计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(5):139−144 [doi: 10.15888/j.cnki.csa.006353]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于知识点与错误率关联的个性化智能组卷模型①

潘婷婷, 詹国华, 李志华

(杭州师范大学 信息科学与工程学院, 杭州 311121)
通讯作者: 詹国华, E-mail: ghzhan@hznu.edu.cn

摘　要: 大数据环境下的个性化学习模型研究是大规模网络学习环境下的研究热点, 本文针对传统的智能组卷策略存在数据训练不足、个性化特点不突出、题库试题知识点分布不均匀等问题, 将大数据运用于组卷之中, 提出了基于知识点权重与错误率关联的个性化训练模型, 优化了抽题的法则并使得个性化特点更精确, 在一定程度上有利于学生对薄弱点和盲点的深入理解与消化. 本文采用将每章节题目的知识点转化为树形进行管理的方法, 并在知识点树中加入知识点错误率元素, 来优化基于知识点的抽题结果, 研究出适合个人学习情况的个性化模拟练习策略.最后将此新研究模型应用于教学教育系统进行实验研究, 研究表明对此关键点的改进更有利于普遍提升学生的整体成绩.

关键词: 大数据; 智能分析; 知识点权重; 树形结构; 个性化学习

引用格式: 潘婷婷,詹国华,李志华.基于知识点与错误率关联的个性化智能组卷模型.计算机系统应用,2018,27(5):139–144. http://www.c-s-a.org.cn/1003-3254/6353.html

Personalized Intelligent Composition of Test Papers Model Based on Knowledge Point Weight and Error Rate

PAN Ting-Ting, ZHAN Guo-Hua, LI Zhi-Hua

(School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 31112, China)

Abstract: The research of personalized learning model in big data environment is a hot research topic under large-scale network learning environment. In view of the shortcomings of the traditional intelligent test paper generating strategy, such as the lack of data training, personalized features are not prominent, and the uneven distribution of knowledge points, etc. This study puts forward a personalized practice model, optimizes the rules of paper organization, and makes the individual characteristics more accurate. To a certain extent, it helps student to understand and digest the weak points and blind spot. In this paper, in order to develop a personalized learning practice strategy for personal learning, the knowledge points of each chapter will be transformed into tree management, and add the knowledge point error rate element into the knowledge tree. Finally, this new research model is applied to teaching and education system for experimental research. Research shows that the improvement of this key point is more conducive to improve students' overall academic achievement in general.

Key words: big data; intelligent analysis; knowledge point weight; tree structure; personalized learning

“Internet+智慧教育”是当今教育界的主流, 且成果斐然.比如网易云课堂、微课网、沪江网、学而思网校等, 它们主要通过记录用户在线参加的课程培训、考试竞赛、试题练习、调查问卷和培训交流等情况, 实

① 基金项目: 浙江省自然科学基金 (LY17D060005)
收稿时间: 2017-08-15; 修改时间: 2017-09-06; 采用时间: 2017-10-12; csa 在线出版时间: 2018-04-23

现对用户学习情况的全程跟踪管理和对用户学习需求的全面掌握, 通过大数据分析, 实现个性化推荐课程, 本文在基于这样一个大数据时代背景下, 也对个性化学习中的一个分支——个性化试卷进行了研究, 致力于研究出适用于每个学生的个性化模拟试卷, 提高学生对知识点的掌握水平.

智能组卷策略是各校、各大型学习网站研究的重点, 本文将从模拟练习入手进行研究. 模拟练习的组织除了要关注题目的组织需符合考试大纲、知识点均匀分布、难易度符合约束条件、题型分布合理等外, 还需依据平时学习者的学习过程以及做题的错误率. 文献[1]提出了一种带权重的树形知识点管理策略, 该方法可以较好地解决智能组卷过程中知识点的选择问题[1], 但没有加入知识点错误率元素, 因此对知识点的深化学习需要进一步研究.

目前有许多抽题算法的研究, 如随机抽取法[2], 回溯试探法[3], 遗传算法[4], 蚁群算法[5]及鱼群算法[6], 大部分算法都能实现章节比例合理分配, 题型分值符合大纲要求等要求, 但也有各自的局限性.

比如, 随机抽取是由用户定义一些抽取条件即约束参数, 由计算机不断地随机从试题库中抽取个体, 直至满足抽取要求并不断循环往复的一个简单的过程, 此算法过程比较简单、易实现, 但较为呆板, 无法满足现如今多变的题库要求[7].

