基于 Hadoop 的个性化书籍推荐系统的设计与实现

软件工程二班　13270212　毛烨辉　指导教师：张秋余　研究员

摘要

基于 Hadoop 平台的个性化书籍推荐系统是一个能够根据用户兴趣爱好并向用户精准推荐书籍的人机交互系统。该系统使用了 Hadoop 分布式文件系统（HDFS）存储数据。数据的清洗和算法的执行依赖 MapReduce 分布式计算框架。系统每天定时在 Hadoop 平台上进行分布式计算，然后保存运算结果到数据库中，并向用户推送最新的信息和资源。

该系统主要采用了基于协同过滤的推荐技术，能够及时地分析用户的收藏情况，并能向他们推荐具有类似收藏经历的其他用户收藏的书籍。同时还能对用户的收藏行为进行预测，并反馈出该顾客还可能对哪些类型的商品感兴趣，从而把用户的潜在需求转化为真实需求，有效地提高了站点的访问量。

关键词：Hadoop；个性化；推荐系统；书籍；协同过滤算法

Abstract

Personalized Book Recommendation System Based on Hadoop platform is a human-computer interaction system which can recommend books accurately according to the user's interests and interests. The system uses the Hadoop distributed file system (HDFS) to store data. Data cleaning and algorithm execution rely on the MapReduce distributed computing framework. The system computes periodically on the Hadoop platform, and then the results are saved to the database and the latest information and resources are pushed to the user.

The system mainly adopts the recommendation technology based on collaborative filtering, which can analyze the user's collection in a timely manner, and can recommend other books which have similar collection experience to other users. At the same time, it can predict the user's behavior and feedback what kind of goods the customer may be interested in, thereby transforming the potential demand of the user into real demand, and effectively improving the amount of the site's access.

Keywords: Hadoop; Personalise; Recommendation system; Book; Collaborative filtering algorithm

一、引言

互联网的出现和普及给用户带来了海量的信息，并且信息的数量和种类还在不断剧增，用户需要花费大量的时间才能找到自己想要的信息，尤其体现在琳琅满目的各类书籍上面。这种浏览大量无关信息的过程无疑会使淹没在信息过载问题中的用户无所适从，大大降低了用户的体验度。为了解决这些问题，个性化推荐系统使用协同过滤等技术。协同过滤推荐技术与目前流行的社会化网络研究有交叉点，比较著名的有文档推荐系统 Tapestry、图书推荐系统 Amazon.com 等。[1] 协同过滤技术又是最基本的数据挖掘技术。数据挖掘的一个关键问题是数据量。典型的数据挖掘问题包括一个大的数据库，需要从中提取有用的信息。[2]

该系统不仅满足了用户对个性化推荐的需求，帮助用户快速决策提供了参考依据，还能和用户之间建立密切关系，让用户对推荐系统产生依赖。其次，该系统还在一定程度上解决了互联网信息超载的问题。

二、系统需求分析

个性化图书推荐系统采用 B/S 架构，从服务器端获得最新的信息和个性化推荐结果，在用户浏览器终端进行展示。计算平台将兴趣图谱想象成对人们与他们的兴趣之间的关系进行建模的一种方式。[3] 我们还可以利用用户搜索的书本名称、作家、类别等信息建立统计语言模型。[4]

基于 Hadoop 的个性化书籍推荐系统功能有：

1. 各类书籍的热门推荐；
2. 新书上架推荐；
3. 书籍的各类排行榜单；

4. 签约作家推荐；
5. 热门作家推荐；
6. 用户注册及登录；
7. 好友关注及取消；
8. 用户收藏书籍；

三、系统设计

（一）系统开发平台及相关技术

系统使用 Linux 操作系统，搭载在 Apache Hadoop 平台上，并以 B/S 作为系统的基本架构，将 Eclipse 作为基本开发工具。Linux 是一套免费使用和自由传播的类 Unix 操作系统，是一个基于 POSIX 和 UNIX 的多用户、多任务、支持多线程和多 CPU 的操作系统。Hadoop 是一个由 Apache 基金会所开发的分布式系统基础架构。Hadoop 实现了一个分布式文件系统（Hadoop Distributed File System），简称 HDFS。[5] MapReduce 是一种计算模型，该模型可将大型数据处理任务分解成很多单个的、可以在服务器集群中并行执行的任务。这些任务的计算结果可以合并在一起来计算最终的结果。[6] Sqoop 是一款开源的工具，主要用于在 Hadoop(Hive)与传统的数据库间进行数据的传。[7] Python, 是一种面向对象的解释型计算机程序设计语言。[8]

