[文章编号] 1003－4684(2019)05-0078-05

基于 Spark 的电信用户画像的研究应用

华 满，邵雄凯，高 榕
（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 为达到精准推荐，给用户提供个性化服务的目的，通过以 Spark 为核心的大数据技术，对电信运营商的数据进行挖掘、分析、聚类、建模，从而发现不同用户的个性化需求。实验结果表明，此方法能够较好地对不同用户的行为进行画像，同时经过优化后的 PK-means 聚类方法准确率有明显提高，与传统的数据处理模式相比，运算速度得到极大提升。

[关键词] 用户画像；Spark；精准推荐；k-means；数据降维
[中图分类号] TP391 [文献标识码] A

传统企业在处理数据分析问题时，往往面临着数据来源多样、数据量巨大、数据计算复杂、速度慢等诸多问题。本文以 Spark 大数据框架为基础，并结合用户画像技术，将繁杂的数据抽象出来，改变过去依靠单机运行或者磁盘运算的模式，对用户的行为进行深度分析，更加全面高效地了解用户的需求。用户画像[1]是推荐领域的一项重要的技术，最早应用于互联网电子商务行业，旨在通过挖掘用户的浏览轨迹，使用大数据、机器学习等方法为用户生成个性化的标签，从而达到画像的目的，然后依据这些标签给客户推荐他们喜爱的产品。目前，百度、阿里、腾讯都已经建立了自己的大数据营销平台，从而更加精准地服务用户。随着以 Spark 为首的相关大数据技术的进一步发展，大数据技术已经逐渐从互联网应用到各种不同的传统产业当中。孟巍[2]等基于大数据技术对如何为电力用户进行用户画像进行了研究与总结，使得电网企业能够更多了解用户；单晓红，张晓月[3]使用携程酒店的在线评论等数据，运用本体技术从用户信息、酒店信息和用户评价信息维度构建用户画像，用于提升用户维护和服务工作的精准性。本文的电信用户画像系统是根据某省电信运营商精准营销的实际需求，运用分布式存储技术、并行计算技术等大数据技术对整个系统进行设计并实现。以某省电信运营商的脱敏数据作为数据源，将用户的多方面数据进行汇聚梳理，实现电信用户画像系统，从多个维度构建电信用户的用户画像，准确把握用户特点及需求，进而挖掘出数据中隐藏的价值和普遍规律，将其应用于用户商品推荐，广告精准营销和用户个性化服务。

1 用户画像模型

用户画像的模型大致分为四个阶段：原始数据获取、数据预处理、数据建模、为用户画像。与传统方式不同的是，本文在 Spark 框架下展开研究，能够更加高效地访问不同来源的数据且运算效率明显提高，流程见图 1。

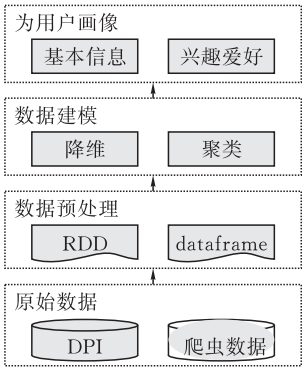

图 1 用户画像流程

1.1 Spark 大数据技术

本文中使用的大数据处理框架是 Spark[4]，与之对应的是 Hadoop[5]，这两个框架最大的区别在于 Spark 是基于内存的计算，而 hadoopd 的计算需要借助磁盘 I/O，所以在相同的数据量规模下，Spark 框架的计算效率是 hadoop 的几倍至几十倍。

