基于协同过滤算法的老年人课程推荐系统研究

杨新月 杨开亮 * 郭抒菡

（深圳职业技术学院，广东 深圳 518055）

摘要：近年来，老年教育领域快速发展，推出了各类丰富的课程。但基于个性化的课程不能得到充分利用，老年人很难找到为自己量身打造的课程。供给质量仍严重滞后于老龄群体终身学习需求。为提升优质老年教育资源开放共享的水平，助力老年人享受快乐生活，本文以老年教育供给改革为切入口，结合大数据和数据挖掘算法，采用推荐系统，从现有老年人课程资源中挖掘优质课程，让优质课程使用率达到最大化，使老年人更方便的“智慧乐学”。

关键词：Hadoop；Spark；推荐算法；老年教育；应用与分析

近年来，老龄化人口比重持续增大，据深圳市第七次全国人口普查数据显示，深圳 60 岁及以上人口高达 94.07 万人，占到 5.36%，与第六次全国人口普查相比，比重提高 2.36 个百分点。汹涌的银发浪潮需要和谐的疏导，为丰富老年人的精神文化生活，让他们老有所学、老有所乐，亟须推进老年教育工作。老年群体日益增长的对优质、高层次老年教育需求与老年教育供给之间矛盾，已经严重制约老年教育发展。

老年教育机构虽然开展了各类丰富的课程，但基于个性化的课程不能得到充分利用，老年人很难找到为自己量身打造的课程。供给质量仍严重滞后于老龄群体终身学习需求。

为扩大优质老年教育资源开放共享，助力老年人享受智慧生活，本文以老年教育供给改革为切入口，结合大数据和数据挖掘算法，采用推荐系统，从现有老年人课程资源中挖掘优质课程，让优质课程使用率达到最大化，使老年人更方便的“智慧乐学”。

一、基于协同过滤算法的老年人课程推荐系统设计

本文通过大数据平台 Hadoop、Spark 等领域的技术，建立老年人生理、心理、行为特征海量数据分析平台，以数据为驱动，通过推荐算法，能为不同的老年学员推荐合适的课程，从而不断创新老年教育方法、理论、教育模式、课程内容，完善当前老年教育领域的“基础设施”，更好地推动老年教育的发展。

（一）理论基础

1.Hadoop

Hadoop 是一个开源的分布式系统基础架构，它提供了海量数据的处理能力。该框架主要有以下几个重要组成部分，分别为 HDFS、MapReduce、HBase、Hive 等组件。

HDFS 适合部署在廉价的机器上，且有着超大数据集（large data set）的应用程序。MapReduce 是一个编程模型，可以并行处理数据，即使不熟悉分布式编程机制，也能在 Hadoop 平台上运行，它能自动地将计算任务分配给集群内的服务器里执行 。

2.Spark

Apache Spark 是一种大数据分析引擎，它可以高效的处理 Hadoop 上的分布式数据 。Spark 默认情况下处理的过程数据存放到内存中，后续的运行作业利用这些结果进一步计算。据 Spark 官网数据，内存中读取的数据的速度比 Hadoop MapReduce 快 100 多倍。Spark 以 Spark Core 为核心，还支持 Spark SQL、Spark Streaming、Spark MLlib 和图计算 Spark GraphX 等。

MLlib 是 Spark 中提供机器学习函数的库，MLlib 由一些通用的学习算法和工具组成，包括分类、回归、聚类、协同过滤、降维等，同时还包括底层的优化原语和高层的管道 API。Spark MLlib 实现了交替最小二乘法（ALS）来学习这些隐性语义因子。

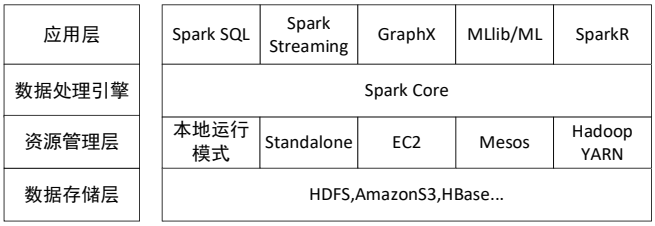


图 1：Spark 结构图

3. 推荐算法

协同过滤是一种应用非常广泛的推荐算法，主要功能是预测和推荐，协同过滤推荐算法主要分为两类，基于用户的协同过滤算法（user-based collaborative filtering，User CF）和基于项目的协同过滤算法（item-based collaborative filtering，Item CF）。

