文章编号：2095-2163(2023)03-0242-05 中图分类号：TP311.13 文献标志码：A

基于大数据技术的电商用户画像可视化系统设计与实现

梁肇敏，梁婷婷
（南宁学院 人工智能学院，南宁 530200）

摘 要：随着信息技术的发展，以及移动智能终端的普及，产生了海量的数据，推动了大数据时代的到来。运用大数据技术为各行业既带来了机遇，也带来了挑战，而电商行业亦是如此。如何利用大数据技术在海量数据中挖掘有价值数据，构建用户画像来提高企业的营销能力是各企业研究的重点。本文利用大数据技术构建用户画像系统，将数据进行预处理，根据实际企业业务架设大数据系统构建出用户画像模型，为实现精准营销和个性化推荐服务提供有效支撑。

关键词：大数据技术；用户画像；数据挖掘；电商

Design and implementation of e-commerce user portrait visualization system based on big data technology

LIANG Zhaomin, LIANG Tingting
(College of Artificial Intelligence, Nanning University, Nanning 530200, China)

【Abstract】 With the development of information technology, the popularity of mobile intelligent terminals has produced massive data, which has promoted the arrival of the era of big data. The application of big data technology has brought opportunities and challenges in various industries, and the e-commerce industry is also the same. How to use big data technology to mine valuable data in massive data, and build user portraits to improve the marketing ability of enterprises is the focus of research by various enterprises. This paper uses big data technology to build a user portrait system, preprocess the data, and build a user portrait model based on the big data system according to the actual enterprise business, providing effective support for the realization of precision marketing and personalized recommendation services.

【Key words】 big data technology; user portrait; data ming; online retailers

0 引 言

互联网飞速发展的现代社会，信息技术正从各个方面影响着企业商业模式的改变。在企业的商业活动中，客户是商业活动的中心，只有客户在商业活动中产生消费行为，企业才能获得利润和发展[1]。所以企业为了能够让客户在自己的商业活动中消费，就需要投入很多成本来吸引客户。在当代社会，电子商务日益发达，更多的人喜欢在线上购买自己想要的商品。考虑到人们在线上各种各样的消费行为，随即产生了大量的数据，同时商家在线上也会提供大量的商品信息。这样导致用户将无法精确选择自己感兴趣的商品，企业也无法对海量用户进行精确广告投送，从而产生了网络上信息过载的问题[2]。为解决上述问题，需要对数据信息进行精准的推荐，用户画像则是实现推荐的重要环节[3]。研究发现，利用客户产生的行为数据构建用户画像，可以让企业全面了解客户的喜好，摒弃传统单一的商品信息投送策略，进而转向精准推荐投送，就能让企业花费最小代价找到契合客户，客户也能找到自己感兴趣的商品，这对企业的发展和提高企业与用户的沟通效率等都有积极的意义[4]。

本系统是基于电商平台进行设计和开发，是面向注册会员的偏好、行为习惯和人口属性的画像还原，同时也包括对商品信息的画像还原。提供用户喜好和商品特征帮助营销平台提升营销的精准度，也有利于个性化推荐系统快速准确地为每个用户推荐相关的商品。

基金项目：广西高校中青年教师科研基础能力提升项目(2021KY1800)；南宁学院校级科研项目(2020XJ10)。

作者简介：梁肇敏(1985-)，男，讲师，主要研究方向：智能推荐、计算机技术；梁婷婷(1983-)，女，教授，主要研究方向：信息检索、计算机应用技术。

通讯作者：梁婷婷 Email：whnmyt@qq.com

收稿日期：2023-01-14

1　需求分析

本系统设计并实现基于大数据的电商用户画像系统是为了将电商客户数据收集起来，深度挖掘出其价值，从而应用在企业的各种营销活动中。所以系统应该满足：能从网速抓取客户网上基本属性、行为等数据，全方位地分析用户数据，构建用户画像，并且能以良好的交互方式展现给使用者[5]。分析可知，系统应该具有以下功能：

（1）服务端数据采集功能。系统自身的客户数据作为用户画像数据源的一部分，以文件的形式从各个子系统传输到服务器的指定存储目录，系统进行统一的汇总分析。

（2）网络爬虫功能。电商网络产生海量客户行为数据，这对构建用户画像系统具有重要意义。但是网上客户行为更多是一些网页链接地址数据，所以需要爬虫技术根据客户访问的网络链接爬取链接所指向的文本内容数据，从而分析出行为偏好。而面对网络海量的电商客户数据，利用分布式爬虫技术是很好的解决方案。

