教育数据的挖掘、分析、应用

□ 魏顺平

【摘 要】本文围绕教育数据挖掘这一主题，简要分析了教育数据挖掘的内涵及价值；介绍了教育数据挖掘的基本过程、主要数据来源、所用技术与工具以及主要应用领域，这些应用领域包括E—Learning数据挖掘、E—Management数据挖掘和E—Research数据挖掘；最后，通过具体案例，涉及学习路径分析、学科知识结构分析等教育数据挖掘任务，来说明各种不同工具包括SSAS、SPSS在开展教育数据挖掘时的用法及效果。

【关键词】 大数据；教育数据挖掘；学习分析；工具；案例
【中图分类号】G40—057 【文献标识码】A
【论文编号】1671—7384（2013）10—0018—04

在过去的10余年，随着教育信息化工作的大力推进，特别是数字化校园建设和网络高等教育的大力推进，在教育领域已经部署了众多的软件系统，并且在这些软件系统中存储着海量的教育数据。如何利用这些教育数据，使这些数据转变为信息、知识，并为教育决策、教学优化服务，这便是教育数据挖掘（Educational Data Mining）所关注的内容。近年来，教育数据挖掘在国内外得到了迅速发展，下面对教育数据挖掘的兴起、内涵、价值、过程、方法及应用领域进行概要介绍，以使大家对“教育数据挖掘”有一全面的认识。

内涵与价值

自2005年以来，国际上许多计算机应用相关会议均设置了有关教育数据挖掘（Educational Data Mining）的研讨会，如AAAI'05、AIED'05、ITS'06、AAAI'06、AIED'07、UM'07以及ICALT'07 等会议。2007年，欧洲技术促进学习协会（EATEL）在希腊克里特岛举办第二届欧洲技术促进学习会议（EC-TEL07），期间举办了“Applying Data Mining in e-Learning”研讨会（ADML'2007）。在这一研讨会之后，该领域研究者组成国际教育数据挖掘工作组（http://www.educationaldatamining.org），创办在线学术期刊《教育数据挖掘杂志》，并从2008年开始每年召开“教育数据挖掘国际会议”(EDM conference)。2011年，国际教育数据挖掘协会（IEDMS）成立。

EDM2008会议论文集在其前言中对“教育数据挖掘”的描述是，“教育数据挖掘是一个将来自各种教育系统的原始数据转换为有用信息的过程，这些有用信息可为教师、学生、家长、教育研究人员以及教育软件系统开发人员所利用”。通过这一描述我们不难发现，教育数据挖掘致力于开发出一系列数据挖掘方法，将这些方法运用于挖掘来自教育系统的独特数据，能够更好地理解学生及其所在的教育系统。这里的“来自教育系统的数据”既包括师生在使用互动学习环境（如网络教学平台）时产生的数据，也包括学校在开展教育教学管理过程中产生的数据。

教育数据挖掘也可被看作嵌入已有教育系统的一个新的模块，并与教育系统中的各种要素产生良性互动，最终实现改进学习的目的。它在教育系统中所处的位置及作用如图1所示。

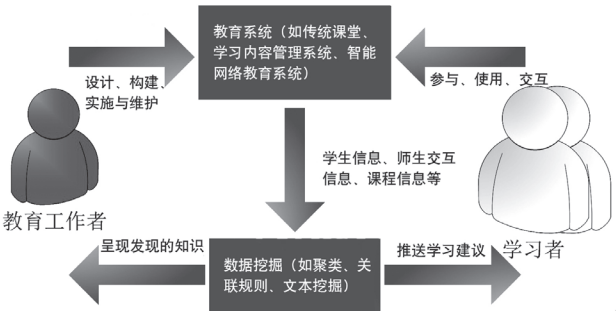

图1 数据挖掘在教育系统中的位置与作用

针对不同的人群，教育数据挖掘有其特定的价值。

对于学习者而言，教育数据挖掘的作用体现在：向学习者推荐有助于改进他们学习的学习活动、学习资源和学习任务，向学习者推荐好的学习经验等。这些建议可以通过分析这些学习者完成的行为以及与之相似的学习者完成的行为来取得。

