基于流式计算的实时用户画像系统研究

姜红玉,汪　朋,封　雷
(中国电子科技集团公司第十五研究所,北京 100083)

摘　要:大数据环境下,基于海量数据,针对用户画像的精准度和实时性问题,对实时用户画像系统进行了研究工作,提出了一种采用流式计算思想的实时用户画像系统架构。从整体角度梳理分析了用户画像的体系结构,利用消息队列中间件 Kafka 实时采集不同维度的用户数据,利用大数据分析和机器学习技术构建了相对精准立体的用户画像数据标签体系及用户画像模型,应用 Flink 框架和数据挖掘技术对多源流式数据进行实时计算处理,深度分析用户,挖掘用户的特征及需求,进而刻画出精准的用户画像,提供精准的个性化信息服务。该架构能准确对用户进行全方位、高精度的画像构建,结果具有较高的实时性和精确度,从而能达到快速且准确地了解用户需求、利用数据服务用户和业务发展的目的。

关键词:用户画像(user profile);流式计算;实时;Flink;大数据;标签
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2020)07-0186-08
doi:10.3969/j.issn.1673-629X.2020.07.039

Research on Real-time User Profile System Based on Stream Computing

JIANG Hong-yu, WANG Peng, FENG Lei
(The 15th Research Institute of China Electronics Technology Group Corporation, Beijing 100083, China)

Abstract: Under the background of the big data, we carry out the research on real-time user profile of massive data and propose a real-time user profile system architecture based on stream computing for the problem of accuracy and real-time. Analyze the architecture of user profile from holistic perspective. With message queue middleware Kafka, the user data from different dimensions in real time is collected, and relatively accurate stereoscopic user profile data labeling system and user profile model are constructed through big data analysis and machine learning techniques. Flink framework and data mining technology are used to process real-time multi-source streaming data for in-depth analysis of users, mining user characteristics and needs, and then depicting accurate user portrait, so as to provide accurate personalized information services. This architecture can accurately construct the user's image with high accuracy in all directions. The results have high real-time performance and accuracy, which can realize the purpose of quickly and accurately understanding user needs, using data to serve users and business development.

Key words: user profile; stream computing; real time; Flink; big data; label

0　引　言

在移动互联网时代,用户网络交易产生的数据量正在爆炸式增长。特别是在快速发展的物联网时代,精细化运营已成为企业的重要竞争力量,“用户画像”的概念也应运而生,时代的特征为企业构建用户画像提供了丰富的数据来源。用户画像作为大数据的基础,将海量用户数据抽象出一个标签化的用户模型,有助于精准、快速地分析用户的行为习惯等重要信息,能为用户分析和用户群体分析提供充分的数据基础,奠定了大数据和物联网时代的基石。

国内外进行用户画像研究与实践的学者日趋增加,刘海鸥等梳理研究了国内外用户画像成果,揭示了用户画像建模的方法[1];王永瑞研究了百度地图的用户信息多维度分析,给出了基于用户画像进行分析的方法[2];宋美琦等对用户画像进行了研究,对用户画像的内涵、研究内容与技术方法和应用价值展开了述评[3];袁莎等分析研究了开放互联网中的学者画像技术[4];施晓光研究了用户画像在用户价值提升中的研究与应用[5]。无论是在传统行业还是当今互联网行业,用户画像研究都具有强大的发展潜力和渗透力。

用户画像技术的系统性和完整性决定了实现的难度,相比于一般的用户行为分析方法,用户画像分析方法更加完善、更具系统性,能更好地满足用户和企业的需求。因此,当前众多企业运用各种大数据分析挖掘

收稿日期:2019-08-10　　**修回日期**:2019-12-11
基金项目:中国电子科技集团重点科研项目(JY201802850)
作者简介:姜红玉(1992-),女,工程师,CCF 会员(B2169M),研究方向为微服务、大数据技术等。

技术进行用户画像的研究与应用。

本系统主要的目标是基于流式处理技术,实时综合收集繁杂的海量用户信息,应用数据挖掘技术对这些海量数字信息进行清洗、聚类、分析,逐步抽象数据形成标签,运用这些标签将用户形象具体化以形成用户画像。通过用户画像为企业提供充足且丰富的信息基础,帮助企业快速找到更全面的反馈信息,如准确的用户群和用户需求,从而为用户提供有针对性的服务,更好地帮助企业实现“千人千面”的运营。

1　概　述

用户画像(User Profile),即用户信息标签化,企业通过收集与分析用户社会属性、生活习惯、互联网行为等主要信息之后,完美地抽象出一个用户的商业全貌。交互设计之父 Alan Cooper 首先提出了用户画像的概念,用户画像是真实用户的虚拟代表,是建立在一系列真实数据上的目标用户模型[6]。用户画像的构建核心是用“标签”标记用户,标签是通过分析用户信息而得到的一个高度精准的特征标识。David Travis 认为一个完整的用户画像需要满足 7 个条件,即 PERSONA,P(基本性,Primary research)、E(移情性,Empathy)、R(真实性,Realistic)、S(独特性,Singular)、O(目标性,Objectives)、N(数量,Number)、A(应用性,Applicable)[7]。

