Software Engineering and Applications 软件工程与应用, 2019, 8(3), 141-148
Published Online June 2019 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2019.83017

Hans汉斯

Design and Research of User Behavior Analysis Platform Based on Big Data

Hanqing Cao[1], Quanbin Li[2*]

[1]School of Science and Literature, Jiangsu Normal University, Xuzhou Jiangsu
[2]College of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou Jiangsu
Email: *liqbzy@163.com

Received: June 2nd, 2019; accepted: June 17th, 2019; published: June 24th, 2019

Abstract

With the rapid development of diversified business models and market segments in the Internet era, many industries are facing the double challenges of high customer cost and high loss rate. There are security risks in third-party services, which cannot be customized according to their own users. On this basis, this paper focuses on the analysis of data flow design of user behavior analysis platform and architecture design based on large data environment, as well as the technical lines in the design process. Different from the full-time user-behavior analysis platforms such as hot search and Baidu Index, the system has the characteristics of traceless burial point, distributed service, visual analysis, cloud service (SAS) service, etc. It has the frontier of technical means and wide application prospects.

Keywords

Traceless Buried Point, Distributed Services, Visualization Analysis, Cloud Services (SAS) Services, User Behavior

基于大数据的用户行为分析平台设计研究

曹汉清[1]，李全彬[2*]

[1]江苏师范大学科文学院，江苏 徐州
[2]江苏师范大学泉山校区物电学院，江苏 徐州
Email: *liqbzy@163.com

收稿日期：2019年6月2日；录用日期：2019年6月17日；发布日期：2019年6月24日

*通讯作者。

文章引用: 曹汉清, 李全彬. 基于大数据的用户行为分析平台设计研究[J]. 软件工程与应用, 2019, 8(3): 141-148.
DOI: 10.12677/sea.2019.83017

摘　要

随着互联网时代多元化商业模式和细分市场的快速发展，众多行业面临着高昂获客成本和高流失率的双重挑战，第三方服务存在安全隐患，无法根据自身用户量身定制。在此基础上本文重点分析了基于大数据环境下的用户行为分析平台数据流程设计和架构设计，以及设计过程中的技术线路。区别于热搜、百度指数等专职分析用户行为平台，系统具有无痕埋点、分布式服务、可视化分析、云服务(SAS)等特色，技术手段前沿，应用前景广泛。

关键词

无痕埋点，分布式服务，可视化分析，云服务(SAS)，服务用户行为

1. 引言

随着大数据应用不断的普及和发展，人们之间通过网络来进行信息交流逐渐频繁，如何能够有效地对用户行为进行分析是目前网络行为分析的主要难点之一。刘柳等人[1]为适应 5G 网络中新的网络运营情况，结合 3GPP 和 ITU 等国际标准会议相关研究成果，提出基于业务感知和用户行为分析的服务调度系统。董茂伟[2]基于内容推荐与协同过滤推荐技术，混合推荐融合的在线群组推荐方法，构建了基于用户行为分析的群组推荐原型系统。王菲[3]通过对用户行为数据的汇总、剖析和深度理解，针对网约车，设计实现了基于网约车用户行为分析系统。陈炜[4]运用大数据技术对用户进行行为分析，以学生在网络中心和教务处中的日常信息为基础，从多个维度对学生的行为数据进行分析，从而辅助校园网的管理工作。廖志芳等四人[5]首先提出 LRF 用户行为重要度度量方法对开源软件开发中相关的用户行为重要性进行度量，并构建了一个完整的开源软件开发过程中用户行为分析的模型(简称 OUBA-Model)。刘闯[6]围绕着移动用户行为分析和机器学习展开，研究如何利用机器学习技术对移动用户数据进行分析和挖掘，并且提出一种基于 DPC (Density Peak Based Clustering)的 Kmeans k 值自适应算法(简称 DPCK-K-means)。

平台统计地域分布、系统环境、访客来源、访问分析等数据，并进行可视化展示，用户可直观获取所需行为信息。具有无痕埋点、分布式服务、可视化分析、云服务(SAS)等特色，与传统专职分析用户行为系统相比，可视化展示宏观指标，满足基本数据分析需求；技术门槛低，使用与部署较简单，只需要嵌入 SDK，极大程度避免了因需求变更、埋点错误等原因导致重新埋点的复杂工作；系统之间的耦合度降低，从而系统更易于扩展。技术专业，实用性强，拥有广阔的应用空间，见表 1。

