

项目反应原理在远程教育中的应用研究

陈乙雄¹ 符云清¹ 祝伟华²

(重庆大学网络教育学院 重庆400044)¹ (重庆大学软件学院 重庆400044)²

摘要 现代远程教育以计算机网络为基础来实施教学的各个环节,有着信息化水平高的特点。这一特殊的教学环境非常有利于项目反应原理(又称IRT,Item Response Theory)发挥优势,提高教学质量。本文结合作者从事远程教育教学系统设计开发工作的经验,对当前IRT的研究成果进行分析比较,最终确立了IRT应用在远程教学考试系统中的理论模型、推导方法以及系统设计方案,并对IRT的发展进行了有益的探索。

关键词 项目反应原理,远程教育,考试系统

Study on the Application of Item Response Theory in E-learning

CHEN Yi-Xiong¹ FU Yun-Qing¹ ZHU Wei-Hua²

(College of Net Education, Chongqing University, Chongqing 400044)¹ (College of Software, Chongqing University, Chongqing 400044)²

Abstract Based on computer network, e-learning possess remarkable features of Internet technologies. It is very helpful for us to take advantage of Item Response Theory, or IRT, in the field, and to promote educational quality. By analyzing current achievements of IRT research, this study proposes the scheme of applying IRT to examination system in e-learning, including it's model, parameter estimation and test flow. Meanwhile, this study will also discuss the theory in different practical situations, and make contribution to the development of IRT.

Keywords IRT, Item response theory, E-learning

1 引言

远程教育是我国构建终生教育和全民教育体系的重要组成部分,其持续健康发展的关键在于教学质量的不断提高,而考试系统的设计工作是其中一项重要内容。当前,保证教学质量的关键环节是考试。从某种意义上讲,考试的模式在很大程度上决定着教育的模式。在实际教学中,尽管学到知识、提高能力是学习的最终目的,但通过考试、取得高分对学生而言同样具有很大的现实意义。因此,只有让分数尽可能反映出学生的真实水平,使两者有机地统一起来,才能真正走向素质教育。此外,和传统教育相比,在远程教育这种新教学模式下的学生往往分布很广,差异性较大,学习过程中不再像传统课堂那样被动地接受知识,而是有了更多的交互和更大的主动性。所以实现“个性化学习”也成为了进一步提高远程教育教学质量的重要手段,是当前远程教育技术的主要发展方向。结合以上分析,可以认为,远程教育对考试环节有着更高的要求,为其提供一种高效、并且体现学生个性化需求的考试系统是当前迫切需要解决的问题。而基于IRT的考试系统则能较好地解决这个问题。该技术充分挖掘了信息技术和网络技术在远程教学中应有的潜力,其具备的自适应性突出了学生在网络课堂的主体地位,同时也使考试结果能更加充分地反映学生的真实水平。因此,尽快将IRT应用到远程教育考试系统的设计开发工作中,对推动我国远程教育的发展,建立更加完善的终身学习体系具有重要意义。

限性,项目反应原理进展得较为缓慢。但近年来,计算机技术的突飞猛进有力地推动了项目反应原理加速发展,使之逐渐成为了教育测量理论研究的重点。应该说,项目反应理论的出现是教育测量理论的一场重要变革。它较好地解决了经典测验理论中一直困惑我们的一些难题,使得我们在考试设计、考生能力估计、测验等值、测验的自适应性等方面的研究都取得了很大的进展。以IRT为理论基础的计算机自适应考试系统(CAT)在世界各国已经引起了广泛的关注,并逐步在社会各个方面得到推广:在教育界,有美国的研究生入学考试(GRE)以及工商管理类研究生入学考试(GMAT);在职业资格认证方面,有全美护士国家委员会资格考试(NNCLT);在企业界,Novell公司在其一系列认证考试中,也成功地应用了CAT系统。在国内,对项目反应原理的研究与应用也得到了教育部门的普遍关注,如全国大学英语四、六级考试就在进行相关的研究和探索。在远程教育领域,目前有相当一部分教学系统的考试部分依然采用的是传统的联考考试方法:即题库建成后,需要输入一些查询参数,系统根据这些参数抽出最适合要求的试题来生成试卷。该方法虽然得到了较多的应用,但随着远程教育无论在教学模式,还是技术手段等方面的不断更新,以及对个性化学习的要求,以IRT为基础的具有自适应性的考试必然是未来的发展方向。