用户画像的本质是深入分析客户,掌握具有实用价值的数据,找到目标客户,根据客户需求制定产品,并利用数据实现价值变现[8]。用户画像的分析刻画非常重要,主要体现在四个方面。第一,精细化运营,将用户群体划分成更为精细的粒度,针对细化的特定群体,通过线上推送、线下活动等手段,以激励、关怀、挽回等策略进行营销;第二,用户分析,借助用户画像更透彻地了解用户,分析不同用户画像族群的特性;第三,数据挖掘分析,用户画像是很多数据产品的基础,在其基础上可以构建个性化推荐系统、广告投放系统、搜索引擎,提升服务精准度;第四,企业管理分析和竞争分析,影响企业发展策略。

用户画像研究正处于蓬勃发展阶段,未来对于用户画像的研究应在精准场景方面加以延伸和创新,着重突破“用户”束缚,强化数据来源和数据质量的同时拓展和改进相关的数据挖掘方法,更加有效地实现多源数据的融合,从而构建更加多源、更加精准的用户画像。

2　流式计算

不同的大数据应用场景,有各自的解决方案。对于数据先存储后计算,对计算处理的实时性要求不高,同时有着非常大规模的数据且计算模型复杂的应用场景,适合使用批量计算框架。但是大多数应用场景中,数据往往动态产生,可以直接进行计算,实时性要求严格,需要在较短时间段内甚至是实时处理完成。同时在处理过程中还要考虑容错、拥塞控制等问题,保证数据处理的每一个环节都正常,确保数据不会丢失且不会被重复处理。针对这些问题,产生了流式计算框架这一解决方案。

流式计算主要是指对数据流进行实时计算,按时间点连续小批量传输大量数据,持续流入预先定义好的流式计算逻辑,提交到流式计算系统,在线系统可以实时获取计算结果进行实时展现。实时计算作为一类计算模型,主要针对流数据进行实时处理,实时计算模型可有效缩短全链路数据流时延、实时化计算逻辑、平摊计算成本,最终能够有效满足实时处理大数据的业务需求[9]。

相较于传统的批量计算,流式计算主要有三方面特点。一是,流式数据的到达、处理和输出都是持续不间断的。二是,流式数据具有瞬时性,只保存或者输出计算分析结果和部分的中间数据。三是,流式数据有明显的时间偏倚性,随着时间流逝,流式数据中所蕴涵的价值不断衰减,最近到达的流式数据通常都比早先到达的流式数据更具知识价值。

当前主流的流式计算框架有 Spark Streaming,Storm,Flink。Apache Flink 作为低延迟、高吞吐、统一流、批处理的高性能大数据计算引擎,很好地支持了流计算的场景,正成为实时流式数据处理应用的首选数据处理框架[10]。

3　用户画像体系结构

用户画像体系结构层次从下层到上层分为五层,即数据源、数据接入层、数据存储与处理层、服务层和业务应用层,如图 1 所示。

(1)数据源层。

用户画像来自于庞大而丰富的用户数据,数据源是构建用户画像的首要工作。构建实时用户画像系统首先以业务视角规划设计标签体系的整体架构,核心围绕用户,从各种数据源实时采集用户的多源数据,以用户唯一标识贯通来自各个平台、系统、渠道的数据,构建数据存储于大数据开发平台上,包括结构化的业务数据、埋点采集的用户行为数据等。

(2)数据接入层。

数据在产生阶段通常是源源不断、持续地产生的。实际情况中流式数据一般通过网络接口提供,而非通过格式化的文件进行交互。这就要求对流式数据有一个接入过程,汇聚为大数据集群内一个较统一的形式,

然后提交大数据集群进行处理[11]。数据接入层,基于大数据体系,提供接纳流式数据的能力,为上层应用分析处理流式数据奠定了基石。数据接入过程可以综合采用分布式海量日志采集、聚合和传输系统 Flume,高吞吐量的分布式发布订阅消息系统 Kafka 等技术,将从数据源采集到的数据存储到大数据开发平台中。