如图 2 所示，Apache Spark 作为 Spark 的核心

[收稿日期] 2019－06－04
[第一作者] 华 满(1994－)，男，湖北黄冈人，湖北工业大学硕士研究生，研究方向为大数据与云计算

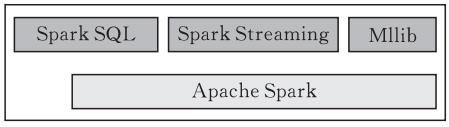

图 2　Spark 应用模块框架图

组件,是从事 Spark 相关计算必须具备的组件,其主要功能是将分布式数据抽象为弹性分布式数据集(RDD),实现了应用任务调度、RPC、序列化和压缩,并为运行在其上的上层组件提供 API。SparkSQL 是 Spark 来操作结构化数据的程序包,可以使用 SQL 语句的方式来查询数据,Spark 支持多种数据源,包括 hive、hbase 数据库等。Spark-Streaming 应用于实时流式计算,主要将实时产生的数据流进行快速运算,实时产生输出;而 MLlib 则是为程序提供大数据环境下的机器学习训练方法,包括数据的预处理、预测、优化、分类、聚类等.在 Spark 方面,本项目中主要使用 Spark core, Spark SQL 和 MLlib 对数据进行计算处理。

实际项目中,某电信运营商在全国一天的 URL 原始日志数据就达到了 56 TB,如此庞大的数据规模下,传统的关系型数据库已经无法完成海量数据的高效率读写,且关系型数据库一般拥有固定的表结构,无法根据程序需要灵活读写数据。Hbase 是一个开源的、分布式的非关系型数据库(NoSQL),具有存储空间可动态扩展、数据分布式存储、更加安全高效的优点,并且支持列式存储,能够灵活地读写不同领域的数据[6]。HBase[7]作为在 HDFS 上开发的面向列的非关系型分布式数据库,利用 HDFS 作为其文件存储系统,通过 Zookeeper 协同服务。

Hbase 的结构说明见图 3。

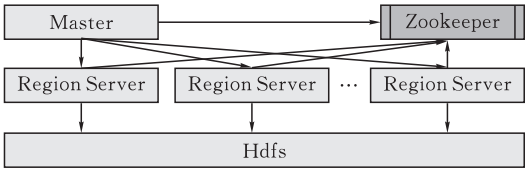

图 3　Hbase 数据库结构框架

HBase 具有与 HDFS 相同的架构模型,采用一主多从的模式,即一个主节点,多个从节点。不同的是:HBase 增加了 Zookeeper 作为协调各个节点的枢纽,使配置数据同步更新,Zookeeper 服务器还负责维护 HBase 集群的健康状况。

1.2　数据预处理

数据来源方面,DPI 数据来源于运营商,而爬虫数据源于笔者通过分析用户的行为轨迹,抓取的网页数据,获取其中的内容后持久化所得。本文中使用如图 4,在用户数据的筛选部分使用 Spark SQL 的 Filter 算子实现,对数据进行过滤筛选操作:

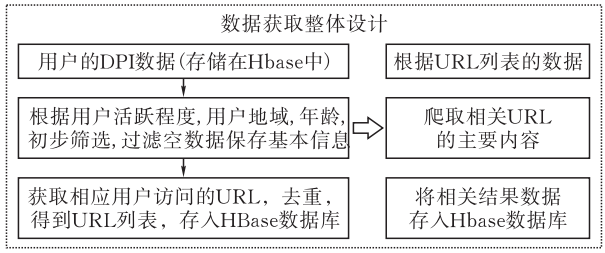

图 4　初始数据获取

在数据预处理中,包含数据结构化、分词、去停用词、将词表转化为向量,这其中分词采用 Hanlp 分词工具,停用词表的功能是去除没有意义的连词、代词、标点符号等,而将词语转化为向量应用的是 Word2Vec 模型,与 TF-IDF 相比,此模型在挖掘语义潜在关系上具有较大的改进。

1.3　文本聚类

文本聚类是用户画像中进行标签转化的重要一步,简而言之,通过聚类将用户访问的不同内容聚类成一定数量的类别,将每一种类别的内容提取出关键字作为这类网页的标签。本文使用的聚类算法主要基于 Spark 机器学习库,是在大数据的基础上对数据进行聚类训练实现的[8],使得大规模数据的并行计算与机器学习算法相结合。聚类是指在一个数据集中选取出 k 个簇,使得每个簇中的数据点所具有的特征相似,但与其他簇中的数据点特征尽可能不同。在数据挖掘中常用来找到数据的不同类别。对于给定的样本集,按照样本之间的距离大小,将样本集划分为 k 个簇。让簇内的点尽量紧密地连在一起,而簇间距离尽量大。用数学表达式表示,假设簇划分为($\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k$),则本文的目标是最小化平方误差(SSE):

