The adaptive and intelligent testing framework: PersonFit

Komi Sodoké
CAMRI Lab, UQAM
sodoke.komi@courrier.uqam.ca

Gilles Raîche
CAMRI Lab, UQAM
raiche.gilles@uqam.ca

Roger Nkambou
GDAC Lab, UQAM
nkambou.roger@uqam.ca

Université du Québec à Montréal (UQAM) C.P. 8888, Montréal (QC), Canada

Abstract

*E-learning has advanced considerably in the last decades allowing the interoperability of different systems and different kinds of adaptation to the student profile or learning objectives. However, some of its aspects, such as E-testing are still in their early age. As a consequence of this delay, most of the actual E-learning platforms only offer basic E-testing functionalities. By making efficient use of well known techniques in artificial intelligence, theories in psychometry and standards in E-learning, it could be possible to integrate adaptive testing functionalities in the actual E-learning platform. This is one of the goals for the platform that we developed named **PersonFit**. In this paper we will present some of its architectural elements and the algorithms used.*

1. Introduction

Although evaluations and assessments are important parts of the learning life cycle, in most E-learning platforms, E-testing functionalities remain in the early development stage [7]. In this paper we will focus on the building of an adaptive E-testing framework. The main contribution of this paper is the dynamic sequence of the items in adaptive E-testing.

2. From classical to adaptive testing

In their traditional format, tests or assessments are the same for all the examinees. They usually have a predetermined duration and fixed number of questions that have various levels of difficulty. The examinee's mark, which is generally the weighted sum of the scores obtained in all the questions, is used as competence criterion or to infer the examinees ability level in the concerned field. This testing format has some limitations and poses various problems of reliability and efficiency [6]. For instance, a high skill examinee could face some very easy items during the test or on the other hand, a low skill examinee could face some very hard items. In both cases, that situation could lead the concerned learner to a lack of challenge and motivation that could have a substantial impact on his test outcome. The precision of the test also is not the same in the whole range of the ability scale especially in the extreme values of the scale. Despite these problems, in most of the E-learning platforms, the tests often have the traditional testing format and they are a simple online "paper and pencil" version.

Questions and tests adaptation has been proposed as solutions for the traditional test problems. Therefore, some attempts have been made in order to integrate adaptive questions or tests functionalities in existing E-learning platform. For instance, Moodle [9] offered adaptive question functionalities based on the appearance and the scoring. Also, relevant works has been done on exam question recommender systems [5]. However, in both cases, the sequence of the items in the test didn't change according to the learner objective or profile. In the PersonFit platform adaptive testing, the first question, the overall sequence of the questions in the test as well as the end of the test can vary from one examinee to another according to their proficiency. Its implementation requires some challenges. One of them consists of finding computational models allowing to estimate and to compare in an objective way the proficiency of examinees that received different questions during a test. By taking into account these requirements, two models have been chosen: the Bayesian network and the item response theory (IRT). In the next section we will provide a short overview of the IRT.

3. Item response theory (IRT)

Item response theory is a set of related psychometric models that provides a foundation for scaling persons and items based on responses to assessment items [1]. The person parameter usually is the proficiency or cognitive ability within a specific domain and it is represented by the Greek letter θ. An item can be a test question.

Much of the literature on IRT focuses on its models. Those models are usually functions relating person parameter θ and item parameters to the probability of a discrete outcome, such as a correct response to that item. Among the available models, we use the three parameters logistic (3PL) model because its give us enough tools for our implementation. Those three item parameters are the discrimination (**a**), the difficulty (**b**) and the pseudo-guessing (**c**). The later represents the chance for a low level examinee to find by guessing the correct response to the item. The conditional probability for a person with an ability θ to get a correct response to an item i (a_i, b_i and c_i) is:

$$p_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-Da_i(\theta - b_i)}}$$

where D is a constant which value is 1.701.
IRT provides strategies for:

- Estimating the learner ability θ;
- Estimating the item parameters from data;
- Ascertaining how well the data fits a model;
- Investigating the psychometric properties of assessments.

One of the major contributions of IRT is the use of the concept of item and test information that could be estimated for a given value of θ. In PersonFit, we use the adaptive version of the MAP [2] method that ensures to have an average good value of θ within a good response time.

