基于项目反应理论自适应考试系统的设计与应用

于海霞,刘竞杰,王家骐
(1. 安徽工贸职业技术学院 计算机系,安徽 淮南　232007)

摘　要: 在个性化教育中,传统的考试形式正面临着越来越多的问题,统一的试题内容并不适用于各个层次的学生,考试成绩无法准确衡量学生的能力. 个性化学习系统中的自适应考试系统(PLCAT),部分地解决了传统考试形式所面临的问题. PLCAT 考试系统是基于项目反应理论的 Web 自适应考试系统,学生可以随时随地使用 PLCAT 系统进行测试,系统能够根据学生的能力自动选择适合学生的试题,并在考试结束时给出个性化的评价. 实践证明,PLCAT 系统可以提高考试效率和提高对学生评价的精度,更重要的是,它为个性化教育提供了一种更加有效的测验途径,也为贯彻现代化教育理念,提高学生自主学习能力、创新能力提供了一种新思路.

关键词: 个性化学习; 自适应考试; 项目反应理论
中图分类号: TP391.9　　**文献标识码:** A　　**文章编号:** 1673-162X(2010)03-0044-05

Design and Application of Computerized Adaptive Testing System Based on Item Response Theory

YU Hai-xia, LIU Jing-jie, WANG Jia-qi
(1. Department of Computer Science, Anhui Vocational and Technical College of Industry and Trade, Huainan, Anhui　232007, China)

Abstract: In personalized education, the traditional examination forms are facing increasing problems, a unified examing paper do not suitable for all students with different levels, therefore, testing score can not measure students abilities accurately. Personalized Learning of Computerized Adaptive Testing System (PLCAT), partially solve some problems of traditional examination. PLCAT testing system is a Web CAT system based on Item Response Theory (IRT). Students can be tested anytime and anywhere by PLCAT. PLCAT system can select questions automatically for students according to their ability and conclude a personalized evaluation for them when examination finish. Practise shows that PLCAT system may improve testing efficiency and students evaluating accuracy. What's more important is it provides a more effective testing way for personalized education. Meanwhile provides a new thought for implementing modern education concepts, increasing self-learning ability, and innovation ability.

Key words: personalized learning; computerized adaptive testing; item response theory

现代教育理念主张个性化教育,强调学生在学习中的主体地位,注重对学生能力的培养,这就要求采用一种新的评价考核方法,来检验学生的学习效果. 在传统教育中,大多是采用统一的纸制考试的形式来实现上述目标. 传统的考试,采用统一的试题内容,并不适用于各个层次水平的学生,无法真正考查出学生对知识的掌握程度; 固定的考试时间,也降低了考核的效率.

为了克服传统考试中的不足,在个性化学习系统中,采用了基于计算机自适应考试(Computerized Adaptive Testing, CAT)理论的考试系统(PLCAT),作为检验学生学习情况的辅助工具. 在自适应考试系统

收稿日期: 2010-05-10　　**修回日期:** 2010-07-18
基金项目: 安徽省高等学校省级教学研究项目(2008Jyxm565)基金资助.
作者简介: 于海霞(1975—),女,内蒙古呼伦贝尔人,安徽工贸职业技术学院计算机系讲师,硕士.

中,将自适应技术、计算机技术和教育技术相结合,测试过程始终围绕学生的能力进行,系统自动地去适应参加测试学生的具体情况,根据学生的能力从题库中自动选择难度适中的试题,考试时间也根据学生的答题情况确定. 这种考试,针对性强,更加突出了学生的主体地位,满足个性化需求,并且提高了考试的效率和可信度,能给予学生更准确的评价,也便于教师实施个性化教学.

1　计算机自适应考试原理

CAT 测试是建构在现代测验理论——项目反应理论(Item Response Theory, IRT) 基础上的,题库建设、参数估计、试题选择和评分都是以 IRT 为指导进行的系统. 被试所做的每一试题(项目) 不但要评定被试对该试题所代表的知识掌握的情况,还决定着下道试题的挑选. CAT 测试能不断地根据题目的各方面信息和被试答题情况估计其能力值,然后从题库中选取符合被试能力的下一道题目进行测试,直到达到预定的测试精度要求,即可以结束考试.