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2$$

其中 $\boldsymbol{\mu}_i$ 是簇 $\boldsymbol{C}_i$ 的均值向量,同时也称为质心,如

$$\boldsymbol{\mu}_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$$

在 K-Means 算法中输入数据集:$D=\{x_1,x_2,\cdots,x_m\}$,聚类的簇数 k,最大迭代次数为 N 输出为每个数据点的簇划分:$\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k\}$

使用的 K-Means 算法流程如下:

1)从数据集 D 中随机选择 k 个样本作为初始的 k 个质心向量$\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_k\}$;

2)对于不同的迭代次数 $n=1,2,\cdots,N$;

a)将簇划分 $\boldsymbol{C}$ 初始化划分为 $\boldsymbol{C}_t(\boldsymbol{t}=1,2,\cdots,\boldsymbol{k})$;

b)对于 $i=1,2,\cdots,m$,计算样本 x_i 和各个质心

向量 $\boldsymbol{\mu}_j(j=1,2,\cdots,k)$ 的距离：将 x_i 与各个质心中距离：最小的标为 d_{ij} 所对应的类别 λ_i，此时更新；

c)对于 $j=1,2,\cdots,k$，对 c_j 中所有的向量，更新每个簇的质心；

d)如果所有的 k 个质心向量都没有发生变化，则转到步骤 3)；

3)将所有的数据划分为 k 个不同的簇输出 $\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k\}$。

在文本聚类中，首先将收集到的数据进行过滤清洗，结构化，然后将文本进行分词，去除其中没有任何意义的词语和标点符号，并结合 Word2Vec 算法，将文本转化为向量。根据新华社的统计，目前常用汉字 3500 余字，这些汉字组成的词语达 50 多万。随着数据的增多，转化的向量维度也会快速增长，这不仅会降低算法的执行效率，消耗运算资源；同时，冗长的向量表达还会掩盖算法对文本中重要特征的提取，导致聚类的准确性降低。PCA 算法(Principal Component Analysis)，是一种数据分析方法，也称为主成分分析算法。PCA 通过线性运算将原始数据变换为一组低维度的线性表示，可用于提取数据的主要特征分量，实现高维度数据的降维[9-10]。本文通过 PCA 算法对转化成向量过后的数据的降维处理并聚类，提出一种新的聚类方法 PK-means，其在凸显主要特征的同时，减少向量计算的开销，实验结果表明，与原始的 K-means 算法相比，前者在准确性和运算效率上皆有明显提高。

1.4 用户画像标签提取

在用户画像标签提取中，将处理好的数据使用 PK-means 聚类算法训练生成模型后，依次遍历 k 的个数，$k\in(10\sim100)$，根据平方误差和(SSE)以及轮廓系数(Silhouette Coefficient)，找到最合适的 k 值。然后根据聚类的结果提取出同一类别的所有词语中心词。取得最佳聚类结果后，用户画像具体展示设计见图 5。

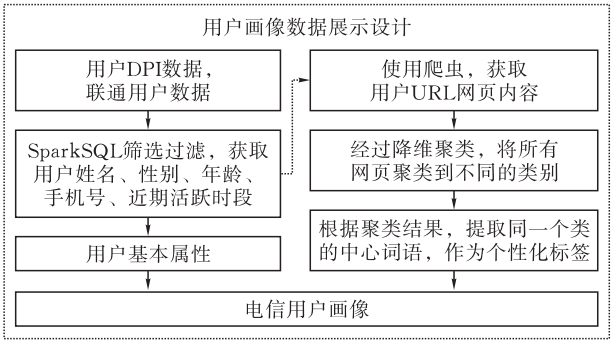

图 5 用户画像展示

用户画像标签展示的具体处理步骤如下：

步骤一：根据运营商的平台的接口，获取到 DPI 数据，运营商用户数据，使用 Spark SQL 获取到用户的姓名、性别、年龄、籍贯、用户访问的 URL 等数据，结构化后转化成 DataFrame。

