基于 Lambda 架构的医学图书推荐系统设计与实现*

邱煜炎 吴福生

(蚌埠医学院图文信息中心 蚌埠 233000)

〔摘要〕 设计并实现基于 Lambda 架构的医学图书推荐系统，阐述系统架构与关键技术，采用 ALS 算法建立离线计算推荐模型，结合读者在线行为进行图书实时推荐。实践证明基于 Lambda 架构的推荐系统准确率较高。

〔关键词〕 Lambda 架构；推荐系统；大数据

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2019.03.016

Design and Implementation of Medical Book Recommendation System Based on Lambda Architecture *QIU Yuyan, WU Fusheng, Library and Information Center of Bengbu Medical College, Bengbu 233000, China*

〔**Abstract**〕 The medical book recommendation system on the basis of Lambda architecture has been designed and implemented. The paper elaborates on the structure and key techniques of the system. An offline computation-based recommendation model has been established by adoption of ALS algorithm to recommend book in real time according to readers´online behavior. Practice shows that the recommendation system based on Lambda architecture is of higher accuracy.

〔**Keywords**〕 Lambda architecture; recommendation system; big data

1 引言

大数据环境下医学文献数字化迅速推进，文献总量急剧增长[1]。为满足读者对文献高效查找、精准推荐和快速响应的需求，专业化推荐系统应运而生。推荐系统[2]帮助用户评估其所有未看过的产品，通过分析用户的基本信息、兴趣爱好和历史行为主动推荐符合喜好的项目。目前推荐系统已在电子商务、电影、音乐网站领域取得显著成绩。

〔收稿日期〕 2018-12-07

〔作者简介〕 邱煜炎，硕士，讲师，发表论文4 篇。

〔基金项目〕 蚌埠医学院自然科学基金面上项目“基于读者行为的医学专业馆藏资源推荐算法比较研究”(项目编号：2017BYKY1762)。

2 相关研究情况

传统单机环境下的推荐系统无法满足大数据规模资源的存储与计算需求，Hadoop 平台能够处理海量数据。对于推荐内容的计算，大量的学者将推荐系统和 Hadoop 进行集成，肖强[3]等改进传统的协同过滤算法，使之适应 Hadoop 平台上的分布式计算；李文海[4]等基于 MapReduce 模型实现关联规则算法，构建分布式电子商务系统；奉国和[5]等采用 Hadoop 平台以及 Mahout 引擎技术改进协同过滤算法，提高推荐系统的准确率。Hadoop 平台解决海量数据计算的问题，但其还存在诸多缺陷，最主要的是 MapReduce[6]计算模型延迟过高，无法满足实时、快速计算的需求，因而只适用于离线批处理的

应用场景。Spark[7]在设计上充分吸收借鉴 MapReduce 的精髓并加以改进，同时采用先进的 DAG 执行引擎，以支持循环数据流与内存计算，因此在性能上比 MapReduce 有大幅度提升，从而迅速获得学术界的广泛关注。何胜[8]等提出一种以文献“混合关联”为主要内容的图书馆文献推荐方案及实现算法，基于 Spark 技术开展实证研究，优化图书馆文献推荐效果和提高统计计算性能。Lambda[9]架构由 Storm 项目发起人 Nathan Marz 提出，集成 Hadoop、Kafka、Storm、Spark、HBase、Redias 等各类大数据[10]组件，提供混合平台。Lambda 架构，见图 1。其具有高容错、低延时和可扩展的特点。本研究利用 Lambda 架构技术特点，融合历史数据离线计算、分布式日志采集等技术构建推荐数据及时反馈的医学图书推荐系统。实验结果表明该系统具有高可靠性和稳定性，能够满足大数据下低成本、快速响应和精准推荐的需求。