（3）数据存储功能。构建用户画像系统所需要的数据具有海量、且数据类型多样性等特点，这就需要系统具备存储海量多样性数据的能力，以满足海量多样性数据存储的需求。

（4）数据处理与分析功能。从不同的数据源传输过来的数据并不能直接加以利用，更多是掺杂着许多“脏数据”，需要系统首先对数据进行去重、数据标准化等预处理，此后才能利用各种算法模型去分析和挖掘出有价值的数据，进而构建电商用户画像。

（5）数据可视化功能。构建用户画像后，提供给用户使用。系统需要提供生动、直观的报表、图形等可视化界面，让用户更容易获取和理解用户画像的分析结果。

2　系统总体架构设计

2.1　系统体系结构设计

基于大数据技术的电商用户画像的总体架构如图 1 所示。由图 1 可知，系统主要分为数据源层、数据获取层、数据存储和处理层、数据访问层、数据服务和展现层。系统中的每一层都对数据的处理有不同的任务和分工，根据整体架构图的设计，对系统的各个功能层进行全面的功能说明和描述。

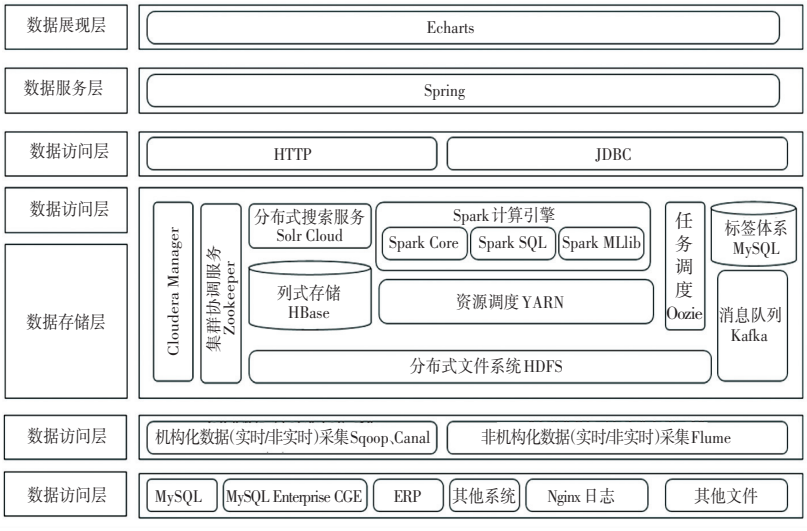

图 1　系统整体体系架构图

Fig. 1　Overall system architecture diagram

（1）数据源层。用户画像系统的最低层是数据源层，该层以实现数据的 ETL（Extract&Transform&Load）过程和利用 Hadoop 进行数据存储，不管是海量数据的存储、还是其他形式的数据，都应持久化在本层，并通过 Hadoop 的高可用、高可靠，保障数据使用的效率与存储的安全性。

（2）数据获取层。该层主要是对数据库 MySQL、MySQL Enterprise CGE、ERP 或者其他系统中的数据进行采集，通过 Sqoop 导入到大数据平台中，Canal 是将数据实时采集到大数据平台中，而 Nginx 日志和其他文件通过 Flume 进行采集。

（3）数据存储层和数据处理层。这 2 层对数据的操作都是基于 Hadoop 平台来完成，主要是对用户原始数据和中间数据进行适当的处理和存储。体系

里分布式文件系统 HDFS 负责对系统相关数据和用户数据进行存储。系统产生的标签存储在传统关系数据库 MySQL 中。不论是系统产生数据、还是从外部源爬取的数据都会以大文件形式存储在 HSFS 和 MySQL 中,随后再导入到 Hive 中进行数据分析。Hive 分析结束后,会将分析的用户画像结果存储在 HBase 中,为最终数据可视化提供数据支撑。

(4)数据访问层。该层主要是为上层可视化或其他应用提供数据接口服务,使前端平台能对其进行有效连接。

(5)数据服务和数据展示层。该层主要是利用 SparkMLlib 相关算法库进行复杂的标签计算后,通过简洁清晰的界面展示给用户。

2.2 构建用户画像流程

用户画像构建流程如图 2 所示。就是将数据通过一系列操作转换成映射的方法,以数据为驱动力推动用户画像的构建。先是将用户的性别、邮箱、地址等属性数据和行为、兴趣爱好等行为数据进行收集,然后利用相关机器学习算法对用户数据做标签化转换并进行用户画像建模,最后根据用户画像模型进行分析和可视化[6]。

