应用型本科高校智慧图书馆平台构建①

郑自辉[1] 朱梦园[2] 施航海[2]
[1]（厦门理工学院信息中心 厦门 361024）
[2]（厦门理工学院图书馆 厦门 361024）

摘 要 随着信息技术的高速发展，海量数据和非结构化数据爆炸式增长，给图书馆带来了冲击，同时“大数据”中又蕴藏着巨大的价值，如何满足读者的需求，提高读者的满意度，给读者提供更好的服务，是一个值得研究的问题。智慧图书馆就是为满足这种需要应运而生发展起来的数据处理技术，通过各种数据分析方法，展现业务的效果与业务之间的关联，从读者数据中找出读者行为规律。只有在调查分析基础之上，采取切实有效的具体措施提高服务质量，才能紧跟时代步伐，保证图书馆服务质量的不断攀升。

关键词 大数据；网络平台；智能推荐；ILAS；核心数据库

Construction of Intelligent Library Platform in Application-oriented Universities

Zheng Zihui[1] Zhu Mengyuan[2] Shi Hanghai[2]
[1]（Network Information Center，Xiamen University of Technology Xiamen 361024）
[2]（Library，Xiamen University of Technology Xiamen 361024）

Abstract With the rapid development of information technology，massive data and unstructured data have increased explosively. This has brought shock to the library. At the same time，there is huge value in “Big Data”. How to meet the needs of readers，improve the reading experience，and provide better services to readers is a problem worth studying. The Smart Library is a data processing technology developed to meet this requirement. Through various data analysis methods，the relationship between the effect of the business and the business is revealed，and the behavior of the reader is found out from the reader data. Only on the basis

①本文系福建省中青年教师教育科研基金项目 JAT170436 的研究成果之一。

of investigation and analysis，and taking concrete and effective measures to improve service quality，can we keep up with the pace of the times and ensure the continuous improvement of library service quality.

Keywords Big Data；Network Platform；Intelligent Recommendation；ILAS；Core Database

1 引言

随着数据库技术的迅速发展以及数据库管理系统在图书馆的广泛应用，图书馆积累了大量的读者对资源的历史访问数据、图书借还数据等。这些数据中隐藏着许多重要的信息，人们希望能够对其进行分析，以便更好地利用这些数据为读者服务。目前的图书馆管理系统无法预测读者的信息需求，缺乏挖掘数据中隐藏的知识的手段，所以在图书馆这样一个知识的海洋中，读者很难找到所需的信息资源。学校的绩效管理要求使图书馆必须关注各类IT系统运行的效果，通过各种数据分析方法，展现业务的效果与业务之间的关联。图书馆的网络化和数字化[1]，为图书馆积累了大量的读者数据，包括读者的身份信息、借阅信息、网上阅读习惯、检索习惯等，图书馆如果想完全分析读者数据，就得从读者数据中找出读者行为规律。

2 设计架构

2.1 总体架构

在智慧图书馆中利用Hadoop的集群[2]特征，将智慧图书馆中需要巨大计算能力的各个模块的计算和存储要求扩展到Hadoop集群中的各个节点上，利用集群的并行计算和存储能力来进行相关数据分析和挖掘工作。系统采用分层的设计思路，在底层，使用Hadoop来存储、分析和处理巨大的数据量，在高层通过接口直接透明地调用底层的技术和存储能力，如图1所示。

其中，每一层的业务功能不同。

（1）数据采集层：负责从各类数据源中提取、导入数据，主要产品包括——动态采集SDK、日志提取分析工具、外部数据导入工具、其他数据提取工具等。

（2）数据存储层：负责将预处理后的数据进行存储，主要由可进行横向扩展的Hadoop集群构成，另外辅之以关系数据库作数据中转、元数据存储、供某些软件使用等。

（3）分析挖掘层：负责图书数据建模、挖掘、评估和发布，核心是实现两类数据挖掘的算法和模型。一类是抽象的数学算法及模型，另一类是在此基础上针对图书信息的专业算法和模型。

（4）业务应用层：负责将分析挖掘结果的可视化展现形式，集成到相应的图书系统功能中。

另外，在数据采集层和数据存储层之间，由ETL工具负责数据预处理任务；在分析挖掘层和业务应用层之间，由可视化展现工具负责分析挖掘结果的可视化展现任务。

2.2 数据采集

智慧图书馆首先需要收集各种网站的图书信息、图书排行、图书评论等数据，它们可能是结构化的，也可能是半结构化或非结构化的；既可能来自网站内部的各业务系统，也可能由外部提供；既可以是静态的（如属性数据），也可以是动态的（如行为数据）。而图书数据采集产品就是根据业务需要，将这些数据采集到智慧图书馆中[3]。