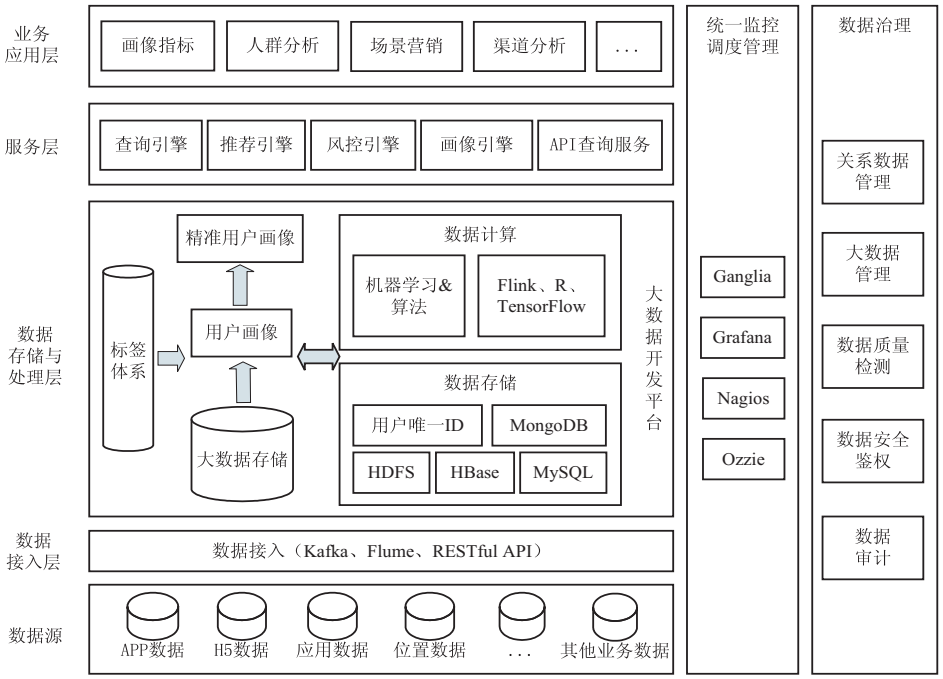

图 1 用户画像体系结构

(3)数据存储与处理层。

使用 HBase 存储用户的行为数据,服务于 Hadoop 或 Flink 的计算;使用 HBase 存储用户的画像数据,以进行在线业务查询和分析。

在大数据的基础上,进行各类标签的研发,例如事实标签、业务标签、统计类标签、算法类标签。然后根据标签体系建模产出画像数据。产品和运营人员可以利用丰富的画像数据,根据用户主题执行数据分析和数据挖掘的相关任务。分类、聚类、主成因分析、关联分析等传统的数据挖掘方法是用户画像分析的主流,且这些方法也在不断完善中。深度学习已经成为互联网大数据和人工智能的一个热潮[12],深度学习和神经网络及其衍生算法也是大数据分析的新趋势。在实际应用中,会遇到较为复杂的问题,仅仅通过算法难以解决,可以综合利用用户画像标签规则和算法去建模从而达到很好的效果。

(4)服务层。

用户画像可以直接和间接反映用户需求,为应用的设计提供客观有效的数据基础和决策依据。基于已建设完善的用户画像模型,结合具体业务,可以设计实现各种服务引擎,比如查询引擎、推荐引擎、画像引擎等等,以支撑应用层基于用户画像的各种应用。

(5)业务应用层。

作为平台级应用程序,用户画像是许多企业服务和推送的信息基础。用户画像可以定性和定量地描述用户,通过对用户性质的抽象和概括,对用户数据的统计分析与计算,实现对核心用户价值的挖掘。标签或者画像投入应用,或对接至下游业务系统,能够产生很大的业务价值。比如广告投放、个性化推荐、渠道分析等场景。广告投放应用场景基于一系列人口统计相关标签,如性别、年龄、兴趣爱好等,根据这些特征标签达到广告的有效宣传;个性化推荐技术可以推动业务增长,在当今拥有大用户量的场景下,研究新增用户的特征、核心用户的属性是否有变化等,需要辅以用户画像配合来解决调研的效用低问题。在渠道分析方面,对渠道人群进行画像验证,通过分析画像结果进行策略制定,对各个渠道的量进行重新分配,同时调整商品的定位,进而帮助企业理解用户的人群特征、消费偏好等,帮助企业分析用户群体,优化市场定位和差异化产品策略。

数据治理也是大数据分析应用取得成功的核心要素,数据的全生命周期包括数据采集、数据处理、数据传输、数据存储与使用。数据的整个生命周期中,必须确保数据的规范性、唯一性、一致性、完整性、准确性和关联性,提高数据的可用性和分析结果的正确性。

设计较为完善的用户画像系统在实际应用中也会产生一定的波动,可以建设相应的监控系统应对这一难题,监控各类标签的使用与效果,对画像的质量进行监控,进而统计出标签,替换掉不恰当的标签,同时,根据实时业务调整业务规则与算法,增添新的标签。借

助监控系统更进一步推进标签体系的规划设计,逐步沉淀出一套精华版标签集合。

4　用户画像构建

为避免形式化的用户画像,用户画像的构建需要技术人员和业务人员共同参与。用户画像系统的构建依托海量数据,其过程可分为数据采集与处理、数据标签体系建设以及模型构建三部分。首先实时采集和整合各个渠道用户的静态和动态数据,以便用户的静态信息和动态信息相关联。然后使用统计、分类、聚类分析等方法对用户信息进行挖掘分析,给用户建设标签体系。最后在此基础上构建用户画像模型,细分用户并勾勒出用户及用户群体的画像,从而更加精准地推断出用户真实需求。

4.1　数据采集与处理

在大数据环境中,一切智能化应用和分析都是建立在数据基础之上,这就需要能够收集到用户的所有相关数据,并且拥有丰富的标签以及自然语言理解的能力[13]。数据采集与处理是构建用户画像的基础,构建用户画像是为了还原用户信息,只有基于客观、真实、全面的数据,才能够生成有效、精准的画像。因此,所有采集的数据来源必须保持客观真实性。用户的特征属性可以是事实属性或抽象属性,也可以是自然属性或社会属性等,具有多方面性。这些属性都可以清楚地描绘一个用户的画像特征。

