



文 / 吕红胤 于晨阳 苏涵 连德富 颜凯

近年来,有部分国内高校开始了数据挖掘的探索。比如,浙江大学通过对资产的归纳、整理,最终形成权威、全面的资产数据,并提供数据查询和分析服务。这些数据分析的结果能够提高教室、实验室等资源的利用率。复旦大学对特定群体的学生进行数据分析,发现了来自不同区域、不同背景学生成绩的显著差异性。清华大学开展了对优秀学生长成长追踪,观察成长路径的研究。以上可以看出,虽然这些高校已经开始通过简单统计和相关分析进行数据挖掘,但是这些数据没有完全打通并协同分析,而且分析缺乏深度性和广度性。

针对这些智慧校园建设的不足,电子科技大学利用大数据技术,研究设计了统一的校园大数据分析决策平台,并首次提出教育大数据一体化平台的概念。该平台不仅可以自动化地接入来自于校内的业务系统数据、资源使用数据、网络日志数据,而且能够接入来自于校外的互联网数据。

在数据治理和隐私保护之后,进行数据的标准化,存放在数据中心,同时通过应用程序接口的方式,进行数据认证式的共享,成功地解决了全校各个管理服务系统分割独立,数据共享困难的问题。更突出的是,一体化平台中的计算中心周期性地调用数据接口,进行包括挂科预警、贫困生挖掘等的大数据分析与预测,将分析与预测的结果提供给最上层的各类应用服务。电子科技大学运用教育大数据一体化平台提升高校管理服务水平和质量,既是一个集成不同部门、层次信息数据的系统性工程,又是一个面向世界一流大学创新管理体系。

教育大数据一体化平台体系架构

教育大数据一体化平台是数据整合、清洗与分析挖掘的开放平台。教育大数据一体化平台的架构如图1所示,它由三大中心组成,分别为数据中心、应用中心和运维中心。数据中心旨在打破数据孤岛,以80%的通用性对接现有的校园业务系

统,整合教务处、学生处、图书馆等部门的业务数据,同时对整合后的数据进行清洗、管理与隐私保护,形成标准化的校园数据。应用中心不仅提供隐私化的数据认证接口与大数据算法服务,而且整合如学生画像、教师画像等多样化的校园管理应用,并不断横向扩展,形成“一平台、多应用”的服务体系。运维中心采集系统应用日志以及机器数据并为其建立索引,构建强有力的搜索、分析和可视化能力,以提供日常运维支持。

通过三大中心的协同工作,一体化平台打破了学校各职能部门的数据孤岛,不仅可以为校园师生提供便捷、高效、精准的个性化引导和服务,而且能为学校开展精准化、智能化师生服务提供总体解决方案。基于该平台,电子科技大学已开发了多项成熟的应用服务,如学生画像、精准资助、失联告警、教师画像、科研探索、财务评估等系统。本文将“学生画像”系统和“教师画像”系统为例介绍教育大数据一体化平台的应用创新。