在整个系统中，我们采用自主研发的数据采集系统，来采集豆瓣、当当等网站的书籍排

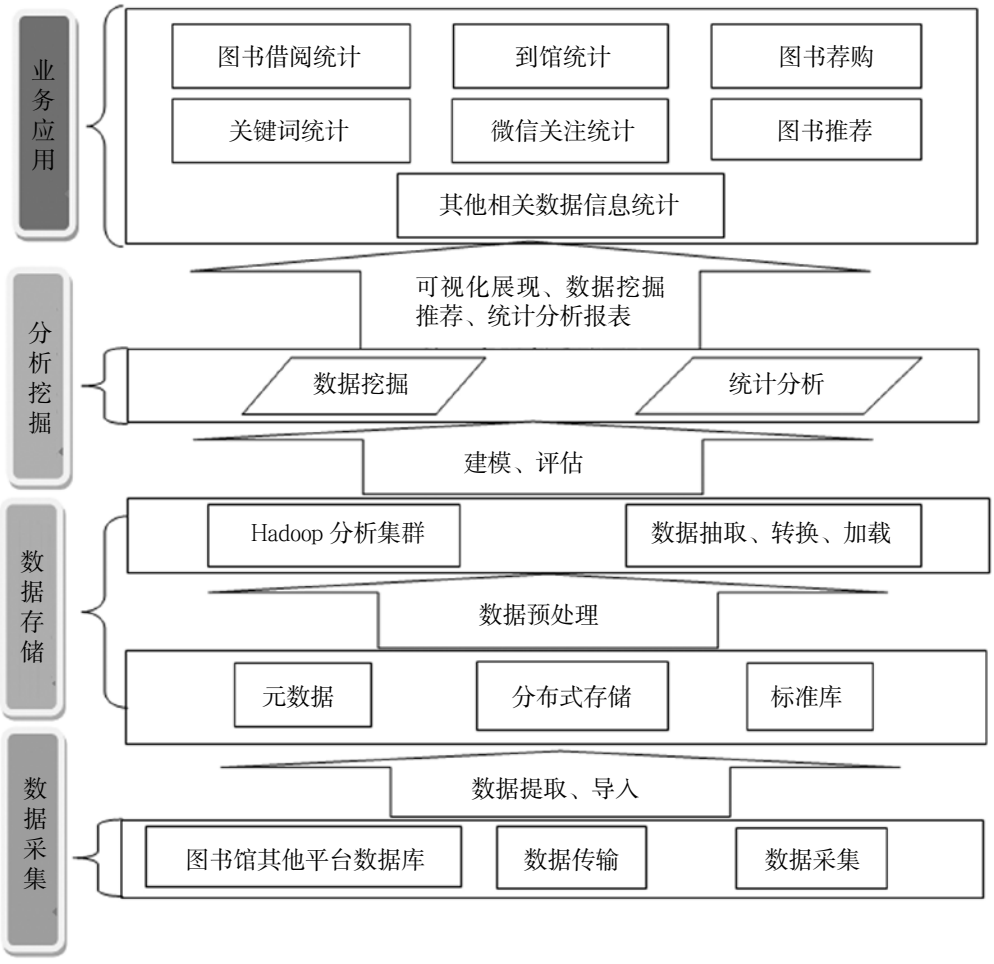

图 1　总体架构

行榜、图书评论等信息。我们采用任务池的方式多进程多并发采集网站数据，并增加爬虫的二次控制层，对爬虫进程进行智能控制，确保采集工作顺利进行，如图2所示：

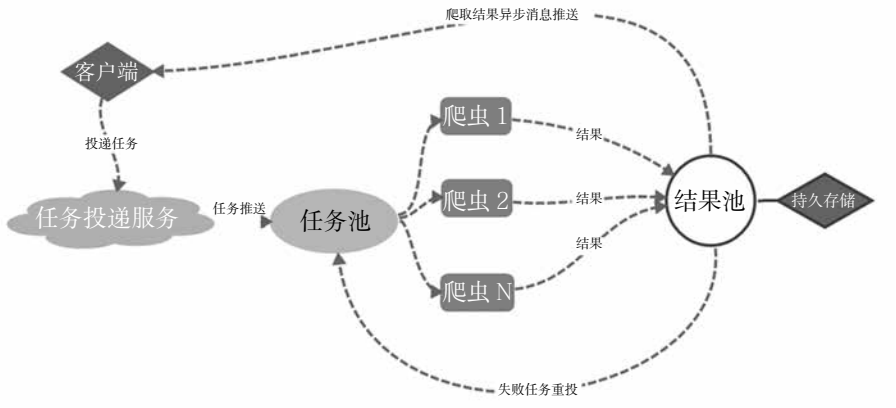

图 2　爬虫框架

2.3 数据存储

Hadoop集群通过将数据分配到多个集群节点上并进行并行处理，因此尤为适合对大数据的存储和分析。Hadoop集群通过添加节点数量来有效地扩展集群，因此具有极好的可扩展性；Hadoop软件都是开源的，也不必购买昂贵的高档服务器，因此具有很好的性价比。Hadoop集群将数据分片发送至多个节点保存，因此具有极高的容错性。