（二）系统结构

系统采用 SSM 框架[9]进行架构、开发，用一种业务逻辑、数据、界面显示分离的方法组织代码，将业务逻辑聚集到一个部件里面，在改进和个性化定制界面及用户交互的同时，不需要重新编写业务逻辑。系统结构模块图如图 1 所示，详细模块描述如下：

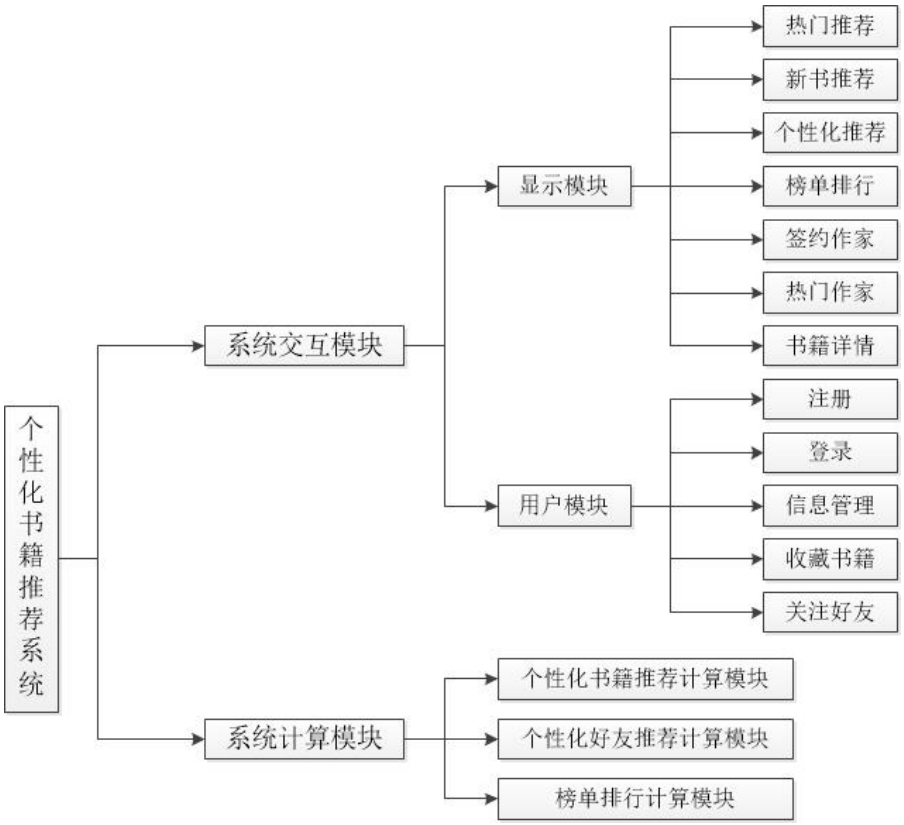

图 1 系统结构模块图

1. 个性化书籍推荐系统主要分为系统交互模块和系统运算模块。
2. 系统交互模块划分为四个小模块：终端展示模块、收藏管理模块、好友管理模块和用户管理模块。四个小模块又划分为不同的功能。

3. 系统运算模块划分为四个小模块：个性化书籍推荐、个性化好友推荐、榜单计算和评分计算。四个小模块又分别统计不同数据，产生各类计算结果。
4. 终端展示模块：用于展示推荐结果和书籍信息。游客和用户都可以浏览各类图书的热门推荐、新书上架、各类榜单、签约作家、热门作家和书籍基本信息。除此之外，用户还可以浏览系统为用户个性推荐的书籍。
5. 收藏管理模块：用于用户对收藏书籍的管理。用户不但可以对感兴趣的书籍进行收藏，而且可以对书籍进行评分和写读后感。用户对于不再感兴趣的书籍可以进行取消收藏操作。收藏的结果对个性化书籍推荐结果产生影响。
6. 好友管理模块：用于用户对好友的管理。用户可以对其他用户进行关注，关注后可以查看好友收藏的书籍及其读后感。用户也可以对已关注的用户进行取消关注操作。关注的结果对个性化好友推荐结果产生影响。
7. 用户管理模块：用于用户管理。游客可以使用注册功能，成为一名用户。用户可以进行登录操作，从而查看个性化推荐结果和收藏关注等一系列功能。此外，用户可以编辑并查看个人信息。
8. 个性化书籍推荐：系统每日定时对用户收藏数据进行分布式运算，产生个性化书籍推荐结果，并将结果保存至数据库中。
9. 个性化好友推荐：系统每日定时对用户关注数据进行分布式运算，产生个性化好友推荐结果，并将结果保存至数据库中。
10. 榜单计算：系统每周定时对书籍数据进行统计，产生各种统计排名结果，并将结果保存至数据库中，用于前台榜单数据展示。例如，书籍的新鲜度排行、书籍的收藏量排行等。
11. 评分计算：系统每周定时对书籍评分数据进行重新计算，并将计算结果保存至数据库中，在前台会展示最新的数据结果。