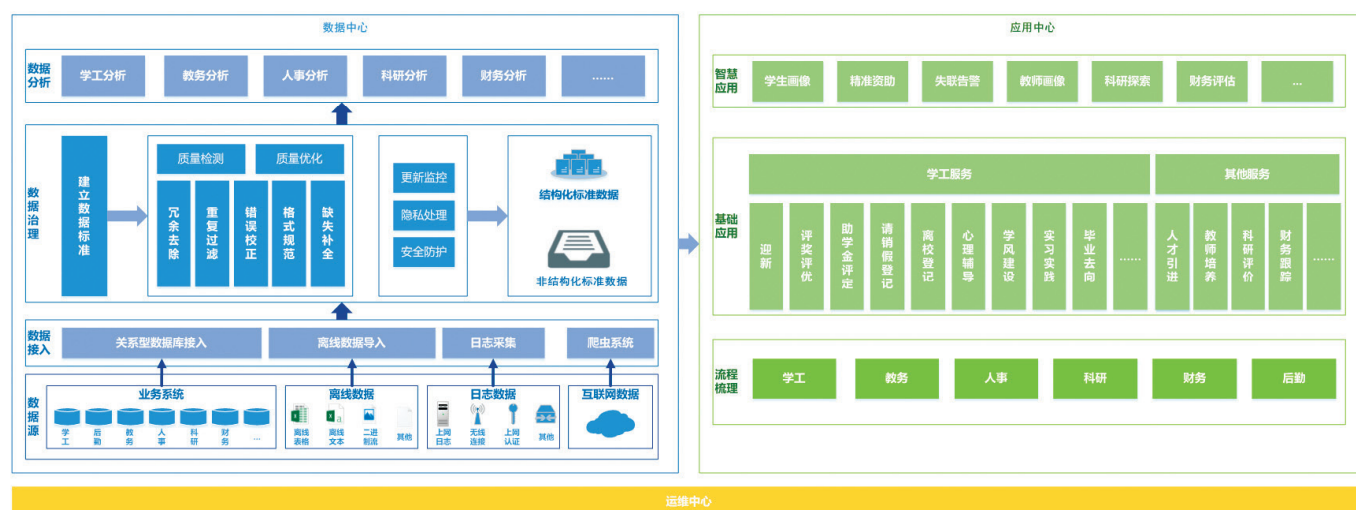


图1 教育大数据一体化平台体系的架构

学生画像：让学生得到精准的帮助

学生画像利用数据中心提供的标准化行为与内容数据，包括课程成绩、图书馆借阅、一卡通消费等，运用大数据分析手段，量化学生在规律性、努力程度、学习技能、经济状况、社交关系等多维度的特性，揭示学生成长轨迹，基于预测模型对学生的学业成绩、就业倾向、心理状况等进行预测，从而可以为学校对学生进行个性化与精准化的教育管理与引导提供重要依据。下面针对以下几个方面进行进一步阐述。

1. 学业成绩预测

成绩特别好、特别差和成绩突变的学生是教育者最关心的人群。直观地说，学生成绩和他们的基础知识掌握情况有密切关系。特别地，基础知识的掌握好坏可以影响到学生在相关课程中的成绩。为此，借助以往课程成绩信息，基于矩阵分解的降维技术，分析学生对具体基础知识的掌握程度，并获得每门课程所含有的知识体系。基于这些信息，该系统不仅可以预测出每个学生在其他课程的得分，也可以预测出这个学生在其他课程的挂科可能性。这个算法背后的思想是通过分析课程之间在知识体系上的相关性来进行成绩预测或者挂科预测。比如某学生在以往课程中

曾修过微积分Ⅰ但成绩不好而且挂科了，那么当他在修读微积分Ⅱ的时候，预警系统就会预警该学生的挂科可能性比较大，提示教育者及早发现问题并进行干预，尽可能避免挂科问题。因而，挂科预警实现了从后置性应急管理转变为前置性预警引导。

由于课程成绩数据更新频率低，挂科预警无法实时更新预测结果。然而，根据我们研究发现，学生成绩也和自身的学习以及生活行为习惯特性密切相关。学生在校园内的行为习惯的变化是可以实时监测的，对于预测成绩的变化特别有价值，所以为实时地预测学生成绩提供可能。该学业成绩预测系统借助了两个影响成绩最显著的行为特性：努力程度和生活规律性对学生进行刻画。用学生去教学楼打水、出入图书馆的次数量度其努力程度。该次数可以反映学生上自习或者上课的频率，间接反映了学生花在学习上的时间。因而，去教学楼打水次数高的学生、频繁去图书馆的同学成绩较好。用学生出入宿舍、吃早饭、洗澡等行为习惯数据衡量其生活规律性。生活规律性与心理学中大五人格中的尽责性密切相关，而心理学的研究表明，尽责性高的学生，表现为自律、细心、有条理性，有更好的学习成绩。因而为研究行为规律性和成绩的相关性提供了理论支

持。在实证研究中，我们发现，按时吃早餐的学生往往会显出更好的成绩。这背后的原因可能是早饭进餐这个事件在个体层面具有很强的随意性，因而对个体的自律与自控能力有较高的要求。

除了努力程度和生活规律性这两个指标外，学生的图书借阅历史也和成绩密切相关。因为借阅的图书可以反映出学生的兴趣爱好，不同的兴趣爱好会对成绩有不同的影响。研究发现成绩好的学生借阅专业方向的进阶读物，而成绩差的同学喜欢借阅诸如小说的各类课外读物。此外，量化了每本书和成绩的相关性，使得可以更加精准地进行图书推荐。

