基于ELK的用户访问行为分析技术

陈楠　陈东辉　邓莉
（国家气象信息中心，北京 100081）

摘要：针对国家气象业务内网用户访问行为记录，基于ELK（Elasticsearch日志检索+Logstash日志收集+Kibana查询展示）流量日志处理技术，建立了气象业务内网日志采集和智能分析系统，实现了对访问用户的行为跟踪和针对不同类别用户的访问热点分析、趋势分析和对比分析，通过对单个页面停留时间、点击流、相关点击量等指标结合关联、聚类分析等数据挖掘算法实现对网站流量日志的深度分析，为建设更加专业和有针对性的气象大数据服务平台提供参考。通过在国家气象业务内网建设中的应用，证明该系统有助于发掘共享用户的真正需求，对全面提升精细化、专业化气象数据服务能力有积极的作用。

关键词：网站分析，ELK，用户行为分析，气象服务

DOI：10.3969/j.issn.2095-1973.2018.01.024

Technology of User Behavior Analysis Based on ELK

Chen Nan, Chen Donghui, Deng Li
(National Meteorological Information Centre, Beijing 100081)

Abstract: According to the users' logs into the National Meteorological Service Network, and based on ELK flow log processing technology, the meteorological network log data collection and intelligent analysis system is established. It realizes the analysis of user behavior tracking and data access for different types of users' access to the hot trend analysis and comparison based on a single page retention time, click stream model and the related click index. At the same time, it combines with correlation, clustering analysis and other data mining algorithms to achieve the depth analysis of the site traffic log. It provides a reference for the construction of more specialized and targeted meteorological data service platform. Through applications in the construction of the National Meteorological Service Network, it is proved that the system is helpful to discover the real needs of the users, and has a positive effect on improving the service ability of the refined and specialized meteorological data.

Keywords: website analysis, ELK, user behavior analysis, meteorological service

0　引言

为满足现代气象业务现代化和集约化发展需要，预报司提出并制定了气象业务内网系统建设任务（气预函〔2012〕94号）。国家气象业务内网于2012年7月启动建设，目前WEB2.5版运行使用。国家气象业务内网结合业务场景，对观测业务、天气业务、气候业务、卫星业务、人工影响天气以及气象信息业务建立六大业务专栏，方便用户有针对性的获取业务信息。同时，国家气象业务内网基于CIMISS统一数据环境，提供基础数据和产品共24大类近600种数据定制下载、FTP下载、接口访问等服务。

分析用户使用网站的行为，可以了解网站的使用情况，挖掘用户的潜在需求，提升用户体验，对网站建设具有重要意义。国家级气象业务内网栏目众多、功能庞大、用户覆盖全国四万多气象行业工作人员。目前该网站每天有约1.9万条约800 MB大小的访问日志需要实时处理和展示。因此，国家气象业务内网的日志分析系统需要有处理海量数据的搜索、分析能力。基于这些需求，国家气象业务内网建设使用ELK实时日志分析系统。本文主要研究基于ELK（Elasticsearch日志检索+Logstash日志收集+Kibana查询展示）搭建的网站日志分析系统进行用户行为分析。使用该系统分析日志总结规律性特征，让网站开发人员更详细清楚地了解用户行为习惯，发现问题并指导和服务于气象类网站的开发和经营策略，进一步提高气象网站的服务质量，努力打造更加优质的气象行业内部业务服务平台。

收稿日期：2017年6月22日；修回日期：2017年11月14日
第一作者：陈楠（1988—），Email：chennan@cma.gov.cn
资助信息：国家气象信息中心青年科技基金课题（NMICQJ201609）

1　基于网络日志的用户行为分析综述

随着互联网和移动互联网的普及和发展，用户利用网络获取信息的行为也越来越复杂。1948年的（英国）皇家社会科学信息会议标志这现代用户信息行为研究的开端。网站分析（Web Analytics）的定义

