数据管理引擎

基于数据融合与信息交换的标准规范，针对不同来源的多格式数据建立统一的数据模型，基于高性能的大数据统一管理引擎实现多源异构数据的高聚合带宽存储读写、分布式统一计算和复杂查询处理。支持业务部门根据自身业务需要获取相关的原始数据或初级处理数据，进行个性化处理。

（1）存储层负责对异构数据类型的统一存储，自下而上又分为设备层和存储虚拟化两层。设备层由定制化存储设备组成，设备间通过高速互联网络进行互联。统一分布式存储对物理设备进行虚拟化，将多个设备进行整合，提供统一访问视图，同时在资源虚拟化方面实现对物理设备的弹性增减。

（2）数据管理层完成对数据生命周期

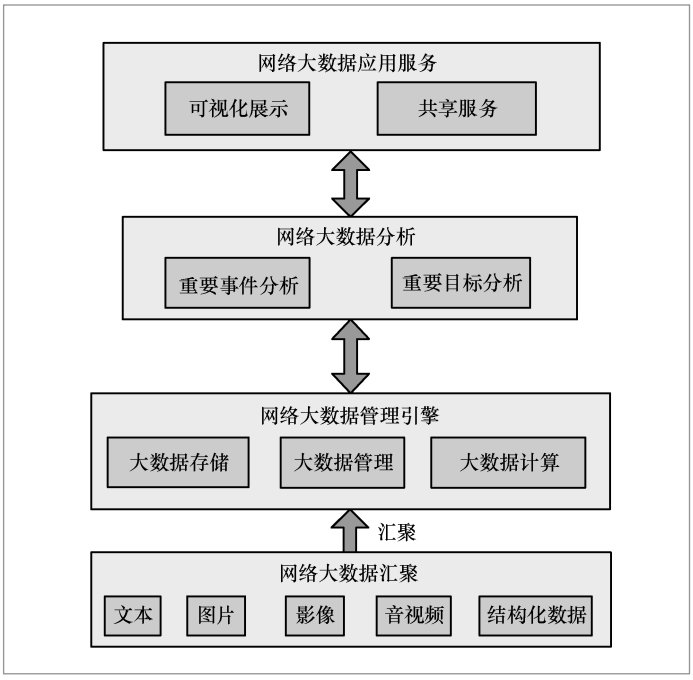


图1 基于大数据引擎的目标关联分析系统架构

的管理，包括数据的接入、存储、组织和维护。针对丰富的异构类型数据，需要对结构化及非结构化数据采用不同的组织、存储和访问方式，需要自动将数据分发到相应的存储组件，同时做到对应用透明。除此之外，海量数据所面临的HA（high available，高可用性集群）、QoS以及性能加速要求也由管理层完成。

（3）计算服务层完成对大数据的处理计算和高效访问，包括并行计算框架和大数据统一访问接口。并行计算框架兼容主流的计算框架，包括MapReduce、MPI、BSP、BOT等。大数据统一访问接口对统一存储的大数据进行在线、离线、随机、批量等多种访问支持，提供多种访问接口，支持上层的分析功能[9,10]。

3.1.3 重要事件分析

针对关系到国家安全的重大社会问题，对多来源网络大数据进行突发敏感事件的主动发现、特定事件和主题的智能跟踪，分析事件的演化传播趋势和大众观点倾向，对潜在的危机进行趋势预测，对危机情况进行及时预警。

3.1.4 重要目标分析

基于遥感影像和监视视频，提取陆海空重要目标（如车辆、舰船、飞机等）的属性、方向、速度和位置等信息，通过多源数据的融合、对比和分析，掌握目标的动态变化趋势，辅助发现可疑目标。

3.1.5 可视化展示

对各类手段所获取的文字、图像、音视频等多元信息素材进行合理编排组织，支持各类信息的二三维联动、分层分类分级综合展示，支持基于数字地球、信息主题、时间轴等多种展示模式[6]。

3.1.6 共享服务

将各类有用信息按照时效性要求和预定的分发规则，通过检索、下载、订阅、广播等多种方式向各级用户提供服务，向用户分发基础资料、事件资料、目标资料、综合资料以及应用软件工具等。