3 项目反应原理

IRT的主要思路是:将学生的行为特征与其潜在特质或能力水平的关系,在一定条件下,用数学模型加以表述,并通过这一模型对项目参数,测验误差等概念重新进行解释,同时引出了信息函数这一重要概念来刻画测验有效性。IRT有三个基本假设:①一维性假设。IRT假定在测验过程中,考生只

2 国内外现状

IRT的发展是以克服经典测验理论的不足为起点的。在发展初期,由于数学上的复杂性和当时计算机技术水平的局

有一种能力被测量,对题目正确回答的概率仅与所要测量的能力水平有关。虽然实际考试中影响考生对题目反应的因素可能不只一个,但大多数考试都是为了考察单一特性而设计的。比如外语考试,设计的出发点必然是考察学生的语言能力,而不会去考虑学生其他特征,因此 IRT 的单维性假设是合理的。② 局部独立性假设。每一个考生对试卷各题目的回答是统计独立的。假设有 n 道题,令 $i=1,2,3,\dots,n$,以 u_i 表示二分变量得分(答对为1,答错为0), $P_i(\theta)$ 为某一能力水平 θ 上考生对题目 i 的答对概率, $Q_i(\theta)=1-p_i(\theta)$ 为答错概率,则这一能力水平 θ 上考生答对题目的概率分布为:

$$P = \sum_{i=1}^n p_i^{u_i}(\theta) Q_i^{1-u_i}(\theta) \quad (1)$$

由上式可以看出,该假设假定考生对某道题目的反应不会受其他题目的影响,每道题的反应情况是彼此独立的。③ 项目特征曲线假设。将上述考生正确回答题目概率的二项式分布在能力量表上进行回归,即可得到项目特征曲线。当获得大量考生的题目分数和能力分数时,就可以通过各能力水平上的分数分布得出具有不变性的曲线^[1]。借助计算机将该曲线与数学函数进行拟合,便可以建立项目反应模型,以此作为系统设计的基础。

4 基于 IRT 的考试系统设计方案

目前,大部分的网络考试系统在考试过程中,呈现给考生的都是同样的试卷,这就使得对单个考生而言,不可避免地会出现难度过高或过低的题目。难度过高或过低的题目对反映考生的水平意义并不大。它只能反映出考生高于或低于某个水平,而不能体现其实际水平。对难度过高的题目,考生会采取猜测甚至作弊的方法作答,这只会增大考试的测量误差;难度过低的题目,则会耗费考生一定的时间和精力。而基于 IRT 的网络考试系统则能在考试过程中根据考生答题的情况动态地从题库中抽取难度相当的试题,去逐步逼近考生的真实水平,直到进入可以接受的误差范围为止。实践证明,当试题难度与考生水平相当时,考生的积极性更高,考试的结果也更有效。为此,本系统的设计将围绕动态施测来展开,可分为以下几个步骤:

4.1 选取适当的项目反应模型

本系统采用的是 Logistic 三参数模型(见图1),该模型的项目特征曲线描述了考生的能力水平与正确回答某一题目的概率之间的关系。横轴为考生的能力水平 θ ,纵轴为答对的概率 $P(\theta)$,考生的能力水平通常可视作连续变化的,从而得到一条连续且光滑的曲线。

