基于 web 的通用试题库平台的设计与开发①

Design and Development of a Web-Based General Question Library Platform

贾华国　刘　志　(浙江工业大学 软件学院 浙江 杭州 310023)

摘　要： 介绍了一种基于 web 的通用试题库平台的设计与开发方法，并针对在线试题库平台中存在的包含图文的试题录入困难以及基本遗传算法在组卷应用方面的不足，提出以图文混排和基于改进遗传算法自动组卷的解决方案，取得了很好的应用效果。

关键词： 通用试题库 改进遗传算法 图文混排

1　引言

目前计算机辅助教学已经成为现代高等教育的一个重要组成部分，作为主要的研究方向之一，网上试题库与自动组卷系统发展迅速。建立网上试题库是使各个高校优质试题实现共享的基础。自动组卷系统则是将人工智能技术与人类教育专家的组卷知识和经验结合起来，由计算机完成试卷内容的设计，使得在不同的组卷条件约束下，生成的试卷能够满足用户的需求。近年来，针对试题库建设与自动组卷系统，人们提出多种解决方案[1-3]，主要有两方面的难题。

网上试题库建立中碰到的主要难题是如何在线将包含各种公式，图片，文字的各门课程的试题方便合理的录入到试题库中。目前大多数的在线试题库仅能支持包含少量公式试题的录入，同时录入试题库后，公式相对文字的位置不变性的效果保存的并不理想，即无法做到图文混排。

自动组卷研究中碰到的主要难题是如何保证生成的试卷能最大程度的满足用户的不同需要，并具有随机性和科学性。同时由于在网络交互式条件下，对组卷时间提出更高要求。目前主要的组卷算法有随机搜索法，回溯试探法，遗传算法等[4-6]，其中遗传算法以其简单、鲁棒性强、全局寻找以及不受搜索空间限制性条件约束等优点日益受到关注，但基本遗传算法应用到组卷上存在一定的不足，比如编码方式不合适，会有早熟现象等。

针对这两个问题，文中设计并实现以图文混排技术和改进遗传算法为基础的通用试题库平台。

2　系统介绍

2.1 系统需求分析

系统可支持多门课程建立题库系统，通过建立新的课程名称、课程结构，在各知识点下还可建立更小的知识章节等操作，用户可以根据实际情况来构建课程体系，在此基础上，可以通过本地或在线进行试题的录入，其中，公式和图片被自动上传到服务器上，公式和图片在试题中的位置以及文字格式的保持问题，本文采用图文混排技术来解决。

因相同试题存在同构性，即很多试题内容相同，只是结构不同，系统需要支持对此类试题的鉴别。本文通过对要录入的试题部分关键字进行搜索，显示相似试题供用户查看来确定是否已存在该试题，如不存在，则可继续录入。

每个用户只需管理自己负责的知识点，故系统需支持对权限的管理。由系统管理员对每一门课程建立管理员，然后该门课程管理员再建立下一级管理员，通过层层赋值，将题库中的各个知识点维护工作落实到最终用户。

课程体系与用户权限都已分配完毕后，用户可进入系统对试题进行管理，如更新试题以及根据关键字对试题进行搜索，同时可进行组卷操作。组卷过程中

① 收稿时间:2008-09-15

可根据需求调整自动组卷的约束条件，系统组卷主要采用改进遗传算法来生成试卷。针对以上需求，可将系统设计为如下三个模块。

2.2 系统设计

1) 系统管理模块

包括系统的一些常规管理：用户管理，权限管理及密码管理。用户分为普通用户、教师用户、题库管理用户、新开课程用户、系统管理员用户。普通用户只能进行组卷，不能修改试题库的内容。教师用户能对试题库中自己负责的知识点进行维护，题库管理用户对所负责课程的题库进行维护，同时将课程的知识点维护权限分配给教师用户。新开课程用户能够建立课程结构。系统管理员用户，具有权限分配和创建题库管理用户与新开课程用户的功能。

2) 题库管理模块

是系统的核心模块之一。包括在题库中增加，修改，删除课程结构以及对试题的录入，修改，删除，查询等操作。其中对课程结构的操作与试题录入是该模块的重点。

实现通用的试题库系统，对课程结构进行增加来适应各门课程的需要是必不可少的。试题录入的功能，则需要图文混排技术的支持。因为对于不同课程，一道试题不仅包含文字，而且还有数学公式，以及使用 word 自选图形画出来的矢量图。为保持试题中公式、图片相对于文字的位置不变性，需要图文混排技术的支持。

