



基于大数据融合算法的 DNS 日志分析系统

廖明, 陈明, 周冀, 向小华, 李芳, 焦叶芬
(中国移动通信集团重庆有限公司, 重庆 404100)

摘要: 总结了目前行业对于 DNS 日志的大数据分析存在的问题, 并设计了一种基于大数据融合算法的 DNS 日志分析系统, 该系统融合了备案信息库、网站库、域名库、IP 地址库、CNAME 域名库以及安全信息库六大信息库数据, 具有强大的日志分析性能, 分析内容范围全面, 分析深度高, 同时具有灵活的报表定制功能, 为电信运营商提高管理效率、优化网络安全、提升本网率提供决策依据。

关键词: DNS 日志; 大数据; 信息库; 用户行为分析

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019065

DNS log analysis system based on big data fusion algorithm

LIAO Ming, CHEN Ming, ZHOU Ji, XIANG Xiaohua, LI Fang, JIAO Yefen
China Mobile Group Chongqing Co., Ltd., Chongqing 404100, China

Abstract: The existing problems of big data analysis of DNS logs in the current industry were summarized, and a DNS log analysis system based on big data fusion algorithm was designed. The system integrates six information databases, such as record information database, website database, domain name database, IP address database, CNAME domain name database and security information database. It has powerful log analysis performance, comprehensive content analysis, and classification. With high analysis depth and flexible report customization function, it can provide decision-making basis for telecom operators to improve the management efficiency, optimize network security and enhance the network rate.

Key words: DNS log, big data, information database, user behavior analysis

1 引言

随着互联网的飞速发展, 移动终端和智能手机的普及使得用户能够更加快速方便地获取信息, 同时人们对移动网络服务的质量要求越来越

高。面对信息爆炸式增长产生的流量, 运营商的收入没有实现跨越式提升, 反而面临数据流附加值被互联网公司“吸”走, 沦为“管道”的挑战。为应对上述挑战, 电信运营商开始从战略层面对大数据展开部署, 通过对海量数据深度分析来挖

收稿日期: 2018-12-17; 修回日期: 2019-03-28

基金项目: 中央高校基本科研业务费专项资金资助项目; 国家自然科学基金资助项目 (No.61501042)

Foundation Items: The Fundamental Research Funds for the Central Universities, The National Natural Science Foundation of China (No.61501042)



掘其核心价值,将数据转化为精准营销、精细化运营以及开拓市场的生产力^[1-3]。

电信运营商常见的大数据应用主要有信令分析系统^[4-6]和 DPI 分析系统^[7-8]。其中,信令监测系统采用高阻跨接、分光、端口镜像或交换机内部收敛等方式获取信令链路上传递的各类原始信令和协议数据,进行信令协议解码、分析处理、网络和业务性能指标统计,从而再现业务接续全过程,帮助维护人员掌握并分析网络和业务运行情况,及时发现和排除网络中的故障和隐患。参考文献[4]全面分析了信令采集与监测系统的研究现状与进展情况。参考文献[5]结合现有信令监测系统功能架构和业务需求,论述了信令监测系统的发展趋势,提出了基于大数据的信令监测系统建设方案,并讨论了系统建设中的关键技术。参考文献[6]在传统信令监测系统的基础上结合大数据特征,设计了信令监测系统建设方案,讨论了其方案的关键技术和功能架构,并详细介绍了其方案在网络优化、业务经营、网络安全 3 方面的应用。而对于基于 DPI 的大数据分析系统,现有的监测技术大致有基于端口数据分组的监测技术、基于应用层特征的深度分组检测(deep packet inspection, DPI)技术和基于流量行为的深度流检测(deep flow inspection, DFI)技术^[7]。参考文献[8]针对基于正则表达式匹配算法的深度数据分组检测方法,提出了基于优先级的迁移边压缩方法,融合相同目的状态最多的迁移边,从而减少存储空间需求。参考文献[9]分析了基于 DPI 的大数据分析技术原理、分类以及在网络中的部署方式,然后以实例的形式说明了通过 DPI 技术能够获取到的各类数据,并结合移动大数据的当前应用领域,指出了 DPI 数据的适用范围,并提出了 DPI 数据在移动大数据的应用过程中面临的挑战。

尽管目前基于信令分析系统和 DPI 分析系统的大数据应用受到了业界广泛的研究,然而对于信令分析系统,其信令数据庞大,存储成本高,

同时信令数据复杂,分析投入成本巨大且分析效率低^[10]。而对于基于 DPI 的分析系统,可以采用软件实现,优点是系统更新简单,但是基于软件实现的系统不能满足网络高速数据分组处理的吞吐量要求^[9]。硬件实现的 DPI 系统具有效率高等优点,但是系统升级难、成本高,且系统识别知识库需要占用大量内存,硬件片内存储空间有限,制约了知识库冗余,系统拓展性不够。因此,针对上述分析系统的不足,本文提出了一种基于大数据融合算法的 DNS 日志分析系统^[11]。在该系统中, DNS 日志相对较小,便于转储且分析相对较容易; DNS 日志包含的信息量大且数据集中,与其他数据的相关性强同时 DNS 日志分析成本低,因此 DNS 日志是运营商利用所掌握的数据进行深入分析、挖掘市场价值性价比比较高的切入点^[12]。