回溯试探法改进了随机抽取法, 它以深度优先的方式进行问题的搜索[8]. 此算法根据约束条件优先在试题库中随机抽取题目, 在题目组织时根据约束条件对选取的题目进行取舍, 如果该题不满足约束条件, 就废除最近的一次操作, 从某个回溯点重新往下搜索[9]. 回溯法很容易陷入死循环, 且极不稳定.

基于遗传算法的组卷是目前应用最广且效率较高的一种多约束条件优化算法.主要采用复制、交换和突变三种操作来求解问题的最优解, 具有鲁棒性、全局寻优、智能搜索等特点, 因此广泛应用于大型题库的自动组卷[10,11]. 但由于遗传算法易出现早熟的现象[12], 另外遗传算法中参数的确定没有普适的方法, 所以需要大量的实验研究来提升搜索性能[13].

综上, 本文根据知识点的考核要求需全面覆盖及在此基础上根据知识点的薄弱及掌握情况来智能抽题进行深入探讨, 并加入个性化因素, 结合优化的遗传算法实现知识点与错误率关联的个性化智能组卷策略.

1 个性化组卷模型

本文个性化抽题模型的流程从数据准备开始, 量化题目难易度, 量化题目知识点权重, 知识点权重包括基本权重以及错误率, 然后进行抽题, 并通过约束条件进行抽题限制, 最后实现即符合大纲又满足用户个人学习特点的练习.

为了保证抽题难易度适中, 不至于太难或太简单, 本模型通过计算公式 (1) 和计算公式 (2) 计算出题目难易度, 将难易度分为 5 个等级, 并通过区间分层将难易度进行量化, 量化值为 1, 2, 3, 4, 5, 表示为不同的难易度.

为了保证知识点均匀分布, 本模型采用树形结构管理知识点, 首先定义叶子权重, 通过计算公式 (3) 完成父级权重量化过程, 为了提高错提率高的知识点选取概率, 通过计算公式 (4) 增加权重, 这样通过知识点权重的判断可以在实现知识点均匀分布的基础上提高抽题概率.

通过难易度值以及知识点权重值控制抽题的概率, 然后根据约束条件, 进行题目抽取的约束, 并最终抽取符合约束条件的题目组合, 最后进行个性化训练或大规模考试. 将最终的考试结果进行分析来进一步优化题库, 优化题目难易度、以及知识点错误率, 为接下去的组卷信息提供实时更新, 组卷流程如图 1 所示.

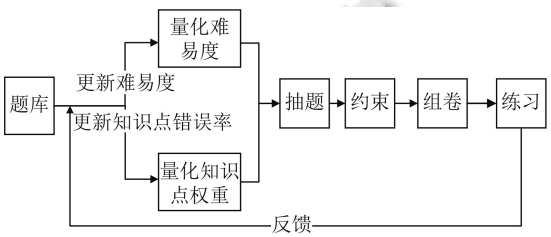

图 1 大数据环境下个性化组卷流程

2 个性化智能组卷步骤

2.1 数据收集与分析

如今已迈入大数据时代, 大数据与传统教育数据的本质区别体现在采集来源和应用方向两个方面. 传统教育数据注重体现学习者整体的学业水平, 而大数据则更关注每一位学习者个体的微观表现, 大数据时代的在线教学能够实现实时跟踪教学, 全面记录及分析掌握数据, 和可视化学生的不同学习特点、学习需求和学习行为[14]. 大数据在教育领域中应用深广, 可以为不同的学生建立属于自己的学习模型和适合他的个性化学习路径.

数据收集是一切结果的来源, 个性化学习的研究需要从收集到的海量数据中, 分析出学习者学情变化的规律, 依靠学生的学习行为分析出这些行为隐含的关联, 并预测出学习者接下去的学习行为及其学情发展趋势. 如表 1 所示是 C 语言程序结构部分知识点, 在系统的题库中每道题目都绑定了所属知识点, 每个学生都会有自己的做题数据信息, 通过分析每个学生所做的题目中的知识点错误情况, 即可得到该学生的知识点掌握情况, 如表 2 所示是某一位学生的知识点错误率表.