最后，基于努力程度、生活规律性、基础知识以及兴趣爱好这些特性，设计多任务迁移学习算法来进行未来成绩的预测。该算法不仅通过多任务特性考虑了特征相关性在学院之间的差异性，而且还通过迁移学习特点考虑了不同学期之间相关性的变化。得出学生的成绩预测分析，能很好地反映出学生成绩的未来走势。

2. 职业倾向预测

数据分析发现学生家庭经济状况、技能掌握情况、兴趣爱好是影响学生就业去向的关键因子。家庭经济状况主要通过分析学生消费数据，建立起消费数据时间序列，度量消费的波动性、周期性与冲动性

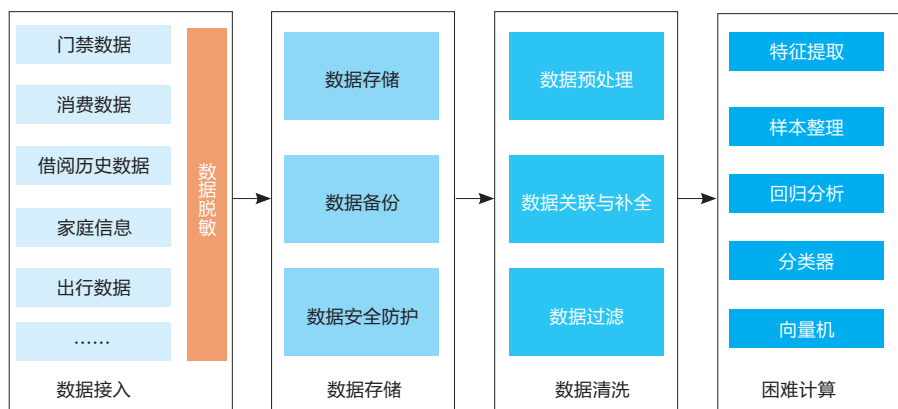


图2 精准资助系统

等。通过分析学生的课程成绩，判断出学生所掌握的专业技能来预测职业选择。最后，不同职业倾向的人的图书借阅偏好会表现出较大差别，如考研的学生往往倾向于数学等考研科目的书，出国的学生则对国外文学和历史、雅思和托福类书籍情有独钟。通过对以上三类数据的收集、分析和挖掘，就能很好地对学生未来就业倾向进行精准的预测，有利于就业指导教师为毕业生提供个性化的就业指导。

3. 社交网络构建与挖掘

社交网络的构建是通过分析学生之间校园行为轨迹相似性来实现的，特别是统计学生在地点共现（短时间内出现在同一地点）的频率，分析共现的显著性。同时通过大数据的挖掘与分析，给不同个体赋予独特属性标签，如专业、性别、民族等，并以此分析出其个性化的社交需求，对其社交圈进行刻画。以此为其定制与之适应的社交网络推荐与信息推送服务，并将个体数据分析结果展现给学校学生管理工作为其提供更好的管理工作参考。系统为学生匹配有相同行为习惯及兴趣爱好的个体与其建立社交关系，以此构建具有共同文化认可的社交群体，更好地服务于学生个体的社会交往需要，对学生成长起到良性正面的促进作用。

4. 精准资助认定

准确定位扶助对象是实施“精准扶贫”管理服务的前提，客观、动态和多维度大数据整合库，是实施“精准扶贫”的基

础。对贫困生判定的影响因子主要有以下几类：（1）学生家庭信息，包括学生家庭成员组成、家庭成员就单位、成员学历、家庭年收入、负债金额等基本家庭信息。（2）历史资助信息，收集学生以往获得的资助信息，为其建立基础的数据库，以便于查询学生是否获得资助、资助金额以及经济困难情况。（3）在校一卡通消费数据特征，包括一卡通平均单次消费金额、单次充值金额及充值间隔、月消费总额、逐月消费变化趋势、消费时间段规律等。国内某大学就是利用这方面数据判别贫困生，将1个月消费次数在60次以上，月消费金额在200元以下的学生认定为贫困生。这种单一的判别方式误差会很大，影响判定贫困生的因素很多，应该综合考虑。（4）资助获得后消费习惯改变的数据信息，如在获得资助之后出现冲动消费及大额消费的数据信息。（5）他人的客观评价，收集来自于辅导员及周围同学日常评价并转换为量化数据。通过对以上数据的收集和处理分析，构建起家庭经济困难学生专项大数据库，在大数据分析的基础上，建立起经济困难学生精准识别系统如图2所示。