2 行业现状

目前行业对于 DNS 日志的大数据分析,只是单纯地分析 DNS 日志或者结合一个数据库进行分析。单纯地分析 DNS 日志,可以了解 DNS 的服务质量以及 DNS 请求次数等信息;而 DNS 日志结合认证信息库分析用户行为,可以了解用户的属性; DNS 日志结合 IP 地址信息库分析用户请求资源的调度结果,如参考文献[13-14],统计流量流向,为本网率提升提供依据; DNS 日志结合备案库分析用户访问域名的合法性,并能分析网站所属公司包含的所有域名; DNS 日志结合网站漏洞库和病毒域名库分析互联网网站和网络的安全性,了解网站和网络存在的安全问题; DNS 日志结合网站/应用库分析了用户的访问网站/应用类型、所属行业的信息。目前行业分析存在的问题如下^[15-17]。

(1) 融合数据库较少

缺少定制化分析工具和方法,互联网内容安全性和合法性管理困难。

(2) DNS 服务质量难衡量

DNS 解析错误原因分析难、DNS 调度分析困

难、网内重复引入相同内容,造成资源浪费,且优化服务质量难以衡量。

(3) 设备投入成本高

为应对用户访问量大、日志多、精细化分析困难,采用大量设备投入来提升分析能力,效率低且成本高。

针对这些问题本文设计了一种基于大数据融合算法的 DNS 日志分析系统,通过 DNS 日志数据融合多个信息库分析整个互联网内容的合法性和安全性、用户访问内容的流量流向,作为运营商提升网络安全管理、提供调度优化的依据。

3 系统介绍

DNS 日志融合多元数据库信息,实现多维度分析。基于此,运维人员可以挖掘更多有价值的信息,有利于问题故障定位和决策支撑。传统以 DNS 日志或者 DNS 日志融合单一数据库分析问题的视角有限,根据统计结果可以了解域名维度的流量流向,但是更深层次的用户行为偏好访问、热门资源备案情况关联、本网率

问题排查等无法洞察。本系统将联动网站库、域名库、IP 地址库等,利用大数据分析技术,从更高层次、更宏观层面分析问题,从更细化、更微观层面定位问题并解决。与此同时,系统优化了数据库存储结构,即便融合多数据库,日志处理任务更多,但依旧保持处理速度快、性能优的特点。

3.1 系统架构

基于大数据融合算法的 DNS 日志分析系统架构如图 1 (a) 所示,其主要分为数据访问层、业务逻辑层与表示层。通过将 DNS 日志数据与六大信息库融合分析,能够得到更有价值的信息,帮助运营商提升管理效率、优化网络安全,为提升本网率提供决策依据。各模块具体介绍如下。

(1) 数据访问层

由六大信息数据库(域名库、网站库、CNAME 域名库、IP 地址库、备案库、安全信息库)和 DNS 日志组成。该层负责数据采集,并将六大信息数据库和 DNS 日志形成 DNS 融合数据库,如图 1 (b) 所示。再将 DNS 日志融合数据库输入业务逻辑层。