表 1 C 语言程序结构部分知识点

知识点序号	知识点描述
3	数据类型、运算符与表达式，变量、常量的定义
Ⅰ--3.1	变量和常量
Ⅰ--3.1.1	变量赋值
Ⅰ--3.2	数据的类型
Ⅰ--3.3	程序编写基础
Ⅰ--3.4	算数运算符和算数表达式
Ⅰ--3.4.1	各类数值型数据之间的混合运算
Ⅰ--3.4.2	赋值运算符合赋值表达式
Ⅰ--3.4.3	逗号运算符和逗号表达式

表 2 知识点错误率表

编号	知识点	错误率(%)	掌握程度
1	3.1	0	
2	3.1.1	0	
3	3.2	4.30	
4	3.3	3.22	
5	3.4	17.70	
6	3.4.1	0	
7	3.4.2	2	
8	3.4.3	15.90	

2.2 题目难易度属性

一般难度高的题目得分率会比较低, 难度低的题目得分率会比较高, 而客观题的得分情况会比较集中, 主观题的得分会比较分散, 通过下面的公式来计算不同题型的难易度值.

对于其中客观题目的难度等级计算公式为:

$$Degree = 1 - \frac{L}{N} \tag{1}$$

其中, L 为这题正确的正确次数, N 为答此题的总次数.

主观题目的难度等级计算公式设为:

$$Degree = 1 - \frac{\bar{X}}{S} \tag{2}$$

其中, $\bar{X}$为该题得分的平均分数, S 表示该题的分数值.

根据题目难易度计算公式可得知, 题目难易值 $Degree \in [0, 1]$, 值越大则该题越难, 答对的人数越少, 本文将试题的难易度分为 5 个等级并进行量化, 如表 3 所示.

表 3 难易度值

难度等级	非常简单	简单	适中	偏难	非常难
难易度区间	[0, 0.2)	[0.2, 0.4)	[0.4, 0.6)	[0.6, 0.8)	[0.8, 1]
量化值	1	2	3	4	5

2.3 知识点权重

个性化组卷是基于用户行为分析和挖掘而提出来的, 目前大多数的个性化组卷都是通过分析错题库, 算出每道题的错误率, 对于错误率高的题目进行强化训练, 将易错题比例加入到组卷约束条件中, 从而实现个性化智能组卷.但以上方法却忽略了一点, 知识点的掌握是决定学生学习好坏的一个重要指标, 光进行错题强化训练并不能使学生完全了解所学内容, 只有掌握了知识点才能说学生学得好, 因此本文将每道题都关联知识点, 一道题可以关联多个知识点, 且会有交叉知识点, 文章就是从知识点出发, 错的越多的知识点才是学生所不理解的地方, 在智能抽题时要针对错误率高的知识点做强化处理.

在教学大纲中会要求考核的知识点的掌握程度, 一般有以下等级: 精通、熟练、掌握、了解.在本文中将这几个等级量化为 1, 2, 3, 4 权重, 用树形结构表示知识点关系图, 树形结构的知识点管理如图 2 所示.

该树形结构管理的知识点可以有效地解决知识点分布均匀的问题, 首先按照大类分为若干个一级知识点, 在每个一级知识点下定义分类更细的二级知识点、三级知识点, 在选择知识点时可以优先一级, 再选二级、三级. 父级权重需根据子级知识点的权重来设定. 本文计算父级知识点综合权重的计算方法:

$$W_j = \frac{\sum_{i=1}^{n} \frac{w_i}{p_i}}{\sum_{i=1}^{n} w_i} \tag{3}$$

其中, W_j 表示第 j 个知识点, w_i 表示子级知识点中第 i 个知识点的权重, p_i 表示该权重在此子级知识点分支上出现的频率.

以上方法是计算知识点的权重, 当用户还没有做过练习时可以根据初始化的权重来抽题, 抽出的练习题目每个用户的相差不会太大, 因为没有更多的数据来显示学生对于知识点的个性化信息, 并约束抽题, 当数据多了之后, 收集用户每次的数据, 分析错题并根据错题挖掘出学生对每个知识点的掌握情况, 提取个性化信息, 这显而易见的非常简单的. 本文中将错误率引

入抽题之中, 来体现学生的对于知识点掌握的个性化信息, 在基于权重的树形知识点管理结构树上加入知识点错误率元素, 根据树形知识点管理结构, 计算新的知识点权重, 既在抽题时对错误率高的知识点, 选择适当的提高其比重, 从而提高抽取率, 如公式 (4) 所示:

$$W'_j = W_j(1+y_j) \tag{4}$$

其中W'_j是新的权重, y_j 表示第 j 个知识点的错误率, 通过公式 (4) 得到新的权重.在抽题中, 错误率高的知识点, 就是对这些知识点提高了优先级, 重点关注这些错误率高的知识点, 能自动优先选择盲点易错点.