很多，谷歌数字营销专家艾韦纳什·卡希克定义为：“通过分析来自网站及竞争对手的定性与定量数据，驱动用户及潜在用户在线体验的持续提升，并最终转化为你期望的结果”。这里可以看出宏观上的网站分析实际上分为两大类：一类被称为网站内的网站分析（On-site Web Analytics，或称基于网站自身的分析），用以衡量用户的访问行为；另一类被称为网站外的网站分析（Off-site Web Analytics），指在整个互联网的环境中，对竞争对手网站的分析，以及对互联网传播和营销效果的衡量和分析。本文研究的重点是基于网站内用户行为的分析，总结规律与气象网站和移动应用的经营策略相结合，进一步提高气象网站的服务质量，不断优化用户体验[1]。

1.1 用户行为分析的介绍

用户行为分析是网络信息检索技术得以前进的重要基石[2]。用户行为是指用户在使用网络资源时所呈现的规律性。用户行为可以分为信息查询行为、沟通交流行为、休闲娱乐行为、电子服务行为和电子商务服务行为等许多方面。用户行为分析[3]，是指在获得网站访问量基本数据的情况下，对有关数据进行统计、分析，从中发现用户访问网站的规律，并将这些规律与网络营销策略等相结合，指导网站建设、运营和营销。

1.2 重点分析的数据

用户访问网站的行为一般都是围绕着某种需求而主动进行。虽然每天有数以亿计的用户在网络上留下特殊足迹，包括检索信息、网页浏览、社交网站的互动等都具有很多的差异性，但经过数据挖掘和分析，都可以体现出对信息需求和服务的普遍规律。分析的重要指标有：用户的来源地区、域名和页面；用户在网站的停留时间、跳出率、回访者、新访问者、回访次数、回访相隔天数；注册用户和非注册用户，分析两者之间的浏览习惯。

1.3 常见的分析方法

网络用户行为分析的过程相当于是对海量大数据获得有价值信息的一个数据挖掘的过程。按照建模的思路，用户行为分析整个过程包含：需求提出、数据采集、数据分析和结果评估。用户行为分析有用户特征分析、关联分析[3]、分类与预测、异常分析、TopN分析等几种分析方法。网站分析工具帮助管理者收集、预估和分析网站的访问记录，对网站优化和市场开拓都有重要作用。比如屏幕录制工具Clictale，页面热区图工具Crazyegg，点击流工具SkyGlue等[4]，商业级网站分析工具如Webtrekk Q3，通过在网站上面嵌入一段代码，这些工具就可以分析用户最常点击的地方、最少点击的地方、鼠标移动的区域等相关数据。另外，Google Analytics和百度统计都是网站常用的网站流量分析工具，它们不只记录点击流数据，更注重点击流的分析与测量，并尽量将这些结果向Web分析和网络营销引导，致力于提升网站质量。

现有的第三方网站分析工具需要经过互联网传参数，而内网具有内部保密性，因而需要自行搭建整个分析系统，使用现有分析工具的技术思路，在此基础上进行国家气象业务内网的日志分析系统的建设。这对得到应用在内部网站符合气象行业特殊用户群体的网站分析报告具有重大意义[5]。

2 基于ELK的网站分析平台建设及优化

网络日志含有网站最重要的基本信息，它包括业务操作行为、服务器运行和故障、用户访问情况等。网络日志由Web服务器产生，可能是Nginx, IIS, Apache, Tomcat等。通过日志分析，可以获得网站访问量、网页访问排行、网页停留时长、访问网站的用户分布、用户肖像等。当前基于网站日志管理系统有多种搭建技术和方案，如结合Flume+Kafka+Storm+HDFS系统，或Elasticsearch日志检索+Logstash日志收集+Kibana查询展示系统进行网站用户行为分析。综合考虑硬件成本不断降低、实时在线分析的需求加大、开发过程追求简单化等多种因素，国家气象业务内网建设使用ELK实时日志分析系统。