对于教育工作者而言，教育数据挖掘的作用体现在：向他们提供更多更客观的反馈信息，使他们能够更好地调整和优化教育决策，改进教育过程，完善课程开

发，根据学习者的学习状态来组织教学内容、重构教学计划等。

过程与应用

1. 挖掘过程

教育数据挖掘过程与数据挖掘的一般过程是一致的，也由数据准备、数据预处理、数据挖掘、模式解释等几大部分构成，其过程如图2所示。

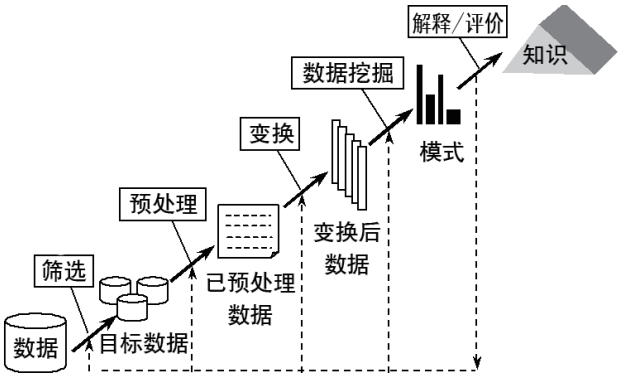

图 2　数据挖掘项目实施流程

（1）数据准备：了解数据挖掘应用领域的有关情况。包括熟悉相关的背景知识，明确使用者需求。

（2）数据选取：数据选取的目的是确定目标数据。根据使用者的需要从原始数据库中选取相关数据或样本。在此过程中，需要利用一些库操作对数据库进行相关处理。

（3）数据预处理：对步骤（2）中选出的数据进行处理，检查数据的完整性及数据一致性，消除噪声，滤除与数据挖掘无关的冗余数据，根据时间序列和已知的变化情况，利用统计等方法填充丢失的数据。

（4）数据变换：根据知识发现的任务对经过预处理的数据进行再处理，主要是通过投影或利用数据库的其他操作减少数据量。

（5）确定数据挖掘目标：根据使用者的要求，确定数据挖掘要发现的知识类型。对数据挖掘的要求不同，在具体的知识发现过程中需要采用的知识发现算法也会有所不同，如分类、总结、关联规则、聚类等。

（6）选择算法：根据确定的任务选择合适的知识发现算法，包括选取合适的模型和参数。

（7）数据挖掘：这是整个过程中很重要的一个步骤。运用前面选择的算法，从数据库中提取使用者感兴趣的知识，并以一定的方式表示出来是数据挖掘的目的。

（8）模式解释：对发现的模式（知识）进行解释。经过使用者或机器评估后，可能会发现这些模式中存在冗余或无关的模式，此时应该将其剔除。如果模式不能满足使用者的要求，就需要返回到前面的某些步骤中进行反复提取。

（9）知识评价：将发现的知识以使用者能了解的方式呈现给使用者。

2. 数据来源

根据教育活动中技术手段的差异，教育数据挖掘的数据来源可分为传统教育数据和远程教育数据。根据教育机构中业务活动的不同，教育数据挖掘的数据来源又可以分为教学数据、管理数据、科研数据等。

因此，从技术手段和业务类型两个维度对教育数据挖掘的数据来源进行分类，可以得到如表1所示结果。

表1　教育数据挖掘的数据来源

技术手段 / 业务范围	传统教育	远程教育
教学	Ⅰ 课堂教学数据	Ⅱ 远程教学数据
管理	Ⅲ 数字化管理系统的数据	
科研	Ⅳ 科研数据库中的数据	

第Ⅰ类数据——课堂教学数据。这一类数据似乎难以获取，但是随着各类智慧校园、智慧教室建设项目的推进，越来越多的情境感知设备、泛在技术系统将在校园、教室中部署，届时课堂教学中也会产生并自动保存海量的数据。

第Ⅱ类数据——远程教学数据。远程教学可基于不同形式的数字化学习环境开展。远程教学数据可来自数字化学习环境所产生的各种日志，既可以是保存在服务器上的日志数据，也可以是保存在客户端上的日志数据。

第Ⅲ类数据——数字化管理系统的数据。这一类数据指的是教育机构使用数字化管理系统录入、保存和管理的数据。这类数据结构良好，可批量采集，是数据挖掘的理想对象。

第Ⅳ类数据——科研数据库中的数据。目前与科学研究有关的信息资料，许多已被转换为数字化形式，存于数据库中，并可通过各类检索系统检索使用。这类数据结构良好，可批量采集，也是数据挖掘的理想对象。

3. 技术与工具

由于教育领域中的数据也逐渐呈现出大数据的“4V”特征，即数据量巨大（Volume）、结构化数据、半结构化数据和非结构化数据并存（Variety）、数据价值密度低（Value）、数据的产生与处理加速（Velocity），这就要求用到的教育数据挖掘技术是复杂

多样的。相关技术涉及内容分析、话语分析、社会网络分析、系统建模等技术以及统计分析与可视化、聚类、预测、关系挖掘、文本挖掘等一系列数据挖掘方法。

常见的工具有很多，如支持对原始帖子进行标注或编码、交叉引用和简短评论的工具，包括Nvivo、Atlasti；支持基本的基于词典的文本分析的工具，如CATPAC、LIWC；专门的内容分析工具，如北京师范大学知识工程研究中心开发的智能化内容分析工具VINCA；专门的社会网络分析工具，如UCINET；用于系统建模的工具，如 Coordinator系统建模工具；专门的数据挖掘工具，如SSAS、Weka、SPSS；等等。

