An Empirical Approach to On-Line Learning in SIETTE

Ricardo Conejo, Eva Millán, José-Luis Pérez-de-la-Cruz, Mónica Trella

Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga. Campus de Teatinos s/n, 29079 Málaga. SPAIN
{conejo, eva, perez, trella}@iaia.lcc.uma.es

Abstract. SIETTE is a web-based evaluation tool that implements CAT theory. With the help of a simulation program, different empirical experiments have been performed with SIETTE with two different goals: a) to study the influence of the parameters of characteristic item curves and selection criteria in test length and accuracy; and b) to study different learning strategies for these parameters. The results of the experiments are shown and interpreted.

1 Introduction

One of the subtasks in an ITS is the evaluation of student's knowledge. SIETTE system [3] has been proposed as a general-purpose web based evaluation system. SIETTE implements *Computer Adaptive Test* (CAT) [5] methodology to improve its performance by reducing the number of questions needed to estimate student's level of knowledge, and is based upon the classical *Item Response Theory* (IRT). SIETTE has been designed as a reusable component to implement a *generic task* [1] for evaluating the knowledge level of a student about certain domain.

Teachers can continuously update the contents of SIETTE question database. This *open architecture* allows the system to evolve and improve its performance over the years. On the other hand, this *on-line* development of question databases is just the opposite of the desired scheme for classical item calibration. Fortunately, the potential great number of students that take the tests provides valuable information that can be used to successively improve teacher's estimations of item parameters.

The main contribution of this paper is an empirical analysis of two issues, namely, the behaviour of SIETTE when using incorrectly calibrated item pools and the feasibility of *on-line* methods for item calibration in SIETTE. The empirical method proposed and implemented uses a program that simulates the behaviour of teachers and students using Monte Carlo techniques.

Item Response Theory (IRT), also known as Latent Trait Theory, was originated in the late 1960s [2]). In a testing context, the *latent trait* is an attribute (*knowledge level*) that accounts for the consistency of test responses. Each question or item is assigned a function (*Item Characteristic Curve*, ICC) that represents the probability of answering to it correctly given the student's knowledge level $\theta \in (-\infty,+\infty)$. Let us represent this probability by the expression: $P(U_i=1|\ \theta)$ or simply by P_i. One of the

G. Gauthier, C. Frasson, K. VanLehn (Eds.): ITS 2000, LNCS 1839, pp. 605-614, 2000.

main problems in IRT theory is to find out the ICCs. It is usually assumed that ICCs belong to a family of functions that depend on one, two or three parameters. These functions are constructed based on the normal or the logistic distribution function. In the three-parameter logistic model the ICC is described by:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7 a_i (\theta - b_i)}}, \qquad \textbf{(1)}$$

where c_i is the guessing factor, b_i is the difficulty of the question and a_i is the discrimination factor. The guessing factor is the probability that a student with no knowledge at all answers the question correctly. The difficulty represents the knowledge level in which the student has equal probability to answer or fail the question, besides the guessing factor. The discrimination factor is proportional to the slope of the curve. If the discrimination factor is high then students with level lower than b will probably fail and students with lever higher than b will probably give the right answer. Assuming that the ICC belongs to this family, the problem of calibrating questions can be formulated as finding the best estimations for the parameters.

Section 2 of this paper describes the implementation of IRT used in SIETTE and the simulator program, and presents some empirical results obtained for correctly and incorrectly calibrated item pools. Section 3 describes a new on-line learning procedure that improves the behaviour of the system by learning item parameters. Finally some conclusions and open issues are addressed.

2 Simulating the Behavior of SIETTE

In this section, we will describe the techniques that we have used to emulate the behavior of the SIETTE system. First, we will describe how to simulate a correctly calibrated item pool.

2.1. Student, Item and Test Simulation

SIETTE implements the IRT model assuming that student's knowledge can be represented as a random variable θ that takes integer values between 0 and K_{max}. This simplification implies that only a fixed and finite number of states of knowledge are considered. Simulated students as proposed in [4] are used. Every student is represented by his/her value for θ. The simulation begins with the random generation of a population of N students, i. e., with the generation of N random concrete values for θ. These values are considered constant during the test (that is, no student learning occurs while taking the test). In the simulations described here the population has been generated to be uniformly distributed in 0, ..., K_{max}. However, other distributions have also been used, not yielding significant differences in the outputs.

Each item is represented by its ICC. An ICCs is also given by K values, corresponding to the conditional probabilities of giving the correct answer to the question given that the student belongs to each of the K classes. The simulator uses a set of Q void questions (ICCs), that are assumed to be correctly calibrated. These ICCs are generated by assigning values to the parameters a, b, and c in a continuous logistic function, and taking the corresponding values for the K percentiles. The