基于数据挖掘的学生行为分析系统设计与实现

方洪雨

（西北民族大学榆中校区　甘肃兰州　730106）

摘要：随着数据挖掘技术的快速发展，各大高校逐渐开始重视校园中的大数据，通过挖掘校园大数据的价值，深入分析学生的行为，提高学校教学和管理水平。本文通过学习和调研大数据数据分析相关技术，对学生行为数据进行了科学合理的预处理，提高了数据的质量，然后通过从大量数据中提取学生的特征数据，从而为之选择合适的算法进行数据的 可视化分析。分别采用聚类分析、分类分析、回归分析以及关联分析算法对不同学生的学习行为数据进行深度挖掘，通过对比分析结果，不仅给出学生高效学习的有效建议，而且也给高校提供了高效科学的学生教育管理建议。

关键词：数据挖掘；搭建 Hadoop 集群；行为分析

Design and implementation of student behavior analysis system based on data mining

Fang Hongyu

（Yuzhong Campus, Northwest University for Nationalities, Lanzhou 730106, China）

Abstract: With the rapid development of data mining technology, universities gradually began to pay attention to big data in the campus, through mining the value of campus big data, in-depth analysis of students' behavior, improve the level of teaching and management. In this paper, through the study and research of big data data analysis technology, the students' behavior data is preprocessed scientifically and reasonably, improve the quality of the data, and then by extracting the characteristic data of students from a large number of data, so as to select the appropriate algorithm for data visualization analysis. Cluster analysis, classification analysis, regression analysis and correlation analysis algorithms are used to dig deeply the learning behavior data of different students. By comparing and analyzing the results, effective suggestions are given not only for students' efficient learning, but also for colleges and universities to provide efficient and scientific suggestions for student education management.

Key words: data mining; Build a Hadoop cluster; Behavior analysis

本研究的主要内容：近年来，随着信息技术多方面的快速发展，我国的大数据产业得到了前所未有的发展，大数据分析近几年来已成为许多研究者的研究热点，通过各种数据分析算法例如 kmeans 聚类分析、粗糙集算法、回归算法等，挖掘数据的 潜在价值变得越来越重要，如何更好地利用这些复杂的数据越来越受到人们的重视。通过对数据进行挖掘，帮助人们在海量的数据中找到有用的信息，是解决当前信息超载问题的有效手段。本文主要是通过运用大数据及其相关技术对学生学习行为数据进行多维度[1]、多层次、多角度、全方位的挖掘，通过搭建数据分析平台，对大量的学生学习行为数据进行深度挖掘，最终将结果可视化输出，给目标用户直观易理解的结果。这是一个无论对学生还是学校教育来说都是非常有用的分析平台[2]，对学校来说，这一举措进一步丰富和发展了数字型智慧校园的技术体系，转变了高校的管理方式与理念，使得高校资源的可以充分有效的被应用。对学生来说，通过充分挖掘分析学生行为数据，通过数据可视化可以将学生的长期以来的学习行为数字化、图示化的显现出来，从而对学生的生活规律、学习习惯、学习情况进一步清晰的理解，进而为学生德智体全面发展提供数据分析基础。具体可以带来的好处如下：

（1）更高效化地教学管理。高校在教学管理方面一直做不到面面俱到，无法做到实时跟踪学生的学习情况和掌握学校的教学质量，现今利用大数据分析技术[3]，就可以充分利用学校的现有资源，充分发挥高校的计划、组织、协调、控制等管理职能，进而据分析结果完善其教学评价标准，提高教学质量。

（2）更专业化地进行教育改革。随着互联网的发展，现在学生的学习行为方式也越来越多样化，就比如疫情期间学生们就主要通过网上进行学习，不同于以往的面对面教学方式，所以新的教育改革急切被需要着，通过深度挖掘[4]高效数据中所蕴藏的信息，可以更专业化的进行教育改革。

（3）更好地为学生服务。通过对学生行为的深度数据分析，高校可以实时掌握学生的动态，根据学生最近的学习状态，对学生进行有针对性的劝谏和指导。相比于之前的集体式教育这种方式可谓尽可能的做到了因材施教。

特色和创新点

将传统教育于现代相结合：

将对数据时代中的传统统计[5]方法进行有价值、有参照性的探索和实践。最终对学校网络课程的教育改革提供可参考的意见。

打造专属学校的特色创新型管理模式：

在学习研究各类大数据应用案例基础上，研究各类数据如何与业务更好的结合以帮助高校建设智慧校园，探索创新创业新方向，形成具有各个学校特色的创新型管理模式。

研究思路以及相关技术

搭建数据分析平台对长期积累的数据和实时获取的海量数据进行深度挖掘，以获取高价值信息统计分析方法：利用回归分析、相关分析、主成分分析等方法，确定数据库中数据之间的函数关系或相关关系的算法。一般的数据挖掘算法分为统计学习算法、机器学习算法和基于数据库技术的数据挖掘算法。目前常用的数据挖掘方法[6]有：神经网络法、遗传算法、决策树方法、粗糙集方法、排斥反例方法、模糊集方法等。