4. 应用领域

教学、管理、科研是教育机构的基本活动，将数据挖掘应用于这些基本活动时，由于业务流程、关注对象上的差异，对数据挖掘的需求及数据挖掘在其中的应用方式有所不同。因此，根据数据挖掘应用的业务领域，可以将教育数据挖掘进一步细分为E-Learning数据挖掘、E-Management数据挖掘和E-Research数据挖掘等。

E-Learning数据挖掘指的是将来自各种教学和学习软件系统（主要是网络教学平台）的原始数据转换为有用信息的过程，这些有用信息可为教师、学生、家长以及E-Learning软件系统开发人员所利用，以实现对学生及其所接受教育的更好理解，并可据此采取更有针对性的管理和教学优化措施。典型的E-Learning数据挖掘应用有管理者视角下在线教学绩效评估、辅导教师视角下在线学习形成性评价、研究人员视角下在线学习规律发现模式等。

E-Management数据挖掘指的是将来自各种教育管理系统（主要是教育管理信息系统）的原始数据转换为有用信息的过程，这些有用信息可为教育管理人员以及教育管理系统开发人员所利用，以实现对管理对象（学生、教师）及各种业务流程的更好理解，并可据此优化各项管理工作。一些典型的E-Management数据挖掘应用包括教师绩效评价、人才引进决策、招生决策、就业预测、职业规划、辍学分析、毕业生追踪、课程设置、教育决策支持系统等多个方面，这些应用归纳起来就是“数据挖掘优化教职工管理工作”和“数据挖掘优化学生管理工作”两大方面。

E-Research数据挖掘指的是将来自各种科研数据库（如文献数据库、政策数据库、语料库等）的原始数据转换为有用信息的过程，该过程可提高研究效率，并优化研究成果的呈现方式，产生的有用信息可为教育研究人员所利用，以实现全面、快速、准确了解某一研究领域的现状，并预测未来发展方向。一些E-Research数据挖掘典型应用包括基于期刊论文数据库的教育学科知识图谱构建和基于法律法规库的教育政策文本分析。

典型案例

下面以两个教育数据挖掘案例来介绍教育数据挖掘工具的使用及其效果。

1. 应用SSAS开展学习路径分析

作者选取中央广播电视大学开放教育学生的入学课程《开放教育学习指南》网络课程作为样本来说明应用SSAS开展学习路径分析的做法及效果。这里采用Microsoft顺序分析和聚类分析算法，数据来源则是用户每天浏览课程页面产生的过程数据，如表2所示。

表2 课程浏览样例

浏览日期	学　号	浏览页面所属模块	浏览顺序号
04 27 2010	1032001200001	课程章节	1
04 27 2010	1032001200001	课程章节	2
04 27 2010	1032001200001	体验区	3
04 27 2010	1032001200001	体验区	4
04 27 2010	1032001200001	体验区	5

在构建挖掘模型时，以“浏览日期+学号”作为键值，以“浏览顺序号”作为序列键值，以“浏览模块”作为预测值来构建挖掘结构，并应用Microsoft顺序分析和聚类分析算法，得到各模块（PAGETYPE）之间的转换概率和展示各模块间跳转情况的状态转换图（图3）。

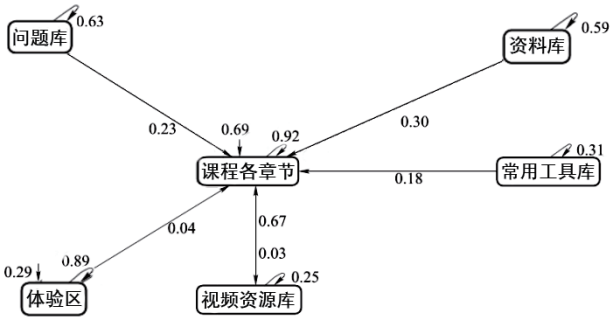

图3　各模块的跳转情况的状态转换图

从各模块直接的转换概率可以了解到，学生在登录网络课程后，最有可能先浏览的是“课程章节”模块，其次是“体验区”模块，极少从其他模块开始浏览。

从图3中可知，学生一旦进入“课程章节”或“体验区”模块，则主要是在本模块活动，而在中间几乎不去访问其他模块直至离开。“问题库”、“资料库”、“常用工具库”等学习辅助模块在课程主体部分即“课程章节”学习过程中几乎没有用到（学生偶尔从这三个

模块进入，然后转入“课程章节”模块，而不是反过来）。这反映了学生学习方法的重要特点，也反映课程的链接设计还有改进的空间。

2. 应用SPSS开展学科知识结构分析

作者选取一项针对农村教育研究现状调查的研究来说明应用SPSS开展学科知识结构分析的过程及效果。这项农村教育研究现状调查以2008年至2012年有关农村教育的核心期刊论文为分析对象。