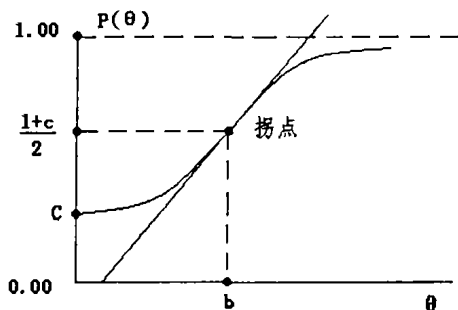


图1 Logistic 项目反应模型

随着能力水平 θ 的增大,答对题目的概率也应相应增大,因此该曲线是单调递增的。但这种递增不是线形的,而是当考

生超过一定的能力水平后,答对题目的概率增长速度变慢;反之,其值的减小速度也变慢。此外,即使考生完全不会该题目,仍有猜对的可能性,所以曲线在 θ 为0时的值并不为0,而是猜对该题的概率 c ,一般把 c 称为猜测系数。故整条曲线是以拐点 $(b, (1+c)/2)$ 为中心对称的曲线。题目的难度可以由 b 来定义,此时在排除猜测原因的情况下答对答错的概率是相等的;曲线拐点处的斜率可以用来表示题目的区分度,因为曲线的斜率越大就越陡峭,考生的能力水平 θ 稍有不同,答对题目的概率就有很大变化,即题目的区分能力也就越强。其具体表达式为:

$$P(\theta) = c + \frac{1-c}{1 + e^{-1.7a(\theta-b)}} \quad (2)$$

选择该模型的主要依据是:① 目前,远程考试过程中计算机技术的优势主要还是体现在客观题方面,而 Logistic 模型在这方面的应用经实践证明较为成熟可靠。主观题的自动评判还有待于人工智能相关技术的突破,这也是作者今后要研究的内容之一;② 虽然有相当一部分应用采用的是较为简便的单参数 Rosch 和双参数对数模型,但三参数的 Logistic 模型可以提供更多的题目信息,而这一点应当是模型选取时考虑的首要因素;③ 系统建设的标准化、规范化也是当前远程教育的重要课题,选择成熟的模型有利于将来与其它的远程考试系统进行资源整合。

4.2 对所选模型进行参数估计

首先,设 $u_{\alpha,j}$ 表示第 α 个考生在第 j 个题目上的反应(答对 $u_{\alpha,j}=1$,答错 $u_{\alpha,j}=0$)。对于不同的 α 和 j 而言, $u_{\alpha,j}$ 的分布是不同的,所以不能用矩估计对模型中的参数 θ 进行估计。但我们可以使用极大似然估计法。根据前面提到的 IRT 局部独立性假设,基于得分矩阵 $U=[u_{\alpha,j}]$ 对 θ 进行推断的似然函数为:

$$L(\theta; a, b | U) = \prod_{\alpha=1}^N \prod_{j=1}^m P_{\alpha,j}^{u_{\alpha,j}} Q_{\alpha,j}^{1-u_{\alpha,j}}$$

其中 $\theta=[\theta_s]$, $a=[a_j]$, $b=[b_j]$, N 为考生人数, m 为题目数, $P_{\alpha,j}$ 为考生 α 答对题目 j 的概率, $Q_{\alpha,j}=1-P_{\alpha,j}$ 为考生 α 答错题目 j 的概率。由于在极大似然估计中,一般都以对数似然函数为对象进行讨论,所以对式两边取对数,得到:

$$\ln L = \sum_{\alpha=1}^N \sum_{j=1}^m (u_{\alpha,j} \ln P_{\alpha,j} + (1-u_{\alpha,j}) \ln Q_{\alpha,j}) \quad (3)$$

然后可用条件极大似然估计(CMLE)估计出未知参数 θ ,即根据考生作答情况估计出其能力值,以此作为随后抽题的依据。

4.3 建立相应的考试系统

整个考试系统主要分为前台和后台两个部分,后台包括试题库、管理维护与统计分析等;前台则向用户提供各种形式的自适应考试,详细设计方案见图2。

在自适应考试过程中,系统必须解决以下三个问题:① 如何开始?即按何种策略选择第一题;② 如何进行下去?当考生完成一道题以后,如何选择下一道题;③ 如何结束?即确定考试结束的终止条件。针对上述三个问题,本系统拟采用如下施测方案:

如图3所示,整个考试是一个逐步逼近的过程。首先,如果考生不是新注册用户,则按照考生最近一次考试所反映的能力水平 θ 作为依据确定初始题目;对于新用户,可以先从题库抽出多道不同级别难度的试题,让考生自行选择与自身水平最接近的一道作为起始题目。然后,根据考生对题目的反应,利用极大似然法重新估计考生的能力水平 θ ,再确定下一道试题。以此类推,考试就可以不断进行下去。最后,由于项目反

应理论所定义的信息函数 $I(\theta)$ 刻画了能力水平估计的精度。随着考试题目的增加, $I(\theta)$ 的值就越大, 对考生能力的估计就越精确。因此, 当 $I(\theta)$ 达到我们所需要的数值时, 考试就可以

结束。当然, 按照迭代法的思想, 我们也可以规定一个误差限 ϵ , 将最后两次作答所反映的能力水平 θ 之差小于 ϵ 作为考试终止的条件。

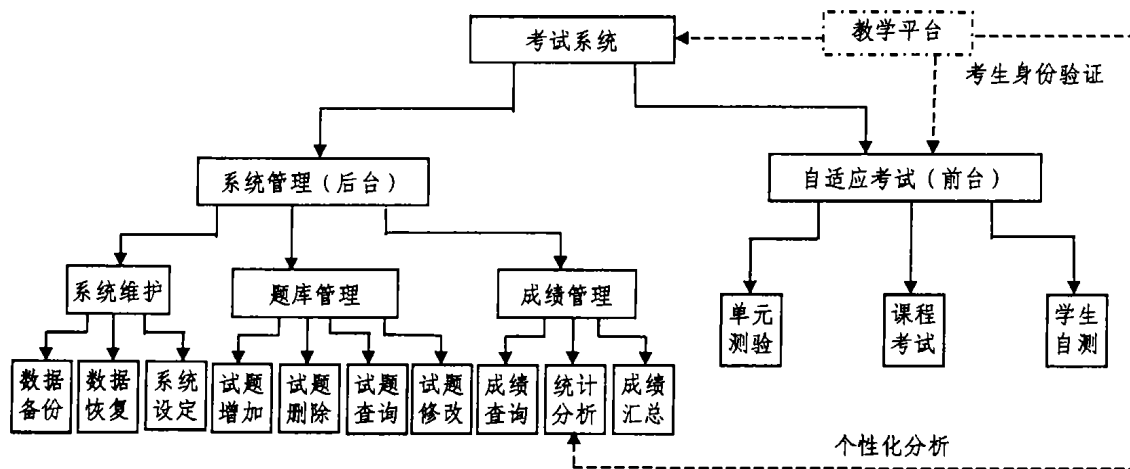


图2 考试系统设计图

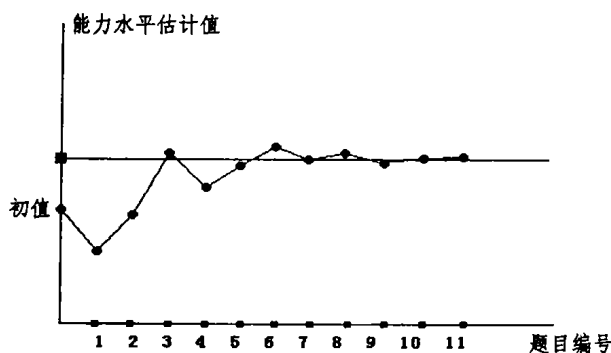


图3 自适应考试施测过程

4.4 系统的接口设计

作为远程教学平台的一部分, 系统在设计过程中应实现与平台其它部件之间的接口。图2上的虚线划分了系统的边界。在实际应用的部署中, 尤其要注意与平台数据库中用户基本信息表在格式上的统一。应尽可能减小系统与平台之间的耦合程度, 增强系统的灵活性和可移植性。

