

# Implementation of Item Recommendation in e-Assessment Framework

## 数字化测评框架下试题推荐的应用研究

Wang Hongrong

College of Educational Information Technology  
South China Normal University  
Guangzhou, China  
[dujiawang@hotmail.com](mailto:dujiawang@hotmail.com)

Wang Dongqing

College of Educational Information Technology  
South China Normal University  
Guangzhou, China  
[scnu\\_wang@scnu.edu.cn](mailto:scnu_wang@scnu.edu.cn)

**Abstract**—Examination culture is changing towards the assessment culture, and this evaluation philosophy migration is advocated in last decade in China, which country has thousands years of testing history. This transformation of assessment philosophy encourages our education practitioners to look deep insight into individuals' acquirement of knowledge, skill and ability when attempting to guarantee the accurate understanding of learning progress, so a robust item bank running behind the learning platform, e.g VLEs, is so indispensable while individuals request for immediate formative assessment. And accordingly, the items, which is for later use, should be systematically and consistently generated or recommended from multi cognitive attributes anyhow, so this paper gives a approach to balancing the contradiction of tagging excessive item attributes and computing efficiency of the recommendation process itself. A hierarchy of meta-level knowledge is designed so as to improve the items' searching efficiency. After that, it's highly emphasized that how the process of item recommendation runs in the self-paced learning activities when learners fail to answer the questions. In our approach, new items will come out continually until individuals successfully passing the interim test, and a fundamental of item recommendation algorithm is discussed. Finally considering the huge improvement space of this algorithm in later action, we only validate the elementary effectiveness of facilitating learning in a relatively simple way. An experience summary of our work is also provided.

**Keywords**—individualized assessment services; item bank; item recommendation

### I. 研究背景

个性化学习环境下的数字化测评, 其理论基础是人们的测评理念由考试文化向评价文化的迁移, 而计算机自适应测试为这种迁移提供良好的实验温床。从考试文化向评价文化迁移, 就是要为考生提供分数同时, 通过考试的各种数据, 为考生提供知识、能力、特长和潜质的评价报告, 让学生了解自己的优势与不足, 使其在学习活动中更有针对性地选择适合自己的学习资源<sup>[1]</sup>。同时, 这种迁移为数字化测评提供丰富的数据, 其隐藏的知识对促进学生学习、优化教学过程以及辅助教育决策均具有重要价值<sup>[2]</sup>。题库(item bank)子系统是数字化测评领域研究内容, 也

是 CAT (Computer Adaptive testing, 计算机自适应测试)领域的研究基础。题库在教学实践中具有独特优势<sup>[3][4]</sup>, 一是表现在教学管理上, 试题库具有高效、经济、灵活和保密的特点; 二是表现在教学测量上, 即由题库生成的试卷具有高质量、可控性等特点。题库子系统的自动组卷策略直接影响到用户对学习环境的体验<sup>[5]</sup>。例如, 题量过少导致的组卷效果差, 试题多次重复出现导致的舞弊等等。这样, 题库子系统的试题结构化, 和题库组卷算法设计、实现与应用成为数字化测评在题库技术研究子领域的必要研究内容。

在试题结构化方面, 人们从数理统计出发, 开发出 Ordering Theory、IRS、Diagnosis 等试题结构化方法, 来达到以最少的题量测评学生最多能力的目的。其中, 郭伯臣等人开发的 KSAT(Knowledge Structure based Adaptive Test, 基于知识结构的适应性测验)系统, 根据学生学习状况与知识结构来推送试题<sup>[6]</sup>。KSAT 方法首先为师生搭建各自的知识点结构树, 以形成补救教学知识点结构的方式, 以达到个性化学习支持服务的理想目标。

在抽题组卷算法方面, 目前学者多倾向于运用随机搜索算法、回溯试探法、优先权组卷算法、遗传算法等支持自动组卷。选题、组卷数学模型的功效因具体题库设计而定。诸如试卷结构、考试限时、试卷整体难度、知识点考查范围、章节考查范围等等约束有时多达数百条。因此, 选题算法设计及其效用评估始终是研究难点。