在整个系统中，我馆使用HDFS来存储文件和数据。HDFS具有很高的数据吞吐量，并且很好地实现了容错机制。HDFS提供了多种访问接口，包括API以及各种操作指令。使用HDFS可以为原始的大数据集提供存储空间，对临时文件进行存储，为数据预处理、数据挖掘过程提供输入数据，输出数据也保存在HDFS中，如图3所示。

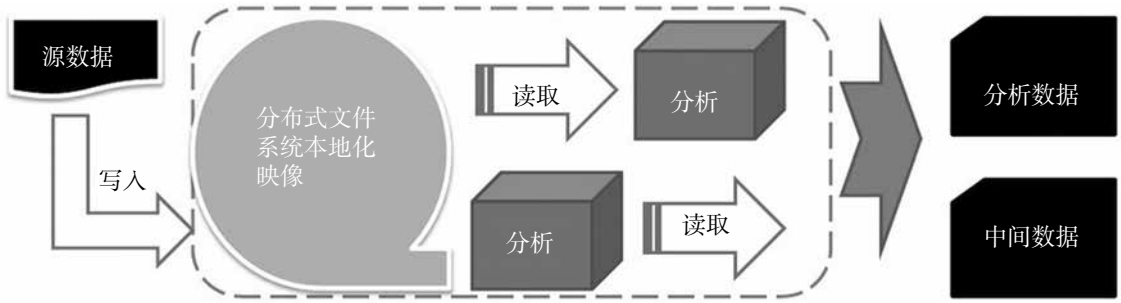

图 3　数据处理访问模式

2.3.1 数据预处理

采集到的图书数据来自多种数据源，大多存在着不完整性和不一致性，无法直接用于数据挖掘或严重影响数据挖掘的效率。因此在进行数据挖掘之前，通过使用数据预处理工具，对原始数据的清理、变换、集成等灵活处理，可以减少挖掘所需数据量，缩短所需时间，并极大提高数据挖掘的质量。