3.2 指导原则

系统建设应遵循如下原则。

- 可扩展性：系统应具备良好的可扩展性，以便将来随着业务和需求的增加扩展系统规模。
- 规范化：形成网络安全系统，建设标准体系，涵盖网络安全业务的全过程环节，为网络安全业务提供规范化指导。
- 易用性：功能定义清晰合理，用户界面简明友好，符合互联网的常用风格和界面规范，操作简单，具备用户交互和自学习能力。
- 开放接口：提供二次开发接口，能够支持二次应用开发。

4 网络大数据分析技术

面向国防安全的网络大数据分析与应用系统涉及一系列传统数据挖掘和大数据分析技术的支持，主要包括网络数据采集相关技术、网页信息抽取技术、内容特征识别技术、事件演化分析技术、网络社区发现技术、倾向性分析技术、音视频获取处理相关技术等。

4.1 网络数据采集相关技术

（1）高速采集技术

针对大规模Web信息采集，主要采用分布式多线程、异步socket通信、本地

DNS缓存机制来实现原始网页的高效并行化采集，在有限的带宽和系统资源的限制下，使采集速度最大化。

（2）噪音过滤技术

在互联网的实际网页中存在大量的噪音链接，这些链接指向的页面并不是系统所需要的，如何消除这些噪音链接而直接定位到目标页面链接是一个具有挑战性的问题。

（3）Web信息重复检测技术

有研究表明，互联网上将近30%的页面是重复的。实际的采集系统都是多机协同工作的并行采集系统，必须要处理好网页的查重才能够保证多机采集不会采集到大量重复信息。

（4）动态网页采集技术

目前网络中有很多链接是通过JavaScript动态生成的，用常规采集器无法实现网页的下载。据统计，有超过73%的网站采用了动态链接生成技术，如果不解决动态链接生成的技术难题，将影响这些网站信息的有效获取。

4.2 网页信息精确抽取技术

实际环境中的网络数据具有海量、格式复杂、变化频率高等特点，这使得信息抽取技术面临巨大的挑战。要想保证后期的数据处理质量，必须有效解决以下3个问题：在线抽取的效率高，包括准确率足够高、速度足够快；面对形态各异的网络数据，抽取方法的适应性强；维护代价足够低。

4.3 内容特征识别技术

面向网络文本的内容特征识别技术包括：面向网络文本的命名实体及其他特定实体识别（如人名、时间、机构名、地址等）；网络流行语的自动识别与分析；面向特定文本内容的指纹特征识别等。

由于网络信息的动态性和非规范性，网络文本中存在大量新的语言特征。需要挖掘那些以前没有出现过的或者很少使用，而最近使用较频繁的词、短语或有确定语义的字符串。

4.4 事件演化分析技术

事件是一个与时间相关的概念，每一个事件都要经历从开始到爆发再到平息消失的过程，这个时间跨度称为事件的“生命期”。不同事件在生命期的各个阶段发生、发展的趋势既有差异又有共性。基于这些共性特征可建立事件演变的模型，通过发现演变过程中的关键点，判断出事件演化的状态和趋势。

4.5 网络社区发现技术

Web、邮件、博客、即时通信等消息传播网络中具有相似特征并紧密关联的文本、事件和网络群体（主体）的聚集现象被统称为网络社区。网络社区分析的主要目的是从多通道网络信息中发现影响社会稳定的群体事件，并挖掘出与特定内容传播行为相关的隐性关系。

4.6 倾向性分析技术

倾向性分析是挖掘网络文本内容蕴含的各种情感、信念、态度、意见和情绪等大众观点信息。目前倾向性分析面临的主要问题是大部分的研究方法和技术手段都与相应的领域密切相关，需要研究跨领域倾向性分析的通用技术手段。

4.7 音视频获取处理相关技术

（1）音视频信息发现与获取技术

基于互联网的音视频信息主被动模式

发现与获取技术主要包括：基于P2P的音视频信息有效发现获取技术、基于Web的音视频信息全面及时准确发现获取技术、音视频信息在互联网上的扩散影响分析及传播情况追踪等。

（2）基于样例的视频内容检测技术

基于样例的视频内容检测技术主要是针对各种画面变化，构造顽健的特征描述算子来建模视频的视觉内容。同时，视觉特征之间的匹配算法和特征聚类技术的设计必须满足快速、准确的要求。另外，因为网络和数据库保有大量的视频内容，视频内容检测系统应该采用高级的特征数据索引架构，以实现在线实时的快速分析和查找功能。

