基于项目反应理论的自适应考试系统的研究与设计

刘洪峰 郭文明 余晓佳
(南方医科大学网络中心 广东 广州 510515)

摘 要 针对基于经典测量理论的传统考试暴露出越来越多的缺点和不足这一情况,提出基于项目反应理论的计算机自适应考试系统。它能根据被试者的能力水平选择相应难度的题目,实现更快、更准确地估计出被试者的能力值。对自适应考试系统中所涉及的几个关键技术(参数估计、参数等值、题目曝光率控制、题库建设等)进行研究与改进,实现了系统的设计与开发。实验结果表明,该系统可以有效地测试出被试者的能力值,达到了预期目的。

关键词 项目反应理论 自适应考试 参数估计

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.10.020

RESEARCH AND DESIGN OF ADAPTIVE EXAM SYSTEM BASED ON ITEM RESPONSE THEORY

Liu Hongfeng Guo Wenming Yu Xiaojia
(*Network Center, Southern Medical University, Guangzhou 510515, Guangdong, China*)

Abstract In light of the situation that the classic measure theory-based traditional examinations expose more and more shortcomings and deficiencies, we put forward the item response theory-based adaptive computer exam system. It can choose the questions with appropriate difficulties according to the faculty of examinees to achieve faster and more accurate estimates on the ability value of them. We studied and improved several key technologies involved in adaptive exam system (parameter estimation, parameter equivalent, questions exposure rate control, item bank construction, etc.), and implemented the design and development of the system. Experimental results showed that the system can effectively test the ability value of examinees and achieves the expected goals.

Keywords Item response theory Adaptive exam Parameter estimation

0 引 言

随着计算机和网络技术的快速发展,将计算机和网络技术应用于教育已经成为一种趋势。基于经典测量理论 CTT(Classical Test Theory)的传统考试已经暴露出很多的缺点和不足。因为 CTT 不论被测试者能力水平的不同,都用相同的试题进行测试,这样就导致对试题区分度和难度的估计严重依赖于被试者的作答情况,对被试者的能力值估计也会依赖于所测试的题目,使得基于 CTT 所测试出来的结果并不能真正代表被试者的能力水平[1]。因此,探索一种新的考试形式已经迫在眉睫。计算机自适应考试系统建构在项目反应理论的基础之上,让考试变得更加高效。由于项目反应理论有关信度理论的先进性,测量的精度也得到了更有力的保证。让人惊喜的是,测验效度的改善和精度的提高并没有带来人力和财力的增长,相反考试变得更加高效了。因为自适应考试的突出特点就是考试的剪裁性,被试者所做的每一道题目对他来说都是最有效的测量,从而可以使被试者要做的题目数量大大减少,有效地节约了考试时间;并且这是在保证测量在一定精度上而达到的[2]。

自适应考试现在还处于起步阶段,现有的自适应考试系统也各种各样,功能参差不齐。本文针对自适应考试系统所涉及的几个关键技术进行了研究与改进,设计并开发了一个稳定和可靠的自适应考试系统。

1 自适应考试系统理论基础

1.1 项目反应理论

项目反应理论 IRT(Item Response Theory)是建立在潜在特质理论基础上的现代测量理论,是该自适应考试系统构建的理论基础。项目反应理论模型提出了被试者对测试内容的反应行为和其潜在的能力特质之间的关系。在 IRT 中应用最广泛的是 Logistic 模型,因为该模型避免了复杂的积分运算,在估计能力和项目参数时要简便得多[3]。

三参数的 Logistic 模型表达式是:

$$p_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (1)$$

其中:$i = 1,2,\cdots,n$;θ 表示考生的能力水平;$p_i(\theta)$ 表示能力水平为 θ 的考生答对试题 i 的概率;a_i 表示试题 i 的区分度;b_i 表示

