国内外教育数据挖掘研究现状及趋势分析

李 婷 傅钢善

（陕西师范大学 新闻与传播学院，陕西西安 710062）

【摘要】教育数据挖掘是一个新兴的、备受关注的研究领域。文章运用文献计量与内容分析法，对国内外公开发表的关于教育数据挖掘的文献进行统计分析，把握其发展脉络及研究现状，探讨研究中的关键内容，并展望该领域未来的研究趋势，为进行教育数据挖掘的研究与实践提供参考。

【关键词】教育数据挖掘；研究现状；关键内容；发展趋势

【中图分类号】G40-057 【文献标识码】A 【论文编号】1009—8097（2010）10—0021—05

引言

对数据挖掘的研究始于20世纪80年代，现在已经成功地应用于商业、金融业和市场营销等领域。教育信息化的发展和网络远程教育的开展，导致教育领域的各类数据迅速增长，如何从海量的数据中挖掘出对教育者和学习者有用的信息，以提高教育管理绩效和学习绩效，这一问题的提出促使教育数据挖掘研究的出现。

信息技术在教育领域的广泛应用促进教育科研信息化的发展，信息化科研（e-research）是指信息技术所“使能的”科学研究实践，其发展经历了以下三个阶段（Halfpenny，2007）：一是对技术的研究阶段，如技术创新、技术应用、数字鸿沟等；二是利用技术开展研究的阶段，如计算机辅助的访谈、技术支持的数据分析、社会性网络分析等；三是技术使能的研究阶段，如数据挖掘、数据处理、数据整合、数据分析、模拟、可视化等[1]。教育数据挖掘正是数字化教育研究的体现，也是教育信息化发展的必然需求。自2005年起人工智能(AAAI)、人工智能教育应用(AIED)及智能导师系统(ITS)等国际会议开展了多次“教育数据挖掘”主题研讨会，2008年在加拿大召开了第一届教育数据挖掘国际学术会议，2009年在北京师范大学举行的第五届高级数据挖掘与应用国际会议首次加入“数据挖掘在教育中的应用”主题，2011年7月将在荷兰埃因霍温举办第四届教育数据挖掘国际会议，并且已经成功创办了专门的电子期刊——教育数据挖掘杂志(JEMD)。

一 教育数据挖掘概述

数据挖掘(DM)，与数据库中的知识发现(KDD)同义，指从大型数据库中提取出有意义的、隐含的、先前未知并有潜在价值的信息或模式的非平凡过程[2]。DM 的研究内容包括基础理论研究和应用研究两大类，其中基础理论研究包含方法、功能、算法以及数据挖掘系统和软件的建设等方面，应用研究的重点不在于数据挖掘技术本身，而在于成果应用，不同领域开展不同的应用研究。

教育数据挖掘（EDM）指应用数据挖掘方法从来自于教育系统的数据中提取出有意义的信息的过程，这些信息可以为教育者、学习者、管理者、教育软件开发者和教育研究者等提供服务[3]。EDM 主要研究数据挖掘在教育领域中的应用，从 EDM 研究领域的角度分析，EDM 研究包括“在教学研究中的应用”和“在教务管理中的应用”两个子类；从数据来源的角度分析，EDM 研究包括“在传统教育中的应用”及“在网络教育中的应用”两个子类，结合这两个方面，可进一步对 EDM 研究内容进行细分，如图 1 所示。

图 1 教育数据挖掘研究内容划分图

数据挖掘技术可应用于招生、就业、后勤、图书馆管理、人事管理、设备管理、师资管理等方面，有助于管理者做出科学的决策。EDM 更重要的意义在于指导和改善学习，提高教学质量，尤其是在网络教育中的应用。网络学习环境不能像传统课堂中通过面对面交流得到反馈，却能够记录学生的大多数学习行为，通过对网络学习系统中的学习者登记信息、日志文件、过程性数据、交互信息及管理数据等进行挖掘，如有多少人访问了该页面、来自哪里、哪些页面是最受欢迎的、用户访问完该页面后下一步可能的访问页面是什么等等，确定学习者个体或群体的特征模型，管理和监控网络学习过程，支持学生的个性化学习，指导教学及课程设计，构建有效的学习模式，改进系统及修改站点、建设适合学习者的资源，进行教与学的评价，为页面推荐和智能化学习提供服务，也可用于网络学习学生流失分析、进行教学决策等。