2.1 ELK 实时日志分析系统的搭建

ELK是由三个开源组建构成的一个实时日志分析平台，包括Elasticsearch（日志检索）、Logstash（日志收集、过滤、格式化）和Kibana（统计查询、可视化展示）。Elasticsearch是一个基于Lucene的全文搜索服务器。它提供了一个基于Restful web接口分布式多用户能力的全文搜索引擎。Elasticsearch是用Java开发的Apache许可条款下的开放源码发布，是当前流行的企业级搜索引擎[6]，具有对海量数据进行快速的实时搜索、稳定、可靠、且安装使用方便等特点。Logstash是一个用于管理日志和事件的工具，用于收集、转换、解析日志并将数据提供给其他模块调用，例如搜索、存储等。Kibana用来进行前端日志展示，它从ElasticSearch中读取数据，用图表等形式进行数据可视化展示，并支持各种查询。国家气象业务内网从建设初期就着手网络日志分析和用户行为的统计，基于现有的服务器使用ELK实时日志分析系统的解决方案进行日志的统计和分析，经过不断的优化升级，当前使用的系统架构如图1所示，基本完成了每日用

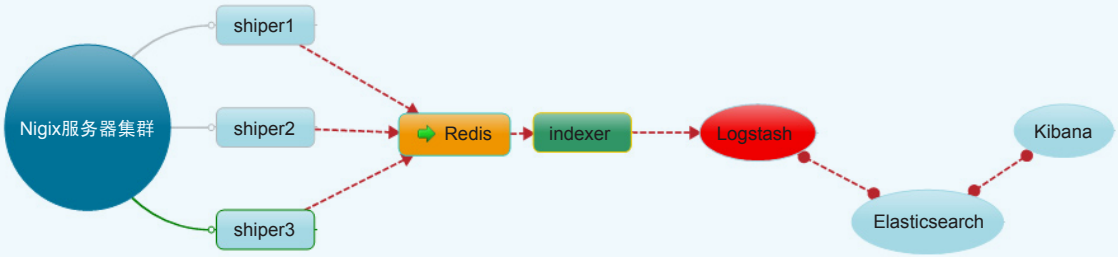

图1 国家气象业务内网ELK实时日志分析系统的架构图

Fig. 1 Architecture diagram of the ELK real time log analysis system for the National Meteorological Inner Service Centre

户的访问行为统计、气象数据下载情况统计、服务器运维状态情况统计等。

2.2 ELK实时日志分析系统的统计查询功能的优化

国家气象业务内网的日志统计分析系统的搭建经过不断优化升级。在建设初期，当用户访问内网的统计分析页面时，前端点击页面系统实时去日志服务器抓取，后台程序统计分析日志，进而前端页面提供给用户的是查询统计结果。这种解决方法虽然实现了业务需求，但大量日志传输占用宽带，且实时计算消耗时间，页面展示速度变慢，用户体验较差。为了提高整体性能，系统架构上在Logstash的shipper和indexer之间增加Redis代理缓存机制。Redis是一个开源Key/Value数据库，用于在索引前队列化日志。为了减少消耗时间，在统计程序上使用了中间表和Filter过滤器插件。通过启动定时任务程序将日志统计查询结果存入中间表数据库中，前端展示从该数据库比较快速的获取数据展示给用户。考虑到增加统计维度会造成中间表结构重新修改和数据重新录入等工作，进而改进使用Filter过滤器过滤海量日志，对其正则解析，并将结构化的日志传递给Elasticsearch存储和查询，这样查询速度增快且具有较强的可扩展性。该系统目前基本满足国家气象业务内网的统计查询，但是日志数的海量增长，Filter正则解析日志占用内存消耗CPU资源增大等问题都是后期需要考虑的问题。如果日志量更大，可以考虑使用hangout来代替logstash，或用kafka来替代redis，从而获得更大的日志吞吐量。