Table 1. Comparison Table of Platform Features
表 1. 平台特色对比表

平台特色对比表				
平台名称 \ 特点	无痕埋点	分布式服务	可视化分析	云服务(SAS)服务
雅虎			√	√

Continued

Coremetrics 网站分析		√	√	
CNZZ 数据专家		√	√	
用户行为分析	√	√	√	√

2. 平台设计

2.1. 功能介绍

2.1.1. “访客地域分布”功能模块

系统通过对 hdfs 中的日志数据信息进行地域分析，利用中国地图的形式进行可视化展示，可直观的获取用户的地域分布信息。本模块还支持数据视图、还原、下载等功能。浏览量的高低与此地区对于网站的需求成正比。可以进一步对网站进行地域推广，以获取更多的访客，大幅度增加商机。如见图 1。

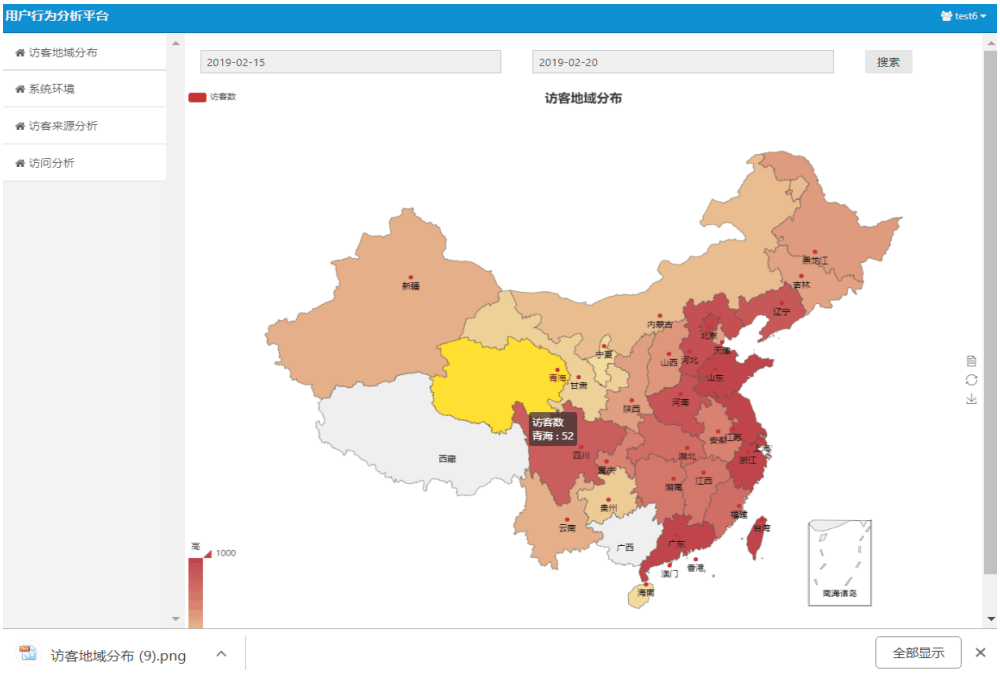

Figure 1. Visitor Geographic Distribution interface
图 1. “访客地域分布”界面

2.1.2. “系统环境”功能模块

系统通过对 hdfs 中的日志数据信息进行系统分析，并利用条形统计图、饼状图形式将用户访问设备、移动手机型号、移动网络设备以及桌面设备系统进行可视化展示，详细的统计了用户的系统设备信息。可以有效的发现网站所需更改的具体问题，从更多的技术功能方面去优化网站，有针对性的调整网站的结构和内容，实现网站的高转化率。如图 2。

2.1.3. “访客来源分析”功能模块

系统通过对 hdfs 中的数据日志信息进行访客来源分析，并进行了头尾相连的可视化展示，头部表示上一级网站，尾部则是本网站，形象的展现出用户的访问来源信息。进一步了解页面的流量动向，有针对性的对网站进行优化，提供内容和制定推广方案。如图 3。