二 EDM 研究的现状分析

1 研究样本的检索及变化趋势统计分析

对国外文献，选取教育数据挖掘相关会议论文集，并且以 educational data mining 为检索词对 Science Direct 外文期刊数据库和 Google 中 2009 年 12 月以前的文章进行检索，筛选与数据挖掘在网络教育中的应用相关的文章。对国内文献，分别以“数据挖掘”和“网络教学”、“远程教育”、“网络教育”等为关键词和索引对中国知网中 2009 年 12 月以前的文章进行高级检索，统计时剔除和主题关系不大与重复的文章，结果如表 1、2 所示。

表 1 国外研究论文总量的时间分布比例

年份 / 总量	2000 及以前	2001	2002	2003	2004	2005	2006	2007	2008	2009	总计
学术论文	12	4	7	9	18	23	28	35	54	65	255
百分比	4.71	1.57	2.75	3.53	7.06	9.02	10.98	13.73	21.18	25.49	100

表 2 国内研究论文总量的时间分布比例

年份 / 总量	2000 及以前	2001	2002	2003	2004	2005	2006	2007	2008	2009	总计
学术论文	0	0	4	5	7	13	16	20	17	21	103
学位论文	0	0	0	1	3	7	8	10	6	7	42
累积量(篇)	0	0	4	6	10	20	24	30	23	28	145
百分比	0	0	2.76	4.14	6.90	13.79	16.55	20.69	15.86	19.31	100

为了更加直观地分析国内外相关研究的趋势，绘制如图 2 所示的分布态势图，该图显示国内外对数据挖掘网络教育应用的研究总体呈上升趋势，关注度逐年增加。国外研究持续增多，并且在 2008 年和 2009 年迅速增加，两年的研究总量近乎相当于前些年的总和，主要因为 2008 开始召开的国际教育数据挖掘会议以及 JEMD 电子期刊的创办。数据挖掘网络教育应用研究在国内发展时间不长，2002 年才开始有学术论文出现，随后几年缓慢增加，该领域的发展与网络教育的发展及数据的来源相关，随着广播电视大学的发展、网络精品课程的开发使用和 2004 年 67 所网络学院的成立，国内研究开始有所增加，2007 年达到了一个小高峰。学术论文代表研究的广度，研究相对较浅，学位论文代表着研究的深度，是相关知识的综合运用，从 2003 年才有该研究的学位论文出现，短短的七年时间学位论文达到 42 篇，其中仅有一篇博士论文。对比国内外研究，国外正处于快速发展阶段，国内开始研究的时间滞后于国外，前些年属于引进探索阶段，现阶段正处发展初期，预计相关文献量的高速增长还将持续很长一段时间。

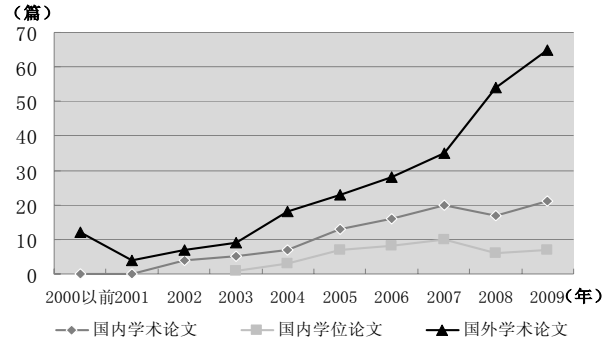

图 2 国内外研究论文分布态势图

2 研究者学科背景统计分析

国外有很多开放的网络教育数据资源，并且大多研究者都具有交叉学科背景，研究者涉及计算机领域专家、教育学家、心理学家、统计学专家等，主要研究力量集中在北美、西欧、澳大利亚和新西兰等地区。而国内 EDM 研究者还没有形成整体力量，基本上都是来自于高校，研究者的学科背景比较单一，其比例如图 3 所示，78%的计算机或相关专业，10%的教育技术学专业，12%的教育科学、管理学、心理学或其他专业。研究者大多数是计算机专业学者，他们熟练掌握数据挖掘技术，但是缺乏教育和心理学理论以及教育数据的来源，导致国内研究相对滞后。