3 系统架构设计

3.1 架构

大数据环境下图书馆网站存在大量的读者隐式行为（如点击、搜索、浏览记录等），不同服务的应用接口对应不同服务器，因此系统日志文件分散存放在各个服务器上。传统 Hadoop 平台无法有效汇总隐式行为日志并做到及时响应。本研究基于 Lambda 架构充分收集读者隐式行为数据，结合显式数据（如读者借阅、预约记录）构建混合模型矩阵以解决数据稀疏问题，离线计算采用 ALS 推荐算法，在传统离线推荐模型基础上实时分析用户行为，以反馈最适合当前用户的推荐列表。基于 Lambda 架构的医学图书推荐系统架构，见图 2。

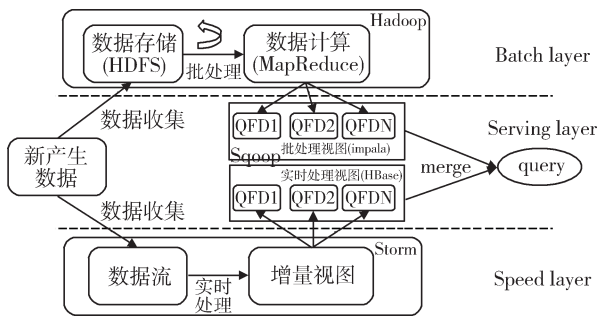

图 1 Lambda 架构

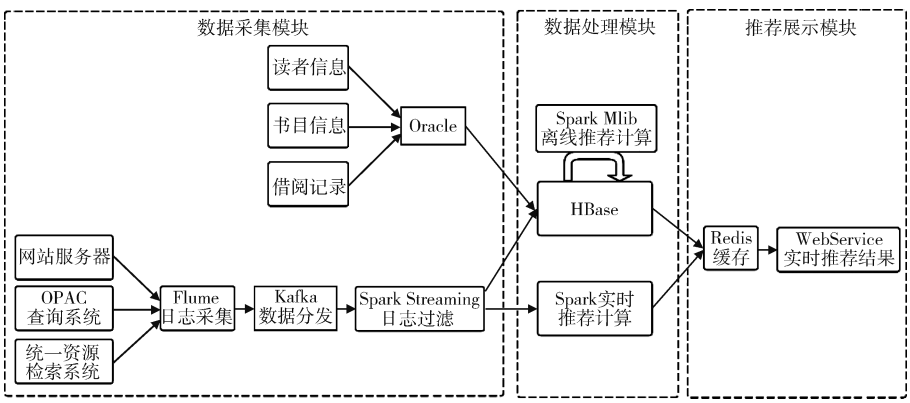

图 2 基于 Lambda 架构的医学专业文献推荐系统架构

3.2 模块

Lambda 架构分为 3 大模块：数据采集、数据处理和推荐展示。数据采集包括两部分：关系型数据库和日志采集系统。关系型数据库通过图书馆管理软件记录读者和书目的基本信息以及读者借阅记录信息。日志采集系统利用 Flume，通过 Kafka 集群的消息分发中间件实现日志数据的统一下发。数据处理模块分为离线处理和在线处理两部分。推荐展示模块将所有用户推荐列表写入 Redis 缓存系统中，缓解图书馆网站系统压力。在离线处理前，将关系型数据库数据通过 Sqoop 工具加载到 HBase 数据库中进行离线推荐模型训练。本设计首先利用用户基本信息，如专业系部和书目信息，还用 Spark 计算

模型进行主题提取，构建基于内容的用户与书目相似度矩阵，融合 Spark Mlib 机器学习库提供的 ALS 算法库[11]，建立基于文献资源内容过滤及相似用户协同过滤的离线推荐模型。