收稿日期:2015-07-30。广东省科技计划项目(2013B090500024,2014A040401026)。**刘洪峰**,硕士生,主研领域:云计算,远程教育。**郭文明**,教授。**余晓佳**,工程师。

试题 i 的难度;c_i表示试题 i 的猜测系数。另外,当猜测系数 c_i为 0 时就变成了双参数模型;当 $c_i=0$ 并且 $a_i=1$ 时就变成了单参数模型。

1.2　项目信息函数

项目信息函数是项目反应理论中用以刻画试题有效性的工具,它是直接反映被试者的得分情况对其能力估计精度的指标。项目信息函数的定义为:

$$I(\theta)=\frac{\left[\frac{\partial}{\partial\theta}P(\theta)\right]^2}{P(\theta)[1-P(\theta)]} \tag{2}$$

由函数定义可知,$I(\theta)$只是 θ 的函数。项目信息函数在测试题目质量高低的过程中扮演着举足轻重的角色,因为它能反映出题目对被试者能力值估计的正确性判断上提供的信息量大小,并且只有当被试者的能力参数接近试题难度参数时,项目信息函数才能取得极大值。同时,根据项目信息函数的定义,可以计算出当信息函数为极大值时的能力参数取值为:

$$\theta_{\max}=b_j+\frac{1}{1.7a_j}\ln[0.5+0.5\sqrt{1+8c_j}] \tag{3}$$

式(3)也是在自适应考试过程中选择后续试题的理论依据所在。总之,项目信息函数是反映试题优劣的一个综合指标,是项目反应理论的重要组成部分。

2　自适应考试系统的功能模块设计

根据该自适应考试系统功能的要求,该系统分为 3 个模块,分别是学生在线考试模块、教师管理模块和管理员系统模块,整个系统结构如图 1 所示。

该自适应考试系统采用 B/S 模式,前台主要负责学生的自主考试,满足学生对所学知识点有更深入了解的需求,同时教师也能根据学生的考试得分情况掌握学生的学习状况。后台主要提供题库中各种试题信息的更新与维护以及试题库的升级等功能。

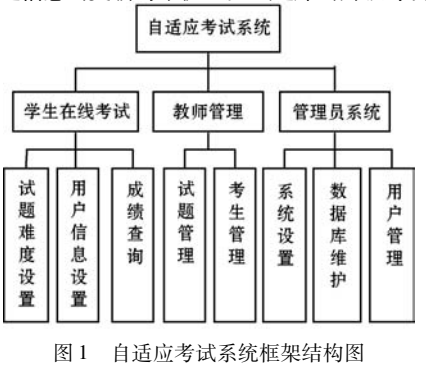

图 1　自适应考试系统框架结构图

3　系统关键技术及实现

3.1　参数估计

在自适应考试中,如何能正确估计学生的能力值是关键之一。目前应用最广泛的是极大似然估计法,但是极大似然估计法有时会出现迭代过程不稳定、无法满足收敛精度等问题,特别是当被试者答对或答错所有题目时,能力估计将无法进行等缺点[4]。所以,下面介绍另外一种估计方法——贝叶斯估计法。

贝叶斯估计法是在贝叶斯公式基础上进行的:

$$g(\theta_j\mid U_j,\xi)=\frac{P(U=U_j\mid\theta_j,\xi)g(\theta)}{P(U=U_j)} \tag{4}$$

在局部独立性假设条件下,能力为 θ_j的被试者对 n 道题目的反应矩阵 U_j的概率就是 $P(U=U_j\mid\theta_j,\xi)$,并且:

$$P(U=U_j\mid\theta_j,\xi)=\prod_{i=1}^{n}P_i^{u_{ij}}Q_i^{1-u_{ij}} \tag{5}$$

对于来自于先验分布 $g(\theta)$的某一被试者产生反应数据 U_j的无条件概率就是:

$$P(U=U_j)=\int_{-\infty}^{\infty}P(U=U_j\mid\theta)\cdot g(\theta)\mathrm{d}\theta \tag{6}$$