步骤二：将用户 ID 、用户访问的 URL 作为爬取用户访问内容的初始数据源，同时对用户信息根据籍贯、年龄、用户是否欠费进行筛选。

步骤三：基于查询、统计等方法找到用户基本属性：用户姓名、性别、年龄、手机号、近期活跃情况，从而得到用户基本属性画像。

步骤四：用户访问的 URL 作为爬取用户访问内容的初始数据源，利用 URL 爬取用户 URL 中的内容存储到 Hbase 数据库。

步骤五：取出用户访问 Url 的内容数据，然后对数据预处理，包括 URL 数据去重，过滤掉爬取内容为空的数据，对已有数据去除停用词，使用 Hanlp 分词工具，用 word2Vec 算法把分词后的数据映射成等长的向量。

步骤六：用 PK-means 算法对向量数据降维、聚类处理，通过多次的反复迭代训练，得出最佳 k 值的聚类结果。

步骤七：将每个聚类下处理后的中文词语取出，提取出中心词，作为该类的用户喜好标签。

步骤八：用户画像展示，通过用户 ID，分别获取当前 ID 用户的基本属性标签和喜好标签。

2 实验过程

2.1 实验平台环境

实验平台的环境由三台服务器组成，一个主节点，两个从节点，所有程序都安装在 /usr/local/目录下面，集群环境和相关组件部署见表 1。

表 1 实验平台环境

IP 地址	192.168.2.114	192.168.2.20	192.168.2.21
节点	主节点	从节点	从节点
主机名	bigdata	worker1	worker2
运行系统	CentOS 7	CentOS 7	CentOS 7
环境	Hadoop，Spark，scala，Hbase	Hadoop，Spark，scala，Hbase	Hadoop，Spark，scala，mysql，Hbase
内存	64GB	64GB	64GB
硬存	1TB	1TB	1TB

生产环境软件版本声明：

Hadoop 2.6.0-cdh5.5.4

scala 2.11.4

java 1.8.0_171

HBase1.0.0-cdh5.5.

Spark 2.2.0

2.2　模型训练

本文的数据源来自于两个方面，其中一部分来自于运营商内部的 DPI 日志数据，包含了用户浏览内容的记录，用户获取信息的时间内容等；另一部分是根据用户访问的 URL，制定爬取策略所得，即使用用户浏览网页访问的初始 URL 编写爬虫，爬取相关 URL 的内容。爬虫部分，本文使用的是 node.js，由于 node.js 主要应用于前端页面，同时也对编程逻辑有很好的支持，所以对获取页面内容具有极大的优势。训练流程如下：

1）数据获取：从原始数据中筛选出包含用户 imei 的数据，用户访问的 URL 数据，用户的基本信息，过滤掉关键字段为空的数据，并将结果存储 hbase 数据库。

2）数据预处理：将用户访问的内容数据进行分词、去停用词，然后转化为向量。使用 Hanlp 分词工具对数据内容进行分词，用空格分离，并建立停用词表，停用词表是指在语句中出现的对聚类没有实际意义的词语和符号。包括代词、连词、标点符号等。

3）聚类模型训练：将处理好的数据按照 $k \in (10 \sim 100)$，取出最优的 k 值，最优的 k 值由误差平方和（SSE）和轮廓系数共同决定（图 6、7）。

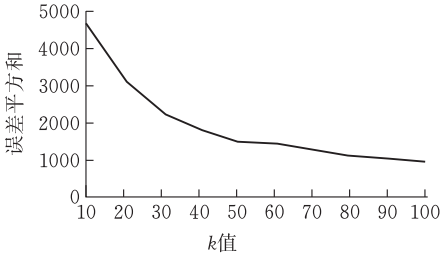

图 6　不同 k 值的误差平方和

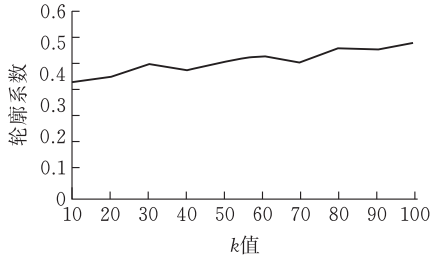

图 7　不同 k 值的轮廓系数

图 6 中的纵坐标为每个数据点到质心距离的平方和，称为误差平方和，横坐标表示不同的 k 值，可以看出随着 k 值的逐渐增大，误差平方和也随之减小，但是由于 k 值在靠近最佳聚类个数时更加适应聚类模型，使得误差平方和快速下降，而远离最佳聚类个数时误差平方和减小速度则会减慢，从图中可以看出，最佳 k 值在 50～80 时为最佳聚类个数.而图 7 中轮廓系数反应的是各个数据点与质心的聚集程度，轮廓系数的值越大，表明数据聚集效果越好。结合图 6 和图 7 可以推测出最佳的 k 值为 80。

