# 试卷评价系统算法设计与分析

王美华, 杨德贵
(华南农业大学信息学院, 广东 广州 510642)

摘要: 分析了试题库组织结构和试卷的评价指标, 对现行利用试题库自动生成的试卷的质量评价存在的主观性、后效性等问题进行了分析, 考虑到试题的“知识点不重复”、“题型”、难度系数等多项评价指标, 提出了试卷评价系统的调优算法. 系统重点利用试卷难度系数和对知识点分布的有效控制, 结果表明该算法对试卷的整体质量做出准确的评价, 有效地提高了组卷的效率和质量.

关键词: 试卷质量评价; 试题库; 题库设计
中图分类号: TP311. 52 文献标识码: A 文章编号: 1007- 855X (2006) 01- 0037- 04

## Design and Analysis on Test Quality

WANG Mei-hua, YANG De-gui
(College of Information, South China Agricultural University, Guangzhou 510642, China)

**Abstract**: The structure of test database and index of evaluation are analyzed. To overcome the common troubles, such as subjectivity and aftereffect occurring in the course of automatic generating the test paper from the test database, the best algorithm is put forward and the parameters of evaluation quality of test paper is taken into consideration, like the degree of difficulty. The results show that the algorithm is widely applicable to the test paper banks in scientific, engineering and medical fields, which certainly improves the quality of test paper.

**Key words**: evaluation quality of test paper; test questions bank; test paper database design

## 1 试卷的评价指标体系及参数

成卷系统的一个重要依据是试题库指标体系[1,5,6], 包括试题指标, 如题型、难度、时间、教学要求度等. 成卷系统的另一重要依据是知识库, 根据专家经验分别对每门课程设计出来. 其知识库的知识包括: 题型、大小题时间比例及允许误差、内容权重、难度 – 时间分布、难度档误差、每题时间 – 时间分布、赋分权重等. 这些规则分别规定了成卷所需的各种比例和分布, 如各类内容的权重以及题型选择的优先顺序等. 成卷时, 以知识库中规定的各种比例和分布来确定当前选题的策略.

成卷系统先获取用户的输入参数, 如考试等级、题型、总时间、考试范围、试卷类型等, 转化为成卷所需的大小题时间分布和难度时间分布. 成卷是通过智能推理选出试题, 若试题库题量足够多, 分布较合理, 则按模式选题就可以得到较为满意的试卷. 但若面对知识点题量不是很多, 分布也不尽合理的情况, 如果选题策略不当, 就很易造成选题失败. 成卷系统的选题总策略是: ①先选择大题, 再选择小题; ②先选择针对该课程重要的题型; ③难度选择遵循从难到易, 即优先选择难度高的题.

## 2 试卷质量评价的数学模型

成卷质量是试题库系统的关键, 根据各种可能出现的问题, 设计评价试卷质量指标体系, 智能化控制

收稿日期: 2005- 03- 15. 基金项目: 广东省自然科学基金资助 (项目编号: 04020079), 广东省计算机网络重点实验室开放基金 (项目编号: CN200403).

第一作者简介: 王美华 ( 1970~ ), 女, 硕士, 讲师. 主要研究方向: 智能技术与信息处理, 算法设计.
E- mail: wangmh92@ 21cn. com. cn

与优化成卷质量[2 3 4]. 根据成卷系统的实际运行情况, 归纳出可能出现的问题: 1)知识点覆盖率不够高; 2)出现重题; 3)各时间段的试题分布不够合理; 4)难度或题型分布不够合理; 5)对题量少的课程, 成卷失败的情况出现次数较高. 定义评价试卷的指标为: 知识点覆盖质量、试题不重题质量、难度分布质量、时间分布质量、大小题比例质量、题型分布质量, 并将每一质量评定分为 5级, 好: 90~ 100, 较好: 80~ 89, 一般: 70~ 79, 较差: 60~ 69, 差: 59以下. 成卷系统选题结束后, 可得到该份试卷中每一道题的有关信息, 包括: 试题号、难度、题型、内容号、时间、分数等, 定义各指标的取值为:

1)知识点覆盖质量 *CON*: 用 *CON* 1表示选中知识点的比例, 用 *CON* 2表示选中知识点权重总和的比例.