经过计算,可以求得 θ_j的无条件期望为:

$$E(\theta_j\mid U_j,\xi)=\frac{\int_{-\infty}^{\infty}\theta_j\cdot g(\theta)\prod_{i=1}^{n}P_i^{u_{ij}}Q_i^{1-u_{ij}}\cdot\mathrm{d}\theta}{\int_{-\infty}^{\infty}g(\theta)\prod_{i=1}^{n}P_i^{u_{ij}}Q_i^{1-u_{ij}}\cdot\mathrm{d}\theta} \tag{7}$$

式中包含了积分运算,使用“Hermite-Gauss 近似积分法”后,$E(\theta_j\mid U_j,\xi)$的近似表达式为:

$$E(\theta_j\mid U_j,\xi)=\overline{\theta_j}=\frac{\sum_{k=1}^{q}X_k\cdot L(X_k)A(X_k)}{\sum_{k=1}^{q}L(X_k)A(X_k)} \tag{8}$$

这一算法的一个突出特点是没有迭代计算。其次,公式中 $A(X_k)$值是在采用“贝叶斯估计法”估计题目参数时,经过最后一次 EM 循环(求期望—极大化)调整过的能力节点 X_k的权重。这意味着使用了这些 $A(X_k)$值作为能力参数 θ_j的先验分布,同理 $L(X_k)$值也是以同样方法得到的[5]。

基于贝叶斯估计法的优点,并且为了更加准确地估计出被试者的能力值,这里采用贝叶斯估计法与极大似然估计法相结合的方法来精确估计被试者的能力值。在考试的初始阶段,使用贝叶斯估计法;随着题目数量的增多,在贝叶斯估计法基础上附加一个极大似然估计。这么做的优点是贝叶斯估计提供了被试者较好的能力估计初值,从而可以使极大似然估计的精度大大提高,同时也可以降低贝叶斯估计对能力先验分布的依赖。

3.2　题目参数的等值

根据项目反应理论的原理,参数估计值具有不变性的特点,但是,参数估计值的单位系统具有不确定性。在 IRT 题库建设中最重要的是题目参数的等值问题。在能力参数未知的情况下,同一批题目根据不同被试样组的实测数据分别估计出的两套题目参数,也会由于参照系的不同而有着不同的表现形式,但两者之间一定具有某种线性转换关系[6],这也就是题目参数等值的理论基础。

对于题目参数的等值问题,这里可以利用不同的测试样本中所包含的相同题目(又称锚题)这一特征。由于不同的被试者都对锚题做出了反应,因此锚题中的每道试题都有成对的估计参数值。如果使用的是二参数 Logistic 模型,就分别是区分度参数 a_x和 a_y、难度参数 b_x和 b_y,并且存在下列的线性转换关系:

$$a_y=\frac{a_x}{A}\quad b_y=Ab_x+B \tag{9}$$

这里,A、B 称作等值常数,对于等值常数的求解,利用题目特征曲线的方法。由于题目特征曲线集中了题目各个方面的信息,因此建立在题目特征曲线基础上的等值方法也有着更优良的特点。

这里用一个函数来表示 m 道锚题在不同的两个量表上的题目特征曲线之差求平方后再求和:

$$H(\theta_j)=\sum_{i=1}^{m}\left[P_{ij}(\theta_{xj};a_{xi},b_{xi})-P_{ij}(\theta_{yj};\frac{a_{yi}}{A},Ab_{yi}+B)\right]^2 \tag{10}$$

再令：

$$Hcrit = \sum_{j=1}^{N} H(\theta_j) \tag{11}$$

其中 N 表示被试样组人数，然后通过极小化 $Hcrit$ 函数就能求得等值常数 A、B 的值。

3.3 题目曝光次数控制

题库中的题目由于性能上优劣的不同，造成在自适应考试的过程中，有的题目经常会被选中，显得很活跃，有些题目则相反。活跃的题目主要是一些难度适中和区分度较好的试题，而非活跃的题目则是一些知识内容比较偏僻的试题。对于活跃的题目，使用一次就曝光了一次，如果在短期内频繁地被使用，就可能造成这些题目大面积被曝光，产生漏题现象[7]，所以题目曝光次数必须要严格加以控制。