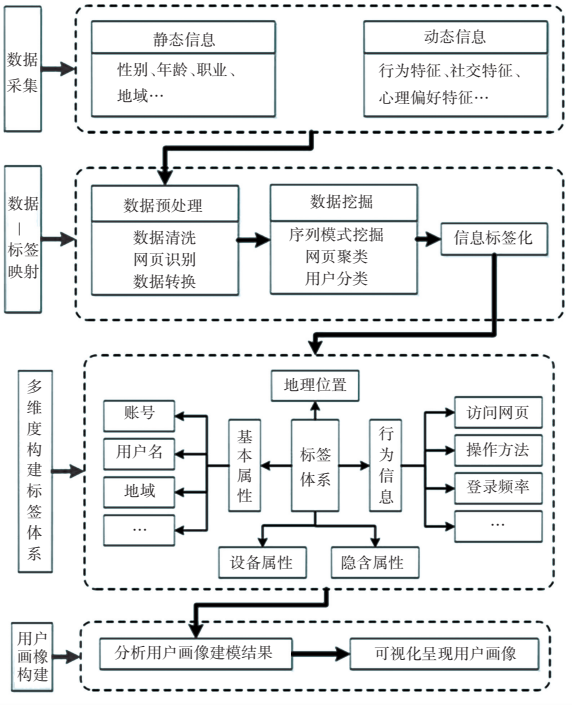

图 2 用户画像构建流程图

Fig. 2 User portrait building flowchart

3 系统功能设计

3.1 标签体系的建立

标签体系是根据已注册用户的偏好、行为习惯和人口属性等不同的领域而建立起来的[7]。标签体现划分需求分析见图 2,按领域可以分为人口属性、商业属性、行为属性和用户价值四类。

按具体的实现方式分为规则标签、统计标签和挖掘标签。

在本项目中,标签体系按照业务类型划分为基础标签和组合标签。按领域可以划分为:人口属性、用户的社会特征相关标签;商业属性、电商平台中购物相关的标签;行为属性、电商平台中的浏览、购买等行为标签。

按实现方式划分为:规则标签、通过匹配标签的属性值实现标签的业务逻辑;统计标签、使用数学统计方法实现标签业务逻辑;挖掘标签、使用数据挖掘算法实现标签的业务逻辑。

按业务类型划分为基本标签和组合标签。其中,基本标签描述基本属性,如性别、民族、职业等。组合标签是多种基本属性组合而成的,如高净值用户就是学历、消费能力、房产属性的组合。

3.2 标签挖掘流程与算法

标签挖掘与算法总体流程如图 3 所示,对于挖掘类型标签开发来说,分为 2 步:

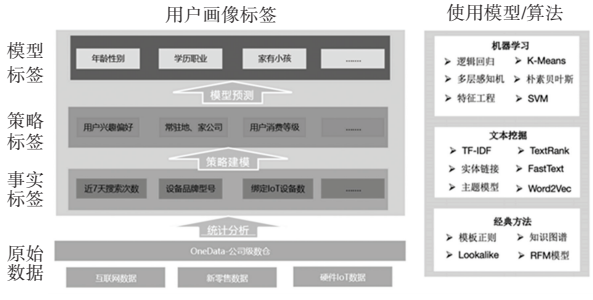

图 3 标签挖掘流程与算法图

Fig. 3 Label mining process and algorithm diagram

(1)构建算法模型。构建算法模型从业务数据中获取算法特征数据(features),此外如果是监督学习算法,需要标注数据(label)。流程包括业务数据、特征工程 、训练模型、最佳模型、保存模型、标注数据、特征转换、特征提取、算法超参数和模型评估[8]。

(2) 模型预测值。加载模型(算法模型提取训练好,保存起来),封装方法(loadModel),如果模型不存在,使用数据训练,获取最佳模型,并保存起来。predictionDF 结合属性标签规则,给每个用户标注上具体的标签值。