4 关键技术

4.1 分布式数据收集

推荐系统需要将各种数据收集到一个中央化的存储系统中，有利于进行集中式的数据处理、统计分析与数据共享。而用户行为是多样化的，包括基本属性数据、访问日志、搜索记录、收藏日志、文献信息等。其收集难点在于数据分散在各个离散设备上，保存于各种传统的存储设备与数据库系统中。针对传统的关系型数据库信息，可以利用 Hadoop 生态系统中 Sqoop 组件，将数据导入到分布式数据库 HBase 中，实现传统数据库与 Hadoop 同步。而针对基于浏览器访问的日志数据，可以借助分布式日志采集框架 Flume 进行数据收集。Flume 系统植入在应用网关处的日志监控可以实时监控日志文件变化，根据偏移量，读取来自联机公共目录查询系统（Online Public Access Catalogue，OPAC）以及统一资源检索系统的最新日志信息，然后将日志记录输出到 HBase 数据库中。

4.2 离线计算推荐模型

利用读者行为分析其对每本图书的兴趣程度，通过 Lambda 平台采集数据，涉及行为包括搜索、点击、预约、正常借阅、盲目借阅、续借、超期借阅。笔者参考相关文献，结合医学专业特点，构建读者图书兴趣模型。针对读者各种行为做如下评分：在关键词搜索列表中，如果对某一图书项目进行点击则表示读者对该图书兴趣为正分，加 1 分。如果点击后并进行在线预约则表示更强的兴趣度，加 2 分。研究发现[12-13]读者对某种图书兴趣度与借阅时间有关，见图 3。借阅时间低于一定阈值（小于 3 天，α 值为 2 天）则表示读者对该图书实质上并不感兴趣，借阅该图书的行为可能属于盲目借阅，减 2 分。读者在一定期限（3 ~ 29 天，β 值为 30 天）内正常借阅，则表示读者兴趣为正分，加 2 分。读者续借某本书则表示读者借阅兴趣达到饱和，加 2 分。期借阅虽然达到时间饱和度，但并不代表读者很感兴趣，超期借阅时间与兴趣也成反向趋势，超期时间在 1 个借阅周期内（β 值为 30 天）为短期超期，减 1 分，此后超期两个 β 值后为长期超期，减 2 分。此外整时间到期归还的借阅历史记录（1 个 β 天数）可以假设读者对图书的兴趣成相反方向，减 1 分。用户行为评分影响，见表 1。

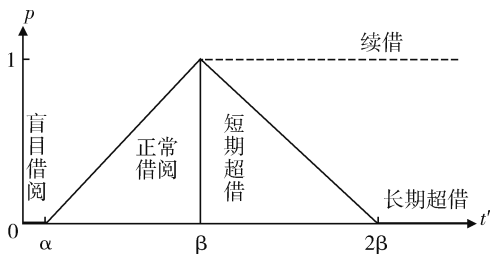

图 3 读者兴趣度与借阅时间关系

表 1 用户行为评分影响

用户行为	加减分值
点击	+1
预约	+2
正常借阅	+2
续借	+2
盲目借阅	-2
短期超期	-1
长期超期	-2
到期归还	-1

初始化评分值为 5 分，根据表 1，构建读者图书评分矩阵，得到读者 - 图书 - 评分 3 元组作为离线推荐模型数据源，利用 Spark Mlib 机器学习库提供的交替最小二乘 ALS[14] 推荐算法进行建模。交替最小二乘 ALS 是求解隐语义模型隐因子参数的优化算法，隐语义模型是奇异值分解方法的一种，具有较好的理论基础，优化一个设定的指标建立最优模型，其核心思想是通过隐含特征联系用户兴趣和物品，利用降维的方法解决评分矩阵稀疏的问题。对于 R（$m \times n$）的矩阵，ALS 旨在找到两个低维矩阵 X（$m \times k$）和矩阵 Y（$n \times k$），来近似逼近 R（$m \times n$），即：

$$R_{m\times n} \approx X_{m\times k} Y^{T}_{n\times k}$$

其中 R（$m \times n$）代表用户对商品的评分矩阵，X（$m \times k$）代表用户对隐含特征的偏好矩阵，Y（$n \times k$）表示物品所包含隐含特征的矩阵，T 表示矩阵 Y 的转置。对于矩阵 X、Y 的计算采用最优化目标损失函数，目标损失函数用均方根误差（Root Mean Square Error，RMSE）定义，如下所示：

