

图书馆对学生学业成效贡献的大数据分析平台构建

杨 亚, 毛海波

(宁波大学 图书馆与信息中心, 浙江 宁波 315211)

摘要: 为探索学生对高校图书馆服务的利用与学生学业成效之间的关系, 为高校图书馆价值主张和服务决策提供理论依据. 通过对图书馆大数据分析存在问题和数据构成进行全面深入的分析, 采用 Hadoop 分布式存储平台与 MapReduce 计算框架搭建统一的图书馆大数据分析平台, 并提出一种图书馆大数据预处理的方案. 最后给出了一种图书馆服务与学生学业成效主题分析模型的构建思路, 为下一步更深入的研究提供理论基础和技术保障.

关键词: 大数据; 图书馆; 学业成效

中图分类号: G250.76

文献标志码: A

文章编号: 1001-5132 (2016) 01-0033-04

高校图书馆因拥有丰富的数字资源和良好的学习环境, 成为学校师生学习和休闲的重要场所^[1]. 学生作为高校图书馆服务对象主体之一, 其学业和高校图书馆所提供的服务必然存在一定的关系, 因此图书馆对学生学业成效贡献研究是指导高校图书馆投入产出的重要组成部分^[2].

大数据技术作为一种新兴的海量数据分析工具越来越受到各行各业的重视^[3]. 互联网企业 Google 及 Facebook 之所以取得令人瞩目的成绩, 其核心的本质就是其公司记录和分析了用户的行为大数据, 从而精确掌握用户行为并形成预判^[4]. 由此可见, 大数据技术对于海量、高速发展的数据具有很好的分析和管理能力, 它被用来研究图书馆用户行为的“大数据”是最佳选择, 研究成果可以为高校图书馆服务决策提供客观的依据, 具有重要理论研究意义和实际应用价值.

1 大数据技术

大数据技术是一系列收集、存储、管理、处理、分析、共享和可视化技术的集合^[5]. 大数据的关键技术有很多, 如借鉴生物界的进化规律演化的随机化搜索方法已被人们广泛应用于组合优化、机器学习、信号处理、自适应控制等领域^[6]; 再如分布

式技术包含分布式文件系统、分布式数据库、分布式计算框架等, 其已经全面运用于各类大数据应用中^[7]. 大数据价值的完整体现则需要多种技术的协同, 总的来说可以归纳为以下几种:

(1) 分布式技术. 最典型的是 Apache 基金会的 Hadoop 大数据分布式处理软件框架, 主要延续了 Google 分布式文件系统 GFS 的开源思想、分布式计算框架 MapReduce 和分布式数据库 BigTable 的实现机理, 开发了自有的产品——HDFS 分布式文件系统、MapReduce 分布式编程框架和 HBase 分布式数据库^[8].

(2) 大数据预处理技术. 大数据的一个重要特点是多样性, 这就意味着数据来源极其广泛、数据类型极为繁杂, 这种复杂的数据环境给大数据处理带来了极大的挑战. 所以在分析大数据前, 首先必须对海量数据源进行预处理, 以保证数据质量及可信性.

(3) 大数据挖掘技术. 数据挖掘是整个大数据处理流程的核心, 因为大数据的价值产生于挖掘过程. 数据挖掘就是从大量的、不完全的、有噪声的、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的过程^[9]. 大数据挖掘的算法众多, 其中

以分类算法、聚类挖掘算法、关联挖掘算法、序列挖掘算法最为主流。

2 图书馆大数据分析面临的问题及数据构成

2.1 图书馆大数据分析面临的问题

大数据时代的图书馆读者行为大数据具有数据海量、类型复杂、处理速度快和价值密度低的特点,要对这些大数据进行统一的存储与分析,对图书馆来说有着极大的挑战性。以下是3个亟待解决问题:(1)数据的存储能力。学生在图书馆的所有行为活动都将产生数据,数据量已由TB级升至PB级,而且还在源源不断地增加,数据量的增长速度已远远大于存储能力的增长速度。(2)数据类型复杂繁多。学生对图书馆的利用行为数据不仅仅是简单的二维表格式存储的结构化数据,还有以文本、图片、XML文档、JSON文档、日志文件和音频/视频等半结构化和非结构化数据,关系型数据库已经无法有效管理这些数据。(3)数据处理的实时性。大数据时代强调的是数据处理的及时有效,图书馆要求大数据分析平台可以快速地获取、存储和分析学生的行为数据,为图书馆提供快速的决策支持。