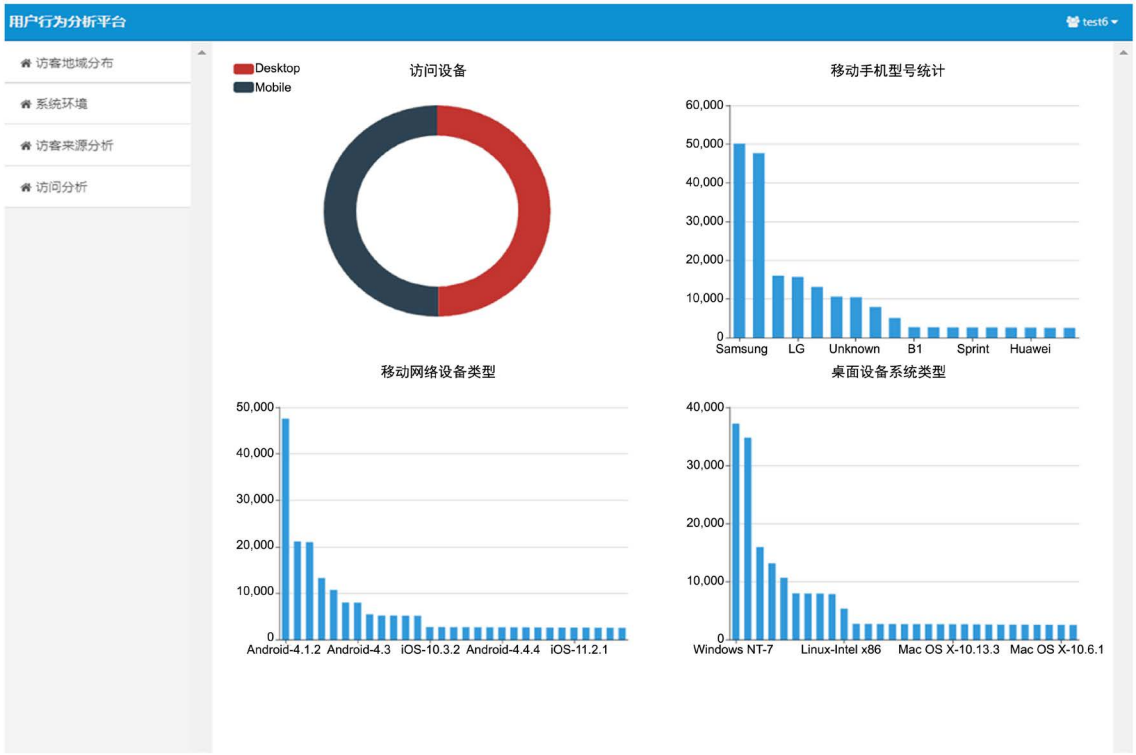

Figure 2. System Environment interface
图 2. “系统环境”界面

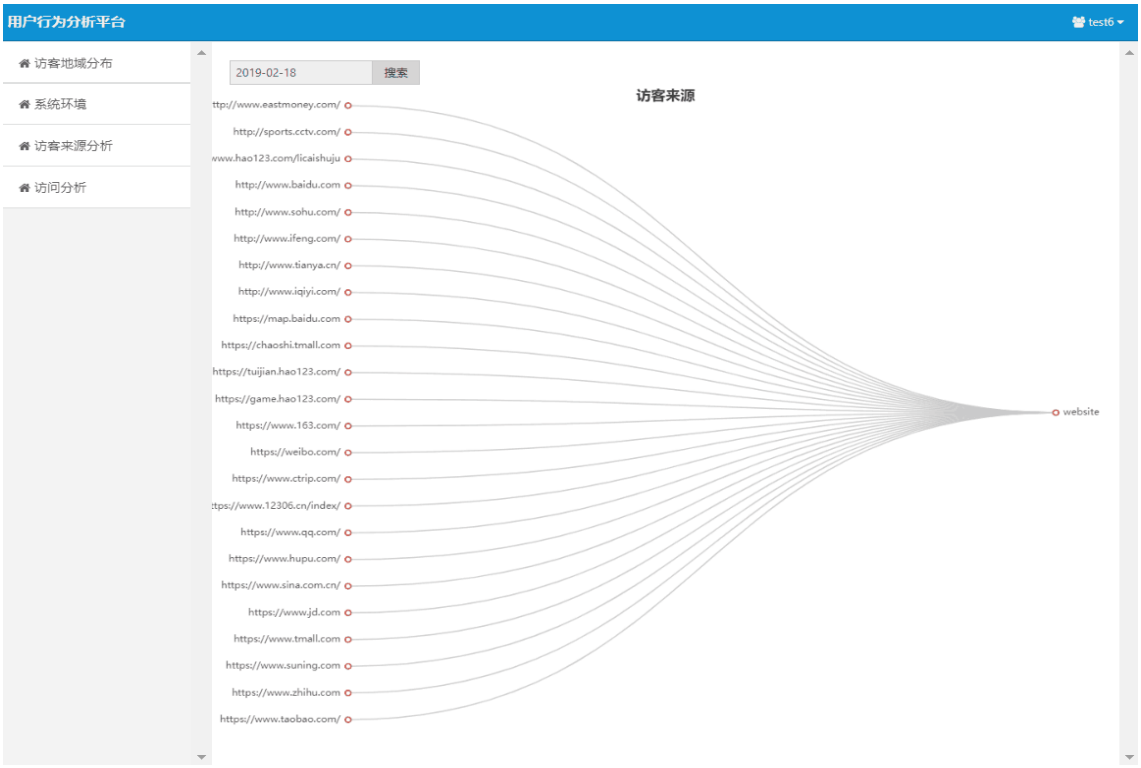

Figure 3. Visitor Source Analysis interface
图 3. “访客来源分析”界面

2.1.4. “访问分析”功能模块

系统通过对 hdfs 中的数据日志信息进行访问分析，以折线统计图、柱形统计图、扇形统计图的形式进行可视化展示。直观的统计出了站点 PV、UV 流量趋势图，以及 Top10 受访页面。本模块还具有开关 PV、UV 显示、区域缩放、区域缩放还原、切换柱状图与折线图、还原、保存为图等功能。正确定位访客的具体行为，以便准确了解网站的被访问情况和受欢迎程度，做到更好地为客户服务，调整相关策略，达到优化的最终目的。如图 4。