3.3 标签体系的存储

系统所涉及标签数据存储在 MySQL 数据库中。对于标签表和模型表,标签表负责存储基础标签,主

要存储标签名称、标签规则、等级等基本信息。模型表存储每个 4 级标签，具体 Spark 应用程序的相关信息。存储标签数据时，也将标签数据同步存储到 Elasticsearch 索引中，方便使用标签进行用户查询，基于 Elasticsearch 为 HBase 表构建二级索引。

3.4　标签模型开发流程

每个标签模型开发流程如图 4 所示。首先是标签管理平台新建标签，然后开发标签模型，最后调度执行。

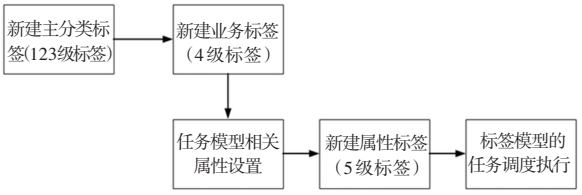

图 4　标签模型开发流程图

Fig. 4　Label model development flow chart

3.5　标签调度

标签调度如图 5 所示。主要是基于 Oozie 实现 Web 管理平台和 Yarn 计算平台的调度，方便计算任务的管理。Oozie 在这里发挥公共协同的作用，所有的标签（模型应用）都需要使用 Oozie 来进行调度、执行标签计算。

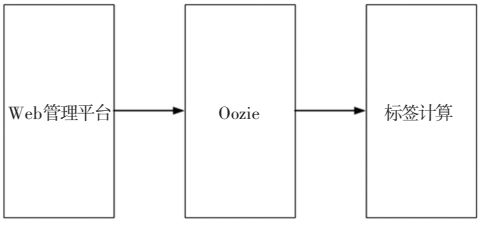

图 5　标签调度流程图

Fig. 5　Label scheduling flowchart

3.6　标签管理平台设计

标签管理平台前端使用 Vue，后端开发使用 SpringBoot，这个管理平台主要负责对标签的创建/查询等进行操作，并且对标签运行状态等进行管理。开发人员可以使用平台并应用 JAR 包上传相应的标签计算，启动标签计算任务，方便标签管理[9]。

3.7　标签模型计算

标签模型计算如图 6 所示。主要负责根据原始表数据以及 MySQL 中的预先设置的标签规则进行相应的计算，比如对规则匹配型、统计型和数据挖掘型等标签有关的计算操作，最终得到用户的标签结果，并将其存储到 HBase 中。这里需要注意的是，在保存到 HBase 的时候，本次的保存一定不能覆盖掉上次计算的标签结果，要将历史的标签数据和新生成的标签数据进行合并操作，这样才能保证数据的完整保存，不会造成数据丢失。本模块是用户画像的核心，主要负责根据原始数据以及标签规则进行相应的计算，比如规则匹配/统计/挖掘等相关操作，最终得到标签结果，将结果存入 HBase 中。

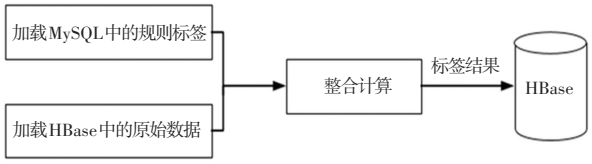

图 6　标签模型计算流程图

Fig. 6　Label model calculation flowchart

4　系统可视化效果实现

用户画像模型封装基于关系型数据库（PG）和大数据平台（Hive、Impala）包含基础标签与分析类知识标签，实现用户特征全貌刻画，即多种封装角度，分用户类型、渠道内容、业务场景进行封装配置。接口数据实时推送，实现用户画像数据实时更新至运营及营销统一视图（WeMeta、WeDate、WeSearch 等）中进行展现，并实时反馈运营及营销信息问题，保证数据应用的时效性。展现 UI 封装依托用户画像，将推荐信息配置应用端进行可视化展现，集中活动运营，实现千人千面的运营效果[10]。

用户数据画像可视化界面如图 7 所示，涉及年龄分布、消费占比、行业区分比例、新增会员信息、消费记录、所属行业分布、用户偏好、精准营销、地区分布可视化功能模块，为企业提供了足够的信息基础，能够帮助企业快速找到精准用户群体以及用户需求等更为广泛的反馈信息。用户画像是在了解客户需求和消费能力、以及客户信用额度的基础上，寻找潜在产品的目标客户，并利用画像信息为客户开发产品。

图 7　用户数据画像可视化界面

Fig. 7　Visual interface of user data portrait

（下转第 249 页）