$$L(X,Y) = \sum_{u,i} (r_{ui} - x_u^T Y_i)^2 + \lambda(|x_u|^2 + |y_i|^2)$$

上式中的 $\lambda \| x_u \|^2 + \lambda \| y_i \|^2$ 是用来防止过拟合的正则化项，通过交替最小二乘算法优化目标损失函数，算法如下：初始化随机矩阵 Q 中的元素值；将 Q 矩阵当做已知的，直接用线性代数方法求得矩阵 P；得到矩阵 P 后，将 P 当做已知参数，再返回求解矩阵 Q；上述两个过程交替进行，一直到误差收敛到可以接受为止。研究发现[15]在不是很稀疏的数据集合上，交替最小二乘通常比随机梯度下降要更快的得到结果。设置 ALS 迭代次数以及相关参数，ALS 算法会对读者 - 图书评分矩阵进行分解，利用隐语义进行表达，计算出隐式因子，填补读者与书目的预测评分，然后训练离线推荐模型。

4.3 实时推荐模型

首先利用 Spark Streaming 技术将 Kafka 集群推送的日志信息过滤出日志点击流，从中抽取读者产生行为对应的图书 ID 和用户 ID。然后根据离线推荐模型进行图书相似度排序，与离线模型进行混合处理，重排序，使得图书馆在线网站可以感知到用户最新行为，提升推荐系统的准确率。笔者随机抽取临床医学专业读者登录图书查询系统，以“麻醉学”作为搜索关键词，显示结果界面左边是 OPAC 系统根据检索词通过图书管理系统展示的麻醉学图书，右边是实时推荐结果，根据系统设定，共计显示 10 条推荐结果，其中前 5 条显示与麻醉学相似的图书，如“麻醉意外”，“麻醉并发症”；后 5 条是根据该读者历史行为显示离线推荐结果，如“临床诊疗学”，“临床医学多用辞典”等。

5 实验分析与结果

5.1 数据来源

以蚌埠医学院 2014 年 6 月 - 2017 年 12 月之间所有医学类（中图法 R 类）图书借阅记录作为实验数据，包含图书 86 022 本，读者 12 030 人，1 806 212次借阅服务，其中借阅 564 399 次、续借 149 507 次、预约 102 505 次。

5.2 实验环境

基于 Lambda 架构的医学图书推荐系统搭建 6 台 Linux 服务器，版本 CentOS6.5；每台服务器配置 8 核 CPU，16GB 内存和 1TB 硬盘。其中 3 台服务器用来搭建 Lambda 平台，另外 3 台服务器分别用来进行数据采集、数据缓存以及前端展示。软件配置，见表 2。

表 2 Lambda 架构软件配置

软件名	功能	版本
Java	Hadoop 支持软件	1.8
Scala	编写 Spark 程序	2.12.6
Hadoop	大数据框架	2.6.5
Spark	分布式内存计算软件	2.3.1
Hbase	分布式数据库软件	2.0.1
Flume	分布式数据采集软件	1.7.0
Sqoop	Oracle 与 Hbase 数据转换	1.3.5
Kafka	分布式日志数据传输	2.1.0
Redis	数据缓存	3.0.7

5.3 结果

为验证兴趣度模型以及推荐系统的有效性，邀请 10 位不同专业的读者进行评价。利用推荐系统结合读者在线行为，为每人推荐 20 本书，对推荐结果进行打分。采用信息检索领域广泛使用的查准率来评价实验效果。评估结果，见表 3。结果显示 10 位读者评价查准率差别很大，均值为 44.4%。虽然系统初始设置为每位读者推荐 20 本专业图书，但由于读者专业、借阅量、借阅行为等不同，实际推荐的数量也不同。通过表 3 可以看出借阅数量对推荐效果影响很大。但是随着读者借阅图书的数量不断增加，推荐效果也越来越好。值得注意的是读者专业对推荐结果影响较大，如临床医学本科生对结果评价明显高于其他专业，说明临床专业学生数量多，借阅行为数据量与推荐准确度成正相关关