2.3　实验结果分析

从实验结果的准确性来看，基于 pca 优化的 PK-means 聚类算法准确性具有明显改善，图 8 为阅读、小说类网页的聚类结果。

host	predict	filteredWords	id
s9-static.jjv	40	言情小说, 都市, 言情小说, 免费, 言情小说, 在线, 阅读, 晋江, 文学, 言情小说, 原创, 言情小说, 穿越	787
xs.sogou.c	40	多多, 看书, 免费, 全本, 小说, 推荐, 最新, 好看, 小说, 阅读, 多多, 看书, 小说, 好看, 小说, 免费, 全本	814
m.138001(	40	138, 看书, 好看, 免费, 小说, 阅读, 138, 看书, 最新, 章节, 全文, 阅读, 在线, 免费, 阅读, 好看, 小说	890
mad.ireade	40	书城, 小说, 出版, 图书, 网络, 文学, 原创, 小说, 免费, 下载, 在线, 阅读, 电子书, 经管, 励志, 名人, 传	1197
www.oldtir	40	校花, 系统, 最新, 章节, 校花, 系统, 弹窗, 免费, 阅读, 旧时光, 文学, 校花, 沈宴, 初暖, 校花, 最新, 盖	1367
mpay.xima	40	有声, 小说, 小说, 有声, 在线, 书电台, 喜马拉雅, fm, 有声, 小说, 有声, 儿童, 故事, 相声, 鬼故事, 电	1385
i.weread.q	40	微信, 读书, 阅读, 不再, 孤独, 微信, 读书, 微信, 阅读, 微信, 小说, 阅读, 免费, 小说, weread, 原创,	1533
m.7xxs.net	40	言情小说, 手机, 小说, 言情小说, 言情, 手机, 小说, 言情小说, 收集, 手机, 小说, 最新, 章节, 最新, 更	1544
yike.allibo	40	在线, 书城, 小说, 免费, 小说, 穿越, 古装, 武位, 言情, 科幻	1545
bookbk.im	40	书城, 小说, 出版, 图书, 网络, 文学, 原创, 小说, 免费, 下载, 在线, 阅读, 电子书, 经管, 励志, 名人, 传	1598
www.biquy	40	值得, 收藏, 网络, 小说, 阅读, 网络, 小说, 小说, 阅读, 小说, 广大, 值得, 收藏, 网络, 小说, 阅读, 网络	1814
adbehavio	40	有声, 小说, 小说, 有声, 在线, 书电台, 喜马拉雅, fm, 有声, 小说, 有声, 儿童, 故事, 相声, 鬼故事, 电	1837
config.k.sc	40	搜狗, 阅读, 海量, 小说, 图书, 免费, 小说, 漫画, 免费, 畅读, 小说, 免费, 小说, 热门, 小说, 言情小说,	1983

图 8　部分聚类结果展示

图 8 中 host 为用户访问的网页 URL，predict 为对数据进行聚类后的类别标记数字，filtered-Words 为对用户访问的网页内容进行过滤后的特征词语，使用 PK-means 算法进行聚类后，准确地将小说阅读类网页聚到了一个类别中。

表 2，表 3 分别为根据用户浏览喜好和个人属性对用户的画像，通过对敏感数据进行加密处理，保障了用户个人信息的安全性。

表 2　用户基本属性

属性标签	描述
姓名	杨＊＊
性别	男
年龄	27
手机号	15BE19C9DB569387A
城市	武汉
活跃时段	20:00—23:00

表 3　用户偏好

聚类标签	描述	聚类标签	描述
1	腾讯网	26	网络安全
14	共享 WIFI	29	Oppo 手机
15	搜索引擎	35	在线视频
18	知识社区	40	小说阅读
21	手机浏览器	49	电脑系统应用
25	手机主题		