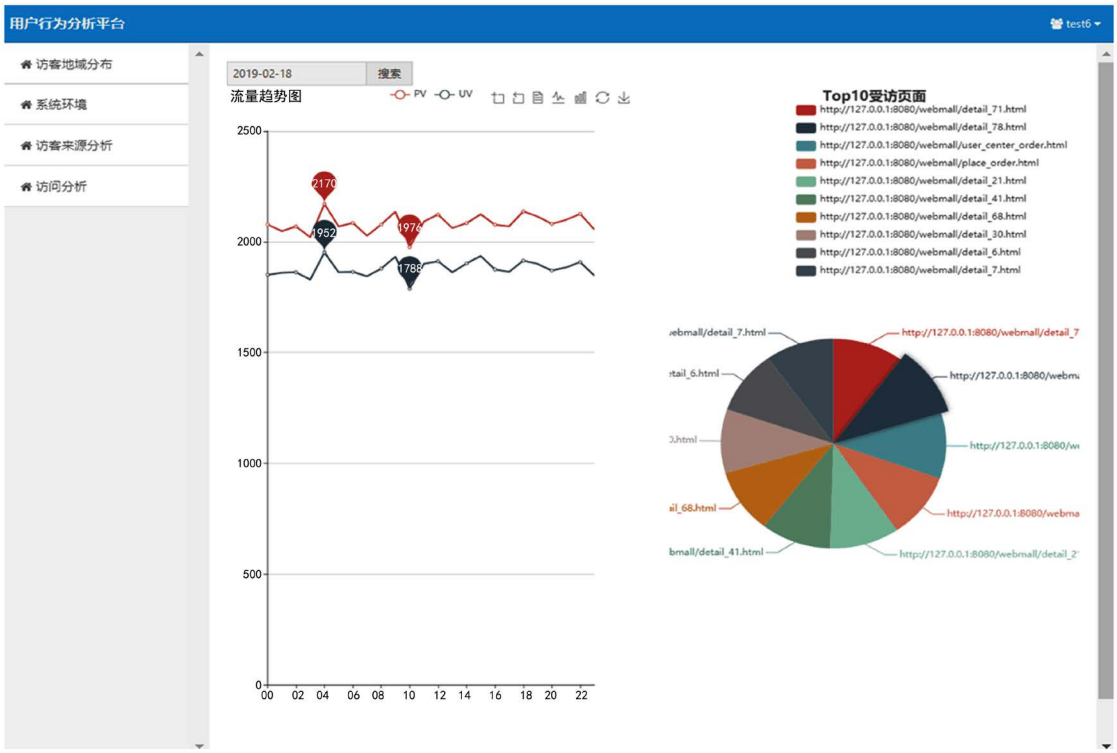

Figure 4. Access Analysis interface
图 4. “访问分析”界面

2.2. 技术路线

系统开发分为四个阶段，数据抓取和提取、数据存储、数据分析以及数据可视化。首先，利用 Javascript、nginx 提取用户日志信息，分类写入到 Mysql 存储，并将数据导入到 Hadoop 空间的 Hdfs 中存储。再利用 HadoopMapReduce、IKAnalyzer 进行数据分析和提取。最后利用 spring mvc、mybatis、quartz、amazeui、echarts 技术进行数据可视化展示。

数据抓取和提取：Javascript、nginx 提取用户日志信息。

数据存储：分类写入到 mysql 存储，利用 sqoop 将数据导入到 hadoop 空间的 hdfs 中。

数据分析：利用 hadoopmapreduce、IKAnalyzer2012(中文分词)进行数据分析和提取。

数据可视化：利用 spring mvc、mybatis、quartz、amazeui、echarts 技术进行数据展示。

如图 5。

2.3. 数据流程设计

系统利用 javascript 收集用户信息再提交到 nginx 中，并对 nginx 中数据信息进行收集，提交到 hdfs

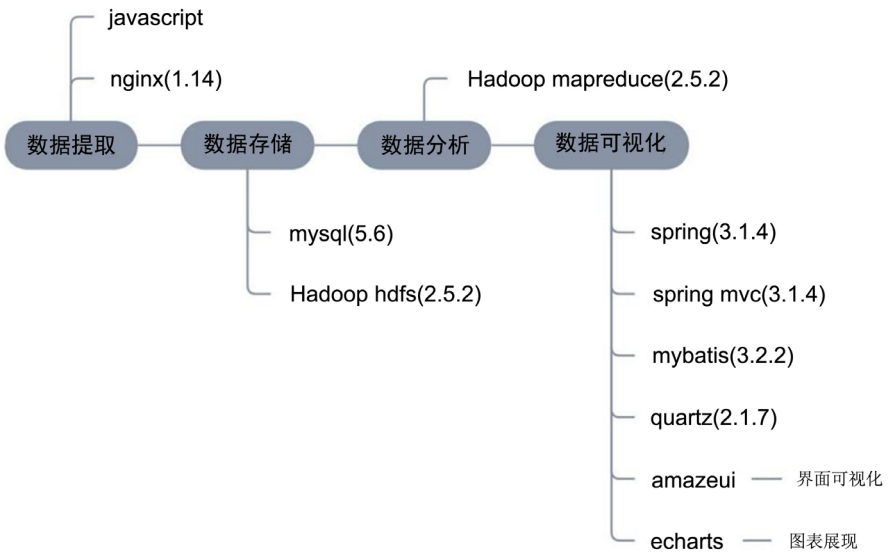

Figure 5. Production steps
图 5. 制作步骤

中进行数据存储，通过 HadoopMapReduce 分析数据信息(包含浏览器分析、访问用户区域分析、站点 pv/uv 分析、访客时间分析)，将分析的结果内容数据信息提交到 Mysql 存储，使用 J2EE 技术展示分析的结果(amazeui, jsp, css, js, springmvc, spring, mybatis)如图 6。

Figure 6. Data flow design renderings
图 6. 数据流程设计效果图

2.4. 架构设计

系统采用两套架构，传统架构和大数据架构，SSM 做外部抓取与展示技术专业，分布式大数据处理热点信息挖掘效率高。Javascript、nginx 提取用户日志信息，分类写入 Mysql 存储，再利用 Sqoop 技术将数据导入 hdfs 空间中存储，并利用 Mapreduce 模型进行数据分析、提取，最后 webserver 提供网上信息浏览。如图 7。

