

医学院校图书馆 OPAC 图书智能推荐系统分析与设计

吕文娟 龚佳剑

(首都医科大学 北京 100069)

〔摘要〕 介绍目前各高校 OPAC 推荐模块, 基于医学院校图书馆的特点, 结合目前流行的推荐技术, 提出利用 Hadoop 平台的 Mahout 算法设计图书智能推荐系统。通过将智能推荐系统嵌入 OPAC, 解决个性化推荐问题, 为读者提供主动智能化的个性服务。

〔关键词〕 联机公共目录查询系统; 智能推荐; 医学院校; Mahout

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2015.09.017

Analysis and Design of the OPAC Book Intelligent Recommendation System for Libraries in Medical Colleges and Universities

LV Wen-juan, GONG Jia-jian, Capital Medical University, Beijing 100069, China

〔Abstract〕 The paper introduces existing OPAC recommendation modules in colleges and universities. Based on characteristics of libraries in medical colleges and universities and in combination with prevailing recommendation technology, it proposes to utilize the algorithm of Mahout of Hadoop Platform to design the library intelligent recommendation system. By embedding the intelligent recommendation system into OPAC, it solves the problem of personalized recommendation and provides readers with active and intelligent personalized services.

〔Keywords〕 OPAC; Intelligent recommendation; Medical college and university; Mahout

1 引言

联机公共目录查询系统 (Online Public Access Catalogue, OPAC), 是图书馆信息管理系统不可分割的一个重要组成部分, Web 2.0 的发展使得 OPAC 系统的读者参与性大大增强。读者除了进行常规的书目检索、预约、续借之外, 还能够方便、快捷地主动参与 OPAC 系统提供的图书评价、图书

荐购、加注标签、建立个人书架等个性化服务。读者贡献的各种数据隐含了对馆藏资源的评价和需求, 蕴藏了丰富、未知、有用的知识, 对图书馆的个性化主动服务、学科文献资源建设决策、优化业务管理等非常有价值。

医学院校图书馆必须服务于学校的学科建设, 以首都医科大学为例, 学校有 8 个国家重点学科, 临床医学、基础医学、口腔医学、预防医学、药学、中医学等 16 个专业, 因此图书馆的馆藏资源涵盖了现代生物医学和生命科学的相关领域。同时医学知识的快速发展, 使得图书馆不断调整馆藏建设, 满足读者对前沿医学信息获取的需求^[1]。对于医学图书馆的读者来讲, 他们有很大一部分正在或

〔修回日期〕 2015-06-11

〔作者简介〕 吕文娟, 馆员, 发表论文 2 篇; 龚佳剑, 副研究馆员, 副馆长, 发表论文 6 篇。

者将要从事医务工作，对文献资源的需求明显带有学科专业性，迫切希望针对其知识需求、帮助其解决具体问题的信息和知识服务。因此，前瞻性、实用型和整合型的医学信息成为热门。OPAC 信息库中保存着大量的读者检索及借阅信息，这些信息能充分揭示读者对馆藏资源的利用和需求^[2]，利用这些信息并结合相关推荐算法向读者推荐相关图书，可以更好地满足读者的信息需求。而推荐系统可以挖掘用户潜在需求，帮助用户发现对自己有价值的信息，同时将信息展现在可能对其感兴趣的用户面前。对图书馆而言，推荐系统还是践行个性化服务的一个新手段。