面对上述问题,传统数据分析工具已不再适合,而Hadoop大数据分布式存储与计算框架能很好地解决了这些问题,因此基于Hadoop建立统一的图书馆大数据存储和分析平台变得尤为迫切。

2.2 图书馆大数据的构成分析

针对目前宁波大学在校的本科学生,将其对图书馆利用行为和学业成效相关数据(2010~2014年)作为此次实验的数据源。经过分析,确定数据源由学生个人信息、学生学业相关数据及图书馆利用行为数据三部分组成,数据类型包括结构化的二维表数据和非结构化的日志数据,数据总量共计6108万条记录。其中,学生学业相关完整数据有120万条记录;图书馆利用行为数据包括从图书馆门禁系统(357万)、阅览室座位管理系统(320万)、图书借阅系统(247万)、数字资源访问日志(5064万)等途径获取学生的图书馆服务数据共计5988万条记录,为图书馆服务对学生学业成效贡献研究提供基础的大数据参考样本。

从上述统计来看,实验涉及到的数据量级已

在十亿字节以上,而且还在随着时间的推移成线性增长,因此定期将相关系统中每天生成的数据导入Hadoop大数据平台成为必然的选择。目前此次实验数据的收集整理工作已经基本完成,初步拟定大数据主题分析的内容见表1。

表1 主题分析的内容

名称	具体内容
学业成效	(1)学业成绩;(2)奖励情况;(3)科研能力
图书馆利用行为	(1)对纸质馆藏的利用;(2)对阅览室的利用;(3)数字资源的利用;(4)其他,如对网络的利用等
学生个人信息特征	学院、班级、专业、学号、高考成绩、性别和生源地(农村或城市)等

高校大学生学业成效主要由学业成绩、奖励情况和科研能力三部分组成。学业成绩主要是学生的在校的各类专业课和选修课的考试成绩积点分,奖励情况是在校期间参加社团活动、校级活动等获得的奖励或荣誉称号,科研能力是指参与科研项目或发表论文等。

图书馆利用行为此内容,并综合已有的相关研究关注的内容,再结合宁波大学的实际情况,考虑选取以下几方面:(1)对纸质馆藏的利用,包括入馆记录(室内阅读情况)和借阅记录等;(2)对阅览室的利用,包括入室记录和在馆时间(座位管理系统记录情况)等;(3)数字资源的利用,包括数据库检索次数和全文下载次数等;(4)其他,如对网络的利用等。

上述是高校图书馆服务对学生学业成效贡献研究中主题分析的相关内容,但在实际的分析建模过程中,可以进行适当的取舍或增加。

3 图书馆成效贡献的大数据分析平台构建

3.1 图书馆大数据分析平台的总体架构

基于Hadoop的图书馆大数据分析平台主要分为两层——大数据预处理层和主题模型构建层,自下往上每层都为上层提供服务。整体的架构设计如图1所示。

3.2 图书馆大数据预处理方案的设计

图书馆大数据预处理层主要是利用Hadoop集群在存储和计算能力的优越性,并结合大数据预处理技术,来对学生图书馆活动及学业数据进行

预处理. 具体的设计方案流程如图2所示.