一般而言，一篇论文有4到5个关键词。多个关键词在同一篇论文中出现，表明这些关键词之间可能存在一定的关系。如果两个或两个以上关键词在多篇论文中共现，则表明这些关键词关系稳定，可视为一种必然关联。因此，作者通过对关键词间关联关系的调查分析来了解我国农村教育研究领域中存在的较为稳定的研究内容组合，从而把握农村教育知识体系。

某个关键词的文献频次是其与其他关键词共现的基础。因此，作者选取前面提出的高频关键词作为农村教育知识体系构建的元素，探索这些关键词的共现关系。这些关键词共篇关系如表3所示。

表3 高频关键词共篇关系（局部）

	教师队伍	教师教育	教师培训	教师素质	教师专业发展
农村地区			1	2	3
农村发展					
农村基础教育		2		2	
农村教师	1	2			10
农村教师队伍			2	3	2
农村教师队伍建设			1	4	
农村教育	5	1	4	5	9

如表3所示，表中的数字表示横纵两个坐标的关键词的共现频次，如“农村教师”和“教师专业发展”的共现频次为10，则表明这两个关键词在10篇论文中共同出现，反映了关于“农村教师”的专业发展得到了多项研究的关注。

基于共词矩阵，进一步构造相关矩阵、相异矩阵。引入Ochiia相似系数法进行计算，将共词矩阵转换成相关矩阵。具体计算公式为：

Ochiia系数=Nij/(Ni*Nj)1/2

其中Ni和Nj分别代表关键词i和j出现的次数，Nij指关键词i和j共现的次数。在所得的相关矩阵中，用“1”减去相关矩阵中的每个数据，得到表示两词间相异程度的相异矩阵。基于相异矩阵，采用SPSS19.0统计分析工具，使用其中的Multidimensional Scaling（ALSCAL）算法进行多维尺度分析，实现关键词聚类，发现学科知识结构，得到图4所示结果。

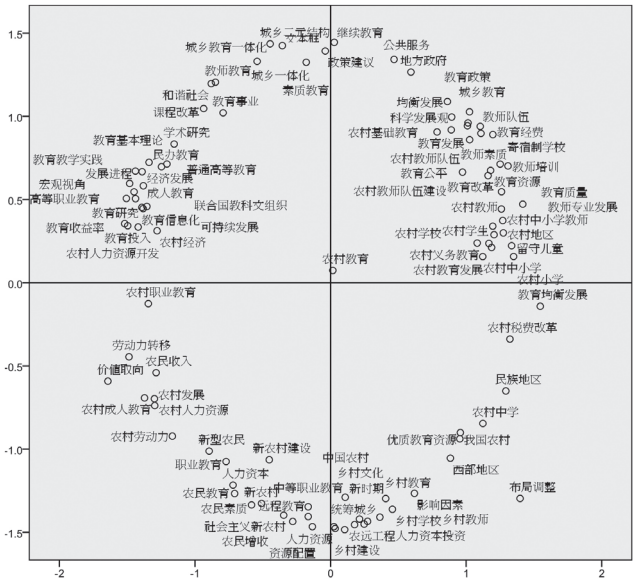

图4 对高频关键词的多维尺度分析

我们能够发现三块较为集中的研究领域，分别是象限1、象限2和象限3、4的交叉处。其中象限1属于“农村基础教育”范畴，关注“学校”、“教师”、“留守儿童”、“教育公平”等问题；象限3、4属于“农村职业教育”和“农村成人教育”范畴，关注“人力资本”、“农民素质”、“职业教育”等问题；象限2属于“农村教育相关理论与政策研究”范畴，关注“基本理论”、“高等教育”、“经济发展”、“教育投入”、“教育信息化”等话题。@

参考文献

[1] Ryan Shaun Joazeiro de Baker, Tiffany Barnes, Joseph E. Beck (Eds.). The 1st International Conference on Educational Data Mining Proceedings[DB/OL].http://www.educationaldatamining.org/EDM2008/index.php?page=proceedings,2012—7—24.

[2] 葛道凯，张少刚，魏顺平 著. 教育数据挖掘：方法与应用[M]. 北京:教育科学出版社，2012.

[3] 魏顺平. 学习分析技术:挖掘大数据时代下教育数据的价值[J]. 现代教育技术,2013(2).

[4] 魏顺平. 在线学习行为特点及其影响因素分析研究[J]. 开放教育研究,2012(4).

[5] 张少刚,魏顺平. 我国农村教育研究现状调查——基于近五年中文核心期刊论文的分析[J]. 天津电大学报,2013(01).

（作者单位：国家开放大学现代远程教育研究所）