为了确保所采集数据的可用性并满足分析目标,用户画像数据可分为静态信息数据和动态信息数据两种类型。静态信息数据即用户的基本属性数据,是指相对稳定的用户信息,主要包括用户注册的基本信息,如姓名、年龄、性别等,这部分数据主要来源于用户填写的个人资料,及由此通过算法模型预测的用户数据。因为采集到的静态信息具有不确定性,不会是完全准确的,所以需要在后面的阶段中通过建模判断、完善。例如,如果用户将性别注册为女性,但通过其行为偏好将其预测为“男性”的概率更大。动态信息数据具有隐蔽性的特点,是指不断变化的用户行为产生的数据,需要通过数据分析和数据挖掘进行提取。

采集到用户的静态信息和动态信息后,在充分保障用户数据隐私和确保用户数据的真实性和有效性的前提下,首先过滤掉与用户特征无关的冗余数据和异常信息;然后把清洗过后的用户数据加工成能够被用户建模使用的数据;最后形成用户画像的有效数据集。

4.2　数据标签体系建设

用户画像数据标签是通过对用户信息分析形成的高度精炼的特征标识,如性别、地域、用户习惯、偏好等,综合所有标签勾勒出该用户的“画像”。构建用户画像数据标签是构建用户画像的关键步骤。

标签是一种相关性很强的关键字,能够表达人的基本属性、兴趣偏好以及行为倾向等某个维度,可以简洁地描述和分类人群。用户画像的结果是通过为用户贴标签的方式来描述用户信息,标签贴的是否准确和全面直接影响到用户画像的质量和结果。因此,精准、细粒度且结构化的标签体系是用户画像的基础[14],建立一套完善的标签系统必须先了解自身数据,从而能够通过该标签系统构建一个全方位的用户画像,甚至更高层次的画像模型。

标签化一般采用多级标签、多级分类,如图 2 所示。例如,第一级标签是基本信息和地理位置等;第一级分类有人口属性,行为兴趣、商业等,行为兴趣又包括运动兴趣、阅读兴趣等二级分类,阅读兴趣又分书籍和杂志等三级分类,书籍又分励志和职场等四级分类。在构建标签时,只构建最下层的标签即可,依据设计能够映射到上面三级标签。上层标签是抽象化的标签集合,一般具有统计意义,但没有实用性。例如,在广告投放应用场景中,用户人口属性标签没有实际意义,但是可以统计包含有人口属性标签的用户比例,用于产品的研究分析。

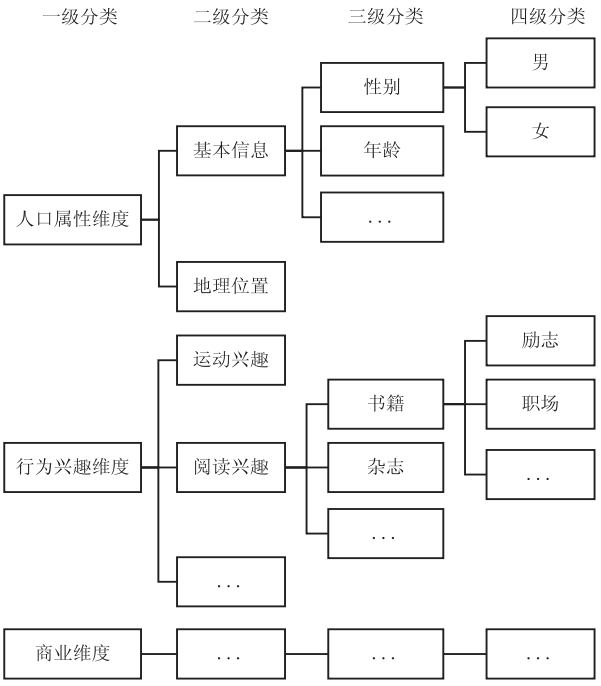

图 2　标签分类体系

大数据时代,标签体系的高效搭建已成为企业的迫切需求。标签体系的建立可以有多种方法,比如人工总计概况、调查问卷等。但是当今面对海量用户打标签的过程就需要借助大数据计算、数据挖掘等技术进行用户特征的提取,使用计算机程序化处理用户的相关信息,从而大幅提高信息获取的精准度和效率。

结合用户静态属性信息和动态行为信息,可以构

建一个相对立体、精准的用户画像数据标签体系。其中,立体指描述用户的标签维度多,精准指描述用户的标签准确,能够准确地描述用户的各种特性。标签体系应当具有原始数据层、事实层、模型层和预测层的层级结构[15],如图 3 所示。