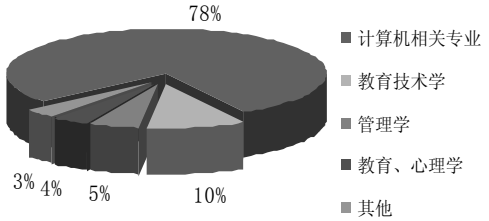

图 3 研究者学科背景统计

3 研究内容类目划分及分析

依据样本研究内容的性质，将数据挖掘网络教育应用的研究划分为“理论探索和方法介绍的描述性研究”、“可行性分析论证研究”“数据挖掘方法的具体应用及数据挖掘过程的实例分析”三个子类，按照这三个研究范畴对国内外学术论文的文献统计结果如表 3 所示。国内外对数据挖掘网络教育应用的各个研究范畴都有所涉及，国内理论描述、可行性分析和应用研究样本的分布差异不显著，理论概述类文章数量相对较多，占 39.8%，这一方面说明了随着研究的发展，这一研究引

起了越来越多的关注，另一方面也在某种程度上说明了研究内容相对浅显，趋于重复。国外理论描述、可行性分析和应用研究样本的分布差异非常显著，应用研究占 63.92%，明显多于理论研究，国外侧重具体的应用研究，这一现象在 EDM 领域是非常合理的，说明国外的研究相对成熟。

表 3 各研究范畴的论文分布比例

具体的研究范畴		理论探索和方法介绍的描述性研究	可行性分析论证研究	数据挖掘方法的具体应用及数据挖掘过程的实例分析研究	合计
国内	论文数	41	33	29	103
	百分比	39.8	32.04	28.16	100
	卡方检验	P > 0.05 差异不显著			
国外	论文数	54	38	163	255
	百分比	21.18	14.90	63.92	100
	卡方检验	P < 0.01 差异非常显著			

通过对样本的进一步分析发现早些年理论描述和可行性分析的文献比较多，而近几年应用研究的文献大量增加，且在核心期刊上的数量居多。理论探索如早期庞先伟（2002）通过对数据挖掘技术、知识发现及资源型学习的认识探讨一种基于数据挖掘技术的资源型学习[4]。可行性分析如Ha等（2002）详细描述了将Web挖掘应用于网络远程教育的可能性，并展示了在网络远程教育中应用Web挖掘的前景[5]，这一篇文章引起了人们对该研究的普遍关注。应用研究大多都是采用一定的数据挖掘方法、选用合适的数据挖掘工具，对来自于一般的网络课程、学习内容管理系统或自适应智能网络教育系统的数据进行挖掘，解决一定的教育教学问题，如孙玉荣等的《数据挖掘在网络教学中的应用》（2009）利用关联分析、序列模式分析、分类分析和聚类分析等分析方法对《数据结构》网络教学数据库的信息进行挖掘，探讨学生的学习习惯，学习兴趣和学习成绩间的关系，为网络分层教学、提升教学质量服务[6]。还有一些基于数据挖掘方法设计学习评价或个性化、智能型教学实用系统的研究，如丁卫平（2009）设计了基于数据挖掘技术的教学评估智能辅助决策平台（TEIA），并对该平台的应用情况进行了分析，结果表明该平台能智能化提取出隐藏在评估数据中有用的规律和知识，为教学评估提供决策支持[7]。

三 EDM 研究的关键内容

教育数据挖掘方法的使用是 EDM 研究最为关键的内容。Zaiane[8]使用挖掘方法评价学习过程，帮助学习者进行网络学习，是目前 EDM 研究中引用次数最多的文章。Romero and Ventura[9]从 EDM 工具、教育数据的来源、EDM 方法几个方面对 1995-2005 年 EDM 的相关文献进行了详细的描述，是了解国外 EDM 发展的权威资料，他们依据任务将用于网络教育系统中的特定数据挖掘方法分为统计和可视化及 Web 挖掘两类。

1 统计和可视化

数据挖掘将描述性数据分析技术本身看作目的，而正式的统计趋向于将基于假设的检验作为最终目标，可视化是将数据信息转化为有意义的、易于理解的图像的过程，虽然它们常常不被认为是数据挖掘技术，但是作为对数据的探测方法，它们可以处理一些通常由数据挖掘解决的问题。蒋玉兰等[10]以宁波电大2000级金融专业为研究对象，采用统计和可视化图形分析学生流失的态势，旨在找出远程开放教育中学生辍学的原因。