$$CON1=\frac{\text{选中知识点数}}{\text{应选知识点数}}\times 100\%$$

$$CON2=\frac{\text{选中知识点的权重总和}}{\text{应选知识点的权重总和}}\times 100\%$$

$$CON=(CON1\times 0.5+CON2\times 0.5)\times 100$$

2)试题不重题质量 REP: 基本保证连续 10份试卷中没有重题出现.

$$REP=\begin{cases}100-\text{重题数}\times 35 & \text{重题数}<3\\ 0 & \text{重题数}\geqslant 3\end{cases}$$

3)难度分布质量 *DIF*: 由于各档误差总和占总的允许误差的比例越大, 难度分布质量越差. 另外, 误差超出允许误差的档数越多, 难度分布质量也越差, 同时考虑到指标值应体现质量评定的等级, 如当 6档难度的误差都超过允许误差时, 难度分布质量应在"一般"等级 (70~ 79)以下.

$$DIF=100-\frac{\text{各档误差绝对值之和}}{\text{各档允许误差之和}}\times 5-\text{误差超过允许误差的档数}\times 5$$

4)时间分布质量 *TIM*: 类似于难度分布质量, 此指标值的推算式子为:

$$TIM=100-\frac{\text{各时间段误差绝对值之和}}{\text{各时间段允许误差之和}}\times 5-\text{误差超过允许误差的段数}\times 7$$

5)大小题比例质量 *TYP* 1: 类似地, 有

$$TYP1=100-\frac{\text{大小题误差绝对值之和}}{\text{大小题允许误差之和}}\times 5-\text{误差超过允许误差的项数}\times 15$$

6)题型分布质量 *TYP* 2 类似地, 有

$$TYP2=100-\frac{\text{各题型误差绝对值之和}}{\text{各题型允许误差之和}}\times 5-\text{误差超过允许误差的项数}\times 6$$

综合以上 6个指标的值, 可得到试卷质量的指标值, 用 *QUA* 表示试卷质量, 加权式为:

$$QUA=CON\times 0.2+REP\times 0.1+DIF\times 0.3+TIM\times 0.15+TYP1\times 0.15+TYP2\times 0.1$$

*QUA* 的值域是 [0, 100], 根据 *QUA* 的值就可衡量出试卷的质量如何.

## 3 试卷质量调优算法

满足下述任一条件的试卷称为合理试卷: (1)试卷质量 ≥80, 且各指标质量均 ≥70; (2) 75≤试卷质量 < 80, 但各指标质量均 ≥75. 当试卷满足这个条件时, 可直接编辑输出, 否则需进行调优. 当试卷不满足合理条件时, 需进行调优, 调优策略如下:

(1)若试卷总体质量 ≥80, 而存在某指标质量 (设为 *X* 1) < 70, 找出除"试题不重题质量"指标外的指标值最大的一个指标 (设为 *X* 2), 按某种方法剔除导致 *X* 1较小的一道试题;

(2)若试卷总体质量 < 80, 但存在某指标质量 > 85, 找出指标值最小的一个指标 (设为 *X* 1)和除"试题不重题质量"指标外的指标值最大的一个指标 (设为 *X* 2), 按某种方法剔除一道试题并重新选取一道加入试卷;

(3)若试卷总体质量及各指标质量均在 75分以下, 则随机抽出两个质量指标, 若其中一个为“试题不重题质量”指标, 则设其为 $X1$, 另一个为 $X2$; 若两个都不是“试题不重题质量”指标, 则设其中一个为 $X1$, 另一个为 $X2$, 然后按某种方法剔除导致 $X1$较小的一道试题并重新选取一道加入试卷;

(4)为了避免死循环, 当对一份试卷调用了 $n$次调优算法后还未达到合理条件, 则停止调优, 并将这 $n$份试卷中试卷质量最高的一份作为最终生成的试卷输出. $n$值应根据系统的运行速度及对时间的要求来确定.