1.1　IRT 原理

IRT 是计算机自适应考试的重要理论基础,基于 IRT 理论的测试模型称为 IRT 模型,IRT 模型定义了被试反应行为与其潜在能力特质或能力水平之间的关系. 这一关系可以用项目特征曲线[1] 表示,见图 1,其中横坐标 θ 为能力水平,纵坐标 $p(\theta)$ 为正确反应的概率.

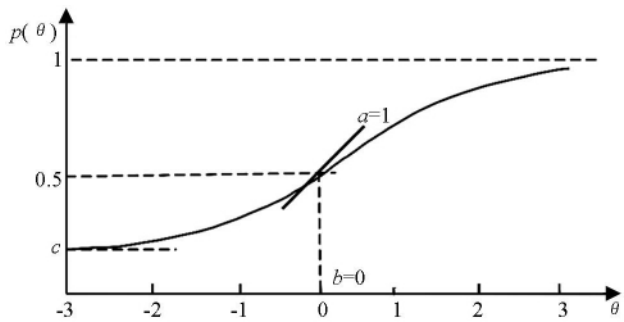

图 1　项目特征曲线图

(1) 项目区分度 a,区分度表示项目对能力高的被试和能力低的被试的区分程度,即曲线拐点处的斜率,斜率越大就越陡峭,$p(\theta)$ 就有很大变化,即题目的区分能力也就越强;

(2) 项目难度 b,即项目答对概率 $p(\theta)$ 所对应的能力参数 θ 值,当 $p(\theta) = 0.5$ 也即曲线斜率达最大时 $b = 0$,此时,表示项目的难度适应大多数被试的能力;

(3) 对选择题或是非判断题这种项目,项目猜测参数 c(即特征曲线的截距) ,表示猜测在正确回答项目时所占的概率,其值越小,表示能力低的被试也可能答对本题.

一般使用 IRT 模型的特征函数有单参数(包括 b 一个参数) 、双参数(包括 a、b 二个参数) 和三参数(包括 a、b、c 三个参数) 三种[2] ,其中三参数的 Logistic 模型的特征函数公式如下式所示,Logistic 模型适用于既有主观题又有客观题的题库.

$$p(\theta) = c + (1 - c)\frac{1}{1 + \exp[-Da(\theta - b)]}(D = 1.702). \quad (1)$$

1.2　IRT 参数估计

正确估计考生的能力是 CAT 得以顺利进行的前提,而从图 1 中可以看出,未知参数有 a、b、c、θ,其中 θ 的值是根据已知的项目参数,采用极大似然函数估计法求出,估计过程如下.

1.2.1　构造似然函数

设有 n 个项目,对给定项目 i,能力为 θ 的被试对该项目的反应为一随机变量 u_i(若回答正确 $u_i = 1$,回答错误 $u_i = 0$) ,由此构造出项目反应矩阵 $\boldsymbol{U}(u_1, u_2, \cdots, u_n)$,相应的回答正确的概率记为 P_i,回答错误的概率记为 Q_i($Q_i = 1 - P_i$) . 以 $L(U_i \mid \theta, a, b, c)$ 表示能力为 θ 的被试对项目 i 的反应为 U_i 的概率,则联合概率为

$$L(U \mid \theta, a, b, c) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}, \quad (2)$$

(2) 式称为似然函数,使似然函数达到最大值时 θ 的取值,称为 θ 的极大似然估计值,即被试的能力参数估计值.