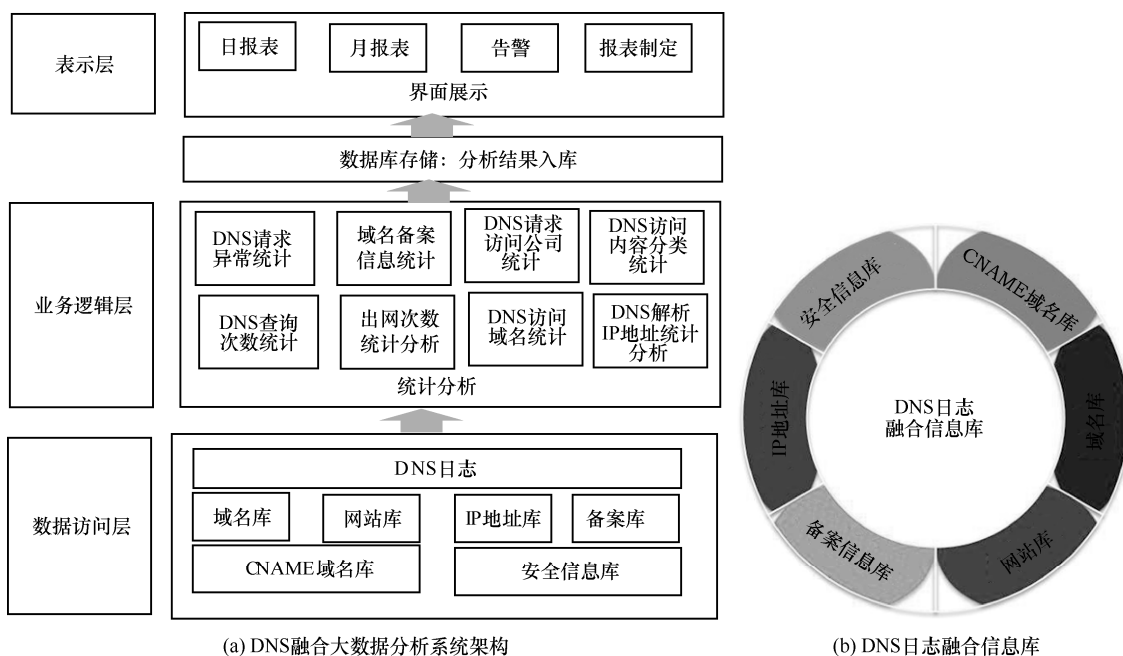


图 1 DNS 融合大数据分析系统



(2) 业务逻辑层

该层主要负责分析处理数据访问层提交的数据。业务逻辑层根据 DNS 日志融合数据库, 对用户行为、用户属性、热点资源、流量流向、域名是否合法、网络是否安全等进行分析, 得到 DNS 请求异常统计、域名备案信息统计、DNS 请求访问公司统计、DNS 访问内容分类统计、DNS 查询次数统计、出网次数统计、DNS 访问域名统计和 DNS 解析 IP 地址统计结果, 然后将该分析结果存入数据库, 并输入表示层。

(3) 表示层

该层用于呈现分析结果给运维人员, 从而为运维人员优化网络提供建议。首先, 表示层根据输入的分析结果形成报表, 展示每日、每月的网

络状态, 如热点资源、流量流向、网络安全性等。如果分析结果显示当前网络处于异常状态, 该层则输出告警信息, 以警示运维人员。该层通过展示业务逻辑层的分析结果呈现网络状态, 为运营商提高网络运营效率、保证网络顽健性提供依据。

3.2 功能特性

DNS 融合大数据分析系统的功能特性主要有以下 5 个方面。

(1) 热门网站类型分析

通过 DNS 日志分析融合网站库(如图 2 所示)和域名库(如图 3 所示), 可以全方位统计分析用户访问网内外内容, 掌握用户访问习惯和趋势; 同时可以了解热门网站类型, 为运营商内容引入提供参考。该系统分别从 3 个维度进行统计: 从



网站名称	网站类型	域名数量	引入域名数量
微信电话本	VoIP业务	2	2
必通	VoIP业务	1	0
易聊	VoIP业务	1	1
有信	VoIP业务	1	1
连连	VoIP业务	508	316
爱聊	VoIP业务	3	3
阿里通	VoIP业务	3	3
飞音网络电话	VoIP业务	4	4
78动漫	动漫	4	2
AC	动漫	1	0
CC漫画图库	动漫	65	46

图 2 网站库



域名	归属网站	归属网站分类	域归属分类	是否引入
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否
weixin.com	微信	即时通信	即时通信	否

图 3 域名库

网站和域名的维度进行统计,根据热门访问网站及类型排序,对出网访问次数多的网站进行统计,对网外的网站/域名解析来源,对网内网站/域名访问解析具体分布;从移动应用维度进行统计,根据热门访问应用及类型进行排序,对出网访问次数多的应用进行统计,对网外的移动应用解析来源,对网内移动应用访问解析分布;从公司的维度进行统计,根据热门公司访问进行排序,统计公司包含域名的出网情况。

(2) 域名备案情况分析

通过 DNS 日志分析融合备案信息库,统计用户访问域名备案情况,了解引入网内的域名是否已备案,提升运营商 IDC 备案管理的效率;运营商可以了解域名所属公司的业务蓝图,为市场开拓提供依据。

该系统可以统计网内引入域名的备案情况,

分析互联网域名的备案情况,统计分析公司包含的网站及域名。

(3) 安全性分析

根据 DNS 日志分析融合安全信息库,可以分析用户安全和网站安全,为运营商保障 IDC 安全和互联网安全提供重要信息。

该系统可以扫描网站的安全漏洞并进行统计分析,快速了解各个网站的安全漏洞,以做出相应的措施;系统支持定期更新应用漏洞信息数据,定期扫描,及时了解用户访问应用的安全隐患,保障本网 IDC 安全;结合病毒域名信息分析哪些用户 IP 地址已中毒,可按病毒的危害程度做出相应的处理措施,保障互联网安全。