2 Web 挖掘

Web挖掘是从WWW资源上获取信息的过程，是数据挖掘技术在Web环境下的应用。依据挖掘对象的不同可以将Web挖掘分为三类：Web内容挖掘、Web结构挖掘和Web使用记录的挖掘，目前Web日志使用记录挖掘在网络学习中的应用研究最多，黄茜[11]通过对学习者在网络教育中留下的日志信息进行挖掘，以实现个性化的网络教育。在Web使用记录挖掘中，网络学习行为采集和学习者的特征分析是关键，王巧玲 [12]、吕莉等 [13]对国内外相关研究进行梳理，王巧玲的硕士论文还实现了基于Web服务的网络学习行为的采集。EDM中的Web挖掘方法可以归纳为以下三组：

（1）聚类、分类和偏差检测

聚类是一个将物理或者抽象对象的集合分组成由类似的对象组成的多个类或簇的过程。分类是通过挖掘数据中的某些共同特性从而对数据项进行分类，用分类或聚类方法划分相似学生群体或个体，以提供相似或个性化的教学。偏差检测是对一些异常或孤立点数据对象进行分析的过程。黄勇等[14]尝试采用决策树分类的D3算法，构造学习者学习能力决策树，对学习者数据库进行分类，将学习者分成学习能力强和学习能力弱的两大类。Ueno[15] 使用在线偏差检测方法分析学习者非常规学习网络课程内容的反应时间数据，指导网络教学。

（2）关联规则挖掘和序列模式挖掘

关联规则挖掘技术用于从用户访问序列数据库的序列项中挖掘出相关的规则，能够揭示学习者访问一些内容的同时会访问哪些内容，借此找出具有相关内容的网页，可更好的

组织课程页面和推荐页面，尽可能缩短相关内容的分布距离，或提供便捷的路径指引。时间序列模式挖掘试图找出页面依照时间顺序出现的内在模式，能够揭示哪些内容能够激发对其它内容的访问，可以用来对学习者的浏览趋势分析，解决远程教育中针对各种层次学生进行因材施教等问题。关联挖掘技术注重事务内的关系，序列模式技术则注重事务间的关系，这两种方法的应用非常普遍。

（3）文本挖掘

Web文本挖掘主要是对Web上大量文本集合的内容进行总结、分类、聚类、关联分析以及运用Web文档进行趋势预测等，是针对非结构化或半结构化的数据集。Web内容挖掘大多是基于文本信息的挖掘，这类方法相对更加困难、复杂，Dringus and Ellis[16]使用文本挖掘策略对异步讨论区进行评价，国内尚未发现文本挖掘在网络教学研究方面的应用。

Baker也对EDM方法进行分类，分为预测、聚类、关系挖掘、人类的判断和模式发现五类。Baker and Yacef[17]对Romero and Ventura的文章中从1995年-2005年运用EDM方法的60篇论文按照Baker的分类法进行归类统计，如表4所示，又将国际教育数据挖掘会议2008年和2009年的文章进行归类统计，如图5所示(有些研究可能使用多种方法，文章被多次统计)。