为了改善这一状况，可以给每道题目赋予一个曝光控制参数 K。然后在考试过程中，当某道题目被选为下个最合适的施测题目时，这时让计算机产生一个 0～1 之间的随机数 $x(0 < x < 1)$，只有当 $x < K$ 时，题目才真正被采用。另外，当题目的曝光率过高时，该题目将被设置为“休眠”状态，暂不使用。显然，K 越小，题目能真正施测的概率也越小，这样题目的曝光次数就会得到很好的控制。

3.4 起点算法与终止算法

表面上看，初始题目的选择对被试者最终得到的估计能力值并不重要，因为自适应考试的特点就是通过被试者对每一道题目做出的反应来逐步估计出其能力值。但是，实际上事情并不这么简单，不同的试题难度起点算法，其影响是多方面的。对于一个被试者来说，假如起点题目选得很难，他将通过不断地答错大多数题目来向自己的能力真值靠拢，心理上会感到挫败，这样会削弱他对后面题目的信心；相对的，假如起点题目选得很容易，他将通过不断地答对大多数题目向自己的能力真值逼近，其信心会激发，但这也有可能使其产生麻痹思想[8]。

为了使上述情况有所缓解，起点算法可以采用给所有被试者施测一个平均难度相等但却包含了难度不等的 m 道题目的测验。这些测验在题库中有很多等值的副本，每一个被试者可以随机选择，但是试题的曝光次数要受到一个参数的控制。被试者做完这个测验后，使用贝叶斯方法估计其能力参数的初值。

而终止算法则采用达到了固定的测验长度与达到了测验估计精度相结合的方法。当然，如果发现有人在考试过程中作弊，考试也可以在人为干预下强行终止。整个考试过程可以用图 2 来表示。

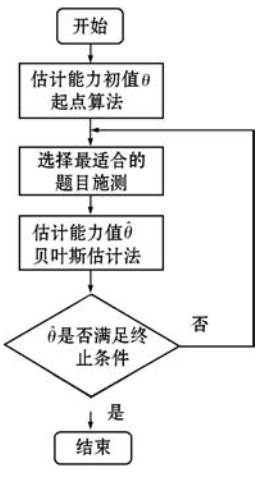

图 2 自适应考试系统流程图

3.5 题库建设和系统实现

在试题库的建设过程中，为了节省人力、物力和节约时间，采用专家估计和似然法相结合的方法来求出试题的难度、区分度及猜测系数。同时，题库中的题目还要保证具有一定的宽度，即试题的考核点要覆盖考试几乎所有的内容；题目又要有足够的深度，即在每一项考试内容上都要有足够多的和难度层次不同的题目来对能力水平不同的被试者施测，这样才能估计出他们的能力值。由于自适应考试是根据“最大信息量”原则从题库中选取后续试题，即每选一道试题都要对题库中所有题目进行信息量的计算，这样会对服务器造成极大负担。根据式(3)可以计算出使项目信息函数取最大时的能力参数取值，因此，在构建题库时，题库表增加一个字段 ability，此字段存放的是最适合此题目的能力值。这样在选取后续试题时只需搜索与被试者当前能力值相匹配的题目即可，从而大大减轻了服务器的负担。题库表的一些字段说明如表 1 所示。

该自适应考试系统使用 Visual Studio 2010 和 SQL Server 2008 作为开发工具，对于后续试题的选择算法，采用遗传算法。因为遗传算法具有其他算法所没有的自适应性、全局优化性和隐含并行性，并且在解决问题时有很强的鲁棒性，所以采用遗传算法来完成后续试题的选取。