Figure 7. Architecture design
图 7. 架构设计

技术优势

平台使用 Sqoop 技术与 Mapreduce 模型，Sqoop 传导数据，Mapreduce 并行运算，系统运行处理高效、便捷。

1) Sqoop 技术

介绍：

Sqoop 是连接关系型数据库和 hadoop 的桥梁，可以将关系型数据库的数据导入到 Hadoop 及其相关的系统中，如 Hive 和 HBase；也可将数据从 Hadoop 系统里抽取并导出到关系型数据库。

优势：

a) 可以高效、可控的利用资源，可以通过调整任务数来控制任务的并发度。

b) 可以自动的完成数据映射和转换。由于导入数据库是有类型的，它可以自动根据数据库中的类型转换到 Hadoop 中，当然用户也可以自定义它们之间的映射关系。

c) 支持多张数据库，如 mysql，orcale 等数据库。

2) MapReduce 模型

介绍：

MapReduce 是一种编程模型，用于大规模数据集(大于 1 TB)的并行运算。概念“Map (映射)”和“Reduce (归约)”，是它们的主要思想，都是从函数式编程语言里借来的，还有从矢量编程语言里借来的特性。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。当前的软件实现是指定一个 Map (映射)函数，用来把一组键值对映射成一组新的键值对，指定并发的 Reduce (归约)函数，用来保证所有映射的键值对中的每一个共享相同的键组。

优势：

a) 开发简单：用户不用考虑进程间的通信和套接字编程。

b) 可扩展性强：当集群资源不能满足计算需求时，可以增加节点的方式达到线性扩展集群的目的。

c) 容错性强：对于节点故障导致失败的作业，MapReduce 计算框架会自动将作业安排到健康的节点进行，直到任务完成。

3. 问题与解决方案

1) 接入第三方服务进行统计存在数据不安全

所有信息暴露给第三方平台，数据存在安全隐患，无法进行深度分析，无法定制化埋点，无法根据自身用户量身定制要求。系统通过开发 js 采集数据信息上传到企业内部服务器中解决。

2) 埋点成本高，且容易出错

埋点的数量随数据的增加而增加，埋点工程量大，开发周期长，耗时费力。系统通过 js 自动监听用户操作收集数据信息解决。

3) 埋点日志量大，通常很难找到自己想测试的埋点

采集数据大，更新快，内容繁多，无法正确定义和获取分析人员需要的数据及相关信息。系统通过 nginx 收集数据信息，并提交到 hdfs 中进行数据存储和分析日志解决。

4) 在版本迭代过程中，埋点漏了或者错了，仍需人力测试，比较花费时间

采集的数据量大，埋点的数量也就大，因此在版本迭代过程中，很容易出现埋点漏了或者错了等问题。系统通过无痕埋点，自动监听用户行为进行数据收集解决。

4. 结论

基于大数据的用户行为分析平台，具有无痕埋点、分布式服务、可视化分析、云服务(SAS)等技术特色。对用户日志信息进行提取后，可视化分析直观呈现了市场发展走势。分布式服务提高了系统的复用性与扩展性，结合无痕埋点，很大程度上提高了网站性能，对需求变更可以敏捷响应。在大数据时代下，用户行为数据是推动经济发展的支撑，企业、事业单位等迫切需求，因此拥有广阔的应用市场。而近年来数据量呈指数型发展，对海量数据进行抓取分析仍然是一个重要的研究方向。

基金项目

2016 年江苏省教育科学“十三五”规划课题(C-a/2016/01/09)；江苏省高校自然科学基金项目(BK20171166)。

参考文献

[1] 刘柳, 李文苡. 基于业务感知和用户行为分析的服务调度系统[J]. 移动通信, 2019, 43(4): 43-46.

[2] 董茂伟. 基于用户行为分析的群组推荐方法研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2018.

[3] 王菲. 基于网约车的用户行为分析系统设计与实现[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2018.

[4] 陈炜. 基于大数据技术的用户行为分析系统的研究[D]: [硕士学位论文]. 西安: 西安科技大学, 2018.

[5] 廖志芳, 李斯江, 贺大禹, 赵本洪. GitHub 开源软件开发过程中关键用户行为分析[J]. 小型微型计算机系统, 2019, 40(1): 164-168.

[6] 刘闯. 基于机器学习移动用户行为分析研究[D]: [硕士学位论文]. 长春: 长春理工大学, 2018.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2286，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sea@hanspub.org