2 OPAC 现状

Web 2.0 的概念提出以后，各图书馆集成系统都推出了新版本的 OPAC。新型 OPAC 从形式到内容都极大地改进了传统 OPAC 检索为主的单一功能，图书封面、作者简介、内容简介等信息丰富了 OPAC 的内容；无缝的电子资源、外部资源链接等扩大了读者获取信息的渠道；各种数据挖掘分析技术提供了有用的参考信息，这些都提升了 OPAC 的实用性。但笔者对国内 10 所 985 高校图书馆的读者调查显示，目前很多读者在 OPAC 上查询馆藏图书时，存在手工检索结果不理想现象，出现信息过载、信息迷失等情况，导致读者丧失明确的借阅目标，往往仅凭主观感觉选择图书进行借阅。对于读者关心的问题“我需要的图书有没有”、“我需要的图书在哪里”、“我想借的图书好不好”，现有的基于关键词匹配检索和类似目录浏览的结果基本无法解决。目前，各高校图书馆 OPAC 推荐服务主要形式有新书通报、书目推荐、借阅排行等，见表 1。由表 1 可以看出，高校图书馆关于图书推荐的应用十分欠缺；并且在国内，还没有代表性的针对医学院校读者特点的推荐系统。

续表 1

清华大学	借阅排行（只更新到 2011 年），新书通告
浙江大学	新书通报，借阅排行，10 大热评，馆员推荐（列表为空）
复旦大学	新书介绍
上海交通大学	新书通报
中国科学技术大学	热门借阅，热门评分，热门收藏，热门图书
南京大学	Book + （包括按学科分类的热门借阅和新书推荐以及豆瓣热评）
华中科技大学	总借阅排行，文学类/社科类/自科类/外文图书借阅排行，百名学生读者借阅排行
四川大学	新书通报，借阅排行，10 大热评，馆员推荐（列表为空）
中山大学	新书通报，借阅排行，10 大热评（列表为空），馆员推荐

3 推荐技术选择

3.1 协同过滤推荐

协同过滤推荐在信息过滤和信息系统中正迅速成为一项很受欢迎的技术。与传统的基于内容过滤直接分析内容进行推荐不同，协同过滤分析用户兴趣，在用户群中找到指定用户的相似（兴趣）用户，综合这些相似用户对某一信息的评价，形成系统对该指定用户对此信息的喜好程度预测^[3]。其最大的特点是通过对具有类似行为或爱好的其他用户进行分析，预测出该用户的兴趣爱好，强调的是人与人之间的协作。优点是对推荐对象没有特殊的要求^[4]；缺点是通过寻找相近用户来产生推荐集，在数量较大的情况下，推荐的可信度随之降低。

3.2 Hadoop 平台

Hadoop 原本来自于谷歌一款名为 MapReduce 的编程模型包，后迅速发展成为分析大数据的领先平台^[5]。Hadoop 是一个能让用户轻松架构和使用的分布式计算平台，具有高扩展性、高效性、低成本等特点。很多开发者将协同过滤算法部署于 Hadoop 平台以用于改善传统协同过滤算法在处理大规模海量数据时的效率瓶颈问题，提高算法执行效率。

3.3 Mahout 算法

Mahout 是 Apache Software Foundation^[6] 旗下的

表 1 国内 10 所 985 高校图书馆 OPAC 推荐模块

大学	推荐模块
北京大学	最新上架，教授推荐，学子推荐，馆员分享，借阅榜

顶级开源项目，运行在 Hadoop 平台下，主要包含机器学习、数据挖掘、个性化推荐等算法库，核心的算法是聚类、分类、推荐引擎、频繁项集的挖掘等^[7]。该推荐引擎开源根据用户对项目（书籍、电影、音乐等）的偏好和行为，计算出用户对未评分项目的预测分从而给出相关推荐。

4 推荐系统设计

4.1 系统框架

图书馆的 OPAC 系统一般采用 B/S 模式，系统在浏览器中获取用户的界面操作信息。通过用户界面可以一方面展示推荐信息，另一方面收集用户日志信息。收集到的用户日志信息通过日志系统写入数据库中，作为推荐系统的数据源。推荐系统整体架构，见图 1。