3 基于国家气象业务内网的用户行为分析

国家气象业务内网经过四年多的建设，目前已经建成了满足国省两级用户需求的业务服务支撑平台，其数据服务版块、视频会商版块、资料传输考核等页面成为网站亮点。借助ELK实时日志分析系统，对网络日志进行收集、处理和分析统计，网站整个运行情况都较好的保留和可视化展示。作为气象行业内部网站，国家气象业务内网目前每日的访问量为十万级别，每天处理约1.9万条日志。我们将日志统计结果存储在数据库里，使用访问量PV、IP来源、访问时长、数据下载量等指标进行页面展示。用户通过统计分析栏目查看网站访问情况，从地域维度划分国家用户和省级用户；从时间维度分时段研究网站访问情况等；从数据使用维度查看数据下载量排行和气象数据产品之间的相关性。网站管理人员通过后台管理系统定位具体时间段、IP属性、栏目等多维度相结合查看网站使用情况，结合多种指标进行用户画像描绘、相似用户扩展Lookalike评测和推荐。

国家气象业务内网的用户多为信息获取类，通过浏览网页进行信息的获取，一般表现为点击相关超链接、阅读和浏览网页、对网站提供的信息进行保存、收藏、复制和下载等行为。

3.1 网站定位和目标

国家气象业务内网的建设是在各单位内网建设功能参差不齐，资源杂乱的局面下提出的，它是基于CIMISS统一数据环境的业务产品共享平台系统，建立集约化的数据环境，面向气象内部用户，支持国、省、地、县级四级用户访问的气象产品展示与服务、业务管理的信息共享平台。网站用户群面向气象业务科研和管理人员，业务栏目覆盖了气象中心、气候中心、卫星中心、信息中心、探测中心、气科院等国家级业务单位，汇聚数据服务产品种类超过2000种的综合性大型网站。

3.2 网站用户群体的特征

国家气象业务内网对用户群体进行特征分类，用户群体主要是气象部门内部用户，包括中国气象局职能管理人员，探测、天气、气候、信息、公共服务等业务人员，以及科研人员。从用户对网站的使用率考虑，主要考察网站的点击率、访问量、访问率、点击量、页面停留时间等。从用户使用产品的时间考虑，主要包括用户什么时候使用，这个研究对系统升级、故障处理、并发量统计等有重要的作用。

总体来说，国家气象业务内网是针对专业用户群体的集科研和业务管理多功能的气象内部网站平台。用户根据需求点击与自身科研和业务相关的某一个栏目页面，且使用时间没有一致性，但与气象业务、工作时间、突发天气现象、汛期等有紧密的关联。

3.3 数据获取和用户调研

用户行为的数据搜集和获取主要分为两大类。主动获取包括用户登录网站浏览，从日志获得数据，模拟用户的操作。被动获取包括使用外部调研的方式得到用户对网站使用情况的反馈。数据的收集从网站访问者输入URL向网站服务器发出http请求开始，借助于ELK实时日志分析系统进行用户行为记录分析。网站服务器接收到请求后会在自己的Log文件中追加一条记录，记录内容包括：远程主机名（或者是IP地址）、登录名、登录全名、发请求的日期、发请求的时间、请求的详细（包括请求的方法、地址、协议）、请求返回的状态、请求文档的大小。随后网站服务器将页面返回到访问者的浏览器内得以展现[7]。对包含用户IP地址、访问的URL、访问日期时间、访问方法和请求的数据大小等进行数据挖掘和分析。另外通过用户调研来查看用户对网站访问的总体满意度，方法有很多，比如电话回访、邮寄问卷、网上问卷、专家咨询等。调研的核心应该是如何设计一份有针对性和引导性题目的调研问卷。

3.4 分析方法和结果解读

分析过程主要从需求出发，对用户的数据进行挖掘，包括日志数据过滤、数据预处理、数据发现、数据综合分析等过程，最终以直观准确的方式展示。

国家气象业务内网查看每日用户访问量和IP数量、访问时长，对历史数据进行曲线图分析得到用户访问时间规律。过去一年内网年访问量约为830万次，其中国家级用户访问约为158万次；省级用户访问分布如图2所示。结合实际网站建设工作，得出国家气象业务内网还处在用户从少到多的建设发展和用户积累阶段，每月访问量有缓慢增长的趋势。由于各省份业务工作侧重点的差异，与内网业务结合度高的省份对内网的访问量较高。国家级业务单位中信息中心、气象中心和气候中心对网站的使用率较高（图3），体现出国家气象业务内网的业务支撑和服务平台的建设思路，但真实用户覆盖面还不够不广泛。挖掘潜在用户、提升网站使用率仍是后期建设的重点工作目标。