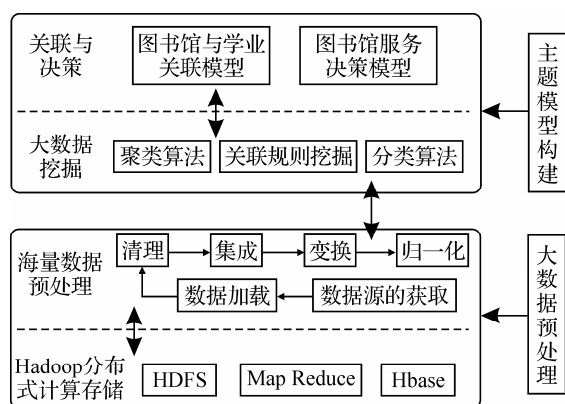


图1 大数据分析平台架构

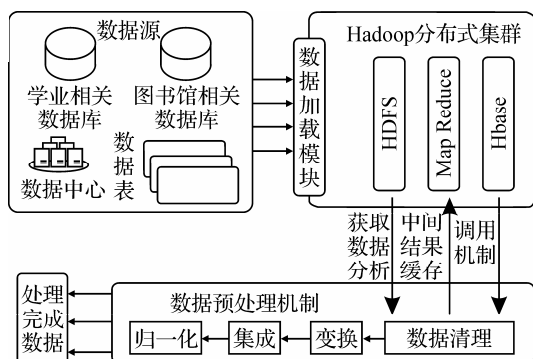


图2 Hadoop 大数据预处理方案

整个方案采用分层设计思想, 底层是 Hadoop 分布式平台层. 目前实验采用 10 台 Linux 操作系统的普通服务器机子, 并分别在每台机器上安装 JDK、SSH、Hadoop 和 Hbase, 搭建 Hadoop 完全分布式运行环境. 整个集群规划为: NameNode: 10.22.102.46, DataNode1~DataNode9: 10.22.102.47~10.22.102.55. HDFS 是分布式计算的存储基础, 主要用于存储学生图书馆利用行为和学业成效相关源数据集, 其采用 Master/Slave 结构, 集群包括 1 个 NameNode 和多个 DataNodes, NameNode 负责整个集群的任务调度分配, DataNode 则是存储实际的数据. MapReduce 过程是把从 HDFS 中待处理的学生图书馆利用行为和学业成效相关源数据集分解成 M 个小数据集进行并行 Map 操作, 输出中间态键值对 $\langle k, \text{value} \rangle$, 然后根据 k 值进行 Group 操作, 形成新的小数据组集 $\langle k, \text{list}(\text{value}) \rangle$, 最后将这些小数据组集分割成 R 个集合, 进行 Reduce 操作后存储到分布式数据库中. Hbase 是个基于列存储的分布式数据库, 数据行有 3 种基本类型: 行关键字、

时间戳和列, 行关键字是数据表的唯一标示. 海量的学生图书馆利用行为和学业成效相关数据通过 MapReduce 计算后, 可以 k 值作为行关键字进行分布式存储, 实现海量数据的存储与管理功能.

在 Hadoop 层之上, 则为相应的大数据预处理模块, 可以透明地调用 Hadoop 底层的计算和存储能力, 包括数据清理、变换、集成及归一化 4 个子模块. 数据清理是删除那些不符合要求的记录. 数据集成是将来自不同应用系统中的数据源合并到一起, 形成一致的数据存储. 数据转换是将学生学业相关数据和图书馆相关应用系统中的数据用一定的格式来表示, 以方便后期做关联挖掘. 归一化处理是把数据值控制在一定的范围内, 保证程序运行时收敛加快. 最终将处理好的数据通过接口或其他方式输出.

3.3 图书馆服务对学生学业成效贡献的主题模型构建思路

图书馆服务对学生学业成效贡献的主题建模主要是利用大数据挖掘算法对相关数据进行深入分析挖掘, 精确发现数据之间的关联关系, 构建相应的主题分析模型. 具体的分析挖掘过程如下: 首先选择合适的聚类算法将学生群体和图书馆服务资源数据按照一定的规则分割成不同的集合, 分析不同集合表现出的特征; 其次利用关联规则挖掘算法对主题分析的内容进行关联挖掘, 分析各项数据之间存在的关系; 最后采用决策树分类算法做一些预测性主题分析.