(4) 流量分析

根据 DNS 日志分析融合 IP 地址库,如图 4 所示,可以从各个角度分析流量流向,为资源引

基础库维护>>IP地址库							
地址库: 全国		新增	导入	导出结果集	删除	保存	取消
IP地址段:				所属运营商		请输入查询关键词	
起始IP	终止IP	所属运营商	所属地区	业务类型	创建时间	修改时间	备注
0.0.0.0	0.255.255.255	IANA	IANA	设备地址	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.0.0	1.0.0.255	亚太互联网络信息中心	澳大利亚	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.1.0	1.0.3.255	电信	中国福建	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.4.0	1.0.7.255	墨尔本Goldenit有限公	澳大利亚	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.8.0	1.0.15.255	电信	中国广东	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.16.0	1.0.31.255	东京I2Tslnc	日本	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.32.0	1.0.63.255	电信	中国广东	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.0.128.0	1.0.255.255	CZ88.NET	泰国	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.0.0	1.1.0.255	电信	中国福建	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.1.0	1.1.1.0	谷歌公司AS55698-AP	澳大利亚	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.1.1	1.1.1.1	亚太互联网络信息中心	澳大利亚	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.1.2	1.1.1.255	谷歌公司AS55698-AP	澳大利亚	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.2.0	1.1.7.255	电信	中国福建	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.8.0	1.1.63.255	电信	中国广东	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.64.0	1.1.127.255	东京都新宿区歌舞伎町	日本	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.1.128.0	1.1.255.255	沙空那空TOT公共有限	泰国	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.0.0	1.2.1.255	电信	中国福建	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.2.0	1.2.2.255	北龙中网(北京)科技有	中国北京	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.3.0	1.2.3.255	CZ88.NET	澳大利亚	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.4.0	1.2.4.7	中国互联网络信息中心	中国北京	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.4.8	1.2.4.8	中国互联网络信息中心	中国北京	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.4.9	1.2.4.255	中国互联网络信息中心	中国北京	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.5.0	1.2.7.255	电信	中国福建	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.8.0	1.2.8.255	网络信息中心	中国	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.9.0	1.2.127.255	电信	中国广东	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入
1.2.128.0	1.2.255.255	CZ88.NET	泰国	其他	2016-09-18 17:02:00	2016-09-18 17:02:00	后台导入

图4 IP地址库



入、调度优化提升本网率提供重要依据。

该系统将 DNS 日志分析与 IP 地址库融合, 可以进行流量流向分析, 分析用户请求域名被调度在网内、网外的具体分布, 并统计多维度的本网点击率; 可以量化考核运营商优化效率, 通过网内请求次数、缓存、CDN、IDC 的请求次数以及网内请求的比例情况; 可以建立引入偏差机制, 对于已引入但仍然被调度出网的请求进行重点分析; 统计分析重复引入资源, 基于 CDN、缓存、IDC 内容以及各个厂商之间引入的内容进行分析, 减少资源重复引入; 设置告警阈值, 基于本网率的变化、DNS 请求次数变化、出网次数变化、网内次数变化等指标产生告警, 帮助运营商迅速定位并及时优化。

(5) 分析结果展示

DNS 融合大数据分析系统可以灵活地定制报表。该系统默认提供多种类型的报表模板, 并支持自定义报表模板功能, 用户可设置导出时间、导出数据库、导出内容等, 灵活定制, 让用户能自主导出各式各样的报表且导出报表效率高。

此外, DNS 融合大数据分析系统还具备 IP 地址反查功能。通过分析指定 IP 地址所服务的域名, 了解内容服务厂商单台设备服务的 CP (content provider); 并且, 该系统还具备日志扫描分析功能。针对 DNS 日志数据量大, 查找、定位问题困难的现象, 日志扫描分析功能将日志按照时间段、源/目的 IP 地址、域名等条件进行扫描并下载, 生

成报表, 方便用户更简单、更快速地定位问题。

4 实施效果

目前该系统已在运营商部署上线, 通过近期的运营数据可以看出提升优化效果明显。从性能上, 日志数据处理效率有了明显提升, 设备成本降低近 50%。原本依靠 DPI 日志采集需要多台设备处理分析热点资源和流量流向, 目前聚焦 DNS 日志采集, 单台 x86 服务器 (配置硬盘 40 TB, 160 GB 内存, 48 核 CPU) 利用大数据分析技术和数据库优化存储技术即可处理每天日志数据量 4 TB 左右, 次日实现分析结果呈现。

从功能上, 满足运营商网络优化运营需求, 对数据关联分析包括如下几个方面。

4.1 关键词: 流量流向, 关联数据库: 网站库、域名库、IP 地址库

综合网站库、域名库以及重要资源库, 从公司、网站和域名维度分析 DNS 请求查询数网内外分布, 网内流量分布如省内、省外、IDC 各占比。公司信息维度的流量分布如图 5 所示, 网站信息维度的流量分布如图 6 所示, 域名信息维度的流量分布如图 7 所示。

4.2 关键词: 出网定位, 关联数据库: 域名库、IP 地址库

出网资源主要集中在国外网站, 如 Google (谷歌)、Yahoo (雅虎)、Apple (苹果) 等, 出网资源

日志分析 >> 全部 >> 综合信息 >> 公司信息										
日期: 2018-03-25	查询	导出结果集	根据域名(公司名,分类)查询					精确搜索		
公司名	DNS查询数	出网次数	网内次数	缓存次数	IDC	CACHE	CDN	网内占比	出网影响度	域名数
总计	58475920802	19204781391	39271139411	14779184375	14779184375	0	0	67.16	32.84	30000000
未知	8769671702	8280458866	489212836	22319692	22319692	0	0	0.84	14.16	20391466
深圳市腾讯计算机系统有限公司	8744852503	99544279	8645308224	1784157662	1784157662	0	0	14.78	0.17	79096
北京百度网讯科技有限公司	5061432477	96033296	4965399181	3907022905	3907022905	0	0	8.49	0.16	344792
深圳市迅雷网络技术有限公司	3278161759	4815224	3273346535	545105113	545105113	0	0	5.6	0.01	5313
浙江淘宝网络有限公司	2004177342	977300461	1026876881	1014182531	1014182531	0	0	1.76	1.67	478814
北京新浪互联信息服务有限公司	2003641896	95357123	1908284773	1704341588	1704341588	0	0	3.26	0.16	10043
北京奇虎科技有限公司	1985493328	22110963	1963382365	146954542	146954542	0	0	3.36	0.04	11239
苹果电子产品商贸(北京)有限公司	1787116764	1296403161	490713603	129323190	129323190	0	0	0.84	2.22	3072
北京字节跳动科技有限公司	1324787705	319886744	1004900961	466427384	466427384	0	0	1.72	0.55	428
北京京东叁佰陆拾度电子商务有限公司	9590971112	5177979	953919133	748843255	748843255	0	0	1.63	0.01	19241
北京爱奇艺科技有限公司	950465691	5510842	944954849	203611766	203611766	0	0	1.62	0.01	2621