协同过滤算法的处理过程如下：

步骤 1：根据用户评分数据，建立用户 – 物品评分矩阵。

步骤 2：计算目标用户和其余用户之间的相似性，根据相似性找到最相似的 K 个用户作为目标用户的相似邻居。

步骤 3：根据用户相似邻居对目标用户未评分物品的评分信息，对目标用户未评分物品进行评分预测，选取预测评分最高的前 N 项作为目标用户的推荐列表，推荐给目标用户。

常用的相似度计算方法：欧几里德距离（Euclidean Distance）、皮尔逊相关系数（Pearson Correlation Coefficient）、Tanimoto 系数（Tanimoto Coefficient）。

（二）模型设计

传统的协同过滤算法在推荐时遇到的突出问题是用户冷启动和用户兴趣随时间变化对推荐结果的影响，从而降低了推荐的准确性。为减少以上问题带来的推荐结果不准确，本文结合实际使用场景，采用了用户特征属性，通过计算用户属性之间的相似度，再结合传统的算法对相似度进一步改善。

用大数据全方位“打理”老人教育，本课题以老年人生理、心理、行为特征大数据为基础，采用基于 Spark 的快速处理能力的协同过滤算法，为老年学员推荐个性化的课程，并为决策者提供适合老年人的教育模式。

1. 建立基于老年人生理、心理、行为特征的用户矩阵

老年人群体有特别的群体特征，可以通过他们的生理、心理、行为特征等数据，挖掘对老年人教育有价值的数据。将老年人生理、心理、行为特征等大量数据存储到数据仓库，建立基于 HDFS 的用户矩阵数据，各数据仓库如下：

老年人生理理特征数据仓库：通过挖掘医院、社康、养老院等机构的就医、用药等数据，判断出老年人的生理健康情况。

老年人心理理特征数据仓库：为构建全面客观的老年人心理健康测量指标，选用感到孤独、认知能力、生活满意度、参与活动等指标去采集老年人的心理健康情况数据。

老年人行为特征数据仓库：老年人生活行为轨迹数据和每日活动情况调查问卷是本研究中的主要内容，同时借助可穿戴设备采集行为数据。在可穿戴设备中设置惯性传感器，通过惯性传感器监测用户的行为数据；根据所述行为数据以及预设策略，确定用户的行为。

表 1 用户特征矩阵

学员 ID	身体健康程度	锻 炼强度	学 历层次	心理健康程度	……	……
201301	6	4	3	3		
201301	7	6	2	4		
201301	2	2	3	5		
201301	9	10	4	9		
……	……	……	……	……	……	……

表 1 中每行的数据是一个学员身心健康状况、兴趣爱好等情况，学员属性特征的相似度越高，推荐的成功率就会越高。

2. 利用 Spark ALS 构建学员 – 课程相似度矩阵。

要计算学员 – 课程相似度矩阵，需对评分矩阵进行降维。ALS 算法可以方便地处理基于矩阵分解的推荐系统数据，在用于隐性数据矩阵分解时，稀疏的输入数据，简单的线性代数运算求最优解，以及数据本身的可并行化，从而使 ALS 算法在大规模数据处理方面效率明显提高 。