本文在数字化测评框架下, 从题库基本功能分析出发, 尝试解决“学生答错试题后的下一步流程如何进行”这一具体实践问题。本文第 II 节简介试题(项目)反应理论(Item response theory, IRT)与信息科学领域的多目标组合优化求解模型, 引出如何为答错试题的学生推送合适试题资源的论题; 第 III 节探讨以知识点层级结构树与教学目标相结合的方式, 实现题库子系统中的数字化测评; 第 IV 节总结了“试题推送”在题库出题、组卷实践中的经验和遇到的问题, 对下一步研究作出展望。

## II. 题库理论支撑与算法基础

题库子系统基本功能有试题编辑发布、自动组卷、试题推送、数据分析、试卷质量评测,辅助教育决策等。其中自动组卷与试题推送是题库的核心功能,均需选取参数、建立数学模型。

### A. 试题(项目)反应理论

IRT 是表征考生与试题之间联系的一种心理测量数学模型,以某些潜在、无法直接测量其数值的学习者变量(如“学习者能力水平参数”、“学习风格参数”),和难度、区分度等试题变量二者作为模型的参数<sup>[7]</sup>。IRT 根据学习者个体的测量误差指标划定个性化的学习指标,它所采用的试题参数不受来自统一样本的因素影响,例如,试题个体的难度、区分度,学习者个体答题时的猜中系数。在实际应用上,IRT 克服了经典测试理论仅对考生分数单一维度作统计分析带来的缺陷,广为人们接受<sup>[8]</sup>。在 GRE、GMAT、CAT-ASVAB 等大规模考试中得到了广泛的应用。

### B. 算法与数学模型

Adema(1996)等人推出的 CONTEST 程序采用了线性规划方法,将测验编制中的心理计量学和非计量学要求转化成线性规划模型中的目标函数和约束条件,为基于 IRT 理论的测验编制开辟了新的研究方向<sup>[10]</sup>。常用的随机选取算法、动态规划算法、启发式算法、模拟退火算法、混沌神经网络算法、遗传算法等,均可概括为多目标组合优化算法<sup>[11]</sup>。所谓“多目标组合优化理论”是针对离散变量的优化问题,又称组合优化问题,是在一定的约束条件下,从一个无限集或几个有限集的无限集中寻找最优解,结果可以是某整数、某集合、某排列或者某图<sup>[12]</sup>。其中利用遗传算法帮助智能选题成为热点,如基于矩阵编码的遗传算法、改进的遗传算法、基于自适应的遗传算法等<sup>[13]</sup>。

人们普遍认为,信息工程领域提供的高级算法向教育测评软件的迁移尚未完全成熟,而数字化测评的用户需求之一,是以较少的题量评测学习者,为学习者个体提供合适的试题。综合考虑,我们在第 3 节提出基于知识点结构的适应性测验子系统。

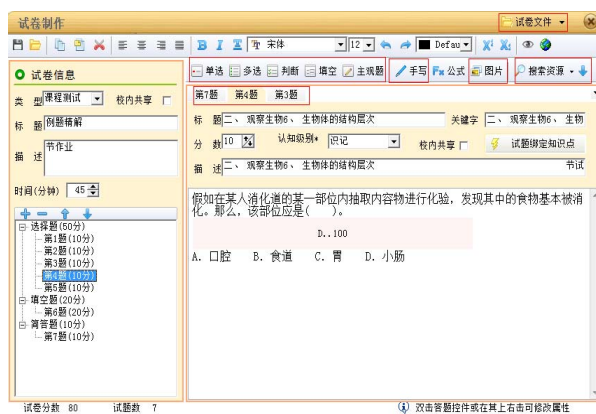
## III. 系统构建与初步应用

本节从评价文化的思想出发,借助计算机自适应测试理论框架,引出数字化测评在支持学习者的自主学习活动的设计与应用方案。

### A. 系统构建

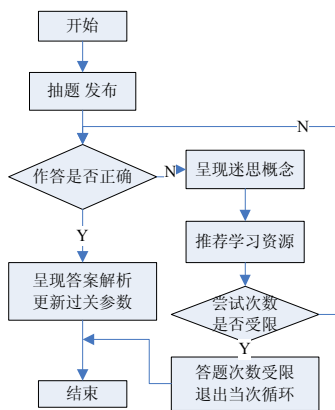
在题库搭建初期,试题难度、区分度等属性标注主观性过强,不能实施对题库信息系统的统计学测试,考虑以正确率等最基本统计量作题库抽题环节参考量,待初期试题得到一定统计、分析,甚至重组后再采纳高级选题算法。作为题库更新的依据之一,可利用 SP 表分析法独有的试题警告系数作为参考依据<sup>[14]</sup>。