它可以实现两个主要功能：一是识别虚假贫困生，能够有效识别家庭情况较好的学生申报贫困生冒领国家资助的现象，对于这类学生取消其资助资格并降低其信用评级；二是发现潜在贫困生，个别

家庭贫困学生由于自尊心较强等因素，往往不主动申请资助，使得这类学生难以通过传统的方式来发现。现在通过精准资助识别系统，能迅速地发现此类学生，采取发放隐性补助的方式进行帮扶（如每月定时向其银行卡中转入一定数额资金），以及实施动态补助等方式，很好地帮助困难资助管理工作对学生资助信息实施动态管理。通过以上流程构建起数据收集、存储、分析和数据挖掘为一体的大数据精准筛选、甄别和定位系统，客观公正且及时、动态和准确地识别校园亟待资助和扶助对象。同时建立学生网络诚信档案体系，坚持精准资助与诚信教育结合，核实学生资助申请材料并将其转化为大数据记录，将申请资助过程中的不诚信行为记入学生诚信档案并做好大数据标识，为管理者切实做好精准资助工作提供有力的技术支持。该精准奖助识别系统已经在电子科技大学等高校推广应用，取得了很好的应用效果。

教师画像：为教师学术发展提供信息

“教师画像”系统是通过打通校园管理层面不同业务系统之间的数据孤岛，实现以教师为主体的数据挖掘，对教师个体及群体的人事信息、科研项目、学科成果及教学状况进行精准刻画，服务于高校人事、科研管理的数据支撑系统。它具有三大优势，一，全面：建立各门类数据桥梁，全盘掌握学校的人事、科研、教学现状；二，高效：打通数据流通渠道，实现信息聚合，提高数据汇总效率；三，前瞻：挖掘数据相关性，发现数据潜在价值，为管理者提供决策依据。

“教师画像”系统应用方案如图3所示，它首先整合校内外数据，内部数据包括高校自身产生的项目、人员、经费、设备等数据信息；外部数据包括各大公开的科研成果数据库，以及各大知名高校的科研人才数据等。然后对这些数据进行清洗、转换、重构提取有效信息并将提取后的信息存入数据仓库；使用关联分析技术

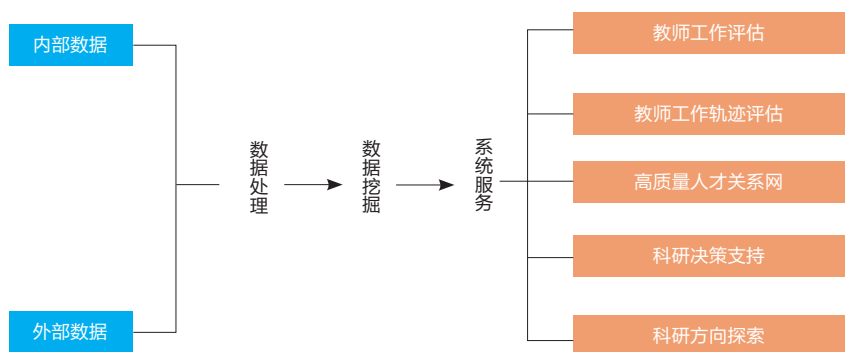


图3 “教师画像”的应用方案

对科研管理系统、财务系统、人事系统以及基于互联网的大型科技文献数据库、专利库等数据资源进行关联分析，找出数据的相关性，提取有价值的信息。将提取的信息应用在教师工作评估、教师成长轨迹分析、高质量人才引进建议、学科前沿研究方向探索、科技评价方法完善等服务上，为解决高校人事管理工作两大核心问题“外引”、“内培”提供建设性意见，为传统的专家定性决策管理提供广泛的、深入的数据支持。该画像系统的主要应用表现在以下几个方面：