数据挖掘的基本原理

数据挖掘主要是利用相关算法对数据进行各种处理，这就要求数据库系统对数据进行存储、索引和查询处理等方面的支持，所以说数据挖掘是借助数据库发掘知识的重要手段，数据挖掘的一个重要的作用就是从大量没有被发现的数据中选取潜在的更深一层的已被忽视的数据并显示其隐藏的结果的过程。将收集到的数据首先进行数据预处理通过可视化分析技术，采用数字、文字、图形、图表等多种数据表现形式来呈现数据，使数据更加清晰更直观地呈现在用户面前，实现数据的充分利用，将大量的数据转换成有用的信息并输出，发现数据背后的隐藏的有价值的信息，给出用户合理的建

议进而给用户的生活带来方便和便利。

通常来说，完整的数据挖掘流程主要分为以下几个步骤：第一步，数据预处理，主要包括数据清洗、数据集成、数据选择、数据变换以及数据归约操作。第二步，数据挖掘，使用智能的方法从大量预处理后的数据中提取和发现模式。第三步，结果可视化显示，对所挖掘的知识或规律，依托可视化或知识表示技术向用户呈现有效、新颖、潜在有用并易被理解的知识。

数据分析平台的系统搭建

（1）操作系统的选择

操作系统一般使用开源版的 RedHat、Centos 或者 Debian 作为底层的构建平台，要根据大数据平台所要搭建的数据分析工具可以支持的系统，正确的选择操作系统的版本。

（2）搭建 Hadoop 集群

Hadoop 作为一个开发和运行处理大规模数据的软件平台，实现了在大量的廉价计算机组成的集群中对海量数据进行分布式计算。Hadoop 框架中最核心的设计是 HDFS 和 MapReduce，HDFS 是一个高度容错性的系统，适合部署在廉价的机器上，能够提供高吞吐量的数据访问，适用于那些有着超大数据集的应用程序；MapReduce 是一套可以从海量的数据中提取数据最后返回结果集的编程模型。在生产实践应用中，Hadoop 非常适合应用于大数据存储和大数据的分析应用，适合服务于几千台到几万台大的服务器的集群运行，支持 PB 级别的存储容量。

使用开源组件的优势显而易见，活跃的社区会不断的迭代更新组件版本，使用的人也会很多，遇到问题会比较容易解决。

（3）选择数据接入和预处理工具

面对各种来源的数据，数据接入就是将这些零散的数据整合在一起，综合起来进行分析。数据接入主要包括文件日志的接入、数据库日志的接入、关系型数据库的接入和应用程序等的接入，数据接入常用的工具有 Flume，Logstash，NDC（网易数据运河系统），sqoop 等。对于实时性要求比较高的业务场景，比如对存在于社交网站、新闻等的数据信息流需要进行快速的处理反馈，那么数据的接入可以使用开源的 Strom，Spark streaming 等。

当需要使用上游模块的数据进行计算、统计和分析的时候，就需要用到分布式的消息系统，比如基于发布/订阅的消息系统 kafka。还可以使用分布式应用程序协调服务 Zookeeper 来提供数据同步服务，更好的保证数据的可靠和一致性。

数据预处理是在海量的数据中提取出可用特征，建立宽表，创建数据仓库，会使用到 HiveSQL，SparkSQL 和 Impala 等工具。随着业务量的增多，需要进行训练和清洗的数据也会变得越来越复杂，可以使用 azkaban 或者 oozie 作为工作流调度引擎，用来解决有多个 hadoop 或者 spark 等计算任务之间的依赖关系问题。

（4）数据存储

除了 Hadoop 中已广泛应用于数据存储的 HDFS，常用的还有分布式、面向列的开源数据库 Hbase，HBase 是一种 key/value 系统，部署在 HDFS 上，与 Hadoop 一样，HBase 的目标主要是依赖横向扩展，通过不断的增加廉价的商用服务器，增加计算和存储能力。同时 hadoop 的资源管理器 Yarn，可以为上层应用提供统一的资源管理和调度，为集群在利用率、资源统一等方面带来巨大的好处。Kudu 是一个围绕 Hadoop 生态圈建立的存储引擎，Kudu 拥有和 Hadoop 生态圈共同的设计理念，可以运行在普通的服务器上，作为一个开源的存储引擎，可以同时提供低延迟的随机读写和高效的数据分析能力。Redis 是一种速度非常快的非关系型数据库，可以将存储在内存中的键值对数据持久化到硬盘中，可以存储键与 5 种不同类型的值之间的映射。

