在线自适应测试系统的设计与实现 *

丘 威[1], 钟治初[1], 黄建妮[1], 张立臣[2]
(1. 嘉应学院 计算机科学与技术系, 广东 梅州 514015; 2. 广东工业大学 计算机学院, 广州 510090)

摘 要: 针对目前计算机自适应测试系统在远程网络测试中存在的局限性,提出了基于 XML 的在线自适应测试系统模型。通过题目自反应理论建立计算机自适应测试模型,并提出了一种约束试卷生成的参数模型,利用增量学习算法设计了组卷策略,解决了远程网络自适应测试系统中计算量大、易造成网络交互阻塞瓶颈等技术问题。最后给出了系统的设计模型,并描述了该系统的实现过程。

关键词: 题目反应理论; 计算机自适应测试; 在线
中图分类号: TP393 **文献标志码**: A **文章编号**: 1001-3695(2008)01-0184-03

Design and application of interaction on-line intelligence examination base system

QIU Wei[1], ZHONG Zhi-chu[1], HUANG Jian-ni[1], ZHANG Li-chen[2]
(1. *Dept. of Computer Science & Technology, Jiaying University, Meizhou Guangdong* 514015, *China*; 2. *School of Computer Science, Guangdong University of Technology, Guangzhou* 510090, *China*)

Abstract: Against the limitations of computer adaptive test(CAT) in remote network test, this paper constructed a new CAT model based on XML. Adopted XML with the relation data model mapping mode, put forward a restriction examination paper created parameter model, and designed group paper strategy using incremental learning arithmetic. This work solved the net flow bottleneck of CAT. Lastly it gave the system design and described the process of implement.
Key words: item response theory(IRT); computer adaptive test(CAT); on-line

教育测试是进行人才选拔和能力评测的主要形式。当前考试的指导理论主要有以真分数理论为代表的经典测试理论和项目反应理论[1]两种。经过了近百年的发展,经典测试理论建立了一系列题目分析的公式,如表示难度的 p 值、表示区分度的题目与测试相关系数、估计分数真值的标准误差及由此推算出来的信度公式等。经典测试理论对建立试卷、考分转换和等值等均有一套较为完整的方法。但此理论仍有不够完善的地方,如考生分数和题目难度有着密切关系,即题目难度是相对考生而言的。如何使得题目参数稳定而不受受测样本影响,出现了项目反应理论。项目反应理论是以受测者回答问题的情况,经题目特征函数的运算,推测受测者的能力[2]。

根据应试者对题目的反应信息量,选择难度与应试者能力相匹配的题目,能够准确、快速地检验被测试者的能力水平,弥补古典测试理论的不足。但由于 IRT 实现技术上需要实时了解被测试者答题情况,并进行大量计算,实际的应用一直受到技术条件限制,一般需要计算机辅助,以 IRT 理论为指导建立计算机自适应测试系统。早期最著名的测试系统 LOGIST、BILOG 等都是单机形式。计算机网络技术的发展为测试理论进行大规模推广提供了技术支持。近年来测试理论的研究与实践应用取得了引人注目的发展,如美国的 GMAT、TOFEL、微软的 MCP 等考试都采取了 CAT 的形式[3]。可见 CAT 代表着今后教育测试的发展方向和重点。本文提出的基于 XML 的在线 CAT 系统模型,解决了实时交互带来的网络带宽问题。

1 计算机自适应测试理论与分析

根据项目反应理论,能力为 θ 的人答对题($u=1$)概率为

$$P(u_i=1|\theta)=c+(1-c)/[1+e^{-a(\theta-b)}] \quad (1)$$

其中:θ 为受测者能力值;a 为题目的区分度;b 为题目的难度;c 为题目的猜测系数;P 表示能力为 θ 的人答对此题目的概率。

作者试题反应理论是教育测试领域中的一个重要理论。IRT 的基本思想[4]是:应试者的某种潜在特质与他们对题目的反应(正确作答的概率)之间存在一定的关系,并且这种关系可以通过数学模型表示出来。IRT 通过数学模型建立起了应试者能力、题目参数以及正确作答的概率之间的关系。

目前最常用的 IRT 模型有 logistic 模型。Logistic 模型是 1957 年伯恩鲍姆提出的一种二级评分 IRT 模型。此模型与实际测验结果匹配较好,分为单参数、双参数以及三参数模型。单参数以及双参数 logistic 模型都是三参数 logistic 模型的特例。应试者的表现情况与这组潜在特质之间的关系可通过一条连续递增的函数来表示。该函数叫做试题特征曲线(item characteristic curve,ICC)。事实上,将能力不同的考生的得分点连接起来所构成的曲线便是能力不同的考生在某一测验试题上的特征曲线。ICC 表示某种潜在特质的程度与其在某一试题上正确反应的概率。这种潜在特质的程度越高(越强),其在某一试题上的正确反应概率就越大。三参数 logistic 模型

收稿日期: 2006-10-19; **修回日期**: 2007-01-29 **基金项目**: 国家自然科学基金资助项目(60474072);广东省自然科学基金资助项目(04009465)

作者简介:丘威(1974-),男,广东蕉岭人,讲师,主要研究方向为 CAI(qiuwei@jyu.edu.cn);钟治初(1964-),男,广东五华人,副教授,主要研究方向为软件工程;黄建妮(1975-),女,广东大埔人,助教,主要研究方向为信息处理;张立臣(1962-),男,吉林长春人,教授,博士,主要研究方向为软件工程.