图5 公司信息维度的流量分布

日志分析>>全部>>综合信息>>网站信息					
日期: 2018-03-25	网站类型(全部)	查询	导出结果集	根据网站名查询	精确
网站名称	公司名	网站类型	DNS查询数	出网次数	网内次数
总计	全部	全部	58475920802	19204781391	39271139411
未知	未知	未知	8769671702	8280458866	489212836
腾讯网	深圳市腾讯计算机系统有限公	门户	7498036926	83873610	7414163316
百度	北京百度网讯科技有限公司	搜索	4467802405	84450561	4383351844
迅雷在线	深圳市迅雷网络技术有限公司	P2P业务	3276327323	4532340	3271794983
淘宝	浙江淘宝网有限公司	电商	1996566133	970313560	1026252573
苹果公司网站	苹果电子产品商贸(北京)有	未知	1763872307	1296321876	467550431
新浪网	北京新浪互联信息服务有限公	门户	1707237826	94256106	1612981720
360安全中心	北京奇虎科技有限公司	安全杀毒	1698042159	21826287	1676215872
爱奇艺	北京爱奇艺科技有限公司	视频	949236872	5468043	943768829
京东商城-京东多媒体网	北京京东叁佰陆拾度电子商务	电商	948298705	4528747	943769958
头条	北京字节跳动科技有限公司	未知	876724195	304566228	572157967

图6 网站信息维度的流量分布

日志分析>>全部>>综合信息>>域名信息								
模式: 按日	日期: 2018-03-25	查询	导出结果集	根据域名(公司名,分类)查询				
域名	DNS查询数	出网次数	网内次数	错误次数	电信引入	本省电信	外省电信	IDC
总计	58475920802	19204781391	39271139411	16465219138	39271139411	19321362915	24491955036	14779184375
www.baidu.com	2631313860	0	2631313860	236193038	2631313860	2631313860	0	2631313860
...	...	...	...	...	...	...	...	...
jiangsudianxin.suiji.look	992806297	992806297	0	176	0	0	0	0
www.taobao.com	981032341	0	981032341	9784982	981032341	981032341	0	981032341
...	...	...	...	...	...	...	...	...
www.sina.com.cn	948297767	1	948297766	7880708	948297766	948297766	0	948297766
...	...	...	...	...	...	...	...	...
www.jd.com	712750704	0	712750704	272543	712750704	712750704	0	712750704
www.360.com	699330479	1	699330478	45505	699330478	0	699330478	0

图7 域名信息维度的流量分布

统计如图8所示,因为这些出网资源未被引入中国电信中。而对于引入资源仍然出现出网情况系统也进行统计分析,如图9所示。以apple.com为例,下钻发现除了地级电信、IDC,主要去向为国外Apple公司和akamaiCDN节点,如图10所示。

4.3 关键词: 行为偏好, 关联数据库: 网站库、域名库、CNAME 域名库

热门访问资源排行榜一定程度反映用户访问的偏好习惯。以2018年3月22日数据为例,如图11所示,排名靠前的网站有腾讯网、百度、迅

域名	DNS查询数	出网次数	网内次数	电信引入	IDC
总计	22520188169	19204781391	3315406778	3315406778	1135930228
...	...	...	...	...	...
jiangsudianxin.suiji.look	992806297	992806297	0	0	0
...	...	...	...	...	...
www.google.com	529088936	529088936	0	0	0
www.yahoo.com	421623160	421623160	0	0	0
...	...	...	...	...	...
www.ieee.org	383234389	383234389	0	0	0
...	...	...	...	...	...
apple.com	162212970	162212970	0	0	0
...	...	...	...	...	...

图8 出网资源统计



域名	DNS查询数	出网次数	网内次数	电信引入	IDC
总计	4441728163	1126321385	3315406778	3315406778	11342459
⋮	⋮	⋮	⋮	⋮	⋮
baidu.com	123931996	61935922	61996074	61996074	0
⋮	⋮	⋮	⋮	⋮	⋮
sohu.com	59871977	39867861	20004116	20004116	0

图9 引入资源仍出网统计

IP	DNS请求数	运营商
23.58.231.11X	7585	美国-Akamai科技公司CDN网络节点
17.253.85.20X	185553	美国-Apple公司
17.253.85.20X	168460	美国-Apple公司
17.253.85.20X	248246	美国-Apple公司
101.227.175.1X	6217793	上海市-电信
58.220.40.22X	6281647	江苏省扬州市-电信
17.253.85.20X	182277	美国-Apple公司
58.220.73.17X	6299417	江苏省扬州市-电信
17.253.85.20X	218895	美国-Apple公司
180.97.218.22X	6229390	江苏省镇江市-电信IDC数据中心
221.230.142.20X	6172228	江苏省泰州市-电信
17.253.85.20X	167262	美国-Apple公司
17.253.85.20X	240727	美国-Apple公司
58.216.109.13X	6272549	江苏省常州市-电信
17.253.85.20X	157388	美国-Apple公司