3) 组卷管理模块

是系统的核心模块之一。主要功能是使用改进遗传算法自动组卷。进行组卷前，用户需在网页上输入试卷的一般性要求，例如：整卷的分数，整卷的难度，知识点分布比例等。相比部分系统需要用户输入试卷中各种题型的题目数量，在本系统中并无这样的要求，算法通过分析题库信息得到试卷中各种题型数量，同时根据用户的输入需求，使用改进遗传算法生成试卷，并在网页上显示。

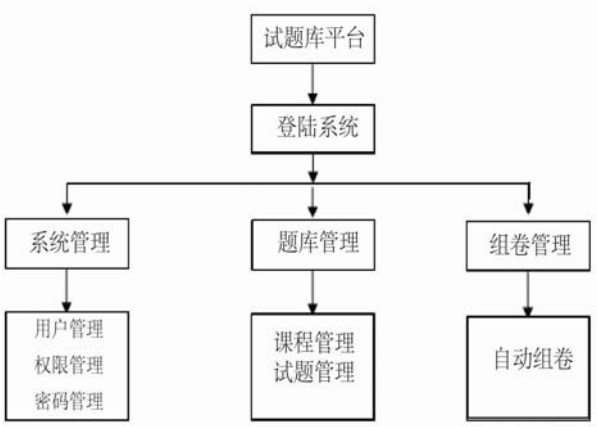

图 1　总体功能结构图

2.3 试题在数据库中的结构

试题的结构是题库的灵魂，是其他内容的基础。以下是系统中试题的结构表格。

表 1　试题的结构

字段名	数据类型	说明
Number	Varchar	试题编号
Question	Text	试题本身
Answer	Text	试题答案
Mark	Int	试题分数
Dif	Int	试题难度
Style	Int	试题类型
KnowledgeID	Char	知识点编号
ChapterID	Char	章节编号
BookID	Char	篇编号
Order	Int	出题次数
ViewLevel	Int	浏览级别

3 系统实现

3.1 通用性实现

实现试题库系统通用性的主要难点在于各门课程课程结构的不一致性。但通过比对，发现各门课程的试题属性是一致的。比如：试题分数，难度，所属章节等。为实现试题库的通用，经分析，本文采取以下做法：

设定“新开课程用户”，该用户根据需要由系统管理员开通。登录后，可在系统中建立新课程的课程结构，用户需要输入课程名，各门章节以及各个知识点，并通过数据库主键映射，建立三者之间的从属关系。同时为每一门课程的试题单独建立一张表，因试题属性的一致性，这些表的字段以及类型可以完全相同。经过这两项操作，不同课程的用户登录，系统可根据需要显示该门课程内容，使试题库能适应多门课程，真正实现跨多课程的通用平台。

3.2 图文混排

①实现难点

为保证题库质量，试题需要被输入到 word 文档

里面请专家评估，通过后将其录入到试题库中，之后在网页上查看。由于网页上文本框的功能有限，实现这样的图文混排有两个基本难点：

1) 保持试题中文本内容的格式

试题包含的文字信息，出于突出重点等原因，部分文字信息经过加粗处理、有下划线或者会略微倾斜，如何保持这些格式的不变性，是试题库系统在试题录入过程中的难点。

2)保持试题中公式，矢量图等相对位置的不变性

试题中包含大量的公式，矢量图等信息，系统不能简单的将这些公式，矢量图当作图片文件上传到服务器，因为每个公式，矢量图都有其在试题中的特定位置，特别是有多个公式的试题，每个公式在文字信息中的相对位置是不能够发生变动的。如何保持这些公式，矢量图在试题中相对与文字信息的位置的不变性，也是试题库系统在试题录入过程中的难点。

②具体实现

本文将 word 文档信息转化成 html 格式信息后，进行保存来解决以上两个问题。

1)对于第一个难点，即如何保持试题里文本内容的格式。Word 文档信息转化为 html 格式信息后，自动将其中的文本格式信息的以 html 语言描述的方式保存下来。

2)对于第二个难点，需要更进一步的处理

将 word 文档信息转化为 html 格式信息的过程中，每一个矢量图类型的公式会被转化成两张后缀分别为.wmz 和.gif 的图，服务器端需要取得其中的后缀名为.wmz 的图片文件。为防止图片文件由于重名被覆盖，使用该文件的上传时间（精确到秒）重命名文件。得到 html 格式的源文件后，按顺序将其中的<v:imagedata src="" >字段中 src 后面的图片来源替换成图片存放的相对路径加文件名，然后将整个 html 源代码存放到数据库中。试题录入的工作就已完成，需要显示该试题时，将数据库中 html 源代码读出，通过浏览器解析显示在网页上。