的题目特征曲线如图 1 所示。

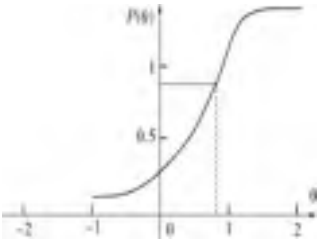

图 1　三参数模型的题目特征曲线

其中:a 参数代表题目的区分度,即特征曲线在拐点处的斜率,它的值越大说明题目对应试者的区分程度越高;b 参数代表题目的难度,即特征曲线的拐点在横坐标上的投影;c 参数代表题目的猜测系数,即特征曲线的截距,它的值越大,说明不论应试者能力如何都容易猜对本道题目。

基于 IRT 的测试系统在实施过程中要求实时反应,所以基于 IRT 理论的测试一般都借助于计算机进行。这就产生了计算机自适应测试系统。根据考生的估算能力值选择合适的题目,不断抽取与受测者能力相适应的题目是 CAT 的基本原则。准确估计被测者的能力水平是 CAT 顺利进行的前提。在测试过程中,IRT 对被测者能力的估计方法一般采用极大似然估计法。通常利用 IRT 题库中题目的最大信息函数来确定所选择的题目。IRT 用题目的信息函数 $I(\theta)$ 来表示题目参数与受测者能力的关系:

$$I_i(\theta) = P'_i(\theta)^2 / [P_i(\theta)(1 - P_i(\theta))] \tag{2}$$

其中:θ 表示受测者能力估计值;a_j、b_j、c_j 分别表示第 i 题的区分度、难度和猜测系数。

对于不同能力的受测者,题目有不同的信息量。信息量取最大值时,它所对应的能力值即是最适合于采用此题目测试的人员的能力值。因此在 CAT 系统中,根据前面推测的能力值,系统搜寻相应信息量最大的题目进行测试。另外,还可以采用 Bayes 方法选取试题。它是以能力估计值在测试后的改变作为选择标准,即选择使得能力估计值在测试后改变最小的题目进行测试。

正确估计受测者的能力是 CAT 顺利进行的前提。为计算其能力值通过对下式进行反复迭代:

$$\theta_{s+1} = \theta_s + \sum S_i(\theta_s) / \sum I_i(\theta_s) \tag{3}$$

其中:$S_i(\theta) = (u_i - P_i)P'_i / [P_i(1 - P_i)]$。直到式(3)的右边很小为止。

初始测试题目的选择一般采用随机进入法,由系统随机选择开始测试的题目。但为了更快地找到符合考生能力的题目,可以从以下几个方面考虑初始题目的选择:选择中等难度的题目,即假设应试者的能力为中等水平,由 CAT 系统在题库中随机选择中等难度的题目作为测试开始点,参数的设置由系统固定为中等水平。由应试者自行决定自己的初始能力水平,系统给出几个选项由应试者选择:初级、中级、高级等。每个选项的参数值都由系统内定。如果想更加准确地得到应试者的能力参数,可以通过预考的方式进行,即在正式测试前,给一定数量与测试内容相似的题目(如 10 道题,这些题目要体现不同的难度系数),系统可以根据应试者预考的结果大体估计考生的实际水平,从而粗略得出考生的初始能力参数[5]。如果是一个连续的网络学习环境,可以根据考生上一次测试的结果确定本次测试的初始能力参数。式(4)用于计算其标准误差,当值小于某个给定值时,考试结束。

$$SE(\hat{\theta}) = 1 \Big/ \sqrt{I_{rr}} \tag{4}$$

CAT 测试终止条件一般有如下几种方式:a)固定测试长度,即固定测试时间或测试题目数量,当时间达到一定期限或当测试题目数量达到一定个数时,测试终止。b)固定能力估计的标准差,当能力估计的标准差小于某一预先确定的值时,测验结束。这种方法能克服 a)的缺点,但如果终止条件定得过严往往会使测验时间过长。c)比较被测试者连续两次估计的能力水平,当比较结果小于某个预先设定的数值时终止测试。这种方法克服了 a)b)的缺点,同时能力水平估计结果与 b)非常接近,但所用的测验试题数目却比 b)少。计算机自适应测试能够用最少的测试题目来估计应试者的能力。在一些自适应测试的应用研究中证明,它只需测试 50% 左右的题量便能对被测者的能力进行准确的估计,有效提高了测量的精度和效率,适合网络自适应测试[6]。