3. 融合用户特征的推荐列表

根据用户特征相似度矩阵，找到生理、心理、行为特征相似的学员，结合基于项目的课程推荐列表，最后生成课程推进列表。

处理过程如下：

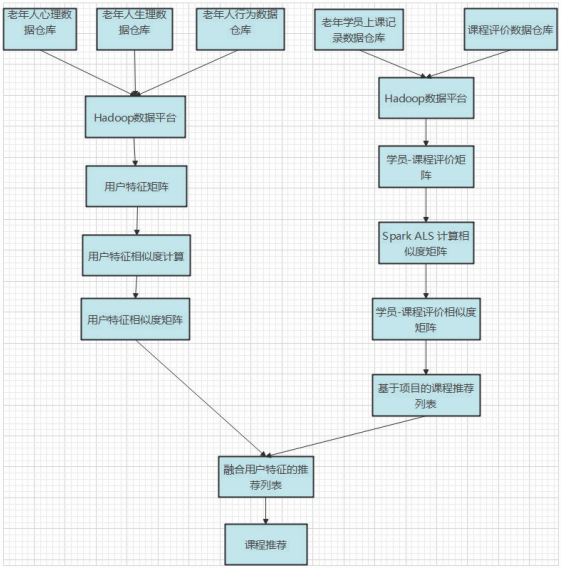


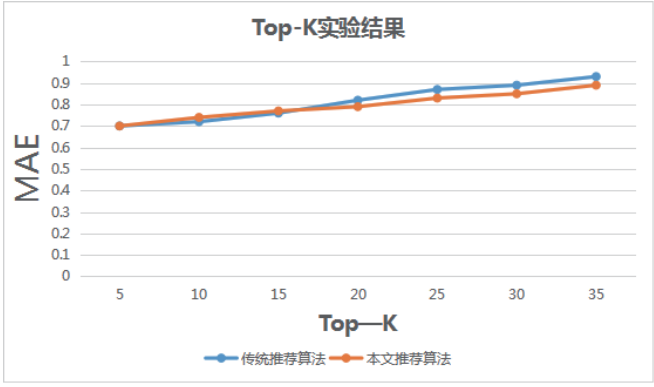
图 2：老年人课程推荐系统处理过程

（三）算法的应用与分析

1. 本文实验数据采用某老年大学的学员档案数据，该数据中包括了 6964 名老年学员的用户基本特征数据，涵盖了身体健康程度、锻炼强度、学历层次、心理健康程度以及 93 个课程。

本文实验选择平均绝对误差（mean absolute error，MAE）作为算法评价标准。MAE 是通过计算目标用户的预测评分值与实际评分值之间的绝对偏差来衡量算法的好坏，MAE 的值越小，即预测评分值与实际评分越接近，则推荐的质量越好。设预测的评分集合为 {p1，p2，…，pn }，对应的用户实际评分集合为 {q1，q2，…，q n }，则平均绝对误差 MAE 为：

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$



二、结论

文章以 Hadoop 平台的强大数据存储及计算能力为依托，通过使用 Hive 建立数据仓库和利用快速通用的计算引擎 Spark，提高了相似度计算的效率。在老年人课程推荐使用场景中为提升传统推荐算法的准确性，通过老年人生理、心理和行为特征的相似度计算，明显提升了推荐效率和准确性。实验结果显示，在用户特征非常多时，不同特征对算法的影响较大，有待优化不同特征对推荐结果的影响。

参考文献：

[1] 滕彩峰 .Hadoop 中的资源调度算法研究及应用 [D]. 成都信息工程大学，2019.

[2] 王建辉 . 基于 Hive 的日志分析系统的实现与优化 [D]. 南京邮电大学，2017.

[3] 刘永增，张晓景，李先毅 . 基于 Hadoop/Hive 的 web 日志分析系统的设计 [D]. 广西大学学报（自然科学版），2011.

[4] 胡德敏，龚燕 . 基于 Spark 的混合推荐算法研究 [J]. 计算机应用研究，2017（12）.

[5] 李俊丽 . 基于 Spark 的倾斜数据虚拟划分算法 [J]. 计算机工程与设计，2021（08）.

[6] 项如 . 基于 Spark 的分布式机器学习平台研究与实现 [D]. 南京大学，2017.

[7] 孟祥武，纪威宇，张玉洁 . 大数据环境下的推荐系统 [J]. 北京邮电大学学报，2015（2）.

[8] 蒋研 . 基于协同过滤的个性化混合推荐算法及模型研究 [D]. 南京邮电大学，2020.

[9]. 黄立威，江碧涛，吕守业，刘艳博，李德毅 . 基于深度学习的推荐系统研究综述 [J] 计算机学报，2018（41）.

[10] 车晋强，谢红薇 . 基于 Spark 的分层协同过滤推荐算法 [J]. 电子技术应用，2015（09）.

课题来源：2021 年深圳职业技术学院老年教育研究课题。