表 4 是在电信运营商的数据集上分别运行的 PK-means 算法和 K-means 算法，该数据集拥有 4878 个样本，从结果可以看出，由于 pca 算法通过降维处理突出了主要特征，同时有效避免了过拟合现象的发生，使得算法准确率均有明显提高。同时，残差平方和也有了一定程度上的减小。

表 4 两种算法在电信数据上的对比结果

算法种类及优化指标	PK-means	K-means
准确率/%	86.80	82.35
误差平方和	1114.19	1175.93

3 结论

本文基于 Spark 聚类的用户画像已经能够完整显示不同用户的基本信息、个人偏好。根据不同用户的个性化需求制定独特的用户服务和营销策略，从而达到精准营销的目的。同时，由于此项目是基于 Spark 内存计算框架，与传统的数据处理模型相比，运算速度有明显提高。使用 PCA 算法对初始数据进行降维处理，也在一定程度上节省了运算开销，同时能够突显数据中贡献较大的特征，与初始 K-means 算法相比，聚类结果明显改善，准确率有一定提高。本文通过对运营商数据的挖掘、分析、建模，提炼出数据中的潜在价值，在移动互联网快速发展的时代，系统具有较高的实用价值和借鉴意义。

[参 考 文 献]

[1] 黄文彬，徐山川，吴家辉，等. 移动用户画像构建研究[J]. 现代情报，2016，36(10)：54-61.

[2] 孟巍，吴雪霞，李静，等. 基于大数据技术的电力用户画像[J]. 电信科学，2017(S1)：15-20.

[3] 单晓红，张晓月，刘晓燕. 基于在线评论的用户画像研究——以携程酒店为例[J]. 情报理论与实践，2018，41(4)：99-104，149.

[4] Zaharia M，Chowdhury M，Franklin M J，et al. Spark：cluster computing with working sets[C]//Usenix Conference on Hot Topics in Cloud Computing，2010.

[5] 朱珠. 基于 Hadoop 的海量数据处理模型研究和应用[D]. 北京：北京邮电大学，2008.

[6] 李绍俊，杨海军，黄耀欢，等. 基于 NoSQL 数据库的空间大数据分布式存储策略[J]. 武汉大学学报(信息科学版)，2017，42(2)：163-169.

[7] Vora M N. Hadoop-HBase for large-scale data[C]// International Conference on Computer Science & Network Technology，2012.

[8] 程国建，赵倩倩. K-means 聚类算法在 Spark 平台上的应用[J]. 软件导刊，2016，15(2)：146-148.

[9] 张媛，张燕平. 一种 PCA 算法及其应用[J]. 计算机技术与发展，2005，15(2)：67-68.

[10] 吴晓婷，闫德勤. 数据降维方法分析与研究[J]. 计算机应用研究，2009，26(8)：2832-2835.

Research and Application of Telecom Portrait Based on Spark

HUA Man，SHAO Xiongkai，GAO Rong

(*School of Computer Science，Hubei Univ. of Tech.，Wuhan* 430068，*China*)

Abstract：Traditional telecom operators are faced up with the situation of failing to accurately understand users under the Internet environment，which leads to a decrease in the number of its users and is thus gradually being marginalized. Meanwhile，the TB-level log files generated by the stock users are stored for a period of time and then cleaned up as useless junk content. A user portrait method for Telecom operators is proposed in this paper. Through Spark-based big data technology，the data of telecom operators are mined，analyzed，dimension-reduced and optimized. Clustering is carried out to discover the personalized needs of different users，so as to achieve the purpose of providing precise recommendation and personalized services for users. The experimental results show that this method can better portray the behavior of different users. At the same time，the accuracy of the optimized PK-means clustering method is significantly improved. Compared with the traditional data processing mode，the computing speed has been greatly improved.

Keywords：user portrait；spark；precise recommendation；K-means；data dimensionality reduction

[责任编校：张岩芳]