由于可能导致试卷不合理的质量有知识点覆盖质量、试题不重题质量、难度分布质量、时间分布质量、大小题比例质量和题型分布质量. 所以调优算法按需调优的质量分为几个模块, 每个模块调优一个质量指标值. 以难度分布质量的调优模块 dif为例说明调优方法. 表 1列出了 3份试卷的难度分布计算数据:

表 1　试卷的难度分布数据

**Tab. 1　Difficulty distributing data of paper**

| NO | 1Y | 2Y | 3Y | 4Y | 5Y | 6Y | 1S | 2S | 3S | 4S | 5S | 6S | 1W | 2W | 3W | 4W | 5W | 6W | JS | YS | D | DIF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9 | 18 | 30 | 27 | 21 | 15 | 4 | 16 | 31 | 26 | 29 | 10 | 4.17 | 1.67 | -0.8 | 0.83 | -6.7 | 4.17 | 18.3 | 15 | 2 | 83.89 |
| 2 | 18 | 18 | 39 | 39 | 18 | 9 | 12 | 20 | 38 | 40 | 16 | 8 | 4.26 | -1.4 | 0.71 | -0.7 | 1.42 | 0.71 | 9.22 | 15 | 1 | 91.93 |
| 3 | 11 | 21 | 32 | 32 | 28 | 17 | 0 | 24 | 27 | 31 | 31 | 21 | 7.8 | -2.1 | 3.55 | 0.71 | -2.1 | -2.8 | 19.1 | 15 | 3 | 78.62 |

其中 NO为试卷号, 1Y~ 6Y 分别为在该份试卷中 1到 6档难度应占的时间, 1S~ 6S 分别为成卷后 1到 6档难度实际占的时间, 1W ~ 6W 分别为各档的误差, 正数表示实际比应占的时间少了, 负数则表示多了. $JS$为各档误差绝对值之和, $YS$为各档允许误差之和, $D$ 为误差超过允许误差的档数, $DIF$ 即为计算所得的难度分布质量.

由表 1可知, 导致 $DIF$ 偏小的原因可能是 $JS$值偏大或 $D$ 值偏大, 所以调优可通过调低 $JS$值和 $D$ 值来调高 $DIF$ 的值, 这两个值都取决于 1W ~ 6W 的绝对值, 所以调优的目标是降低某些误差的绝对值. 从表 1可知, 在每份试卷中, 1W ~ 6W 的值都是有正有负, 即有些难度档出的题所占时间比应占时间少了, 而另一些则比应占时间多了, 若能剔除占时多了的难度档的题目, 补选占时少了的难度档的题目, 则能降低 $JS$的值, 若调整后能令原来误差大于 2.5的难度档的新误差在 2.5以内, 则能降低了 $D$ 值. 所以在调优之前应先找出一个误差大于 0的档和一个误差小于 0的档, 然后在这两档之间调整题目, 若误差大于 0或小于 0的档不止一个, 则找出绝对值最大的. 如表 1中的第 3份试卷, 先找出误差大于 0且绝对值最大的难度档, 即误差为 7.8的第 1难度档, 再找出误差小于 0且绝对值最大的难度档, 即误差为 -2.8的第 6难度档, 然后剔除一道难度为 6的试题, 选取一道难度为 1的试题加入试卷, 由于同时符合被剔除的试题的内容、时间、题型和难度为 1的条件的试题不一定能找到, 所以可以放宽条件, 改变内容、时间、题型之中的一个指标值, 而改变哪一个指标值, 同样是由 $X2$来决定. 此模块的流程图如图 1.

## 4 结 论

文章分析了试题库组织结构的基础, 针对现行试题库存在的通用性、图形处理等问题, 考虑到试题的难度系数, 利用试卷难度和对知识点分布的有效控制, 提出了试卷评价系统的调优算法, 通过将主观的试卷质量评价转化为评价函数, 对自动成卷后的试卷整体质量做出准确的评价, 该试卷质量评价还增加与用户的期望试卷质量的对比, 若有较大出入, 再利用上述信息反馈后调优, 这样有效地提高了组卷的质量. 形成了一个整体的“评价 - 调优 ”模型, 通过在考试服务平台中对高教出版社发行的多门试题库的实际使用表明, 该模型是有效的.