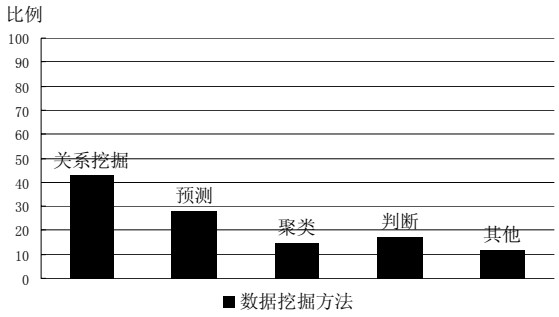

图4 1995-2005数据挖掘方法使用比例

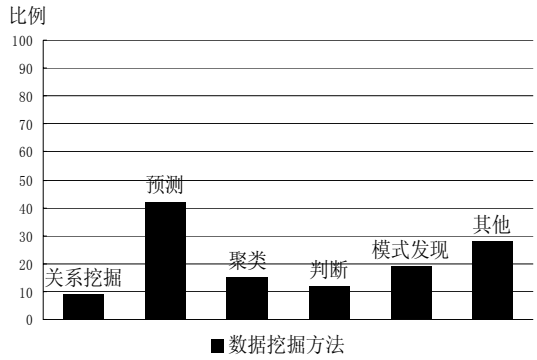

图 5 2008 和 2009 数据挖掘方法使用比例

对图4和图5进行对比分析，近几年对EDM方法的应用重心发生了变化，关系挖掘方法在1995-2005年间占主导地位，但在2008年和2009年下滑，预测占据了主导地位，人类判断和聚类大致保持一致，模式发现位居第二，而在1995-2005年间几乎没有此类方法应用的文章，模式发现能够真正体现EDM研究的价值。另一个趋势是来自于项目反应理论的结构方程模型分析和贝叶斯网络等方法的使用，这些变化反映了国外研究者群体的和研究领域的继续扩大。分析国内研究中使用的EDM方法，基本上一直都使用分类、聚类和关系挖掘。

除了对EDM方法的研究以外，EDM工具、EDM数据的来源、EDM过程也是研究中的关键内容，对这些问题的研究相对固定。目前所进行的研究大多使用广泛的数据挖掘工具，国外也有少数的研究者开发专门的EDM工具使用。随着网络学习人数的增加，数据的来源越来越广泛，早期需要研究者自己搜集收据，现在已经有大量的开放数据供免费使用，陶剑文等[18]、卢永艳[19]对网络教学中可利用的数据来源做了详细的描述，EDM数据的来源包括服务器数据、客户登记信息和代理级数据。曹梅[20]对数据挖掘过程的研究现状进行梳理，数据收集、数据预处理，数据挖掘和结果评价是必要的环节，整个过程是一个不断循环和反复的过程。

四 EDM 的研究趋势

目前国内外研究者对EDM的前景持肯定的态度。教育信息化引起信息量的急剧增长和对信息提取的更高要求，使用计算机进行研究数据搜集、分析和处理的数字化教育研究随之发展；借助数据挖掘技术可以发现数据中隐藏的教育规律和模式，反过来教育科研信息化的进程将极大地促进教育信息化的进程。未来的研究方向主要在以下几方面：

1 EDM 方法的应用研究

可视化是优先选择的方法，传统统计在数据分析方面持续发挥着作用，Web 挖掘成为 EDM 方法研究转向的焦点。Web 是一个动态性极强的信息源，数据库中的数据时刻都在发生变化，所以面向 Web 的数据挖掘研究极具挑战性，Web 使用挖掘是其中最有前途的研究领域。Web 数据挖掘通常有两种应用方式：离线式数据挖掘和在线式数据挖掘，离线式数据挖掘主要通过分析服务器访问日志来发现规律，而在线式具有实时性，难度大，但能做到及时和有效的指导和帮助，目前 Web 数据挖掘的主要方式还是离线式数据挖掘，而 Web 在线式数据挖掘很少有研究，应该加强这一方面的研究。

2 EDM 方法和数据的标准化研究

标准对于任何系统都很重要，EDM 领域也需要整合网络学习环境的普适化的工具和方法。数据库之间采用不同的数据存储类型、对数据的不同定义等问题导致了不一致的现象，数据标准化的主要功能是消除变量间的量纲关系，从而使数据具有可比性。数据标准化、规范化是实现信息集成和共享的前提，在此基础上才能达到信息的准确、完整和及时，没有数据标准化，信息共享就无从谈起，没有信息共享则没有普适化的可以应用到任何教育系统的工具。因此，数据的标准化和预处理任务是必需的。

3 开发易于使用的 EDM 工具

数据挖掘结合人工智能、统计学和数据库技术等多个学科的思想，数据挖掘技术是一种面向应用的复杂技术，应用

难度很大。很多数据挖掘工具封装了挖掘算法、可视化技术等，例如应用于商业领域的 DBMiner、Clementine、Intelligent Miner 等，然而这些工具不是专门为教育领域而设计，对很多教育工作者来说设计的过于复杂，不易于使用，在挖掘方法和数据标准化的基础上，针对教育领域的特点，开发一些专门的 EDM、统计和可视化工具，设计更加直观和易于使用的接口，以帮助教育工作者对于不同层次的教学过程进行分析。