鉴于具体业务需求,构建以知识点结构树为基础的试题推送,确立两类结构:课程知识点结构树与教学目标结构。其中,课程知识点结构树的确定,由学科知识专家 (Subject matter experts, SMEs) 和一线教师根据特定教材版本的具体教学内容,共同确定并填入信息系统,而没有按照统计学与数据挖掘的方法从海量数据中取得数据内在结构,这是对 KSAT 知识结构确定方式的新尝试;教学目标结构,则基于布鲁姆教育目标分类,结合基础教育实际需要,确定为“认知、理解、运用、拓展”四个层级,作为基础数据,填入信息系统。整个过程遵循常用的双向细目表试题标注法,是语言学科领域的“双环目标教学法”的学科扩展的尝试<sup>[15]</sup>。题库系统初期的试题标注是一项繁杂任务。为响应未来的试题个性化推送业务,系统每一道试题前后含有题型、手写输入、公式插入、图片插入、绑定资源、试题序号等 12 项显性信息,试题隐性信息多分布与试卷分析子模块中。系统截图如图一所示:



图一 试卷制作界面截图

信息系统支持学生自主学习之应答情况分析,控制较为复杂。初步设计如下:若学生试题应答失误,首先呈现回答错误试题所考察知识点明细,推送对应知识点教学资源;然后进入二次考察环节,将上次尚未回答成功的试题再次呈现,直至答题次数达到系统预设上限。其中选择题二次(或以上)的推送,将经过选项乱序等简单自动处理;其余客观题型与部分主观题的二次(以上)的推送,由管理员视试题统计结果而人工修改。最后,当学习者应答情况全部满足系统预设值,即满足试题解答数目要求、满足正确率要求后,界面终止抽题与推送。至此完成一次知识点的自主学习,并判定为“过关”。过程如图二所示:



图二 学生自主学习过程

以单元学习目标为例，初步设计的“推送试题”流程：首先供以“随机测试”功能，获取学生该单元试题应答情况，得到第一次错题集合；然后系统呈现“迷失概念”细目、相应课件资源和“试题推荐”，其中新增“试题推荐”之内容来自首次应答的错题集合；接着再次读取学生应答记录并重复上述过程直至应答次数升至系统预设的次数上限；最后全过程单元题解答数目、单元题正确率均满足系统预设值，完成此单元的自主学习，系统中止出题、判定“过关”。上述步骤亦适用于知识点学习目标。特别地，知识点与教学目标、单元与教学目标各自内部之间的“过关”，均设有“起评题数”限制，即仅在学生尝试应答次数达到系统预设上限后，才会启动资源、试题推荐功能。

上述过程发生在每一个 SMEs 定义的知识点结构树中。每当学生完成一次过关任务，该知识点学习状态均由红转蓝，如图三所示：



图三 自主学习“过关”信息细目

由于课程级别试题数目少，课程——目标学习的“过关”更趋复杂。课程——目标学习不可简单判定为：全部单元——目标学习“过关”，则自动判定为整门课程“过关”，逻辑复杂之处在于：新增“掌握率”设置。添加掌握率的设置可阻断“所有单元题均过关则整门课程过关”的业务。算法当中综合考虑单元级别试题应答总数、课程级别试题“起评题数”、单元（甚至知识点）级别试题应答正确率等三个因素，权重比值为 1:3:6，可酌情更改。

## B. 初步应用

2010 年 12 月我们在深圳市某初级中学的两个班进行平台的教学有效性探索。我们依据课程标准搭建了浙教版 7 年级《科学》第四章的课程结构树与教育目标结构，邀

请 SMEs 一起完成《科学》第四章 6 个小节 31 个知识点的试题命制，命题时遵循如下原则：

- 认知等级-分值分布：较高认知要求的知识点考题，其赋分高于较低者的赋分；
- 题型-分值分布：主观题的赋分高于客观题的赋分；
- 考点难度-分值分布：较高难度的知识点考题，其赋分高于较低者的赋分。

为防止相同应试能力学生接受相同的试题，提供“随机自测”以弥补部分简单试题从未被抽取的缺陷。“随机自测”场景出现在学生能力偏离教师预期、但仍不能完成学习任务时的情形。在我们的系统中，除算法本身之外，“随机自测”也帮助他们完成剩余学习任务。