另外，网站访问量在时间上的分布进行检测，连续处理三个月每天约80 M的日志数据量。通过对大量IP每天每小时访问的数据流进行聚合可以得出，内网的使用与业务工作的发生成正相关，工作日成驼峰式分布。而某些监控和传输类页面的访问按照时间均匀分布。

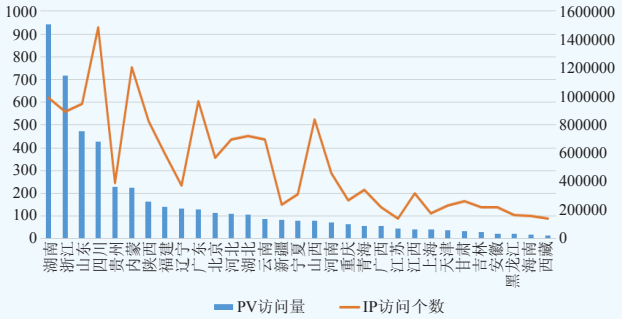

图2 2016年国家气象业务内网省级用户访问情况统计图
Fig. 2 Statistical diagram of provincial user logs into the National Meteorological Inner Service Centre in 2016

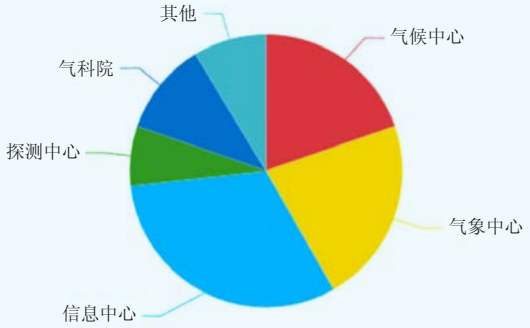

图3 2016年国家气象业务内网国家级用户访问量分布图
Fig. 3 Proportion of user logs into the National Meteorological Inner Service Centre in 2016

国家气象业务内网现有150多个栏目，分别提供业务文档查阅、业务填报、气象产品展示、数据下载等功能。平均页面访问时长是网站分析的重要指标之一，体现了用户与网站的黏性和网页的吸引力。使用ELK日志分析系统对日志进行清洗，对行为轨迹建立点击流模型[8]，对当天同一个IP的所有操作行为合并处理获得访问时长。从表1可以看出，用户用于在线阅读类页面的访问时长较多。基于此项规律，为提升用户使用体验，在2017年初对网站原有的在线浏览流程进行了优化，使用web office空间使文档加载速度加快，展示更加流畅，后台管理便捷。

表1 国家气象业务内网的栏目平均访问时长排行
Table 1 Averaged visiting time of logs into the National Meteorological Inner Service Centre

栏目名称	气象现代化专栏	会议在线	气候模式产品	CMACast目录清单栏目	资料传输质量报告栏目
平均访问时长	18.3 s	14.7 s	13 s	12.6 s	11.1 s

数据服务栏目是国家气象业务内网的重要版块。国家气象信息中心作为中国气象国家级数据中心，负

责承担全国和全球范围的气象数据及其产品的收集、处理、存储、检索和服务。结合内网统计分析平台，可以查阅国家级和省级业务单位用户某个时间端内数据下载量、下载次数等情况（图4）。对下载和搜索信息关键字提取，相似性去重和加权[9]，可以得出不同用户对各类气象数据产品的关注度不同，不同时期用户对数据集的需求也不同（图5）。跟踪分析可知，随着全国汛期的到来，数据服务栏目的访问量和数据集的下载量增长明显，用户对降水产品的需求和关注度增加。特别是CMPAS中国区域地面-卫星-雷达三源融合降水分析产品（CMPAS-V2.1）产品自2017年3月上线以来订单量逐月增长且已经进入TOP5。暴雨数据集从4月开始下载量伴随汛期到来有增长趋势。