2.3.2 数据分析与算法设计实现

数据挖掘系统将设计构建一个庞大的数据云，依托该数据云打造6大方面的系统功能，如图4所示。

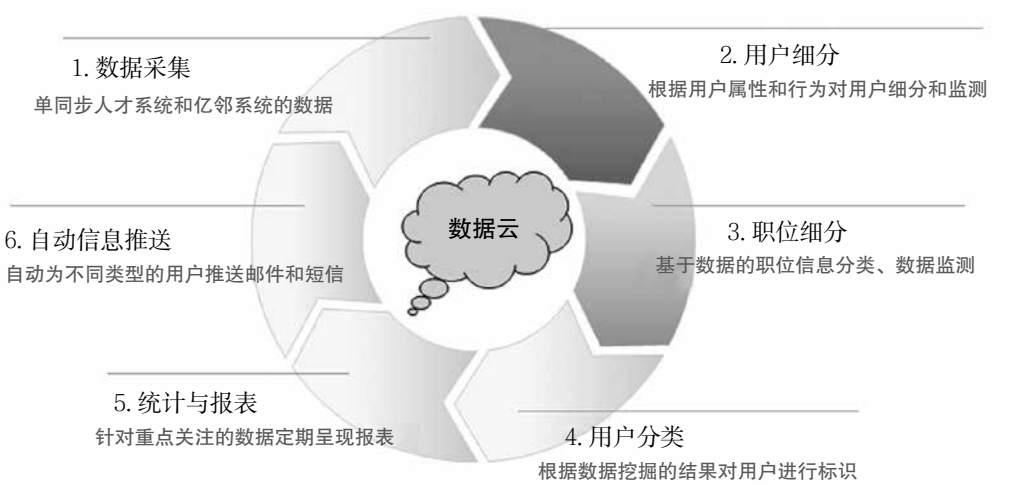

图 4　数据挖掘系统

（1）数据采集：同步来自当当网、亚马逊以及京东商城和数据挖掘结果方面的数据。

（2）基于学生基本属性、行为、标签等对用户进行细分，根据细分规则导出（自动推送）数据、监测数据变化，为数据挖掘和决策分析提供帮助。

（3）基于专业的基本信息、用户行为等对专业进行细分（如计算机专业的学生借阅书籍情况比较良好等），根据细分导出（自动推送）数据，监测数据变化。

（4）用户分类：根据数据挖掘的结果固化数据挖掘规则，定期为用户设置标签，对用户进行分类（如每天早上2点将一周内有登录达到5次的用户标识为“活跃用户”）。

（5）统计与报表：根据细分监测到的数据变化，形成报表。监测数据挖掘的效果，为决策分析提供依据。如近半年内活跃用户数量按月的变化曲线图。

（6）信息推送：调用邮件系统，实现自动信息推送：智能地给用户推荐工作、给企业推荐人才。

数据挖掘是通过分析数据、从大量数据中寻找其潜在规律的技术。利用预测、关联、分类、聚类、时序分析等技术，数据挖掘可以从海量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识。目前，传统的数据挖掘产品在大数据平台上还存在一些局限性，研发一套Hadoop平台下的数据挖掘工具是一项极具挑战性的任务。

在系统中，通过应用PageRank算法、朴素贝叶斯、贝叶斯信念网络、神经网络等算法，进行数据挖掘与分析；同时使用MapReduce将数据挖掘系统中具有大计算量特征的各个子模块的计算任务发布到集群中的各个节点以实现并行计算。MapReduce具有良好的伸缩性和扩展性，通过提供接口使系统快速地实现各种算法的并行计算。

2.3.3 可视化展现

数据挖掘得到的结果，往往具有数据量巨大、关联关系复杂、维度多以及双向互动需求等特征。可视化展现工具以适合人类思维的图形化的方式对结果进行展示，提高了数据的直观性和可视性。可视化展现面向各类客户，通过选择合适的可视化模型，将枯燥的数据转换为令人印象深刻的美丽图形，极大提升了数据的利用价值[4]。

3 功能设计与实现

3.1 图书借阅统计

（1）选择查询分类、查询时间范围，点击查询之后，才去统计数据。

（2）预设分类（分类编号、分类名称）、预设时间单位（天、周、月、年）。

（3）分类名称不能重复，所属分类的编号不能重复。

（4）总体数据：通过借阅记录表找出符合条件的在校学生和教师记录，按照性别分组统计借阅量和借阅人数，进一步计算学生总借阅人数和借阅数量。

（5）借阅情况：通过当前选择的时间的单位作出不同的SQL语句（以全体师生为例）。

（6）时间：按字段分组计算借阅人数和借阅量，并把对应的日期获取出来。通过起始时间和结束时间把中间的这些日期以及起始时间和结束时间放进一个数组，遍历从数据库查询出来的数据，把数据填充进去。

（7）学院借阅排行：通过借阅记录表以学院分组统计符合条件的借阅人数和借阅量，并按借阅量从大到小排列。

（8）分类排行榜：通过学生表和教师表获取符合条件的在校师生读者No.，然后从书的分类表、书籍表、借阅记录表和读者表找出符合条件的记录，按书籍表的二级分类分组统计借阅人数、借阅量等信息。

（9）TOP100热门书籍排行榜：通过学生表和教师表获取符合条件的在校师生读者No.，然后从书籍表、借阅记录表和读者表找出符合条件的记录，按书籍ID分组统计借阅人数、借阅量等信息。