## C. 数据分析

前文亦可归结为两类问题，一是属性与分值分布，二是试题曝光率控制。前者需从“题型——分值、认知等级——分值、考点难度——分值”分布所出现的规律出发，探索实际现象所暗示的规律；后者由上文提供的试题抽取过程可以推断：由于题量有限，若某类试题的实测难度高，将致使被多次抽取，亦即重复出现多次。为此作两种维度的分析，一是以试题属性标注(题型、认知等级)为自变量，观察真实测验结果的对应统计是否与标注值拟合；二是观察同一章节知识结构下新增或修改试题是否帮助学生过关。

(1). 对于属性标注，以知识点“6.1 物理变化和化学变化”所考察的试题为例，发现如下规律：

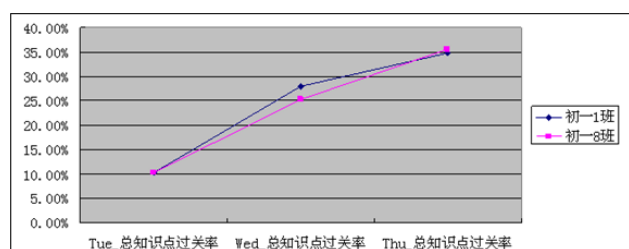
认知等级	试题数	平均正答率	平均收藏次数	
识记	51	65.45%	8.75	填空题30道 选择题62道 实验探究题2道
理解	41	33.00%	7.78	
应用	2	6.00%	2.00	

表一 原有属性标注及其度量值的演变

从 IRT 试题设计、编制理论上讲，试题难度、区分度与此处的认知等级并非同质概念，然而有研究证明二者存在较强关联。本系统尚处于完善阶段，对试题难度的属性标注并不科学，仅借用认知等级证明初步假设：试题难度越高，或者认知等级要求越高，则平均正答率越低。学生是否乐意收藏容易、中等难度，疑惑高难度的试题，则涉及个人学习风格等内部心理因素，暂不进行假设与验证。

(2). 对于新增或修改试题是否能帮助学生过关，需要解析基本算法。本算法的基本原理是通过新增同一知识点下的试题，采取题干数据置换、题型转换等客观题常用设计、编制方法，和对前一轮错误回答率从高到低排列的前 27% 试题的修改与删除等方式，让学生在解答同一知识点结构树时，以新增、修改试题的方法帮助达到最终的过关效果。其中判断知识点过关的条件是：在满足“测评最少题数”后，正确回答该知识点结构树下的全部 80% 试题。

本研究期望这套算法能通过新增试题的方式帮助学生掌握同类知识点。下表为学生知识点过关率演变趋势(见图表一),其中“总知识点过关率”为当天已通过知识点数与当天新增试题与回答错误导致的遗留试题之和的比值。



图表一 知识点过关率演变趋势

#### D. 系统改进

##### (1).系统中存在的问题

在使用平台时发现试题被抽中次数异常。例如:正确率过低的试题题量少、被频繁抽取,正确率过高的试题逐渐从题库中沉积。这说明这两类试题应删除或更新。

最终我们认识到,试题绑定到知识点结构树的最末端节点这一做法,一方面成为试题推送技术的一种可行方案,另一方面却使试题粒度过细,致使较难知识点因题量不足而导致相同试题多次重复出现。特别地,在自主学习情景下学生可调取其答案,也不利于该类题今后在正式考试中的组卷。

##### (2).改进措施:

- 试题更新——信息系统管理员与 SMEs 一起,参考题库报表模块显示的得分信息,首先将不合 IRT 要求的试题返工编辑,如改动题干、打乱试题选项、同一知识点下的题型变换、修改试卷题型-分值分布,知识点-分值分布,认知层级-分值分布等。籍此合理调整试题难度,增加较难知识点与失误较频繁知识点的题量。

- 区分试题功用——自主学习过程中的试题,由于答案解析对外开放,不能加入正式考试场合组卷;自主学习活动较多章节,增加题量、扩展其难度-分值分布、扩展其认知层级-分值分布,扩展其题型-分值分布,籍此避免同样的试题同时在正式考试场合,与自主学习场合出现。一种可行的方法是,在题库系统初期,采纳信息工程领域的选题算法为学习者团体自动组卷;待试题正答率贴近正态分布规律后,在试题某统计参数达阈值后,转而采纳其他数学模型用于评估学习者个体答题能力,从而完成试题资源的个性化推送。