4 特定的 EDM 技巧研究

这一领域的研究者群体中很大一部分是计算机专家，他们缺乏教育和心理学理论，即使通过数据挖掘能够获知学习者的行为，但是不能解释学习者产生这种行为的原因，提不出合理的预测和解决方案。教育领域有其自身的特点，某些具体对象的属性难以用数量方法描述，信息具有隐含性和模糊性，网络教学信息数量大，包括教师教的信息、学生学的信息及交互信息，这些信息很难进行挖掘。数据挖掘在教育系统中的应用，需要考虑教育情境做相应的调整，结合特定的整合教育领域知识的数据挖掘技巧，可以借助教育测评技术、教学理论和教育心理学理论等寻求突破口，在有效且易于使用的数据挖掘工具支持下，将特定的DEM技巧整合到网络学习环境中，使得所有的数据挖掘任务都能成为一个应用，所获得的反馈和结果能够直接被应用到网络学习环境，更好地为教育教学服务。

参考文献

[1]顾小青,李雪.信息化科学研究及其教育应用综述[J].开放教育研究,2008,(8):17-21.

[2]Jiawei Han,Micheline Kamber著.范明,孟小峰译.数据挖掘概念与技术[M].北京:机械工业出版社,2001

[3] Educational Date Mining[DB/OL]. <http://www.educationaldatamining.org>

[4]庞先伟.基于数据挖掘技术的资源型学习[J].现代远程教育研究,2002,(3):39-42.

[5]Ha,Bae,Park.Web Mining for Distane Edueation [J].IEEE,2002, (2).

[6]孙玉荣,罗立宇,黄慧华.数据挖掘在网络教学中的应用[J].现代教育技术,2009,(6):104-106.

[7]丁卫平,王杰华,管致锦.基于数据挖掘技术的教学评估智能辅助决策平台的设计与实现[J].电化教育研究,2009,(4):90-105

[8] Zaiane,O.Web Usage Mining for a Better Web-based Learning Environment[C].Proceedings of conference on advanced technology for education, Banff, Alberta,2001:60–64.

[9] Romero & Ventura. Educational Data Mining:A Survey from 1995 to 2005[J]. Expert Systems with Applications.2007,(33):125-146.

[10]蒋玉兰,周磊.关于开放教育学生流失情况的调研报告[J].宁波广播电视大学学报,2006,(3):50-56.

[11]黄茜.WEB日志挖掘在个性化网络教育中的应用[J].现代教育技术,2004,(5):52-55.

[12]王巧玲.基于Web服务的网络学习行为采集与集成初步设计与实现[D].武汉:华中师范大学,2007.

[13]吕莉,张屹.基于Web服务的网络学习行为采集研究现状[J].开放教育研究,2009,(6):99-104.

[14]黄勇,李玉华.面向知识发现的数据分类技术在网络教学中的应用研究[J].南华大学学报,2006,(6):32-35.

[15]Ueno,M. Online Outlier Detection System for Learning Time Data in E-learning and Its Evaluation[C].International conference on computers and advanced technology in education.2004:248–253.

[16]Dringus & Ellis.Using Data Mining as a Strategy for Assessing Asynchronous Discussion Forums[J].Computer & Education Journal,2005,(45):141–160.

[17]Baker & Yacef,The State of Educational Data Mining in 2009:A Review and Future Visions[EB/OL]. <http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMVol1Issue1_BakerYacef.pdf>

[18]陶剑文,黄崇本.Web Usage Mining在网络教学中的应用研究[J].情报杂志,2006,(5):73-77.

[19]曹梅.知识发现在网络教学系统中的应用研究进展[J].开放教育研究,2008,(12):89-93.

[20]卢永艳.数据挖掘在网络教育中的应用[J].现代计算机,2007,(11):56-60.

An Overall View of the Educational Data Mining Domain

LI Ting　FU Gang-shan

（*The College of News and Communication, Shaanxi Normal University,Xi'an,Shanxi,710062,China)*

Abstract: Educational data mining is an emerging and concerned research domain. Based on literature measurement and content analysis method, the paper analyzes published literature on educational data mining all over the world. By doing this, knows its development and research status, explores its key contents, and forecasts its future research trends, which provides reference for doing research and practice in this domain.

Keywords: Educational Data Mining; Research Status; Key Contents; Research Trend

收稿日期：2010 年 4 月 28 日　　编辑：宋树