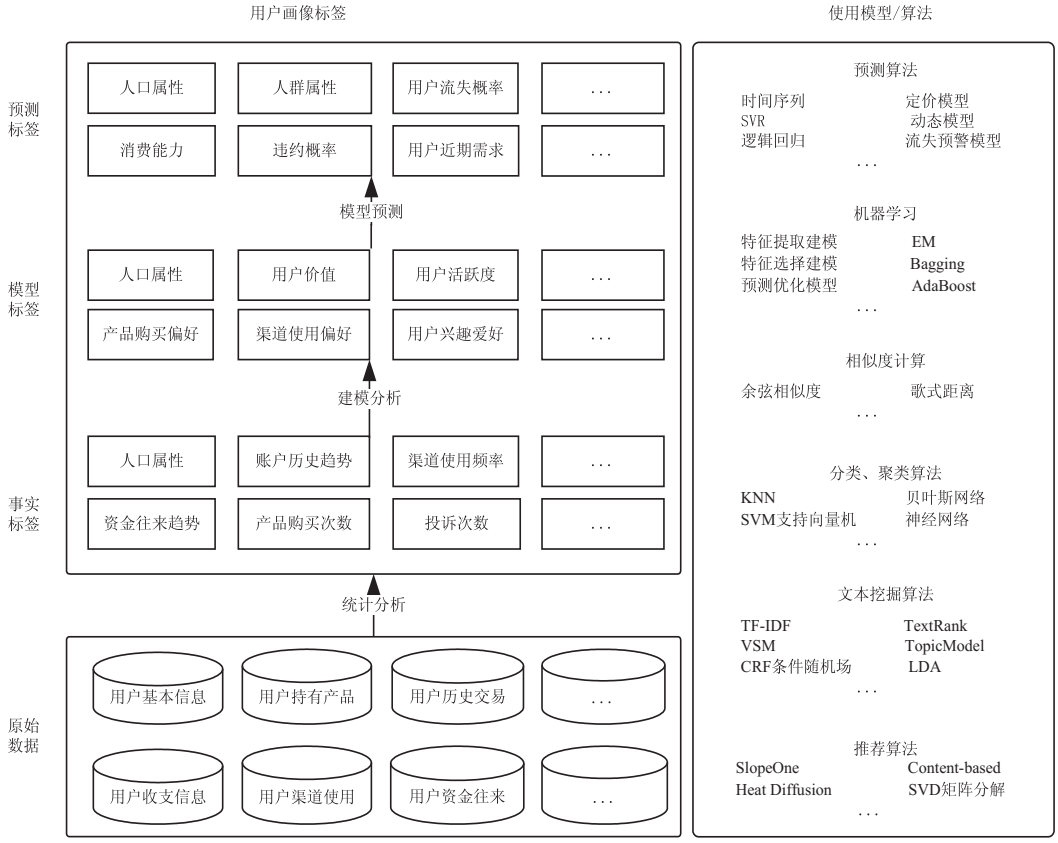

图 3 用户画像标签体系

用户画像标签模型分 4 个层次,用来描述标签的加工和计算过程。最下层是原始数据,从数据提取维度来看,标签数据可以分为事实标签、模型标签和预测标签[16]。

事实标签即既定的事实,从原始数据中直接获取,比如人口属性、渠道使用频率等都是通过用户的原始数据获得;事实标签层主要用于校验原始数据层,从而将准确无误的数据传输到模型标签层进行预测建模。

模型标签,没有与之相对应的数据,需要首先定义规则,然后建立模型计算出标签实例;根据事实标签层传输的数据,建立模型、提取特征偏好、获得用户的行为信息。根据模型标签可以为用户建立不同的标签体系,分析、加工数据,然后利用数据发现用户的潜在信息。

预测标签,参考已有的事实数据,对用户的行为或者偏好进行预测而得出的标签信息。预测标签层的建立需要通过数据挖掘和机器学习等方法,对用户特征和用户行为进行标签化。根据预测出的画像标签可以预测群体用户的忠诚度、流失度等,并探索用户的潜在需求。

用户画像标签模型中的模型标签和预测标签的生成方法有很多,包括统计方法、相似度计算算法、分类聚类算法、推荐算法、预测算法、自然语言处理等。

标签化是对用户最直观的解释,标签体系结构主要是将大数据挖掘后进行聚类分类的信息进行标签化。例如,对于一位科研工作者,用户标签化的结果如图 4 所示。

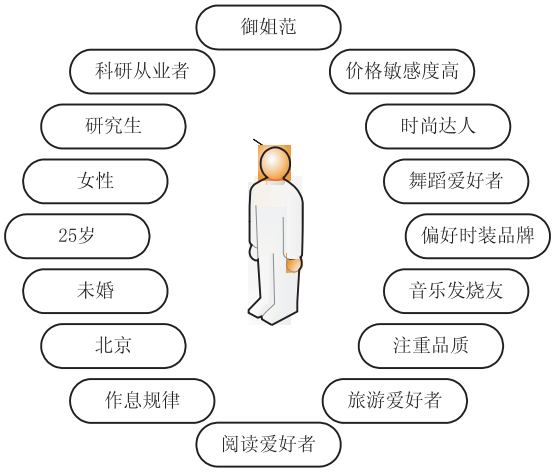

图 4 用户标签化的结果

4.3 模型构建

用户画像模型构建可归纳为“数据信息-标签-方

案”的过程,采集海量用户数据并进行处理成能表达这一个或一类用户的用户标签,然后根据相应场景形成合适且精确的方案,真正带给用户一种“千人千面”的用户体验。优异的实时用户画像系统,具备良好的数据生态,同时能够促进业务和运营的发展。图 5 是用户画像模型系统的技术架构。