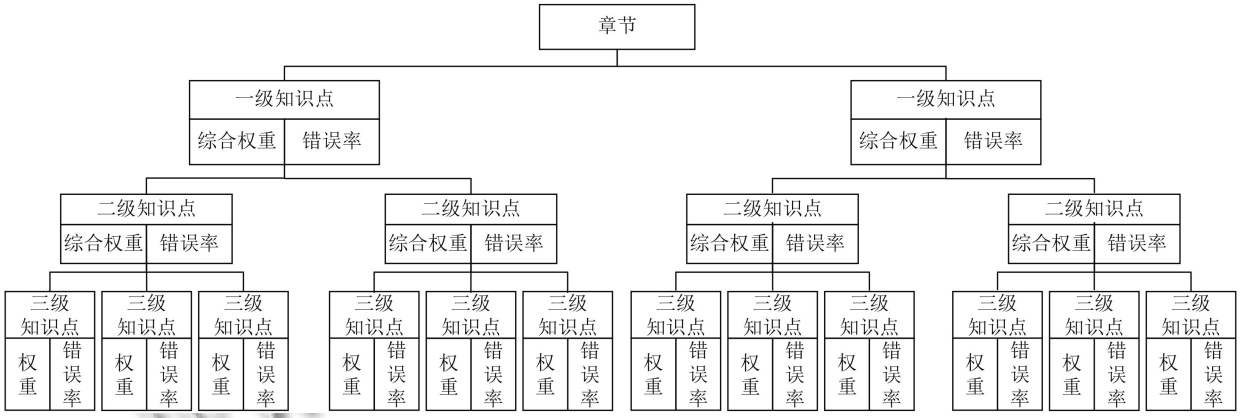

图 2 基于知识点权重与错误率关联的树形知识点管理结构

2.4 约束条件

个性化抽题原则除了有大纲考核的约束条件, 还有通过错题库展开问题和薄弱知识点分析, 最后结合起来进行个性训练, 抽题的约束条件有: 习题总分、答题总时间、知识点约束、题型分布、试卷难度分布, 如表 4 所示.

表 4 约束条件表

编号	约束条件	具体内容
1	习题总分	抽取题目后组成的一套练习的总分数, 一般为系统预设
2	答题总时间	要完全模拟考试.就必须规定答题时间
3	知识点约束	保证知识点分布或选择比例符合考试大纲
4	题型分布	确保题型分布符合考试大纲
5	试卷难度分布	要求组题成功后, 确保整体难度符合要求

根据这五个约束条件, 可以建立一个 P=N*7 目标矩阵, 其中 N 为一套练习的题目数, 如下面公式所示:

$$P = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} (m=7) \tag{5}$$

该矩阵中每一行代表一道选题, 每道选题有 6 个属性由目标矩阵列代表: 题目在数据库中所属 ID 编号=a_{i1}、题目所属题型=a_{i2}、题目分值=a_{i3}、关联知识点=a_{i4}、题目难度等级=a_{i5}、题目最佳答题时间=a_{i6}、题目所属章节=a_{i7}.

当目标矩阵符合约束条件, 才能完成选题, 否则要不断回溯继续优化, 直到所有的约束条件都满足为止, 试卷的约束条件计算表示如下所示.

约束 1. 习题总分约束: $TOTLE = \sum_{i=1}^{n} a_{i3}$为习题总分, 是所有题目的单个分支的总和, 每道题目的分值都是题库事先给定的.

约束 2. 答题总时间约束: $TIME = \sum_{i=1}^{n} a_{i6}$为答题总时间, 和总分一样, 由题库给出每道题的最佳答题时间, 这个值是一个大概范围, 只要在规定值的±10 之内即可.

约束 3. 知识点约束: $POINT(j) = \sum_{i=1}^{n} a_{i4} \times point(j)$/总知识点数, 表示第 j 个知识点的出现概率, 如果第 i 题属于该知识点则 $point$ 取值为 1, 表示出现了该知识点, 否则为 0, $point(j) = \begin{cases} 0, & j = a_{i4} \\ 1, & j \neq a_{i4} \end{cases}$.

约束 4. 题型分布约束: $S(j) = / \sum_{i=1}^{n} a_{i3} \times t(j) TOTLE$, 表示第 j 个题型的分数占比, 如果第 i 题属于该题型则 t 取值为 1, 否则为 0, $t(j) = \begin{cases} 0, & j = a_{i2} \\ 1, & j \neq a_{i2} \end{cases}$.