1. 教师工作绩效自动评估

教师工作绩效自动评估系统通过整合人事、科研、财务、教学等多门类数据信息，采用教师经费效益、经费使用情况、成果影响力、成果转化、同行意见等多维度的评价因素，并支持不同单位结合各单位的实际情况调整评价模型，全方面呈现教师在科研和教学工作的成绩，从而为教师的入职、晋升、聘任、培训和奖惩提供量化决策依据。避免了传统教师绩效评估受到的人为因素影响，使得评估结果更加客观、准确。

2. 工作轨迹评估

传统的教师发展研究主要停留在经验层面，传统的教师信息系统只能看到单一的信息，而“教师画像”是利用大数据刻画教师，基于教师基础信息数据（包括学习经历、海外经历、工作经历、岗位聘任经历、科研项目、学科成果等），围绕教师职业素养、专业知识、专业能力、工作

绩效等多方面构建教师成长轨迹，并分析影响教师的发展因素，从而制定个性化成长方案，如预测发表论文数量、能否入选人才计划、优秀青年教师等。寻求适合教师的个性化发展路线，引导教师可持续发展，实现教师个人与学校发展的“双赢”。

3. 学术圈层研究

搜集学术、社交网络等多门类广泛的数据，如搜索每个文章的合作者，构建合作者网络，挖掘隐藏其中的人才关系。实现以人才为中心的数据整合，构建各学科的学术圈层网络。利用该网络一方面可以为校内教师寻找帮助自己提升的外部老师，另一方面挖掘有潜力的学术新星，帮助高校人事部门有针对性地获悉人才有效信息，成功猎取高质量人才。

4. 科研热点

科研工作不能闭门造车，及时掌握时下国内外的科研热点及难点，结合自身

能力与学科特点进行有效的科研工作对于科研工作者至关重要。而在海量数据中分析当下学科研究的热点及前沿，单凭人力是很难做到的，需要借助于大数据分析技术。科研热点分析首先收集国内外论文数据库、专利申报及项目审批等科研热点数据信息，再对过滤后的海量数据利用大数据算法进行挖掘分析，最后有效预测科研热点，并结合高校学科建设现状与特点，分析各学科前沿研究方向。为科研工作者的科研工作提供有力的科研数据支撑，为其选定符合自身学科特点的科研发展方向提供有效建议，帮助其有效定位自身科研工作努力方向及深度。

目前，一场以云计算、大数据、物联网、移动应用、智能控制技术为核心的“新IT”浪潮风起云涌。大数据技术在教育领域的广泛应用，必将催生教育领域的深刻变革。对此，电子科技大学抓住机遇，迎接挑战，利用大数据技术开创性地构建了教育大数据一体化平台，基于此平台创新性地分别开发了服务于学生和教师的“学生画像”和“教师画像”系统。利用“学生画像”系统提供了精准预测学生成绩，就业倾向预测和指导，助力学生精准资助等个性化、精准化的管理服务；利用“教师画像”系统提供教师工作评估、教师成长轨迹分析、高质量人才引进建议、学科前沿研究方向探索、科技评价方法完善等服务。CEN

（作者单位为电子科技大学大数据研究中心）

曙光公司发布大数据技术解决方案

本刊讯 日前，国内大数据技术综合服务商曙光公司发布其最新的大数据技术解决方案——曙光方舟大数据融合分析平台。

据 IDC 预计，到 2020 年，全球就将拥有 40ZB 的数据量，而且如此规模之下，将是结构化数据、半结构化数据和非结构化数据并存。要从海量的大数据中攫取数据价值，无异于沙里淘金。

“行业用户要快速响应大数据的时代挑战、提高大数据处理效率，一款涵盖大数据全生态系统的标准化大数据工具必不可少。”曙光公司总裁助理王海荣提出。当前各行业在从 GB 过渡到 PB 级数据分析过程中会遇到各种运维管理问题，“方舟大数据融合分析平台”就瞄准了这一“痛点”，致力于提供用户大数据融合、分析的“一站式”解决方案。

此次发布的曙光方舟大数据融合分析平台，在发布前就已有多个客户进行应用，如江苏省通信管理局、徐州公安、陆军指挥学院、镇江国税局等。“我们将进一步推进曙光方舟在更多城市和行业的部署，以建立大数据汇聚、融合、共享到实现智慧城市的完整价值链。”王海荣说。