1.2.2　求解极大似然函数

求似然函数的自然对数在参数 θ 上的偏导,偏导数为 0 时 θ 的值即为要估计的值. 即:

$$f(\theta_i) = \frac{\partial \ln L}{\partial \theta_i} = 0, \tag{3}$$

(3)式是非线性方程,无法直接求解,通常采用 Newton-Raphson 迭代公式求出 θ 的值.

$$\theta_{t+1} = \theta_t - h_t, \tag{4}$$

其中,θ_{t+1}、θ_t 为 $t+1$、t 次迭代的能力估计值.

$$h_t = \frac{f'(\theta)}{f''(\theta)}, \tag{5}$$

$$f'(\theta) = \frac{D\sum_{i=1}^{n} a_i(u_i - P_i)(P_i - c_i)}{P_i(1 - c_i)}, \tag{6}$$

$$f''(\theta) = \frac{D^2\sum_{i=1}^{n} a_i^2(u_i c_i - P_i^2)(P_i - c_i)Q_i}{P_i^2(1 - c_i^2)}, \tag{7}$$

迭代的终止条件为 $|\theta_{t+1} - \theta_t| < \varepsilon$($\varepsilon$ 为某一预先确定的一个很小的正数).

利用上述方法估计 θ 时,是在项目参数 a、b、c 已知的条件下完成的. 在实际应用中,往往是建立题库时,先用专家估计法估计出 a、b、c 的初值,然后在测试过程中利用公式(3)估计出 θ 值后,再分别用似然函数的自然对数在 a、b、c 上的偏导,求出 a、b、c 的值.

2 自适应考试系统(PLCAT)的设计

PLCAT 系统是个性化学习系统中的重要组成部分,学生可以使用该考试系统随时随地地进行测试,来检验学习效果,发现不足;教师也可以利用该系统组织统一测试,将测试成绩作为学期末学生的正式考试成绩.

PLCAT 系统采用基于 Internet 的三层模型,提高了系统的可扩展性、安全性和可重用性. 三层模型将应用逻辑与用户界面和数据访问分离,使系统的维护变得简单.

2.1 PLCAT 系统的主要模块

PLCAT 系统主要包含以下几个模块.

(1) 系统管理模块主要包括权限管理、用户管理、历史信息管理等功能,系统用户分为三种:学生、教师和管理员,不同的用户设置不同的权限.

(2) 题库管理模块是 PLCAT 系统中的主要模块,主要用来完成题库的建立和维护,包括试题录入、试题参数设置及试题的修改、删除和查询等功能. 其中试题的参数主要指试题的难度、区分度和猜测系数等.

(3) 考试管理模块是考试系统中另一重要模块,提供了大部分与测试过程有关的功能,包括学生能力参数估计、试题抽取、考试终止条件判断、考试评分等. 学生能力参数估计是指根据学生的当前答题情况和以前的能力值来估计最近的能力值.

(4) 成绩管理模块主要包括成绩查询、成绩汇总、试卷分析、个性化评价等功能. 个性化评价主要是根据学生的测试结果对学生的学习效果进行评价,提出不足,并给出学习建议.

2.2 主要模块功能的实现策略

2.2.1 题库建设

题库建设是系统的首要工作,高质量的题库应具有优质、量大、等值、动态可扩充等特点.[3] 题库中的题目不仅要包含试题内容,还应包含必要的参数信息,以保证能力评估的准确性. 题库建设主要包括以下几方面内容:选择 IRT 模型、编制题目、确定题目参数和题库动态维护等.

PLCAT 系统的题库采用三参数的 Logistic 模型,题目包含以下主要信息:试题编号、题型、试题内容、难度系数、区分度、猜测系数、所属知识点、分值、最多使用次数、最长反应时间(如果超出指定时间不回答,认为答错)等.

编制题目时,应注重不同知识内容与能力层次、不同难度和不同题型的结合,对编制的题目应组织审查,确保题目的质量.

题目参数的确定主要是指难度系数 a、区分度 b 和猜测系数 c 的确定. 一般有两种方法,一是经过大量

的测试后统计分析; 二是专家评估法. 在 PLCAT 系统中,采用两种方法相结合的方式来确定参数值. 先由教师按经验暂时确定下 a、b、c 的值,然后在此初值条件下采用以下算法.

Step1: 使用 1.2.2 中的公式(3) 估计出 θ 的值;

Step2: 求出 a、b、c 的值;

Step3: 判断结束条件是否满足,如果不满足,转向 Step1; 否则结束,此时 a、b、c、θ 的值即为所求值.