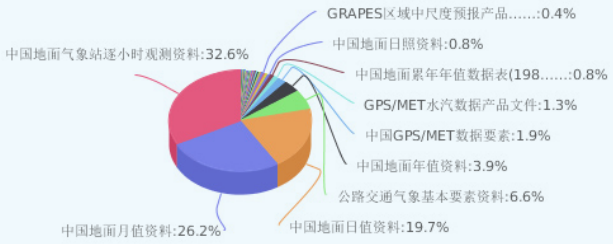

图4 2016年数据下载服务订单量总体分布图

Fig. 4 Proportion of order quantity of data downloaded in 2016

图5 2017年3月气象数据和产品搜索热点图

Fig. 5 Search hot spots on meteorological data products in March 2017

不同业务对日志数据的关注点不同，只有从业务角度进行日志数据分析，才能获得精准可靠的分析结果。针对国家气象业务内网的用户，可以通过IP判断用户所属单位，再将访问频次、页面停留时长进行关联分析获取不同用户使用内网的兴趣爱好。按照聚类分析的思路和模型，对用户进行归类和相似人群扩展，这对数据下载服务栏目、首页快捷功能的建设有指导性意义。由于气象业务与突发天气、灾害预警等天气变化息息相关，通过网站日志分析可以看到诸如视频会商、实况展示、会议在线等页面的访问量与某个时间段有较强的相关性，这对网站做到“好用”具有重大意义。

4 结论

当前，许多气象类网站的建设受到多种条件制约，制度建设还不够完善。网站缺乏灵活性，不能给用户提供更加精细灵活的服务。另外，网站建设人员普遍缺乏主动提供服务的意识，后期开发维护技术人员匮乏。这些问题使得网站用户体验度低、使用过程不流畅。本课题的研究成果用来提升和改善网站使用体验：基于用户访问习惯和时间的关系对国家气象业务内网网站的首页重点业务产品栏目和台风、高温等天气现象结合联动发布；基于用户对气象业务关注度将观测产品和预报产品、预警等信息结合地理信息展示使用,了解用户操作提升网站展示方式；基于对用户访问时间分布的研究，对系统升级、栏目更新等做了细致安排，确保最小程度影响用户对网站和移动APP使用。

转变思路，变被动为主动才能改变现状，在后期网站建设和维护中不断提升自己的品牌价值。抓住用户需求，了解用户普遍行为规律，有助于建设更贴近用户和行业的气象服务类网站，不断推动气象事业的发展。

参考文献

[1] 左军. 基于大数据的网络用户行为分析. 软件工程师, 2014, 17(10): 5-6.
[2] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析. 中文信息学报, 2007, 21(1): 109-114.
[3] 常慧君, 单洪, 满毅. 基于分段、聚类和时序关联分析的用户行为分析. 计算机应用研究, 2014, 31(2): 526-531.
[4] 王彦平. 人人都是网站分析师：从分析师的视角理解网站和解读数据. 北京: 机械工业出版社, 2015.
[5] 郑伟才, 马琰钢, 李建, 等. 基于气象网站访问统计系统设计与应用分析. 电子技术与软件工程, 2014(22):56-57.
[6] 宣明. 企业级海量数据搜索引擎核心技术实现与优化. 广州: 中山大学博士学位论文, 2015.
[7] 张兴科. 数据挖掘在Web日志分析中的应用. 微处理机, 2009, 30(3): 80-83.
[8] 易明, 操玉杰, 毛进. 基于点击流的个性化信息检索研究. 情报科学, 2011(4): 619-623.
[9] 陈墨, 程刚, 王小娟. 基于互联网海量数据的热点分析系统研究. 互联网天地, 2015(9): 30-35.