③效果图

为对图文混排功能的效果进行检验，将包含图文的试题存入数据库，并显示在网页上。如图：

从图中效果可知，包含图片、公式的该题已被录入到数据库，从数据库中读出后，各个相对位置保持不变，即图文混排取得预期的效果。

图 2　试题显示效果图

3.3 本文使用的组卷算法(改进遗传算法)

针对基本遗传算法在组卷应用上存在编码设置不合适及会出现“早熟”现象等不足，提出解决方案，使用改进遗传算法作为试题库的组卷算法进行组卷。

①编码的设置

使用分段实数编码。初始化试卷时，使用数据库中的表示试题的编号作为基因，并且使同一种题型的基因编码放在一个向量上。所有题型的向量中的基因编码合起来就是一条染色体。如下图表示随机生成的初始个体中某一类题型的题目。

表 2　某一类题型的试题向量

02040111011	……	02040111058

分段实数编码的好处不仅在于缩短了编码长度，而且这种编码方法将题型，题数等试卷约束条件体现在编码中，避免遗传操作中出现不符合组卷约束的试卷组合，比如：不存在的试题这个非法解已被消除，部分减少了非法解出现的概率，有利于提高组卷的效率。

②适应度函数的设计

针对自动组卷的四个约束条件(度约束，分数约束，曝光度约束，以及章节分布比例约束)构造染色体满足组卷目标的误差函数为：

$$E = \sum_{i=1}^{4} e_i * a_i$$

其中，$e_i \in \{0,1\}$ 是相对误差，a_i 为该相对误差的误差加权系数并且 $\sum_{i=1}^{4} a_i = 1$。根据专家意见，每个约束条件相对误差的容差需要限制在(%~5%)实际操作中需要将该容差范围适当放大来适应题库的需要。若超出限定

的误差容差的值，则将该约束条件 e_i 的置为 1，否则 e_i 置为 0。对于难度约束和分数约束的相对误差计算，需要计算试卷个体的难度、分数和目标试卷难度、分数之间的差值。曝光度的约束是针对试卷的平均出题次数，需要计算个体的平均出题次数和题库中试题的平均出题次数之间的差值。章节分布比例的约束是针对知识点的分布情况的潜在约束，一般情况下，比较重要的章节，在题库中的题目也会比较多一些，算法将题库中各个章节的比例作为目标试卷的各个章节的分布比例，对每个章节比例上的误差值进行求和，得到最终的差值。根据各个约束的差值是否在误差的容差范围内，决定各个约束 e_i 的取值。

适应度函数和误差函数是成反比的，即个体相对于目标个体的误差越小，该个体的适应度就越大。根据该原则，构造满足组卷目标的适应度函数为：

$$Fitness = (E + \mu)^{-1}$$

其中，E 为误差函数的取值，μ 根据经验值可取为 0.01。

③进化算子设计

常用的选择算子有：轮盘赌选择，适应度比例选择等。本文采用的是轮盘赌选择算子。采用分段单点交叉，即进行交叉运算的时候，每类题型的试题之间都会进行一次交叉，相当于多点交叉。采用单点变异，即进行变异运算的时候，整个个体只进行一次变异，为避免个体中题型数量的改变，变异的时候从题库中选择相应的试题进行代替。在进行交叉和变异运算的过程中，要对非法解进行修正，由于不存在的试题这个非法解在编码阶段已经解决，这里主要解决的是重复试题，当有重复试题产生，需要重新进行运算得到合理解。

④抑制进化过程中出现早熟的策略

基本遗传算法进化过程中出现早熟的最主要原因是种群中多样性的丧失。进化初期，种群中优秀个体和较差个体之间的适应度差距较大，导致优秀个体由于竞争力突出而控制了选择过程，此时，需要将适应度之间的差距适当拉小。进化后期，种群中各个个体之间的适应度差距较小，很容易陷入局部最优解，此时，需要将适应度之间的差距适当拉大，增加种群中个体的多样性。

解决早熟的主要措施有对适应度函数进行尺度变换以及对变异，交叉概率等控制参数的自适应调整等。本文采用对种群中的适应度进行线性变换的方法来防止进化过程中出现“早熟”现象。通过该变化，在进化初期，将种群中个体之间的适应度差距拉小，防止竞争力突出的个体控制选择过程。进化后期，将种群中个体之间的适应度差距拉大，增加种群中的个体多样性，防止进化过程中的局部收敛。

具体做法：将种群中的个体按适应度大小从小到大排序，将每个个体在序列中的位置标记在数组 location[]中，把该个体经过尺度变换后的新适应度置为 location[i]*2。