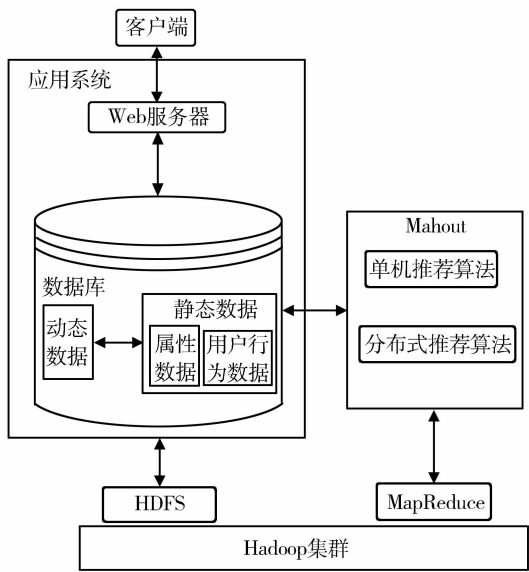


图 1 推荐系统整体架构

4.2 数据库设计

OPAC 系统采用的是 Oracle 数据库，存储读者、图书等固定属性数据、读者以往行为数据和读者检索行为产生的结果数据。这些是上层的图书推荐引擎产生推荐的数据基础。根据具体的情形，读者的偏好数据有许多来源，显性数据有读者的收藏记录、书评、推荐记录等，隐性数据有读者的浏览记

录、检索记录和借阅记录等。数据的重要程度不同，在推荐系统里占的权重就会不同。这些数据都会通过日志系统写入数据库中并与数据库中的推荐引擎结果数据进行整合、过滤后呈现给读者。数据库的主要表结构，见图 2。

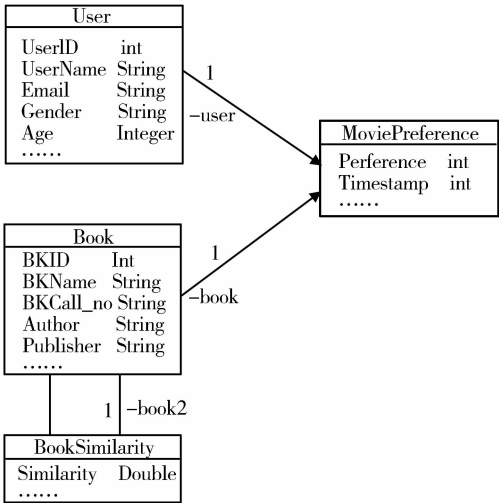


图 2 读者和图书信息数据模型

4.3 产生推荐层的设计

推荐系统最核心的技术是推荐算法的选择，基于同样的数据，采用不同的推荐算法，就会有不同的推荐结果。本文以 Mahout 算法为基础，在其上改进优化实现推荐算法，包括基于读者相似度的推荐算法、基于图书相似度的推荐算法和 SlopeOne 推荐算法。3 种推荐算法的比较，见表 2。

表 2 3 种推荐算法的比较

算法	参数	特点
基于读者相似度	读者相似度矩阵 近邻算法和数量	容易实现，适用于数据量少的情形
基于图书相似度	图书相似矩阵	适用于数据量少的情形
SlopeOne	差值数据	计算快，有大量预处理

基于读者相似度的推荐流程是：读者通过 OPAC 界面发送推荐请求给推荐系统，系统接到请求后，通过推荐引擎得到一个数据模型，通过这个数据模型计算出读者相似性和相似的读者群，然后据此产生推荐的图书 ID，OPAC 系统得到这些图书 ID 后，通过数据库表调取图书基本信息，呈现给读

者^[8]。基于图书相似度的推荐流程与之类似，接到 OPAC 的推荐请求后，推荐系统先获得数据模型，根据数据模型计算出图书间的相似性，产生推荐的图书 ID，然后推荐给读者。这两种推荐策略是最常见和最容易理解的策略，但是当数据量巨大时，它们的计算量也很大，致使推荐效率较差。SlopeOne 算法是对基于评分的协同过滤推荐算法的改进，主要思想^[9]如下：假设系统对于图书 A、图书 B、图书 C 的评分分别为 2、5、5。采用 SlopeOne 算法会得到以下规律：读者对图书 B 的评分 = 读者对图书 A 的评分 + 1；读者对图书 B 的评分 = 读者对图书 C 的评分。