表 1 题库表字段说明

编号	字段名称	字段类型	字段说明
1	ID	int	题目编号(主键)
2	Content	text	题目内容
3	Answer	VarChar(2)	答案
4	a	real	a 参数
5	b	real	b 参数
6	c	real	c 参数
7	Ability	real	适用能力值

4 实验与评价

4.1 题目参数估计检验

题目参数估计是题库建设中的重要环节，也是自适应考试质量能否得到保证的关键[9]。在实验中，采用与现在常用的参数估计软件 Bilog 软件进行对比的方法，分别对取自题库中的 100 道题目进行题目参数估计。该自适应考试系统估计出的区分度和难度分别用 a、b 表示，Bilog 软件估计出的区分度和难度则用 A、B 表示。其中一部分题目的参数估计结果如表 2 所示。

表 2 题目参数估计值

Item	a	b	A	B
1	0.82	-2.29	0.75	-3.41
2	0.81	-2.23	0.72	-3.43
3	0.53	-2.13	0.38	-3.79
4	1.43	0.00	1.05	-0.81
5	1.64	-1.16	1.23	-2.38
6	0.73	0.01	0.53	-0.80
7	0.97	-0.62	0.68	-1.02
8	1.06	-1.22	0.80	-1.50
9	1.24	-1.34	0.99	-1.68
10	0.90	-2.53	1.14	-3.12

从表 2 中的实验数据可以看出,该自适应考试系统的题目参数估计结果是可靠的,基本上实现了 Bilog 软件同样的功能。由于可以对该系统的题目参数估计程序根据实际需求作进一步的修改和完善,因此,它将会更适合复杂的具体应用环境。

4.2　能力参数估计检验

由自适应考试的原理,即根据被试者对所呈现题目的反应数据,动态地估计被试者的能力值,并贯穿于考试过程的始终。这里基于题库中的题目,利用计算机模拟一个考试,并进一步模拟了被试者不同的得分模型。其中一部分的实验结果如表 3 所示。

表 3　能力参数估计值及估计误差

Seq	答对题数	能力值	标准误差
1	4	-0.167	0.533
2	5	0.045	0.491
3	5	0.011	0.497
4	6	0.373	0.494
5	7	1.042	0.801
6	9	2.127	0.603

从表 3 中的数据可以看出,答对题数越多,能力值的估计也就越大。其中第 2 个和第 3 个虽然答对题目的数量一样,但由于其答对的具体题目不相同,所以能力估计值也不相同。除了第 5 个的能力值估计标准误差稍大一点外,其他的能力值估计标准误差都很接近。这些实验结果无论是从理论上还是从实际考试经验上看,都是合理的。

5　结　语

基于项目反应理论的自适应考试系统在提高考试效率的同时着重考察被试者的实际能力水平,对于提高学生的学习自主性和积极性有一定帮助。由于在考试过程中呈现给考生的试题难度与其能力水平相适应,因此每一名考生的答题情况更为可靠,更能充分体现出考生的能力水平。经过改进后的该自适应考试系统,参数估计过程更加稳定,结果更加可靠;起点算法的改进提高了出题速度;同时,对试题曝光次数的控制保证了考试的安全性。今后,如何能更加准确地评估考生的能力值,以及如何改善出题策略,使得考试变得更加高效,是未来研究的重点。

参　考　文　献

[1] 罗永莲,郭玉栋. 经典测量理论在小型专业题库中的应用研究[J]. 计算机应用与软件,2009,26(10):105-106,129.

[2] 刘丽平,王文杰,郭世宁. 计算机自适应考试系统题库的设计与实现[J]. 计算机系统应用,2006,15(3):10-12,16.

[3] Linden W J V D, Glas C A W. Computerized adaptive testing: Theory and practice[M]. Netherlands: Kluwer Academic Publishers, 2000: 101-116.