2　试题组卷算法

设计一个模式优良的试题数据库,需要首先设计出它的实体联系模型。一道试题最重要的特征是它所考查的知识内容,即知识点。对题目难度衡量值的确定和修正应当是对知识层次和智力层次都相当的学生而言,同时它还应当建立在统计的基础之上[7]。为了满足设计功能的通用性,本设计分别实现了两种试题生成模式,即自动选题模式和手动选题模式。在自动选题模式下,需要解决如何在给出一种题型的题目总分数和题目总数的条件下,在试卷总分数、考试时间和卷面难度系数的约束下自动合理地选取试题,生成符合约束条件的试卷。约束条件也称为试卷指标,即一份试卷或一道试题应具有的参数特性。该参数特性包含:a)试卷组成指标,包括总分、题分、题目总数、类型题目数量、考试时间、卷面难度系数、知识点数量、各知识点所占比例等;b)单道试题选取指标,即选取试题库中某道试题所需要满足的条件,包括题型、难度系数、估时(完成该试题所需时间的参考值)、知识点等。

本设计采用一种增量学习算法来实现满足试卷指标的试题的选取。它的基本思想是考生在它的考试状态空间(历史记录)中执行动作(答题),以期获得它的目标。当考生从状态 N 到状态 $N+1$ 转换时,它接收历史记录行为的反馈信息。选题策略的目标是学习一种控制策略来选择一个试题(卷),从而使考生最大化积累反馈信息带来的“回报”。算法如下[8]:

a)初始化工作,系统给出试题的初始值,考生给出答题保留初始值;

b)对于系统的每一次所给定的累加值,循环 c)~g);

c)考生给出系统累加值的第 I 次预测值;

d)考生进行第 I 次答题达标值;

e)动态更新学习率;

f)系统用动态增量—学习算法学习考生的答题达标值;

g)考生是否达到预期的达标值,如果没有则转到 b);否则结束。

设计出符合用户要求和一定约束条件的试卷模式;然后再按试卷模式选取试题组成试卷。组卷过程是在考纲的题分、难度系数、试题覆盖面、题型比例等约束都满足的条件下,根据经验和考试目的,通过对不同的知识点赋予恰当的题型组合;并在此基础上确定各考题的难度系数,最终由具有这些特性的试题构成试卷的算法实现过程。

3　系统主要功能的关键技术实现

目前大部分测试系统都基于 C/S 结构,计算的逻辑主要集中在服务器端。在测试过程中,被测试者每做一道题目都要通过网络与测试服务器进行交互。服务器进行应试者能力的估计和试题的选择后,通过网络重新发布新的题目。这样,一旦用户过多,系统的负载就呈级数增长,网络不堪重负,很容易造成网络阻塞,影响测试的正常进行。现在都采用设置考点、将试题库下载到考点,然后考点通过局域网络的方式进行考试。这样虽然解决了网络阻塞问题,但不能实现完全开放形式的测试。考生必须在指定时间到指定的考点进行测试,测试的时间和地点受到很大限制。这种模式适合正式严格的能力测试,而对于通过远程网络平台进行学习的学生来说,测试的目的主要是考查对知识的掌握程度,并根据测试的结果及时调整自己的学习进度和思路。这样就无法实现真正的远程网络自适应测试,达到辅助学习的目的。为此本文通过引入移动 XML 技术,提出了基于 XML 的 IRT 远程网络测试系统框架;通过 XML 携带题目和测试策略移动到客户端的方式,测试可以异步进行,在技术上避免了网络交互阻塞问题,从而使真正的开放式远程网络自适应测试成为可能。

本文建构基于 XML 技术的跨平台分布性和数据与操作分离的、特性的试题库管理系统,采用在网络环境下的物理上分布、逻辑上分布的分布式数据库结构来设计试题库管理模型。试题文档库的数据交换功能有:a)客户端可根据自己的需求选择和制作不同的试题文档,对试题文档进行编辑和处理。服务器只需发出同一个 XML 试卷文件,数据计算不需要回到 Web 服务器就能进行。这样将大部分处理负载从 Web 服务器转移到 Web 客户端,从而使广泛、通用的分布式计算成为可能。b)由于 XML 具有数据显示与内容分开的特点,利用 XSL 就能对同一个 XML 试卷文档引用不同的样式表。可根据具体的教学环境需要预先定义 XSL 试卷文档的显示样式,得到不同的显示结果,使试卷文档的表现更加合理,最大限度地满足用户的分布化、开放化和个性化需求。c)在客户端能实现颗粒状刷新,即每当一部分数据变化后,服务器不需要重发整个结构化数据,只需发送变化的数据给客户。客户端不需要刷新整个使用者的界面就能显示出变化的数据。