最终拟定图书馆服务与学生学业成效的主题模型主要从以下几个主题分析方向进行构建:

- (1) 各学院在某学年/学期学生整体学业成效情况分别与对图书馆利用行为(学院平均进出阅览室时长、借阅次数、数字资源利用)的相关性分析.
- (2) 各专业在某学年/学期不同班级间学业成效分布情况与对图书馆利用行为(班级平均进出阅览室时长、借阅次数、数字资源利用)的相关性分析.
- (3) 相同专业学业成效相差较大的个人与对图书馆利用行为(个人进出阅览室时长、借阅次数、数字资源利用)的相关性分析.
- (4) 图书馆不同的资源服务(纸质馆藏、阅览室利用、数字资源等)对学生学业成效贡献的比重进行分析.

此外,在上述主题分析中分别加入控制因素(性别、生源地以及高考成绩),分析它们与学生学业成效的相关性,建立相应的主题关联模型,达到指导服务决策与优化资源配置的目的,最终更好的为用户提供服务资源。

4 结语

高校图书馆对学生学业成效贡献的研究,以学生作为研究主体,以学生的学业成效作为关注目标,提出基于 Hadoop 开源平台与大数据技术进行分析与主题建模,探索学生对于图书馆的利用与其学业成效之间的关系。在整个主题模型构建过程中,将一些控制因素加入到图书馆与学业成效的关联分析中,根据分析结果可以反过来指导完善图书馆服务对学生学业成效贡献主题模型的构建,从而更全面地对高校图书馆资源与服务在帮助学生提高学业成效方面的作用进行考察。最终的研究成果可指导图书馆进一步将服务嵌入到学生学习的每一个细节处,并以此为依据提升图书馆服务的质量与水平。

参考文献:

- [1] 王丽艳. 图书馆资源利用对大学生学业成绩的影响研究——以南京三所高校为例[J]. 数字图书馆论坛, 2013(9):27-31.
- [2] ZU H N, ZAKARIA A, ROSNI A K, et al. The impact of environment on engineering students' academic performance: A pilot study[C]. Engineering Education (ICEED), 2011:113-118.
- [3] 杨亮, 雷智雁. 大数据环境下图书馆个性化服务研究[J]. 现代情报, 2014(4):74-77.
- [4] WILLIAMS K, WU J. Scholarly big data information extraction and integration in the CiteSeerX digital library [C]. Data Engineering Workshops (ICDEW), 2014:68-73.
- [5] 朱静薇, 李红艳. 大数据时代下图书馆的挑战及其应对策略[J]. 现代情报, 2013(5):9-13.
- [6] 鲍翠梅. 面向大数据时代的图书馆知识服务[J]. 现代情报, 2013(10):33-37.
- [7] 毛晓燕. 大数据环境下图书馆信息服务走向分析[J]. 图书馆工作与研究, 2014(3):72-75.
- [8] 郭逸重. Hadoop 分布式数据清洗方案[D]. 广州: 华南理工大学, 2012.
- [9] 申彦. 大规模数据集高效数据挖掘算法研究[D]. 镇江: 江苏大学, 2013.

Construction of Big Data Analytics Platform in Library Contributing to Students' Academic Achievements

YANG Ya, MAO Hai-bo

(Library and Information Center, Ningbo University, Ningbo, 315211, China)

Abstract: The main purpose of constructing big data analytics platform of library contributing to students' academic achievements is to explore the relationship between college students using library services and the students' academic achievements. The platform sets up a theoretical stage on which decision making for library services can be evaluated. Through using big data analytics and building a unified big data analytics platform based on Hadoop distributed computing storage platform and MapReduce framework, a big data system is proposed in this paper.

Key words: big data; library; academic achievements

(责任编辑 章践立)