图10 apple.com DNS 请求分布

日期: 2018-03-22	网站类型(全部)	TOPN	查询	导出结果集	根据网站名查询	精确搜索
网站名称	网站类型	DNS查询数	网内次数	出网次数	错误次数	
总计	全部	42979923471	34955604975	8024318496	8044598992	
腾讯网	门户	7341446629	7257945561	83501068	1150256841	
百度	搜索	4767682626	4679106329	88576297	786408157	
迅雷在线	P2P业务	3216368906	3211334174	5034732	458803147	
淘宝	电商	2015117320	1018154534	996962786	315652895	
360安全中心	安全杀毒	2006745439	1982058798	24686641	127500255	
新浪网	门户	1761718631	1641984591	119734040	36359474	
苹果公司网站	未知	1754540185	461796899	1292743286	76118805	
京东商城-京东多媒体网	电商	955727217	950931579	4795638	34781023	
爱奇艺	视频	842819610	837799537	5020073	300684970	
头条	未知	828447724	540574626	287873098	102657307	
搜狐网	门户	700994303	640483331	60510972	21920042	
网易	门户	639188976	635994222	3194754	45627808	
搜狗网	搜索	550135482	545747218	4388264	53669622	

图11 热门网站资源

雷在线、淘宝、360 安全中心、新浪网、苹果官网、京东商城等。

按照热门网站维度，分析运营商重点关注指标——本网率和本省率情况。本网率越低说明用户需要出网访问次数越多，可能带来时延，影响

用户体验。所以对网站进行细分，统计分析流量占比高的游戏类、视频类和信息安全要求高的网银类、证券类的 Top10 关注度最高的网站本网率和本省率大小，如图 12 所示。