（5）选择数据挖掘工具

Hive 可以将结构化的数据映射为一张数据库表，并提供 HQL 的查询功能，它是建立在 Hadoop 之上的数据仓库基础架构，是为了减少 MapReduce 编写工作的批处理系统，它的出现可以让那些精通 SQL 技能、但是不熟悉 MapReduce、编程能力较弱和不擅长 Java 的用户能够在 HDFS 大规模数据集上很好的利用 SQL 语言查询、汇总、分析数据。Impala 是对 Hive 的一个补充，可以实现高效的 SQL 查询，但是 Impala 将整个查询过程分成了一个执行计划树，而不是一连串的 MapReduce 任务，相比 Hive 有更好的并发性和避免了不必要的中间 sort 和 shuffle。Spark 可以将 Job 中间输出结果保存在内存中，不需要读取 HDFS，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。Solr 是一个运行在 Servlet 容器的独立的企业级搜索应用的全文搜索服务器，用户可以通过 http 请求，向搜索引擎服务器提交一定格式的 XML，生成索引，或者通过 HTTP GET 操作提出查找请求，并得到 XML 格式的返回结果。还可以对数据进行建模分析，会用到机器学习相关的知识，常用的机器学习算法，比如贝叶斯、逻辑回归、决策树、神经网络、协同过滤等。

（6）数据的可视化以及输出 API

对于处理得到的数据可以对接主流的 BI 系统，比如国外的 Tableau、Qlikview、PowrerBI 等，国内的 SmallBI 和新兴的网易有数等，将结果进行可视化，用于决策分析；或者回流到线上，支持线上业务的发展，成熟的搭建一套大数据分析平台不是一件简单的事情，本身就是一项复杂的工作，在这过程中需要考虑的因素有很多，比如：稳定性、可扩展性、安全性等。

主要算法简介

K-means 聚类算法是算法中最广泛使用的算法之一，对于大型数据的处理效率较高，特别是在样本分布呈现更明显的聚集现象时更是如此。该算法的主要思想是利用少数服从多数的原则，算法会将数据集分为 K 个簇，每个簇使用簇内所有样本均值来表示，将该均值称为“质心”。将采集到的一系列数据分成两部分，以相邻两个数据的差值最大的几个作为分界点。即聚类后，尽可能将相同的数据聚集在一起，尽量将不同类的数据分离开来。

粗糙集理论是一种有效的处理模糊数据分类方法，它能对不完整资料进行分析、推理、学习、和发现，具有较强的知识获取能力，同时该算法具有很强的非线性映射能力和泛化能力，能够从大量复杂的数据中学习规律因此它在解决非线性和高维模式识别等问题上具有独特的优势，在分类问题和回归问题的处理上都取得了很好的效果。

回归分析算法是统计学中一种经典的分析方法，该方法可以确定两种或两种以上变量间相互依赖的定量关系，借以从一系列散落的点中挖掘出某种规律关系。回归分析研究的主要对象是客观事物变量之间的统计关系，它是建立在对客观事物进行大量实验和观察的基础上，用来寻找那些看似不确定的现象的统计规律的统计。

参考文献：

[1]李步青. 基于大数据挖掘的学生行为分析系统的研究与开发[D].浙江农林大学，2021.DOI：10.27756/d.cnki.gzjlx.2021.000183.

[2]张亚琦. 基于数据挖掘技术的高校学生体能分析系统设计与实现[D].华中师范大学，2020.DOI：10.27159/d.cnki.ghzsu.2020.002634.

[3]张晓燕. 基于数据挖掘的高校学生行为分析系统设计与研究[D].西安电子科技大学，2019.DOI：10.27389/d.cnki.gxadu.2019.000692.

[4]张成勇.基于数据挖掘的学生行为监测与预警系统设计[J].山东农业工程学院学报，2018，35（12）：43-44.DOI：10.15948/j.cnki.37-1500/s.2018.12.021.

[5]孙杨博. 基于大数据挖掘的高校学生行为数据分析系统的研究与开发[D].华北电力大学（北京），2017.

[6]王铮钧.基于数据挖掘的高职学生个性化学习分析系统设计[J].计算机光盘软件与应用，2014，17（04）：229+231.

作者简介：方洪雨（2000-），女，河南南阳人，西北民族大学数学与计算机科学学院学生。