该模块界面如图5所示：

图 5　图书借阅统计模块

3.2 到馆统计

（1）班级：通过学校、学院、专业、班级不同分类层级来统计学生到馆次数。

（2）时间：按字段分组计算到馆次数，并把对应的日期获取出来。通过起始时间和结束时间把中间的这些日期以及起始时间和结束时间放进一个数组，遍历从数据库查询出来的数据，把数据填充进去。

（3）根据时间统计到馆人数，分析出每月最多到馆人次、最少到馆人次、平均到馆人次和高峰期时间段。

（4）统计各学院、各年级到馆情况，计算所占比例并做排行统计。

该模块界面如图6所示。

图 6　到馆统计模块

3.3 图书荐购

（1）馆藏书籍补充建议：推荐理由的值为该书籍（含副本）被借断天数总和与入馆至今天数总和的比值，值越大该种书籍越紧缺。

（2）新书好书推荐：结合豆瓣、当当等网站的书籍排行榜、图书评论等信息，以及用户的阅读偏好，搜索记录等，对图书馆的好书和新书按学科分类进行推荐。

（3）密集书库预借排行：根据密集书库的预借量生成排行，预借数量较多的可直接将书从密集书库调至普通书库，提高借阅率，方便读者借阅。

该模块界面如图7所示。

图 7　图书荐购模块

3.4 关键词统计

（1）通过对检索日志进行行为分析，统计检索关键字，建立关键字词云。

（2）左侧的图中关键词字体大小根据关键词被搜索次数而定，搜索次数越大，字体越大。

（3）默认显示上一个月的搜索关键词排名。

该模块界面如图8所示。

图 8　关键词统计模块

3.5 微信关注

（1）微信关注量统计：通过分析关注厦门理工学院图书馆微信的读者信息，按院系统计关注量，统计每个院系、专业的学生对图书馆微信的关注量。

（2）导入微信关注列表：每次导入关注列表都将数据库原先的清空，再导入上传的EXCEL。上传EXCEL格式为：第一行为列名，第一列为图书馆账号、第二列为微信ID、第三列为微信昵称、第四列为关注时间，列名可变化（如“图书馆账号”改成“账号”也可），但顺序固定不可调整。

（3）关注率计算公式：关注率=关注人数/统计单位的总人数（如列表展示各院关注排名，则关注率=该院关注人数/该院总人数）。

该模块界面如图9所示。

图 9　微信关注模块

3.6　图书推荐

（1）通过分析借阅历史，挖掘同一类读者的借阅行为，分析其阅读偏好，提高图书推荐的精细和准确性；结合读者的专业、班级、成绩信息，进行读者画像构建等；目前的推荐只是做到了同一类书的借阅量推荐[5]。

（2）根据读者所属专业、专业方向、课程、出版时间（或入库年份）、检索日志、本专业方向成绩优秀的读者所借阅过的图书、本专业方向科任教师所借阅过的图书，向读者推荐（其中课程、出版时间（或入库年份）可以在推荐结果中手工选择过滤排除其他数据）。

（3）推荐书目包括：喜好推荐、班级推荐、新书推荐、好书推荐、分享推荐。

（4）个人数据中可浏览本人的借阅历史。

该模块界面如图10所示。

图 10　图书推荐模块

4　结束语

本平台从大量的、不完全的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识，为读者提供个性化推荐，推动图书馆决策的精确性发展，同时为图书馆做绩效统计工作提供了有效的方法。

本平台的构建虽然取得了初步的成功，但是还有许多需要深入研究的工作。例如：在借阅统计模块下，对分类号的分级暂时无法做到完全匹配中图分类法，并且对于复分和仿分没有涉及，这对于我们做统计工作时会出现误差；并且在读者推荐模块中，会经常出现访问限制的情况，目前没有找到原因。这些问题将在以后的工作中逐步完善，保证图书馆服务质量的不断攀升。

参考文献

[1] 魏江辉. 信息时代智能化高校图书馆的构建方法与思考[J]. 科技情报开发与经济，2015（07）：43—44，50.

[2] 廖松博.Hadoop上的PageRank算法优化[D].复旦大学，2013.

[3] 郭竟. 高校图书馆智慧服务体系的构建[J]. 河南图书馆学刊，2015（05）：26—28.

[4] 吴宇芬.“互联网+”高校图书馆智慧化服务模式探究[J]. 农业图书情报学刊，2017（02）：184—187.

[5] 王正勤，牛永芹，颜莉莉. 数字图书馆中的智能推荐技术研究[J]. 图书情报工作，2011（17）：110—113.