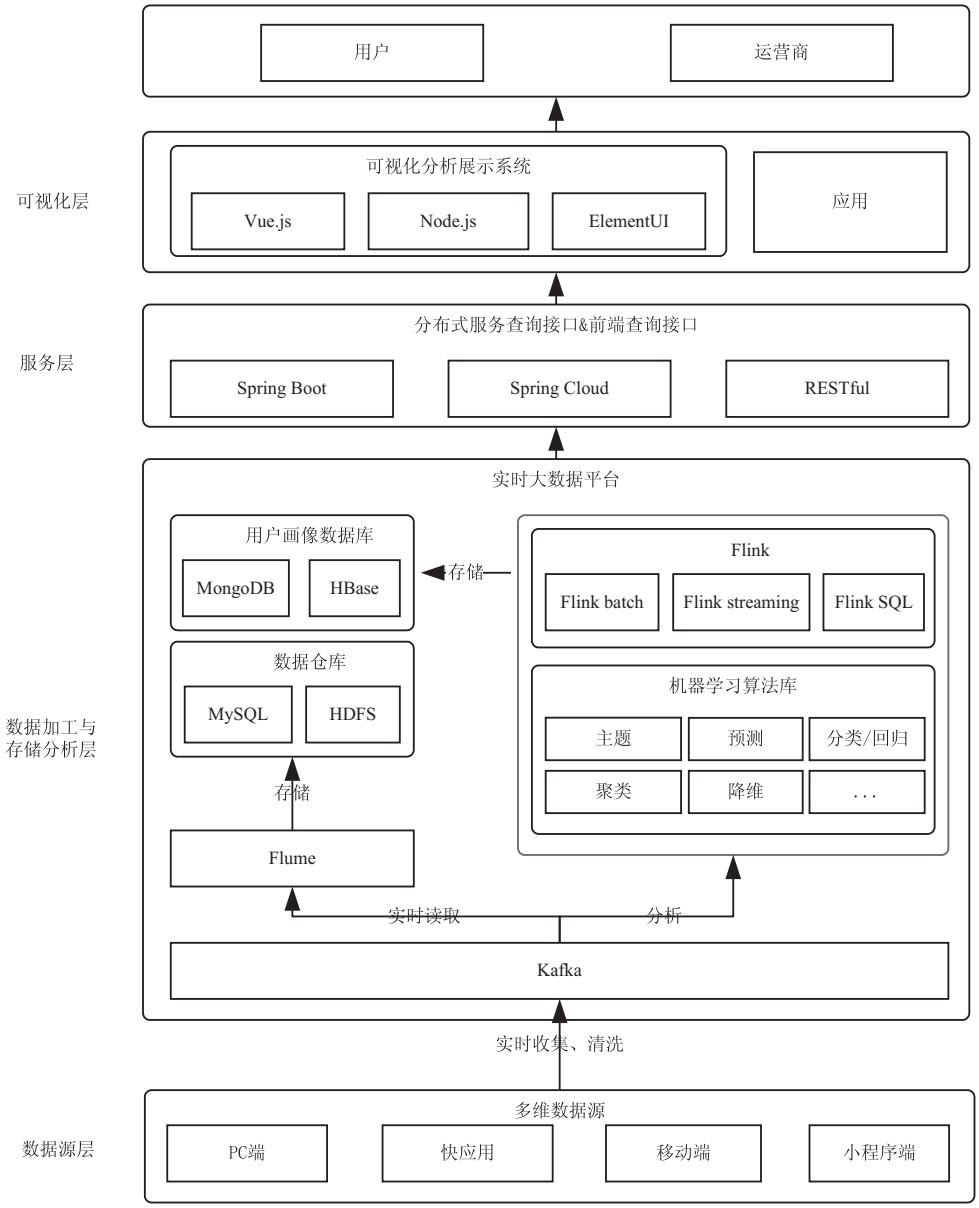

图 5　用户画像系统技术架构模型

实时用户画像系统是一个综合性非常强的系统,架构设计非常关键,采用四层架构,分别是数据源层、数据加工与存储分析层、服务层和可视化层,通过多层设计将特定功能的处理流程沉淀在各层完成。

本系统设计模式采用微服务架构、前后端分离架构设计,其可重用性较高、部署便捷、可维护性较好。前端开发语言可使用 HTML5、CSS3、JavaScript 等,应用 Vue、ElementUI、Bootstrap、和 jQuery 等前端开发框架,ElementUI 和 Bootstrap 框架可以提高开发效率。系统后台开发使用 Java 语言、Spring Boot 和 Spring Cloud 框架,可以让开发流程变得层次清晰。