#### IV. 结论与展望

本文简要介绍面向学习者团体与个体的选题算法与数学模型,在数字化测评框架下,尝试结合课程教学实际需求,以课程知识点结构树与教学目标结构相结合的方式,取得客观数学模型与教学规律、教育教学需求的平衡,帮助实现个性化测评的理想。信息系统底层选题算法是题库子系统的基础。一方面选题算法中的题源应该拓宽获取渠

道,另一方面学生应答界面的试题创建、分发、呈现应该符合学习(内容)管理系统的试卷/试题互操作标准。初步实践表明,单纯以学生应答失误的那部分试题,作为下一步试题推送的题源,并不符合教育理念;而试题在日常教学活动中使用如此频繁,人们却忽视其可共享性、重用性与互操作规范。本次实践中,由于自主学习过程中的试题泄露,致使真实数字测评结构与个人学习水平不符。

与一般智能组卷算法侧重研究多目标组合优化数学建模不同,本文主要工作是利用 CAT 理念初步尝试为学习者个体推送自适应的试题资源,重在提供个体测评服务。虽然二者皆有抽题组卷成分,但是前者侧重于如何按照考务管理员意志实现针对学习者团体的自动组卷,而本文侧重针对学习者个体的试题推送,在试题适应性上二者不重复。认知诊断测评的理论和方将是题库子系统下一步的研究方向。

致谢

本研究得到教育部考试中心“十一五”科研规划课题“技术支持下的多元评价理论与应用研究”(批准号:2009JKS2002)资助。

#### REFERENCES

- [1] Dongqing Wang, Jun Xu, A new method for learning environment design: engaging learning and development by applying multi-assessment. E-education Research, Vol 182, pp. 49-55, June 2008 (In Chinese).
- [2] Jun Xu, Quanbo Liu, Yushun Li, Dongqing Wang, Meilin Shi, Service-oriented grid computing: a new distributed computing architecture middleware, Science Press, 2009. (In Chinese).
- [3] S. Brown, P. Rice, J. Bull, Computer assisted assessment in higher education, Routledge, 1999.
- [4] S. W. Bonham, A. Titus, R. J. Beichner, L. Martin, "Education Research Using Web-based Assessment Systems", to be published in Journal of Research on Computing in Education, Fall 2000..
- [5] J. M. Linacre, S. Chae, U. Kang, E. Jeon, Computer-adaptive testing: a methodology whose time has come, Seoul, South Korea, Komesa Press, 2000.
- [6] Bocheng Guo, KSAT Computer adaptive test system and adaptive instruction remediation, NTCU, 2006 (In Chinese).
- [7] B.B. Reeve, An introduction to modern measurement theory, Division of Cancer Control and Population Sciences, National Cancer Institute, Feb, 2002.
- [8] F. B. Baker., The basics of item response theory, 2nd Edition, ERIC Clearinghouse on Assessment and Evaluation, 2001.
- [9] X.F Chen. A review and outlook for computerised academic test in Taiwan, Journal of Educational Research and Development, Vol 4, Issue 3, December 2007. (In Chinese)
- [10] J.G Zhang., Research of auto-generating test paper algorithm in intelligent tutoring system, Master's thesis, Henan University, Applied mathematics School, May 2009. (In Chinese)
- [11] H. Cui, Research on an automatic test paper generation system based on quantitative model and approximate matching algorithm, Masters' thesis, Jilin University, Software engineer School, October, 2007 (In Chinese).
- [12] N. Wang, Multi-object optimization for test paper construction, Journal of Shenyang Normal University (Natural Science Edition), Vol 2, February 2006 (In Chinese).

- [13] F.R. Wang, W.H. Wang., Q.K. Pan., Intelligent test-sheet composition research based on differential evolution algorithm, Computer Engineering and Design, Vol 8, August, 2008(In Chinese).
- [14] S.C. Zhong., Y.J. Zhong, J.N. Li, W. Wang, C.h. Zhang, The design of web-based intelligent item bank, Technologies for e-Learning and digital entertainment, Lecture Notes in Computer Science, Vol. 5093, pp.278-289, 2008(In Chinese).
- [15] X.M. Jiang, S.Q. Yu, The research on model and strategy of English teaching with hand-held devices, unpublished(In Chinese).