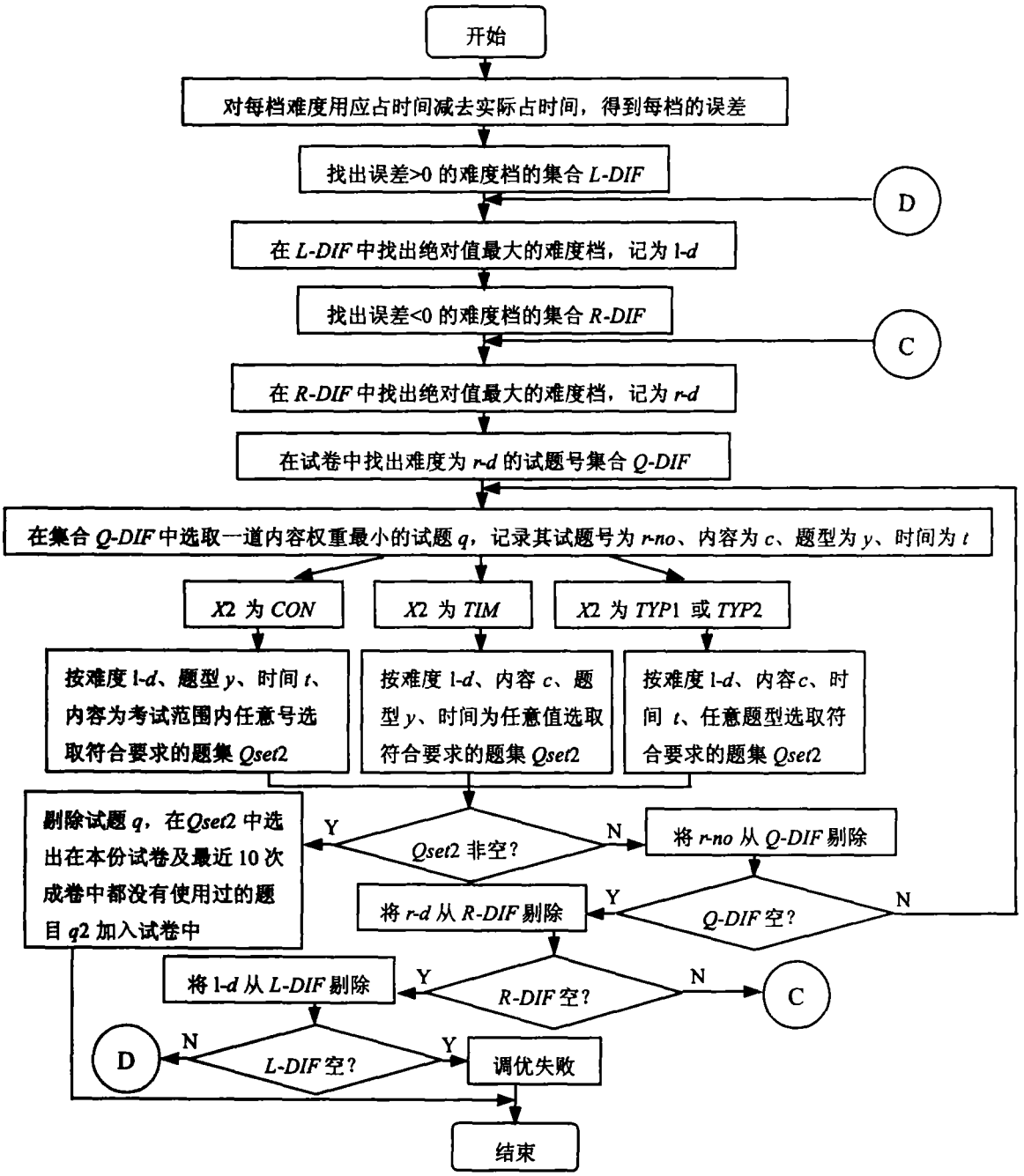


**图1 难度分布调优模块流程图**

**Fig. 1 Flow chart of the optimized model concerning difficulty distribution**

参考文献:

[1] 戴忠恒. 心理与教育测量 [M]. 上海: 华东师范大学出版社, 1987.

[2] 侯晓霞, 王子勇. 试卷质量分析与评价系统的设计 [J]. 计算机工程, 2000, 26(11): 151~153

[3] 王成, 周泽汉, 刘继纯, 等. 考试质量分析程序及应用 [J]. 上海第二医科大学学报, 2002, 22(1): 78~81

[4] 姜华, 胡欣, 李明革. 题库设计与成卷系统 [J]. 东北师大学学报 (自然科学版), 2000, 32(3): 23~25

[5] 田翔, 肖人岳. 一个改进的通用成卷模型 [J]. 计算机工程, 2004, (10): 183~185

[6] 李金平. 考试质量分析 [J]. 江南大学学报 (自然科学版), 2004, (4): 430~434