(1)数据源层。

数据无疑是大数据时代最具战略性的核心资产,拥有高质量的数据是开展先进的数据分析、挖掘数据价值的前提与必要条件。采用流式数据处理,从各个维度的数据源,实时采集所有与用户相关的原始数据信息,通过一系列数据处理,将数据分为用户静态数据和动态数据。然后对采集的数据进行数据统计、分类和清洗处理。数据清洗可以利用 Hadoop 和 Flink 实现设备唯一性识别、行为数据清洗等,过滤掉与用户特征无关的冗余、冲突和异常数据。

流式处理能够实时分析连续的数据流,数据以流的方式进入系统,使用支持高吞吐量、高度稳定的分布式发布订阅消息系统 Kafka 作为数据接入手段。基于

Flume,将经过清洗和加工处理的实时数据发送给 Kafka。实时数据分析组件和数据消费者可以通过 Kafka 得到实时数据。这里通过 Flume 组件获取 Kafka 中的实时数据,并将其存储到 HDFS 中备份以备后续离线分析场景使用,通常情况下,将需要离线进行分析计算的数据存储于 HDFS 中。与此同时,定义 Kafka 的 Flink 消费者去消费 Kafka 中的实时数据,进行实时分析和计算,并深入挖掘用户的行为偏好等信息。使用 Kafka 旨在接收实时数据,但因受制于 Kafka 本身的特点,消息消费端的速度应尽可能快,保持最短的中断时间,从而不让 Kafka 存储太多的数据。

(2)数据加工与存储分析层。

数据的加工与存储分析依托实时大数据平台,实时大数据平台主要包括数据仓库、用户画像数据库和实时计算分析系统几部分,采用 Hadoop 技术框架处理企业级海量数据,分析数据,进行并行化处理。数据仓库主要包括 MySQL 和 HDFS,MySQL 用于业务数据、客户群等元数据管理,HDFS 备份存储实时收集、清洗后的用户数据。用户画像数据库包括 HBase、MongoDB。MongoDB 内存储 Flink 数据梳理后的标签对应的数据的统计结果,便于可视化分析、统计。通过 Flink 数据梳理后的标签写入 HBase,HBase 内存储的数据主要用于用户的实时查询,前端应用可通过服务获取标签数据并可视化展现。

用户画像需要维护对人的静动态、行为统计等大量维度,并且需要频繁变更字段来准确衡量各个维度的数据价值,因此需要使用具备字段弹性扩展能力的数据库来存储用户画像数据。HBase 是水平扩展的、分布式的、开源的有序映射数据库,可被看作弹性扩展的多维表格,通过动态添加列的特性,能够在数据插入或查询之前修改列结构,以支持任意的数据结构[17]。HBase 显著的特点就是无需定义表中的字段,可以直接往表中插入新字段,单表的字段数可达百万个,并且空字段不占用存储空间,适用于表结构经常调整或者字段数目非常多的数据,如用户画像场景。并且,HBase 的访问延迟在毫秒级,能够满足应用的在线调用快速响应。因此,本用户画像系统采用 HBase 存储用户画像数据,借助 HBase 的动态列这一特色功能,刻画用户的上千维度。

实时计算分析系统采用 Flink 技术进行流式数据实时化分析,Flink 和 Kafka 的组合也是比较常见的搭配。在 Hadoop 上应用 Flink 技术,能够处理分布式数据集上的迭代作业,适用于构建大型、低延迟的数据分析应用。使用基于 Flink 集群的数据管理架构,实现实时数据自动化处理,使用 NoSQL 数据库存储标签,用于构建弹性可扩展的实时用户画像系统。

用户画像是对现实世界中用户的数学建模,把用户的一些行为进行量化,用数学的手段来进行统计[18]。实时计算引擎通过监听消息消费队列内的数据,进行实时计算。整个实时计算分析系统会用到很多模型来把用户的基本属性、行为特征、心理特征、兴趣爱好、社交网络大致标签化,比如根据行为可以得出败家指数、品牌偏好、用户活跃度等标签。在数据建模过程中,主要使用机器学习中的聚类(无监督学习)和深度学习技术,使模型能够主动学习用户行为数据并对行为做出判断,从而生成用户标签。同时,可以聚集相同特征的用户,根据用户群体的特性挖掘个性化资源。画像数据存储至用户画像数据库中,以便实时调取使用,为各种精准化服务提供支持。在 Flink 上运行自然语言处理、分类聚类等组件,可以更为实时、精准地得到用户标签,实现海量数据的实时计算。

用户信息及其特征变化迅速,并且用户画像难以 100% 准确地描述一个人,只能做到无限地去逼近一个人。因此,应根据不断变化的基础数据不断修正用户画像,同时,根据已知数据抽象出新的合理标签,进而构建出更加形象、立体的用户画像。

(3)服务层。

用户画像可以看作是业务层面的数据仓库,各种标签数据是多维分析的天然要素,分布式服务查询接口用来打通用户画像标签数据。

服务层采用 Spring Boot 框架和 Spring Cloud 框架开发分布式服务查询接口以及前端查询接口,Spring Boot 作为开发单一服务的框架基础,使用 Spring Cloud 框架实现完整的微服务架构解决方案,包括服务注册与发现、监控等。服务查询接口负责获取实时计算分析后的用户画像数据,提供各种服务;前端查询接口负责调用服务查询接口提供的各种服务,以获取数据并将其封装成 RESTful Web 服务,供前端可视化层使用。前端可视化模块或业务应用通过相应的 RESTful 接口获取数据并将其可视化展示。