从图 12 数据可以看出某一时间段内的趋势变

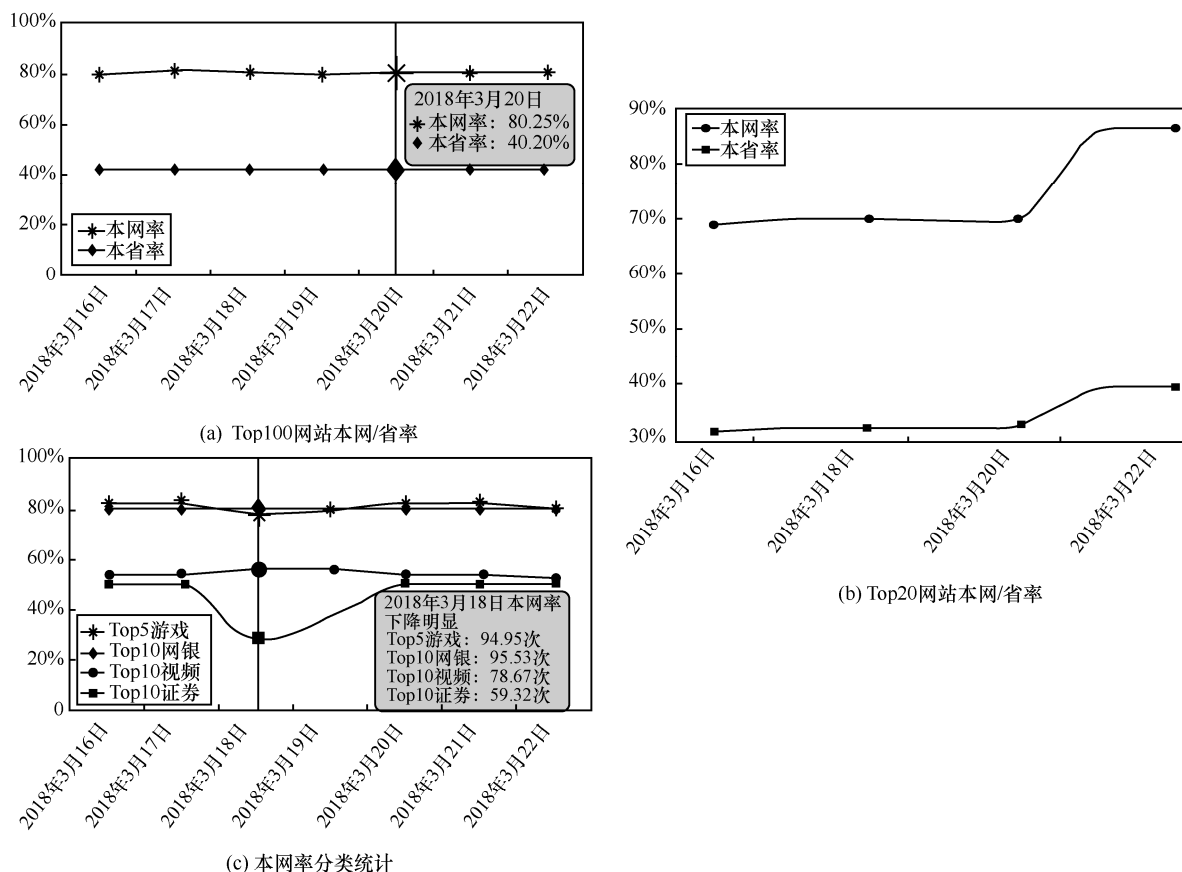


图 12 热点网站本网率统计

化。当本省率和本网率未达到要求或者下降过多时, 定位具体网站进行分析。以 Top10 证券类分析为例, 发现 2018 年 3 月 18 日本网率下降明显。针对这一现象, 获取当日 Top10 的证券类数据和前后日数据进行对比分析, 发现中信证券办公网本网率 6.5% 过低, 继续下钻发现该网站大部分域名资源访问均出网, 中国电信并未引入该资源, 造成本网率低。当该网站访问量过高将造成出网次数增多, 从而出现本网率下降的现象。

4.4 关键词: 安全备案漏洞扫描, 关联数据库: 域名库、IP 地址库、网站库、备案库、安全信息库

DNS 日志结合域名库、备案库信息了解域名备案情况, 提供安全绿色上网环境。IDC 机房对安全备案有着较高的要求, 以域名维度开展关联该域名的公司备案信息分析和漏洞扫描, 将能大

大提升 IDC 的安全性, IDC 热门资源备案信息如图 13 所示。

5 结束语

DNS 融合大数据分析系统具有以下优点: 强大的日志分析性能, 日志分析效率可达 10 GB/min、 800×10^6 条/s, 分析粒度细化到分钟、小时、天、月; 分析内容范围全面, 分析流量流向 (调度分析)、内容热度、内容的合法性、安全性等; 融合六大信息库进行数据分析, 包含的库有: 备案信息库、IP 地址库、安全信息库、网站库、域名库、CNAME 域名库; 支持按用户类型、用户地域分析访问内容, 系统提供针对用户类型 (手机、固网), 用户所在地市进行分析。DNS 融合大数据分析系统分析内容范围广, 分析深度高, 为电信运



备案	公司名	所属网站	本网率	国外	中国移动	中国联通
全部	全部	全部	78.02%	3906937725	170935830	815087685
粤ICP备09000001号-1	深圳市腾讯计算机系统有限公司	腾讯网	98.88%	32643	17003408	20437430
京ICP备09001714号-1	北京百度网讯科技有限公司	百度	98.11%	168902	63797860	15180196
粤ICP备09000001号-1	深圳市迅雷网络技术有限公司	迅雷在线	99.91%	0	1724	1715835
浙ICP备09000001号-1	浙江淘宝网络有限公司	淘宝	51.4%	90055	0	4
京ICP备09000001号-1	苹果电子产品商贸(北京)有限公司	苹果公司网站	26.51%	1249159026	0	39393
京ICP备09000001号-1	北京新浪互联信息服务有限公司	新浪网	99.61%	0	669	2443097
京ICP备09000001号-1	北京爱奇艺科技有限公司	爱奇艺	99.42%	871	5359	5113292
京ICP备09000001号-1	北京京东叁佰陆拾度电子商务有限公司	京东商城-京东多媒体网	99.5%	803	69481	1552087
京ICP备09000001号-1	北京奇虎科技有限公司	360安全中心	97.79%	10502	56685	19703410
京ICP备09000001号-1	北京字节跳动科技有限公司	头条	62.87%	0	0	1
京ICP备09000001号-1	北京奇虎科技有限公司	360安全中心	100%	4753	0	7
京ICP备09000001号-1	北京搜狐互联网信息服务有限公司	搜狐网	93.67%	69162	19918753	20183732
粤ICP备09000001号-1	广州网易计算机系统有限公司	网易	99.46%	15149	0	2766040
京ICP备09000001号-1	北京新浪互联信息服务有限公司	新浪网	84.34%	88254502	166	291263
京ICP备09000001号-1	阿里云计算有限公司	阿里旺旺	91.85%	10912511	0	0
粤ICP备09000001号-1	广东步步高电子工业有限公司	vivo智能手机官方网站	92.36%	7983449	0	1940
京ICP备09000001号-1	北京字节跳动科技有限公司	静态文件域名	98.78%	0	0	2538
京ICP备09000001号-1	高德软件有限公司	高德地图	0.57%	514753	0	825
京ICP备09000001号-1	北京搜狗信息服务有限公司	搜狗网	99.16%	2879	181956	3044068
京ICP备09000001号-1	普联技术有限公司	TP-LINK官方网站	0	383917421	0	0
粤ICP备09000001号-1	深圳市腾讯计算机系统有限公司	微博图片网	100%	0	26	56

图 13 IDC 热门资源备案信息

营商提升管理效率、优化网络安全、提升本网率提供决策依据。

参考文献:

