# 基于大数据的电信用户行为分析系统的设计与实现

宋 曼

（广州城建职业学院信息工程学院，广东省 广州市 邮编 510925）

**摘要：**本文基于电信大数据设计了一个移动互联网用户行为分析系统。该系统具备数据处理、数据统计分析、数据可视化和数据管理等多个功能，为用户提供一站式移动互联网用户行为分析服务。系统提供用户流量分群分析和用户行为特征分析两个场景，对用户上网时段分布、流量特征、服务和应用偏好进行分析研究，满足移动互联网用户行为数据挖掘的需要。

**关键词：**移动互联网；用户行为分析；电信大数据

## 1. 引言

移动互联网用户上网过程中每分每秒都产生着大量的数据，电信运营商为用户提供管道服务的同时也收集了移动用户位置、上网情况、话务记录等海量数据。通过大数据技术从海量数据中分析移动互联网用户上网的共性和个性特征，发现用户流量特征、内容偏好、上网时段分布等行为习惯规律，无论是对于电信运营商有效进行网络资源配置调优，还是对于各行各业商家实现精细化运营、提高用户体验，都具有十分重要的意义和价值。

## 2. 用户行为分析系统需求分析

### 2.1 系统的总体需求

系统总体需求分以下五个步骤，数据通过处理后可以得到有价值的分析结果。具体流程如图 1 所示。

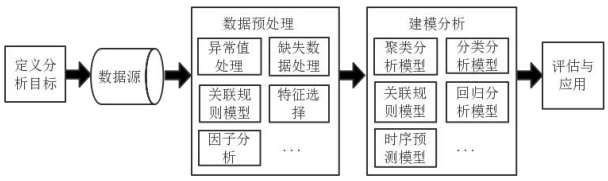


图1 移动互联网用户行为分析过程

### 2.2 系统功能需求

系统的功能需求包括用户分群分析、用户行为特征分析、数据挖掘计算和数据源管理四大功能模块，系统功能结构如图 2 所示。

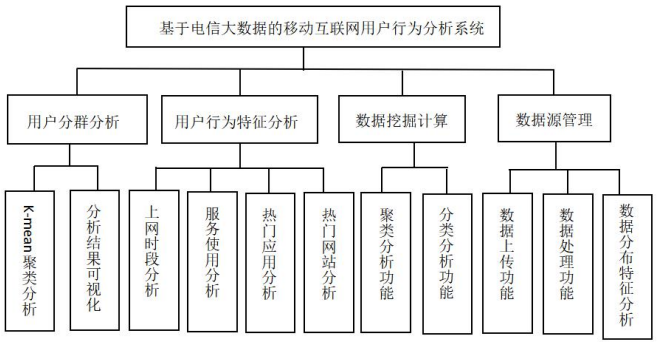


图 2 系统功能结构图

## 3. 用户行为分析系统的设计与实现

### 3.1 系统总体设计

系统总体采用 SSH 框架进行设计。底层采用 HDFS+Spark 组成的分布式集群，通过 SSH 框架处理用户交互逻辑并对结果进行可视化展示。系统采用视图层、控制层、服务层和数据层的分层设计模式。

1. 视图层

视图层通过前端界面接受用户请求操作，将请求发往控制层。视图层通过调用 Echarts 插件将结果等以柱状图、折线图、饼图等进行可视化展示。

2. 控制层

控制层负责接受来自视图层的数据清洗、数据上传等请求。控制层包括用户特征分析控制模块、用户分群分析控制模块、数据管理控制模块、数据挖掘计算控制模块。

3. 服务层

服务层负责响应控制层的请求，对请求进行处理，根据请求对相应的数据模型进行操作。服务层通过远程调用 Hadoop 和 Spark 分布式集群的计算能力，使处理异步化。

4. 数据层

数据层负责特征数据、业务数据和数据源的存储。特征数据和业务数据存储在 MySQL 数据库中，数据源存储在 HDFS 中。业务数据库负责存储数据任务、脚本等系统控制信息。

### 3.2 数据库设计

1. 特征数据库设计

特征数据库负责数据源的分布式模型、分布特征和计算结果等特征数据的存储。主要对数据多维度多级别特征进行存储，提高查询效率。

2. 业务数据库设计

业务数据库负责用户信息、任务状态信息、日志信息、脚本信息等和系统控制相关数据的存储，包含数据挖掘算法信息表、数据源信息表、挖掘任务表、参数表和数据知识流信息表。

### 3.3 系统核心模块的实现

系统的核心模块的实现主要是任务调度管理过程的实现。系统设计了任务状态管理模块和任务脚本管理模块进行任务调度管理。类的设计如下：

（1）ScriptInfo：分为 SparkScript 和 HadoopScript 两个子类。是所有分布式计算任务脚本的父类。存储各项任务的方法以及相关的信息。