(4)可视化展示层。

对企业用户数据分析建模获得的信息需要进行可视化的展现,最终目的将其应用于现实中。数据可视化技术为数据分析提供了更直观的挖掘、分析和展示方法,它是一种表示数据信息的技术,它将不同种类的数据用不同的可视化视图元素描述,从而更容易地向用户展示数据中的信息[19]。

采用前后端分离的架构,前端可视化模块主要采用 HTML5、CSS3 等技术进行搭建,数据可视化效果图展示在前端浏览器网页。数据请求通过浏览器发送到服务器,获取到数据后,使用 D3、ECharts 等可视化组件绘制相应的可视化视图。为了确保系统各模块之间

的低耦合性,系统内部数据通信采用 RESTful 接口的形式。

相比传统的非可视化技术制定的图表,可视化图表能够实时的动态调整,展示出最新的实时用户画像。综合应用数据可视化技术、人的智能以及先进技术的科学计算分析能力,已逐步成为解释复杂数据的重要手段和方法。

5　结束语

伴随着大数据处理技术的不断发展和数据挖掘分析算法的演进,网络用户信息的多维度数据分析已经是当前互联网发展所必须研究的内容。文中首先对用户画像体系结构进行了探讨,给出了构建用户画像的思路,提出了一种实时用户画像系统构建通用技术架构。通过实时采集用户的多源数据,从不同维度进行用户行为分析,对数据进行挖掘分析,确定用户的事实标签,构建标签体系,从而刻画出精准的用户画像,能够秒级分析出用户的消费能力、实时兴趣偏好等,能够更加准确地了解用户需求,更好地利用数据服务用户和业务发展,体现实时用户画像系统的研究价值和意义。计划构建用户画像时,能够为类似的实时用户画像系统提供一个系统性、框架性的思维指导。

参考文献:

[1] 刘海鸥,孙晶晶,苏妍嫄,等. 国内外用户画像研究综述[J]. 情报理论与实践,2018,41(11):155-160.

[2] 王永瑞. 百度地图的用户信息多维度分析研究[D]. 武汉:武汉邮电科学研究院,2018.

[3] 宋美琦,陈　烨,张　瑞. 用户画像研究述评[J]. 情报科学,2019,37(4):171-177.

[4] 袁　莎,唐　杰,顾晓韬. 开放互联网中的学者画像技术综述[J]. 计算机研究与发展,2018,55(9):1903-1919.

[5] 施晓光. 用户画像在用户价值提升中的研究与应用[J]. 移动通信,2019,43(4):70-74.

[6] 田　娟,朱定局,杨文翰. 基于大数据平台的企业画像研究综述[J]. 计算机科学,2018,45(11A):58-62.

[7] TRAVIS D. E-commerce usability:tools and techniques to perfect the on - line experience[M]. London:CRC Press,2002.

[8] 王晓霞,刘静沙,许丹丹. 运营商大数据用户画像实践[J]. 电信科学,2018,34(5):127-133.

[9] 李　博. 阿里云开发者社区[EB/OL]. [2019-07-17]. https://developer. aliyun. com/ask/129410? spm = a2c6h. 13159736.

[10] 余海峰. 深入理解 Flink 实时大数据处理实践[M]. 北京:电子工业出版社,2019:1-47.

[11] 朱进云,陈　坚,王德政. 大数据架构师指南[M]. 北京:清华大学出版社,2016:107-160.

[12] 黄立威,江碧涛,吕守业,等. 基于深度学习的推荐系统研究综述[J]. 计算机学报,2018,41(7):1619-1647.

[13] 万家山,陈　蕾,吴锦华,等. 基于 KD-Tree 聚类的社交用户画像建模[J]. 计算机科学,2019,46(6A):442-445.

[14] 李　娜,范正洁,郝传洲,等. 采用语义分析的标签体系构建方法[J]. 西安交通大学学报,2019,53(1):169-174.

[15] NG T W H,LAM S S K,FELDMAN D C. Organizational citizen-ship behavior and counterproductive work behavior:do males and females differ? [J]. Journal of Vocational Behavior,2016,93:11-32.

[16] 张羽萍. 适应性学习系统中用户与资源画像研究[J]. 科技与创新,2018(24):84-85.

[17] SHRIPARV S. Learning HBase[M]. 周彦伟,娄　帅,蒲聪,译. 北京:电子工业出版社,2015:1-10.

[18] MOBASHER B,DAI H,LUO T,et al. Integrating web usage and content mining for more effective personalization[C]// Proceedings of the international conference on e-commerce and web technologies. [s. l.]:[s. n.],2000:165-176.

[19] SAKET B,KIM H,BROWN E T,et al. Visualization by demonstration:an interaction paradigm for visual data exploration[J]. IEEE Transactions on Visualization & Computer Graphics,2017,23(1):331-340.