- [1] 漆晨曦. 电信企业大数据分析、应用及管理发展策略[J]. 电信科学, 2013, 29(3): 12-16.
QI C X. Big data analysis, application and management development strategy of telecom enterprises[J]. Telecommunications Science, 2013, 29(3): 12-16.
- [2] 刘觅, 杨海川, 李振华. 运营商构建电信大数据生态之道[J]. 电信科学, 2018, 34(1): 109-119.
LIU M, YANG H C, LI Z H. The way for operators to build telecom big data ecology[J]. Telecommunications Science, 2018, 34(1): 109-119.
- [3] 张云勇. 电信运营商大数据发展建议[J]. 电信科学, 2018, 34(1): 103-108.
ZHANG Y Y. Suggestions for big data development of telecom operators [J]. Telecommunications Science, 2018, 34(1): 103-108.
- [4] 黄璇婧. 信令采集与监测系统研究和应用[D]. 上海: 上海交通大学, 2012.
HUANG X J. Research and application of signaling acquisition and monitoring system [D]. Shanghai: Shanghai Jiaotong University, 2012.
- [5] 汤雅妃, 魏进武, 张云勇. 基于大数据的信令监测系统研究[J]. 邮电设计技术, 2014(7): 47-52.

- TANG Y F, WEI J W, ZHANG Y Y. Research on signaling monitoring system based on big data[J]. Design Technology of Posts and Telecommunications, 2014(7): 47-52.
- [6] 胡为艳, 艾民, 周光彬, 等. 基于大数据的信令监测系统的设计与实现[J]. 电视技术, 2016(1): 95-101.
HU W Y, AI M, ZHOU G B, et al. Design and implementation of signaling monitoring system based on big data[J]. TV Engineering, 2016(1): 95-101.
- [7] DHARMAPURIKAR S, KRISHNAMURTHY P, SPROULL T, et al. Deep packet inspection using parallel bloom filters[C]//11th Symposium on High Performance Interconnects, Aug 20-22, 2003, Stanford, CA, USA. Piscataway: IEEE Press, 2003: 44-51.
- [8] 黄昆, 张大方. 一种面向深度数据包检测的紧凑型正则表达式匹配算法[J]. 中国科学, 2010, 10(40): 1062-1077.
HUANG K, ZHANG D F. A compact regular expression matching algorithm for deep packet detection[J]. CJUS, 2010, 10(40): 1062-1077.
- [9] 冉萌, 韩玉辉. DPI 技术在移动大数据中的应用[J]. 邮电设计技术, 2016(8): 33-36.
RAN M, HAN Y H. Application of DPI technology in mobile big data[J]. Design Technology of Posts and Telecommunications, 2016(8): 33-36.
- [10] BECCHI M, CROWLEY P. A hybrid finite automaton for practical deep packet inspection[C]//ACM Conference on Emerging Network Experiment and Technology, Dec 10-13, 2007, New

York, USA. New York: ACM Press, 2007.

- [11] KUMAR S, DHARMAPURIKAR S, YU F, et al. Algorithms to accelerate multiple regular expressions matching for deep packet inspection[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(4): 339-350.
- [12] SNYDER M E, SUNDARAM R, THAKUR M. Preprocessing DNS log data for effective data mining[C]//2009 IEEE International Conference on Communications, June 14-18, 2009, Dresden, Germany. Piscataway: IEEE Press, 2009: 1-5.
- [13] 彭巍, 贺晓东, 陆小铭, 等. 运营商 DNS 海量数据分析技术优化和应用[J]. 电信科学, 2018, 34(3): 138-144.
PENG W, HE X D, LU X M, et al. Optimization and application of operator DNS massive data analysis technology [J]. Telecommunications Science, 2018, 34 (3): 138-144.
- [14] 申志伟, 房秉毅. 一种智能动态域名系统在大型电子商务中的应用方法[J]. 电信科学, 2014, 30(1): 128-135.
SHEN Z W, FANG B Y. Application of an intelligent dynamic domain name system in large e-commerce[J]. Telecommunications Science, 2014, 30(1): 128-135.
- [15] YEN T F, OPREA A, ONARLIOGLU K, et al. Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks[C]// Computer Security Applications Conference, Dec 9-13, Louisiana, USA. New York: ACM Press, 2013: 199-208.
- [16] NICHOLAS D, HUNTINGTON P. Micro-mining and segmented log file analysis: a method for enriching the data yield from internet log files[J]. Journal of Information Science, 2003, 29(5): 391-404.
- [17] JEON K S, PARK S J, CHUN S H, et al. A study on the big data log analysis for security[J]. International Journal of Security and Its Applications, 2016, 10(1): 13-20.

[作者简介]



廖明 (1970-), 男, 中国移动通信集团重庆有限公司高级工程师, 主要研究方向为互联网内容运营管理。



陈明 (1979-), 男, 中国移动通信集团重庆有限公司高级工程师, 主要研究方向为 DNS 与用户访问感知。



周冀 (1978-), 男, 中国移动通信集团重庆有限公司高级工程师, 主要研究方向为互联网内容管理。

向小华 (1986-), 男, 中国移动通信集团重庆有限公司中级工程师, 主要研究方向为互联网质量优化分析。

李芳 (1990-), 女, 中国移动通信集团重庆有限公司中级工程师, 主要研究方向为互联网 OTT 内容。

焦叶芬 (1978-), 女, 中国移动通信集团重庆有限公司中级工程师, 主要研究方向为网络质量管理。