[4] 张淑梅,辛涛,曾莉,等. 2PL 模型的 EM 缺失数据处理方法研究[J]. 应用概率统计,2011,27(3):241-255.

[5] Wainer H. Computerized adaptive testing: A Primer[M]. Hillsdale, NJ; Lawrence Erlbaum Associates, 1990.

[6] 全敏鸣. 基于项目反应理论的计算机化自适应测试系统的研究[D]. 上海交通大学,2010.

[7] 罗永莲,贾玉芳. 项目反应理论在题库建设中的应用研究[J]. 计算机应用与软件,2015,32(1):86-88,152.

[8] 黄伯平,赵蔚,余延冬. 自适应学习系统参考模型比较分析[J]. 中国电化教育,2009(8):97-101.

[9] 王芳,燕雁,赵守盈. 项目反应理论模型应用中需要注意的几个问题[J]. 中国考试,2015(2):20-24.

(上接第 89 页)

通过以上实验分析证明基于障碍距离的兴趣管理方法和混合法相比能够进一步减少系统的网络开销,算法效率稳定,有助于提高协同虚拟环境的可扩展性。

4　结　语

本文提出了一种基于障碍距离的兴趣管理方法。该方法考虑了虚拟环境中存在多个不规则障碍物的情况,对虚拟环境中的障碍物进行多边形建模,使用障碍顶点和障碍边表示障碍物;对传统混合法的兴趣匹配过程进行了改进,通过与障碍边进行相交测试计算实体间的障碍距离,基于障碍距离计算兴趣域,进一步减少不相关信息的传输。实验结果表明,该方法能够比混合法取得更好的过滤效果。本文的下一步工作是如何在此基础上降低算法的计算开销,进一步提高算法的整体性能。

参　考　文　献

[1] Montoya M M, Massey A P, Lockwood N S. 3D collaborative virtual environments: exploring the link between collaborative behaviors and team performance[J]. Decision Sciences, 2011, 42(2): 451-476.

[2] 甘茂华,阮丽娜,李昌国,等. 多人协作虚拟实验室综述[J]. 计算机应用与软件,2010,27(5):130-132,143.

[3] Liu E S, Theodoropoulos G K. Interest management for distributed virtual environments: A survey[J]. ACM Computing Surveys (CSUR), 2014, 46(4): 51.

[4] 周佳文,薛之昕,万施. 三角剖分综述[J]. 计算机与现代化,2010(7):75-78.

[5] Liu E S, Theodoropoulos G K. A parallel interest matching algorithm for distributed-memory systems[C]//Distributed Simulation and Real Time Applications (DS-RT), 2011 IEEE/ACM 15th International Symposium on. IEEE, 2011: 36-43.

[6] Li Y, Fujimoto R, Hunter M, et al. An interest management scheme for mobile peer-to-peer systems[C]//Proceedings of the 2011 Winter Simulation Conference (WSC). IEEE, 2011: 2747-2759.

[7] Yahyavi A, Kemme B. Peer-to-peer architectures for massively multiplayer online games: A survey[J]. ACM Computing Surveys (CSUR), 2013, 46(1): 9.

[8] 梁洪波,朱卫国,姚益平,等. 一种面向大规模 HLA 仿真的并行区域匹配算法[J]. 国防科技大学学报,2013,35(3):84-91.

[9] Denault A, Cañas C, Kienzle J, et al. Triangle-based obstacle-aware load balancing for massively multiplayer games[C]//Network and Systems Support for Games (NetGames), 2011 10th Annual Workshop on. IEEE, 2011: 1-6.

[10] Pan K, Cai W T, Tang X Y, et al. A hybrid interest management mechanism for peer-to-peer networked virtual environments[C]//Parallel and Distributed Processing (IPDPS), 2010 IEEE International Symposium on. IEEE, 2010: 1-12.

[11] 王小乐,刘青宝,陆昌辉,等. 一种处理障碍约束的聚类算法[J]. 计算机应用,2009,29(2):406-408,411.