（2）SparkScript：存储的具体信息包括集群 master 节点 IP 地址、脚本路径、脚本类型、执行脚本的命令行等信息，是 Spark 脚本类，ScriptInfo 类的子类。

（3）HadoopScript：该类负责实现数据在 HDFS 上的上传、下载、删除，以及获取数据信息，是 Hadoop 脚本类，ScriptInfo 类的子类。

（4）ScriptManage：提供对脚本信息进行增删改查等管理，出

现新的脚本时，将新的脚本信息注册到业务数据库中，并修改 SparkScript 脚本信息，是脚本管理类。

（5）JobInfo：任务类，动态存储脚本的相关信息。当发起以该脚本为基础的任务请求时，产生任务的概念，任务类继承 ScriptInfo 中的脚本信息，包括任务的提交用户、提交时间和完成时间、任务状态、日志和任务结果。是 ScriptInfo 的子类。

（6）JobManage：任务管理类主要负责任务管理，主要是将执行的任务线程放置在 Job 队列中进行管理，使任务在对应的平台上执行。

（7）Runner：任务执行类的父类，实现 Runnable 方法，由该类负责创建线程执行对应任务，提供 init(args)和 run()两个虚函数，实现任务执行的方法初始化和执行。

（8）FunctionRunner：Runner 类的子类，是 HadoopScript 对应的任务执行类，执行 target 指向的函数，并将执行结果返回。

（9）RomoteRunner：Runner 类的子类，是 SparkScript 对应的任务执行类，根据 SparkScript 中的脚本信息，通过远程 Shell 命令执行对应脚本，并获取标准输出。

（10）ResultHandler：根据 SparkScript 中的 Schema 信息对标准输出进行解析，将执行状态返回，并将结果存储到特征数据库中。被 RomoteRunner 调用，对远程 Shell 命令的标准输出进行处理。

4. 总结

本文主要描述了移动互联网用户行为分析系统的设计与实现。首先从系统总体架构和系统动态流程两个方面对本系统总体设计进行分解，描述了系统的分层结构以及功能实现的处理流程；然后，详细对系统数据库设计进行了介绍，详细描述了特征数据库和业务数据库表的逻辑结构；最后，详细描述了系统核心模块的主要类的功能。

参考文献：

[1]Jethwa Ketan D,Bishton Mark J,Fox Christopher P. The role of high-dose chemotherapy and autologous stem cell transplant for treatment-nalve patients with peripheral T-cell lymphoma: a systematic review of the literature.[J]. British journal of haematology,2019,178(3).

[2]The role of high-dose chemotherapy and autologous stem cell transplant for treatment-naÏve patients with peripheral T-cell lymphoma: a systematic review of the literature.[J]. Jethwa Ketan D,Bishton Mark J,Fox Christopher P. British journal of haematology.2019(3).

[3]Successful hemihepatectomy following chemotherapy for primary liver lymphoma: case report and review of literature.[J]. Skulimowski Aleksander,Hogendorf Piotr,Poznańska Grażyna,Smolewski Piotr,Strzelczyk Janusz,Durczynski Adam. Polski przeglad chirurgiczny.2018(5).

[4]刘健.电信大数据驱动下用户电话行为分析与应用[D].济南大学,2019.

[5]杨磊.基于大数据的电信公司精准营销研究[D].广西:广西大学,2019.

[6]朱成,刘海强,朱峰,等.电信大数据的数据挖掘关键技术分析与探讨[J].电信快报,2018,No.564(06):25-27.

[7]张扬.大数据背景下的国际电信业务精准营销研究[D].四川:成都理工大学,2018.

[8]汪东升,黄传河,黄晓鹏,等.电信大数据文本挖掘算法及应用[J].计算机科学，2018(12):238-244.

[9]董红羚.基于电信大数据的移动互联网用户行为分析系统的设计与实现[D].北京邮电大学,2017.

[10]基于 Spark 平台的 K-means 聚类算法改进及并行化实现[J]. 吴哲夫,张彤,肖鹰. 互联网天地.2017(01).

**作者简介**

宋曼（1984.03-），女，汉族，湖北天门人，副高，硕士，主要研究方向：大数据技术与应用。

通讯地址：广州市从化区环市东路 166 号 邮编：510925 手机：15914397526

基金支持：

2018 广东省普通高校青年创新人才项目：基于大数据协同过滤算法的电影推荐系统的设计与实现（2018GkQNCX145）；省级

2019 年校级科研项目：基于大数据的电信用户行为分析系统设计研究（Yzk16）；校级