题目的动态维护是指对题目的增加、删除、修改等操作. 要将那些不再适宜的题目删除,增加包含新知识点的题目、修改有错误的题目以及题目参数的修改等.

2.2.2　选题策略

如何根据学生的能力参数,选取适合他的题目,是在测试过程中,始终要关注的问题. 在 CAT 测验中,选题策略主要有二种: 一是传统的选题策略(包括项目的信息函数最大模型,和加权偏离模型); 另一种是分层选题策略(主要以 A-STR 和 BAS 方法为代表). 传统的选题策略在测验安全、测验效率、项目平衡和题库维护等方面都存在许多不足,分层选题策略则部分地解决了这些问题.

PLCAT 系统采用 BAS 选题策略. BAS 选题策略是按 b 分区,a 分层的选题策略,是对 A-STR 改进的一种方法. A-STR 方法是将项目按区分度分层,在测验的早期阶段实施低区分度的项目,测验的后期分别实施中等区分度和高区分度的项目. A-STR 方法解决了项目的区分度问题和项目的平衡应用问题,缓解了测验效率和测验安全之间的矛盾,但是对于每一层中的项目的难度分布它并没有考虑在内. 然而,在实际应用中,项目的难度是在测试中必须要考虑的一个参数,BAS 选题策略正是在这种要求下产生的. BAS 的实质是使得每一层都有一个难度 b 的平衡分布来保证去匹配不同被试的能力值 θ. [4]

测试时,首先判断学生是否是首次登录考试系统,如果是,则先挑选一道中等难度的试题,让学生做答; 如果不是,则根据该学生最近一次考试所确定的能力参数 θ 的值,来选择题目. 后继题目的选择,始终要根据 BAS 策略来进行. 具体流程如图 2 所示.

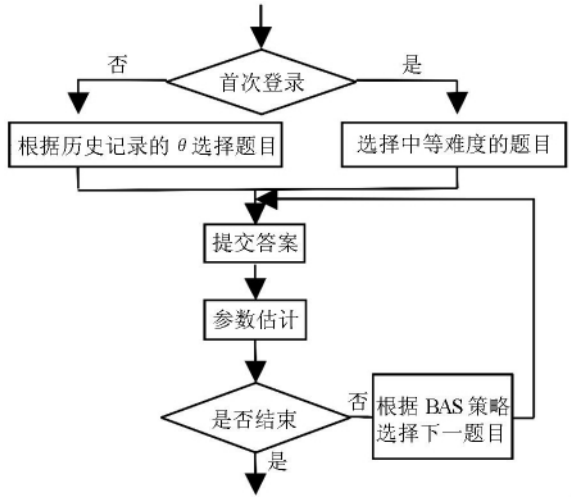

图 2　考试过程的部分流程图

PLCAT 系统设定,选择的题目要覆盖所有指定章节(考生在登录系统时,可以设定考试的范围),且每章的题目数量分布尽量与知识点的重要与否相符合. 为了实现这一目的,系统设定了每一章可选题目数量的最大上限和最小上限.

PLCAT 系统采用了 BAS 分层选题策略,可以有效地控制高区分度题目的曝光率,增加低区分度题目的曝光率,平衡题目的应用,提高测验效率,减轻题库维护的工作量.

2.2.3　测试终止策略

CAT 测试终止的策略主要有三种[5].

(1) 固定测验长度,即当测验项目达到一定数量时,测验自动终止,此方法易于实现,可以对每个测验项目的使用率作精确统计,但是它对不同被试的能力参数的估计精度不同,而且要确定一个合适的长度一般来说并不容易.

(2) 当能力参数估计的标准差小于某一预先确定的值时,测验自动终止. 这种方法能避免固定长度法的缺点,但终止条件定的过严往往会使测验时间过长,降低测量效率.