系。此外研究生由于借阅量相对本科生大，借阅目的性相对较强，准确率普遍较高。

表3 推荐系统评估结果

读者 ID	读者专业	读者年级	借阅数量（册）	推荐数量（册）	成功推荐数量(册)	查准率（%）
U1	临床医学	3 年级	32	18	11	61
U2	临床医学	2 年级	26	14	9	64
U3	口腔医学	2 年级	21	15	6	40
U4	预防医学	1 年级	13	18	5	28
U5	麻醉学	4 年级	52	17	11	65
U6	影像学	3 年级	25	15	7	47
U7	护理学	2 年级	19	16	3	19
U8	内科学	研究生	52	20	12	60
U9	心血管	研究生	47	19	11	58
U10	护理	研究生	42	20	13	65
平均查准率（%）	–	–	–	–	–	44.4

6 结语

本研究设计和实现大数据环境下基于 Lambda 架构的医学图书推荐系统，提出基于读者行为的评分模型，将隐式行为数据应用到评分模型中，优化模型结构。实验证明在充分采集用户行为数据后，推荐系统的准确率和召回率有明显提升。鉴于 Lambda 架构获取的数据具有数据量大、实时性强、多样性的优势，下一步将引入书本信息及用户基本信息的特征，设计多模型数据处理方式，进一步提升推荐效果。

参考文献

1 韩翠峰. 大数据带给图书馆的影响与挑战［J］. 图书与情报, 2012, 32（5）: 37 – 40.

2 李树青. 个性化信息检索技术综述［J］. 情报理论与实践, 2009, 32（5）: 107 – 113.

3 肖强, 朱庆华, 郑华, 等. Hadoop 环境下的分布式协同过滤算法设计与实现［J］. 现代图书情报技术, 2013, 29（1）: 83 – 89.

4 李文海, 许舒人. 基于 Hadoop 的电子商务推荐系统的设计与实现［J］. 计算机工程与设计, 2014, 35（1）: 130 – 136, 143.

5 奉国和, 黄家兴. 基于 Hadoop 与 Mahout 的协同过滤图书推荐研究［J］. 图书情报工作, 2013, 57（18）: 116 – 121.

6 Dean J A G S. MapReduce: simplified data processing on large clusters［J］. Communications of the ACM, 2008, 51（1）: 107 – 112.

7 Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin. Spark: cluster computing with working sets［C］. Boston: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 2010.

8 何胜, 熊太纯, 柳易君, 等. 基于 Spark 的高校图书馆文献推荐方案及实证研究［J］. 图书情报工作, 2017, 61（23）: 129 – 137.

9 Wikipedia. Lambda Architecture［EB/OL］.［2018 – 08 – 13］. https: //en. wikipedia. org/wiki/Lambda_ architecture.

10 林子雨. 大数据技术原理与应用（第 2 版）［M］. 北京: 人民邮电出版社, 2017: 286.

11 王全民, 苗雨, 何明, 等. 基于矩阵分解的协同过滤算法的并行化研究［J］. 计算机技术与发展, 2015, 25（2）: 55 – 59.

12 景民昌, 于迎辉. 基于借阅时间评分的协同图书推荐模型与应用［J］. 图书情报工作, 2012, 56（3）: 117 – 120.

13 江周峰, 鄂海红, 杨俊. 基于时间上下文信息的借阅次数评分模型与应用［J］. 图书情报工作, 2014, 58（S2）: 220 – 223.

14 李改, 李磊. 基于矩阵分解的协同过滤算法［J］. 计算机工程与应用, 2011, 47（30）: 4 – 7.

15 Wang J. A Novel Data Distortion Approach via Selective SS-VD for Privacy Protection［J］. International Journal of Information and Computer Security, 2008, 2（1）: 48 – 70.