约束 5. 试卷难度约束: $DIF = \left(\sum_{i=1}^{n} a_{i5} \times a_{i3}\right) / TOTLE$ (n 为总题数), 表示整体难度, 值越高则难度越大, 值越小则难度越小[15].

3 个性化组卷模型实现

此模型通过优化遗传算法来实现, 改进后的遗传算法将题目的属性分为习题总分、答题总时间、知识点约束、题型分布、试卷难度分布 5 个维度.

步骤 1: 将题库的题目进行预处理, 首先将题目按照题型进行分类, 并按照不同题型对题目进行编码.

步骤 2: 初始化试题, 用 $Q=\{Q_1, Q_2,\cdots,Q_n\}^{\mathrm{T}}$ 表示题库中的试题, 试题总数为 N, 选出的目标题数设为 M, 每个试题 Q_i 的都是维数为 v 的向量, 即表示有 v 个属性.

步骤 3: 确定练习结构, 包括练习时长、练习总分值、练习难度、练习中每个题型的数目, 对于每种题型都有分数约束, 通过分数约束计算出每种题型所需的题目数量, 从而保证每个题型的总和分数满足练习总分约束条件.

步骤 4: 根据图 2 建立的分级树形知识点管理链表, 根据考核的要求指定三级知识点的权重, 根据式 (3) 分别计算出二级知识点和一级知识点的综合权重 W_i, 初始化知识点错题率为 $y_i=0$, 并通过公式 (4) 计算新的权重 W_i'.

步骤 5: 每类题型中依照章节进行二次分类, 每章的知识点依照权重 W_i' 降序排序, 对每个一级知识点下的二级知识点按权重降序排序, 对每个二级知识点下的三级知识点按权重降序排序.

步骤 6: 根据每种题型待选的题目数 D, 随机抽选选择 $1.5D$ 个知识点加入到待选试题中, 最后组成 M 道目标题目数.

4 模型测试与分析

为了验证本文所提出的模型, 实验选取了某校 iStudy 通用实践评价平台上的《C 语言程序设计》选修课程与学习本课程的 120 名学生为研究对象, 其中 569 道题库数据集, 首先将这些数据集应用于基于知识点权重与错误率关联的抽题模型和基于遗传算法的组卷模型[8], 对生成的知识点分布进行对比分析. 经多次试验, 最终选取种群规模为 60, 交叉概率 $Pc=0.6$、变异概率 $Pm=0.05$, 进化代数为 100, 图 3 是根据学生学习两个月和学习四个月的成绩数据生成的知识点分布图, 其中 (a)、(b) 是根据本文提出的方法生成练习知识点分布图, (c)、(d) 是基于遗传算法生成练习的知识点分布图.实验发现本文的模型生成的试卷在知识点分布上差异性比较大, 且知识点分布比较聚集, 这是因为根据学生知识点掌握程度来选择的, 而后者的模型在知识点分布上基本相同, 且比较分散没有强化性, 因此本文的模型能够针对学生的个性化差异提供不同的个性化练习.

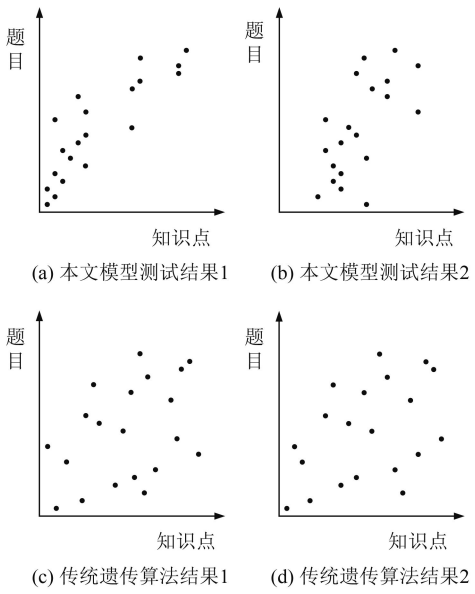

图 3 最终知识点分布

由此可见本文模型更能根据学生的特点生成适合于不同学生的个性化练习.在此研究下, 我们将本文模型应用于该门课程的选修学生中, 首先进行一次测验, 根据测验结果将学生平均分为两组, 这两组的学生测验的平均分相同, 然后用非本论文的基于遗传算法的普通组卷模型生成给一组的学生进行练习, 另一部分学生则用该模型进行练习, 经过八个月的测验后, 发现运用本模型进行练习的那部分学生的平均成绩的变化率自第五个月开始有较明显的提升, 直至第八个月他们的平均成绩已经达到超过十分之差, 实验结果如图 4 所示.