(3) 满足迭代终止条件,即比较被试能力参数最后两次估计值,当这个值之差小于预先给定的数值时,测试自动终止

PLCAT 系统采用上述(1)、(3) 二种方法的综合,满足任何一个条件,测试即终止. 在系统中,根据需要设定测试的最多试题数和最长时间,从而避免了测试时间过长、效率低下的问题.

2.2.4　评分策略

为了减少评分过程中的难度和提高测试的效率,PLCAT 系统目前采用了自动评分和人工评分两种方

式,对于那些有固定答案的客观题,在测试过程中由系统自动完成评分工作,而对于那些主观题,如编程题、画图题、论述题等,需要人工评分,教师根据事先设定好的评分标准,给出一个相应的结果,这个结果仍然是以 θ 表现出来的. 测试时,如果包括主观题和客观题二种题型,则最后的成绩将按主观题得分和客观题得分二部分给出.

PLCAT 测试的最后结果是学生能力参数估计值 $\theta(\theta \in [-3,3])$,不易于学生理解,因此要转换为传统测试中易于理解的分数. 系统设定最终的测试结果仍然采用百分制的形式,并采用如下的转换公式: $S = \frac{40}{3}\theta + 60, S \in (20,100)$ 变换后 S 的值,基本上符合所有能力水平的学生情况.

2.2.5 个性化评价策略

个性化学习的一个主要特点是强调学生的主体意识,提高学生自主学习的能力. 因此,在测试之后给出合理、有针对性的评价是必不可少的.

PLCAT 系统中个性化评价主要从以下几个方面进行:

(1) 学生对知识点掌握的情况,指出学生对哪些章节掌握的比较薄弱. 在测试的时候记录下学生回答错误的题目所在的章节,测试结束时统计出错较多的章节.

(2) 对学生今后的学习方向,需要重点学习的知识点给出建议.

(3) 绘出考试成绩的变化曲线图,根据学生历次考试成绩,绘出成绩变化图,让学生了解自己的学习情况,近期的学习效果,从而督促学生今后学习,改变学习方法,提高学习成绩.

3 结束语

PLCAT 系统是一个基于 Web 的自适应考试系统,是个性化学习系统中的子系统. 通过实际应用证明,PLCAT 系统在提高学生自主学习能力,帮助学生提高学习效果和成绩上具有非常大的作用,它不同于传统的考试系统; 在这个系统上进行考试,没有时间和空间的限制,考试过程完全是围绕考生的能力进行的,而不像传统考试那样,都是千篇一律的试卷内容和固定的考试时间. PLCAT 系统对学生学习能力的评估不是建立在他答对题目的绝对数量上,而是建立在他能正确回答试题的难易程度上,难度不同的题目得分不一样. 因此,可以通过给每个学生建立个性化的考试来达到更为准确的知识、能力和水平的测量. 这样就避免了传统考试中学生成绩与实际能力不相符的现象,具有更准确的评价和更高的效率. 但是,PLCAT 系统也存在一些不足之处,如在确定题目初始参数方面,还不能给出一个准确的标准; 题目选题策略上,虽然部分解决了题目的曝光率和题目分布的平衡问题,但是在题库的分区数目和每个区的分层数上没有一个很好的衡量标准; 个性化评价方面还有待于进一步完善.

参考文献:

[1] 刘丽平,王文杰,郭世宁. 计算机自适应考试(CAT)系统题库的设计与实现[J]. 计算机系统应用,2006,23(3):10-12.
[2] 邵晨辉,陈玉泉,徐良贤. 基于项目反应理论的机助自适应考试系统[J]. 计算机工程,2000,26(11):161-163.
[3] 王飞. 基于 Agent 的计算机自适应考试系统的应用研究[M]. 南京: 南京工业大学,2004: 14-50.
[4] 张忠华,谢小庆,郑日昌. 计算机自适应性测验(CAT)选题策略的新进展[J]. 心理发展教育,2002(4):92-94.
[5] 张琳,朱春鹤,赵奕. 个性化学习中的机助自适应测试系统设计[J]. 上海海事大学学报,2007,12(4):63-64.

[责任编校: 张永军]