（3）视频文字识别技术

由于部分网络信息以视频的形式存在，同时视频中的文字也对图像的内容进行了描述，可以作为视频类别判断的依据，因此视频帧中的文字检测也是网络视频监控的内容之一。

（4）视频人脸识别技术

人脸检测是特定人物图像检测的基础。人脸识别主要包括面部关键特征提取、姿态校正、光照补偿算法、人脸识别等。其中，人脸识别核心算法是人脸识别成功的关键，面部特征配准、人脸表示和判别特征分析是关键环节。

（5）音频分析与分类技术

音频分析与分类技术是在连续的音频信号流中，找出音频特征发生突变的信号点，把变化出现的地方作为分割点将音频流切分开，从而将连续音频信号分割成长短不一的音频例子。通过比较音频例子与已知音频的相似性，将每个音频例子归类到不同音频类别，对其进行中级语义标注，确定其分类。通过提取音频文件的时域、频域、时频域特征来分析音频片断的语义含义，可以直接从音频流中发现监控信息。

（6）文本音视频综合处理技术

文本信息与音视频信息的综合处理技术快速提取与匹配多媒体信息中的文本信息，挖掘文本与音视频节目之间的相互关系，利用文本与音视频节目进行相互表示和描述，从而将文本处理技术和音视频分析处理技术进行有机融合，更加全面深入地分析处理音视频信息。具体包括文本音视频综合检索技术、文本音视频综合分类技术、文本音视频综合过滤技术等。

5 结束语

面向国防安全的网络大数据分析与应用系统，将在现有独立分散的各类网络信息搜集处理系统基础上，集成基于多种来源的各类媒体格式的数据分析工具，提供统一开放的多通道网络大数据搜集、综合处理分析、危机预警和共享分发平台，弥补我国成体系、成规模、一体化建设网络信息搜集分析平台的不足，大大提升面向国防安全的网络大数据利用能力。

参考文献

[1] Marx V. Biology: the big challenges of big data. Nature, 2013, 498(7453): 255~260

[2] 李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊，2012，27(6)：647~657
Li G J, Cheng X Q. Big data research: the major strategic areas for future science and technology, and economic and social development——research status of big data and scientific thinking. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647~657

[3] 中国计算机学会大数据专家委员会. 中国大数据技术与产业发展白皮书（2013），2013

CCF Task Force on Big Data. White Paper on Big Data Technology and Industry Development in China(2013), 2013

[4] 中国计算机学会大数据专家委员会, 中关村大数据产业联盟. 中国大数据技术与产业发展白皮书(2014), 2014
CCF Task Force on Big Data, Zhongguancun Big Data Industry Alliance.White Paper on Big Data Technology and Industry Development in China(2014), 2014

[5] Batini C, Cappiello C, Francalanci C, *et al*. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), 2009, 41(3)

[6] Johnson C, Moorhead R, Munzner T, *et al*. NIH/NSF Visualization Research Challenges Report. Los Alamitos: IEEE Computing Society, 2006

[7] Jin X L, Wah B W, Cheng X Q, *et al*. Significance and challenges of big data research. Big Data Research, 2015, 2(2): 59~64

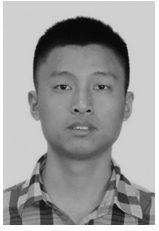
[8] 杨小牛, 杨志邦, 赖兰剑. 下一代信号情报侦察体系架构: 大数据概念的应用. 中国电子科学研究院学报, 2013, 8(1): 1~7
Yang X N, Yang Z B, Lai L J. The structure of the next generation SIGINT reconnaissance: application of the big data. Journal of CAEIT, 2013, 8(1): 1~7

[9] Das S, Sismanis Y, Beyer K S, *et al*. Ricardo: integrating R and hadoop. Proceedings of the SIGMOD, Indianapolis, Indiana, USA, 2010: 987~998

[10] Wegener D, Mock M, Adranale D, *et al*. Toolkit-based high-performance data mining of large data on MapReduce clusters. Proceedings of the ICDM Workshop, Miami, FL, USA, 2009

作者简介

许洪波, 男, 博士, 中国科学院计算技术研究所副研究员、硕士生导师, 主要研究方向为互联网挖掘与搜索、大数据分析与计算等。



陈波, 男, 中国科学院计算技术研究所研究实习员, 主要研究方向为大数据计算。

收稿日期: 2015-11-05
论文引用格式: 许洪波, 陈波. 面向国防安全的网络大数据分析与应用系统. 大数据, 2015038
Xu H B, Chen B. Network big data analysis and application systems for national defense security. Big Data Research, 2015038