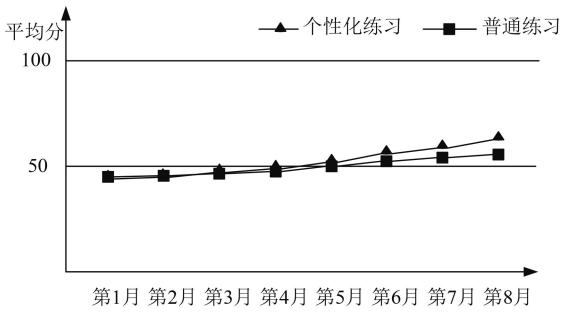

图 4 学生练习平均成绩变化图

由实验结果可知, 将大数据环境下基于知识点权重与错误率关联的个性化学习模型从知识点和知识点错误率着手, 对题库中的题目进行筛选组织, 从而实现了针对不同学习者的个性化练习, 提高了学生的学习成绩, 对提升学习效率有显著帮助.

5 结语

智能组卷是教学系统中重要的辅助学习工具, 提供一个针对用户学习特点不同的练习系统, 为用户最终的优异成绩打下坚实的基础是本模型的出发点.在本模型中通过知识点权重与错误率关联构建了基于知识点的个性化智能抽题练习策略, 测试结果表明, 相较于传统的策略, 此策略更能凸显学生的薄弱知识点范围, 并实现有针对性的训练.如何将学生的各种学习行为与本策略相结合, 从而完成更好的用户体验并提高抽题质量, 有待于进一步研究.

参考文献

1 鲁萍, 何宏璧, 王玉英. 智能组卷中分级带权重知识点选取策略. 计算机应用与软件, 2014, 31(3): 67–69.

2 周文胜, 潘中柱. 一种实用的随机组卷算法的设计思想. 湖南科技学院学报, 2005, 26(11): 299. [doi: 10.3969/j.issn.1673-2219.2005.11.111]

3 李大辉. 基于广度优先回溯算法的试题搜索算法. 大庆石油学院学报, 2006, 30(3): 100–101, 110.

4 全惠云, 范国闯. 基于遗传算法的试题库智能组卷系统研究. 武汉大学学报 (自然科学版), 1999, 45(5): 758–760.

5 李东, 王虎强. 基于动态蚁群遗传算法的士兵个性化学习. 计算机系统应用, 2015, 24(11): 204–208. [doi: 10.3969/j.issn.1003-3254.2015.11.034]

6 任剑, 卞灿, 全惠云. 基于层次分析方法与人工鱼群算法的智能组卷. 计算机应用研究, 2010, 27(4): 1293–1296, 1300.

7 胡泊, 刘欣. 基于改进随机选取法的自动组卷方法研究. 海军工程大学学报 (综合版), 2013, 10(3): 78–81.

8 孟祥娟, 王俊峰, 曹锦梅. 利用遗传算法实现试题库自动组卷问题. 计算机系统应用, 2010, 19(1): 180–184.

9 孙蓓蕾, 陈高云. 基于多策略的个性化智能组卷的研究. 成都信息工程大学学报, 2016, 31(3): 261–264.

10 吕健. 试论计算机自动组卷的常用算法. 电脑知识与技术, 2011, 7(8): 1802–1803.

11 唐启涛. 基于改进的遗传算法的智能组卷算法研究. 计算机技术与发展, 2014, 24(12): 241–244.

12 Yuan XH, Cao L, Xia LZ. Adaptive genetic algorithm with the criterion of premature convergence. Journal of Southeast University, 2003, 19(1): 40–43.

13 Li Y, Li SH, Li XR. Test paper generating method based on genetic algorithm. AASRI Procedia, 2012, (1): 549 –553. [doi: 10.1016/j.aasri.2012.06.086]

14 杨雪, 姜强, 赵蔚. 大数据学习分析支持个性化学习研究——技术回归教育本质. 现代远距离教育, 2016, (4): 71–78.

15 鲁萍, 王玉英. 多约束分级寻优结合预测计算的智能组卷策略. 计算机应用, 2013, 33(2): 342–345.