基于以上的规律，可以对读者 A 和读者 B 的评分进行预测：对于读者 A，他给图书 A 的评分为 3，那么可以推测出他对图书 B 和图书 C 的评分均为 4；对于读者 B，他给图书 A 的评分为 4，给图书 C 的评分为 3，根据第 1 条规则可以推断出他给图书 B 的评分为 5 分，而根据第 2 条规则他给图书 B 的评分为 3。出现这种冲突时，根据制定的规则进行平均，所以给出的推断是 4。这就是 SlopeOne 推荐的基本原理^[10]，但是在一些特殊的情况下，尤其是一些新读者或者新上架的图书，就需要根据读者的专业和图书的分类号，进行定向推送^[11]。

5 推荐系统嵌入 OPAC

5.1 与借阅历史相关的书目

推荐系统会给出与用户目前的借阅偏好最相近的书目^[12]，“与您的借阅历史相关的书目”包括了封面、书名、著者、出版社和推荐指数，读者点击封面或书名可以对该书进行检索^[13]。对于一些读者由于借阅记录不足、无法产生推荐的，系统会根据读者的专业、年级推荐热门借阅书目^[14]，见图 3。

5.2 一起被借的图书

一起被借的图书是基于图书相似度算法实现的功能^[15]，出现在每本书的详细信息页的右侧，推荐系统给出当前该书被借阅的数量，点击书目的封面或书名可以直接对其进行检索，见图 4。



图3 与借阅历史相关的书目

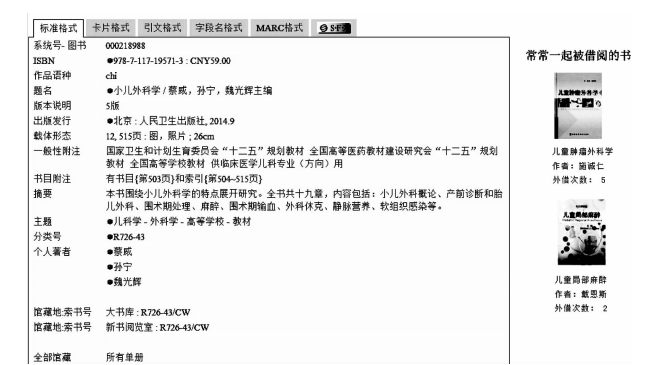


图4 通常一起被借阅的书

6 结语

本系统将数据挖掘技术与现有图书馆 OPAC 进行结合^[16]，使其具有智能推荐图书的功能，为读者提供个性化的检索推荐服务，大大提高了 OPAC 的用户体验。通过本文的研究，可以发现利用目前最流行的 Hadoop 平台的 Mahout 算法作为图书智能推荐系统的核心技术是非常可行的，为图书馆资源建设和决策提供数据支持，提高 OPAC 服务水平和质量^[17]。推荐系统中的读者反馈模块采集的数据暂时未投入到改进读者偏好中，正在探索读者的隐性反馈如何与用户的显性反馈（借阅、评分）相结合，从而更加精准地描述用户偏好。

参考文献

1 库睿. 医学院校图书馆特色馆藏建设的思考 [J]. 内蒙古科技与经济, 2014, 315 (17): 144 - 148.