服务器端用 XML 语言编写,使用 Microsoft XML-parser 作为 XML-解析器。用 XML 语言来描述题目的数据结构并以非常自由的格式存储,同时使用 XML 语言来分解原题目内容的语义。本系统的数据结构主要有两类:a)测试,描述某一测试的属性;b)题目,表示某一具体题目的属性。其 DTD(文档类型定义)的格式定义如下:

(a)测试的各元素定义

```
〈ELEMENT TEST(CDATE|STARTDATE? |ENDDATE? |) * 〉
  〈! ATTLIST TEST
    ID CDATA #REQUIRED ! 8 位数字长的一个惟一的 ID 号
    TITLE CDATA #IMPLIED ! 测试的名字
    MAXSCORE CDATA #REQUIRED ! 最高分值
    MINSCORE ADATA "0" ! 最低分值,缺省为 0
    PASSSCORE CDATA #REQUIRED ! 及格分值(测试通过分值)
    TIMELIMIT CDATA "0" ! 测试时间,缺省 0 为无限
    〉
```

例如,〈TEST ID = "20020112" MAXSCORE = "100" PASSSCORE = "60"〉…〈/TEST〉。〈IRT 参数〉语法为如下:

```
〈! ELEMENT IRT_PARAMETER EMPTY〉
  <! ATTLIST IRT_PARAMETER
    DISCRIMINATION CDATA "1.0"
    ! 在 IRT 中使用的题目的区分度
    DIFFICULTY CDATA "0.0" ! 题目的难度系数
    GUESSING CDATA "0.0" 〉! 题目的猜测系数
```

例如,〈IRT_PARAMETER DISCRIMINATION = "1.5" DIFFICULTY = "-0.8"/〉。

(b)题目的各元素定义

```
〈! ELEMENT QUESTION (CDATE|CATEGORY|IRT_PARAMETER
|CONTENT|HINT)〉
  〈! ATTLIST QUESTION
    ID CDATA #REQUIRED
    ! 8 位数字长的一个惟一的 ID 号
    TITLE CDATA #IMPLIED ! 题目的名称
    MAXSCORE CDATA #REQUIRED〉! 题目的最高分值
```

例如,〈QUESTION ID = "20020101" MAXSCORE = "10"〉…〈/QUESTION〉。

网络用户远程登录到测试网站,系统为每个考生生成一个专用登录助手。该助手负责为考生提供测试引导、信息交互等服务。一旦考生登录成功,该助手就由管理助手派遣,导航到考生客户端,并且负责考生与系统之间的沟通。

4　结束语

系统在反复论证的基础上,组织专家和科技人员进行认真的调研,针对各种考试方式开发出实用的考试平台。目前,本系统已经基本建成,正在试运行阶段,基本功能都已具备,但在远程自适应等方面有待进一步的研究和开发。以项目反应理论为基础的自适应测验是根据每个学员的不同情况,用几组不同的试题来测量学员能力水平的一种测验。自适应测验比常规测验具有更高的效率。它可以用比常规测验更少的试题而获得可与之相比或更佳的测量效果。

参考文献:

[1] 毕忠勤,陈光喜,徐安农. 计算机自适应测试系统的算法研究[J]. 桂林电子工业学院学报,2004,24(6):50-53.

[2] 张华龙,龙华. 计算机自适应考试技术在网络教育中应用[J]. 东华大学学报,2003,30(3):76-80.

[3] BRUSILOVSKY P. Knowledge tree: a distributed architecture for adaptive e-learning[C]//Proc of the 13th International World Wide Web Conference. New York: ACM Press,2004:104-113.

[4] 余民宁. 试题反应理论的介绍[J]. 研习资讯,2004(1):98-120.

[5] 吴志新. 基于 XML 的计算机自适应测试技术的应用研究[J]. 微机发展,2005,15(2):137-139.

[6] BRUSILOVSKY P. Developing adaptive educational hypermedia systems: from design models to authoring tools[C]//MURRAY T, BLESSING S, AINSWORTH S. Authoring tools for advanced technology learning environments: toward cost-effective adaptive, interactive, and intelligent educational software. Dordrecht: Kluwer Academic Pulishers, 2003:377-409.

[7] MURRAY T. Metalinks: authoring and affordances for conceptual and narrative flow in adaptive hyperbooks[J]. International Journal of Artificial Intelligence in Education, 2003,13(2-4):197-231.

[8] HENZE N. Personal readers: personalized learning object readers for the semantic Web[C]//Proc of the 12th International Conference on Artificial Intelligence in Education. Berlin:Springer, 2005.