⑤种群更新策略

采用最优保留策略，即对于每一代的最优个体，直接拷贝进入下一代。同时为保持种群中个体的多样性，最优个体同样参加变异和交叉操作。

具体做法：使用上一代中的最优个体来代替下一代中最差的个体。同时设定算法的终止条件不再是单一的有个体到达满意适应度就结束，而是计算种群中最优个体的适应度和种群的平均个体适应度之间的比例(end)来决定算法是否结束，本文的做法，当 end>=0.99 的时候或者迭代次数到达规定次数时(100 次)，算法结束，取出其中的最优个体作为问题的解，若超出 100 次迭代，则认为本次组卷失败。

4 实验结果分析

因图文混排效果实验图已在前文出现，故这里主要针对改进遗传算法的效果进行分析。表 3、表 4 以及图 2、图 3 为在默认的用户需求(难度为中等，分数为 100 分，对应知识点和题型分布需符合题库的题量分配)下系统进行自动组卷得到的 2 次典型结果。其中交叉、变异概率分别为 0.4、0.2。经测试，连续运行 100 次改进遗传算法总共花费 50.391 秒，且组卷全部成功，即运行一次的花费大约在 0.5 秒。因默认的为难度中等，分数 100 分的试卷，由表中数据可得，试卷中中等题占多数，同时分数误差为 3%，且章节分布大致符合题库中题量比例，故改进遗传算法在试卷难度、分数、知识点分布及运行速度上可满足的组卷约束要求。同时，通过对比图 2、图 3 中两试卷的题号可知，两套试卷仅有一道试题相同(04080413014)，即改进遗传算法满足不能连续出相似试卷的约束。

表 3　试卷中各种难度所占分值

试题难度等级	容易	偏容易	中等	偏难	难
试卷 1	11	25	37	20	4
试卷 2	2	23	58	10	4

表 4 为试卷中各章的分值分布，由于目前题库中第六章，第十一章，第十二章的试题还没录入，故此次组卷实验不包含这些章的试题。

表 4　试卷中各章节所占分值

章节	1	2	3	4	5	7	8	9	10	13	14	15
卷 1	3	23	9	0	9	11	17	2	4	8	3	8
卷 2	3	17	2	17	6	11	11	0	0	5	9	16

04080413011 04070212006 04080413010 01020511004 01020712005 03050612004
01010112002 04080413014 02030611002 01021211002 06140512002 01021221001
04080623001 05130923007 04080423002 05130423003 04070223001 04070523002
04090123001 02030124016 01020522004 03050625001 03050125001 03050623003
07150632003 04080633001 04100134007 07150834001 05131033010 02030133020
04070832007 01021244001

图 2　试卷 1 中试题题号

06140313003 03050114006 04080413014 04070712002 04080613002 04080413015
01021213006 03050613005 05130813006 02040414001 01010113008 01021222001
04070522006 06140423001 07150123002 07150124001 04080323007 02030323003
05131423001 01021224004 01021223001 04070822002 07150621005 07150722003
06140332004 02040133006 01020333007 07150632002 04070135002 01020933005
07150632001 02040143001

图 3　试卷 2 中试题题号

由实验可知，改进遗传算法在运行时间，以及对组卷各项要求的满足度上，可以支持在网络交互式条件下进行自动组卷。总体来讲，使用改进遗传算法能够获得满足组卷要求的试卷，且提高了组卷的成功率、缩短了遗传算法收敛的时间。

5　总结

基于 web 的通用试题库平台已成功应用到大学物理课程中，能够在线对试题进行管理，同时改进遗传算法也取得比较好的效果，但在试题检索功能中还需进一步改进，减少人工干预，增强试题查询的匹配度，提高检索效率。

参考文献

1 闭应洲,苏德富,陈宁江.基于矩阵编码的遗传算法及其在自动组卷中的应用.计算机工程,2003,29(6):73－75.

2 范明虎,孙斌.通用试题库管理系统的设计与实现.计算机工程与设计, 2007,28(9):2185－2188.

3 何春华.基于遗传算法的自动组卷系统的设计与实现[硕士学位论文].武汉:华中科技大学,2006.

4 王友仁,张砦,崔江,姚睿,储剑波.智能组卷系统的建模与算法研究.系统工程理论与实践, 2004,(9):85－89.

5 全惠云,范国闯,赵霆雷.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.

6 路景.基于改进遗传算法的智能组卷研究[硕士学位论文].长沙:中南大学, 2007.

7 Simon Brown.白雁等译. jsp 编程指南.5th ed,北京:电子工业出版社, 2004.