- 2 张炜, 洪霞. 基于 OPAC 读者行为的知识发现研究 [J]. 图书馆论坛, 2011, 30 (1): 17-19, 49.
- 3 协同过滤简介及其主要优缺点 [EB/OL]. [2015-05-08]. http://zh.wikipedia.org/wiki/Slope_one
- 4 奉国和, 梁晓婷. 协同过滤推荐研究综述 [J]. 图书情报工作, 2011, 55 (16): 127-130.
- 5 Hadoop [EB/OL]. [2015-05-15]. <http://baike.baidu.com/view/908354.htm>.
- 6 Apache Software Foundation [EB/OL]. [2014-08-04]. <http://www.apache.org/>.
- 7 陈嘉恒. Hadoop 实战 [M]. 北京: 机械工业出版社, 2011: 292.
- 8 杨杰, 陈恩红. 个性化推荐系统应用及研究 [D]. 合肥: 中国科学技术大学, 2009.
- 9 Jiawei, H, Micheline K. 数据挖掘概念与技术 [M]. 范明, 孟小峰, 等, 译. 北京: 机械工业出版社, 2001.
- 10 项亮. 推荐系统实践 [M]. 北京: 邮电出版社, 2013.
- 11 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): 130-137.
- 12 黄岩. 移动互联时代数字图书馆发展要述 [J]. 医学信息学杂志, 2013, 34 (7): 72-77
- 13 陈晶. 网络环境下医院科研人员信息查询行为及图书馆个性化服务研究 [J]. 医学信息学杂志, 2013, 34 (8): 63-67.
- 14 陈进, 刘宝杰. 从未被借阅图书数据分析医科大学图书馆 LIB 2.0 的应用 [J]. 医学信息学杂志, 2009, 30 (9): 76-78.
- 15 李宁, 马路. 国内外高校图书馆电子资源服务策略研究 [J]. 医学信息学杂志, 2014, 35 (2): 71-74.
- 16 季汉珍, 练晓琪, 周建伟. 新型信息媒体技术应用于医院图书馆服务探索 [J]. 医学信息学杂志, 2010, 31 (10): 35-39.
- 17 冯研, 刘薇薇, 张兵兵, 等. 国内图书馆数据挖掘研究及应用的文献计量分析 [J]. 医学信息学杂志, 2011, 32 (6): 57-60.

2015 年《医学信息学杂志》编辑出版重点选题计划

2015 年本刊将继续以“学术性、前瞻性、实践性”为特色, 及时追踪并深入报道国内外医学信息学领域前沿热点, 反映学科研究动态、展示学科应用成果、引领学科发展方向。现对 2015 年度编辑出版重点选题策划如下:

一、医药卫生体制改革与医药卫生信息化

1 医药卫生信息规划与发展战略; 2 公共卫生、区域卫生、基层卫生信息化建设; 3 健康与社会保障数据服务; 4 居民健康卡示范工程和远程医疗系统建设; 5 药品供应与监管信息化、药品供应保障信息系统建设; 6 国外医药卫生信息化建设最新技术、成功经验。

二、医学信息技术

1 健康大数据研究与应用; 2 医药信息标准化建设及信息互联互通; 3 物联网、智慧医疗技术与实现; 4 移动医疗服务与健康管理; 5 各类医学信息系统信息互通与操作衔接; 6 医学机构知识库构建技术与方法。

三、医学信息研究

1 医学信息学科发展及研究内容的衍生、变化; 2 医学科技创新体系和发展战略; 3 医学科技监测与舆情监测; 4 医药卫生数据整合、信息共享研究与实践; 5 生物医学数据挖掘与利用、知识发现技术与实现; 6 竞争情报方法、策略在医药卫生领域的应用。

四、医学信息组织与利用

1 医学数字图书馆发展趋势与标准建设; 2 医学数字知识资源环境构建及其管理系统建设; 3 泛在化医学知识服务与决策咨询服务; 4 医学信息资源组织的关键技术与发展方向; 5 医学信息服务模式创新及其评估; 6 医学图书馆区域合作及资源共享模式研究。

五、医学信息教育

1 医学信息专科、本科、研究生教育及继续教育体制改革与模式创新; 2 医学信息素养的培养与教育; 3 医学信息职业岗位教育与培训; 4 国外医学信息学教育的先进经验及其借鉴。

(《医学信息学杂志》编辑部)