（三）数据存储设计

该系统将 MySQL 用作于数据持久化存储的数据库。MySQL 是一种关系数据库管理系统，关系数据库将数据保存在不同的表中，而不是将所有数据放在一个大仓库内，这样就增加了速度并提高了灵活性。MySQL 所使用的 SQL 语言是用于访问数据库的最常用标准化语言。由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点，一般中小型网站的开发都选择 MySQL 作为网站数据库。

以下是数据库详细设计：

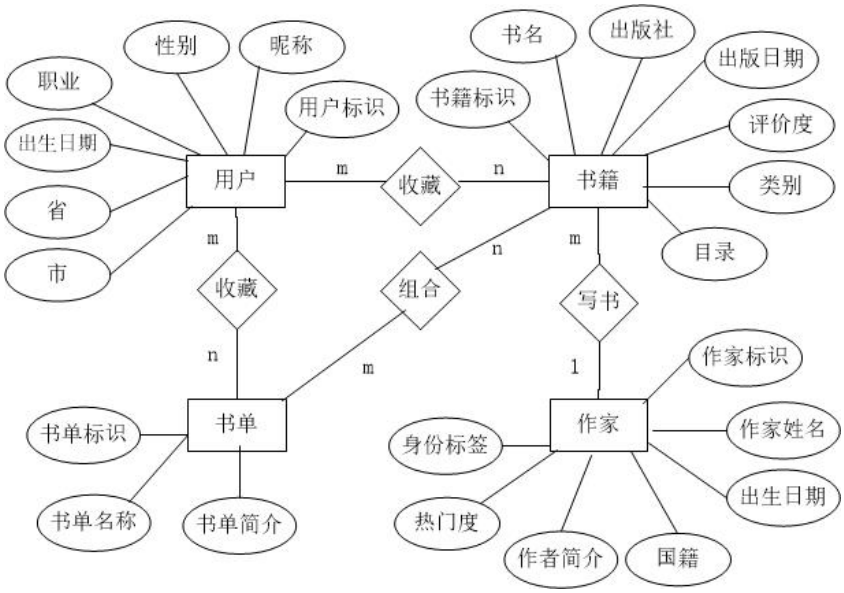

图 2 E-R 图

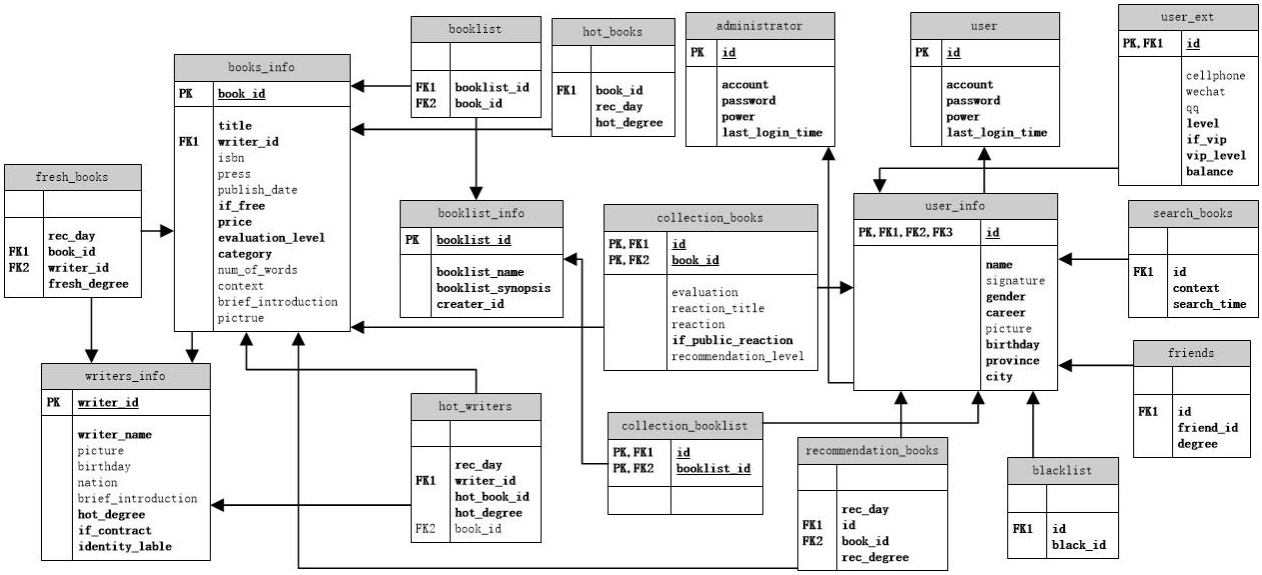

图 3 数据库模型图

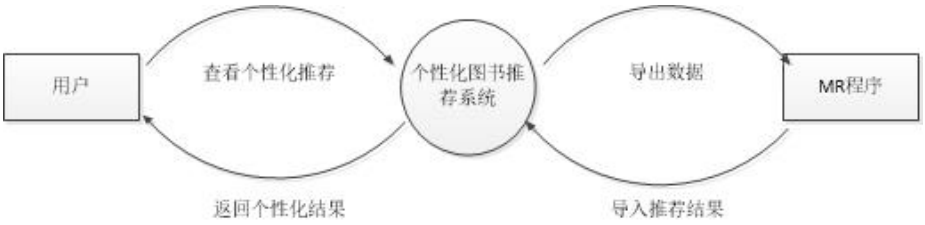

图 4 个性化书籍推荐系统顶层图

四、设计总结

本次毕业设计 Hadoop 的个性化书籍推荐系统实现了个性化书籍推荐、个性化好友推荐、用户管理等功能。系统还对终端展示页面做了改进，提高了用户体验度。由于该系统需要运行在 Hadoop 平台上，我不仅学习了分布式集群的搭建和 Linux 的基本命令，还学习了 MapReduce 编程和协同过滤推荐算法[10]。通过毕业设计，巩固了我的专业知识，提高了实践应用能力，为今后的学习和工作打下了坚实的基础。

在本次毕业设计期间也遇到了诸多问题，比如前后台数据交换，分布式平台的搭建等问题。通过张秋余老师和崔略老师的悉心指导和同学们的热情帮助，最终解决了遇到的问题。在此，谨向老师们致以衷心的感谢和崇高的敬意。