结论 将项目反应原理应用于远程教学平台考试系统的设计开发, 关键在于项目反应模型的选取, 提出相应的参数估计方法, 并最终给出实施考试的算法。在模型的选择上, 本课题采用了 Logistic 模型。虽然目前人们在这方面还提出了其

他一些数学模型, 但有些模型已经过时, 有些仍处于实验阶段。而 Logistic 模型在理论和实践上都得到了较为充分的验证, 采用该模型能大大降低系统开发的风险, 也有利于与其他系统的集成, 发挥资源整合的优势。参数的估计问题是本课题的难点。极大似然估计法虽然能在多数情况下有效地估计出未知参数, 但该方法也还存在一些缺点: 如不能处理特殊反应模式, 伴随参数可能不存在充分统计量, 结构参数不一定是相合估计量等。而其它估计方法也有一些自身缺陷: 边际极大似然估计(MMLE)与 EM 算法虽然有助于消除伴随参数的影响, 但收敛速度太慢; 边际贝叶斯估计虽然可以解决特殊反应模式的问题, 但使用时必须先给出参数的先验分布, 如果这个先验分布一旦偏离实际情况较远, 则得到的结果可能会远离真实值。如何在不同条件下选择最佳的方法, 还有待进一步研究。

参考文献

- 1 万勇, 于全训. 传统测验理论与项目反应理论(J). 山东师大学报, 1991, 66(3)
- 2 俞晓琳. 项目反应理论与经典测验理论之比较(J). 南京师大学报, 1998, (4)
- 3 吴军其. 计算机自适应测验软件模型研究(D). 华中师范大学, 2001
- 4 罗芬, 丁树良, 胡小松, 万宇文, 甘登文. 基于IRT若干参数估计方式的比较(J). 江西师范大学学报, 2003, 27(1)

(上接第106页)

网络效率低下, 甚至威胁到整个网络的可用性。

面向对象网络体系结构, 采用面向对象的思想对网络协议进行设计和开发, 并对不同层次中协同完成某种功能的对象进行有机组织, 是对新型网络体系结构研究的有益尝试。但它保留了层次的限制, 增大了网络系统的开销。

基于角色的网络体系结构和服务元网络体系结构, 同属于非层次型网络体系结构。它们根据网络事务或功能进行模块划分, 能够保证机制的一致性, 并避免网络功能的冗余重复, 提高网络效率。但二者能否实际投入使用仍取决于其后期研究和开发工作的进展情况。

参考文献

- 1 Braden R, Clark D, Shenker S, et al. Developing a Next-Genera-

- tion Internet Architecture, Proc ACM SIGCOMM 2000, 2000
- 2 Clark D, Tennenhouse D. Architectural Considerations for a New Generation of Protocols. Proc ACM SIGCOMM 1990, 1990
- 3 Tanenbaumk A S. 计算机网络(第3版). 北京: 清华大学出版社, 1998
- 4 Boecking S. 面向对象的网络协议. 北京: 机械工业出版社, 2000
- 5 Braden B, Faber T, Handley M. From Protocol Stack to Protocol Heap: Role-based Architecture. Proc ACM SIGCOMM 2003, 2003
- 6 鲁珂, 赵继东, 曾家智. 基于角色的网络体系结构与TCP/IP层次网络体系结构的比较研究. 计算机科学, 2004, 31(11): 53~54
- 7 Zeng Jiazhi, Xu Jie, Wu yue, et al. Service Unit Based Network Architecture. Proc PDCAT'03, 2003
- 8 曾家智, 徐洁, 吴跃, 等. 服务元网络体系结构和微通信元系统构架. 电子学报, 2004, 32(5): 745~749
- 9 Clark D, Wroclawski J, Sollins K, et al. Tussle in Cyperspace - Defining Tomorrow's Internet. Proc ACM SIGCOMM 2002, 2002