参考文献

[1] 曾子明 著. 信息推荐系统, 北京：科学出版社, 2013.

[2] (葡) Luis Torgo 著. 数据挖掘与 R 语言[M], 北京：机械工业出版社, 2008.

[3] (美) Matthew A. Russell 著. 社交网络的数据挖掘与分析[M]. 北京：机械工业出版社, 2015.

[4] 董启文 著. 基于语言处理技术的蛋白质结构和功能预测若干问题研究[M]. 博士论文,2007.

[5] (美) Tom White 著. 华东师范大学数据科学与工程学院(译). Hadoop 权威指南（第 3 版）[M]. 北京：清华大学出版社，2015.

[6] (美) Edward Capriolo Dean Wampler Jason Rutherglen 著, Hive 编程指南[M], 北京：人民邮电出版社, 2013.

[7] Katbleen Ting & Jarek Jarcec Cecbo 著, Apache Sqoop Cookbook, O'Reilly Media，2013

[8] (挪) Magnus Lie Hetland 著. Python 基础教程（第 2 版 • 修订版）[M]. 北京：人民邮电出版社, 2014.

[9] 赛奎春. JSP 工程应用与项目实践[M]. 北京：机械工业出版社, 2008.

[10] (美) Sandy Ryza Uri Laserson Sean Owen 著. Spark 高级数据分析[M], 北京：人民邮电出版社, 2015.