We choose the IRT and the Bayesian network because they are both probabilistic models. Its make them convenient for the probabilistic nature of some of the data that are processed in PersonFit. In addition, IRT has been successfully used in some computer adaptive testing system [6]. One of the great advantages offered by the IRT is that the estimated value of the learner proficiency is independent from the items used for the computations. If when the learner ability is known in a specific concept, IRT doesn't provide mechanism to make inferences about his proficiency in other related concepts. The Bayesian network comes as a handy complement of IRT for that purpose. They are also proposed to manage uncertainty in student modeling [4].

4. The PersonFit platform

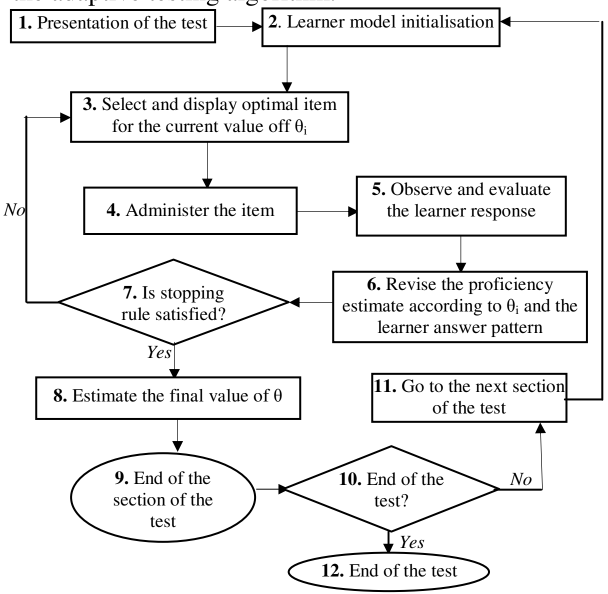
The PersonFit framework has two mains models: the domain knowledge model and the student model.

The domain knowledge model is elaborated from a learning perspective of the concepts in the concerned domain. For that purpose, a taxonomy of the concepts that will be evaluated in that domain is defined. That taxonomy is used to create categories which are physical representations of the concepts. Each category is used to store the items relevant to evaluate the concerned concepts. In addition, an IMS [7] compliant manifest file is placed in the root folder for each domain to describe it content.

The student model mainly contains the learner's cognitive state. It's consists of the long-term knowledge state, the short-term knowledge state and the episodic memory. First, the short-term knowledge state is a temporary structure that principally holds and describes the learner estimate proficiency θ, its standard error (S_θ) and the misfit of the answer pattern (L_z) for the current evaluation session. Their values are updated at each response given by the learner and they are used to select the next item during the test. At the end of the evaluation session the contents of the short-term knowledge state is used to update the long-term knowledge state. That long-term knowledge state represents the learner knowledge state as interpreted through the results of all of the evaluations that he had taken. It is represented as an overlay [4] of the domain knowledge. A Bayesian network is used for its implementation. Each node of the network represents a specific concept of the concerned domain and it expresses the acquisition probability by the learner. The episodic memory stores the traces of the learner evaluations and his actions during the evaluation sessions. It is physically stored as a log file.

5. Adaptive testing logic used in PersonFit

Figure 1 presents a simplified flow chart illustrating the adaptive testing algorithm.



1. Presentation of the test
2. Learner model initialisation
3. Select and display optimal item for the current value off θ_i
4. Administer the item
5. Observe and evaluate the learner response
6. Revise the proficiency estimate according to θ_i and the learner answer pattern
7. Is stopping rule satisfied?
No
Yes
8. Estimate the final value of θ
11. Go to the next section of the test
9. End of the section of the test
10. End of the test?
No
Yes
12. End of the test

Figure 1. Adaptive testing algorithm flow chart

The step 2 of the algorithm is about the learner model initialization. It aims to estimate the initial value of the learner proficiency named θ_i at the beginning of the current evaluation session. That initial or prior estimated value θ_i is used to select the first item of the adaptive test. If the learner proficiency in the field concerned by the evaluation or some of the concepts related to that field are known, θ_i is obtained by using an inference in the learner long-term knowledge state. But, if that's not the case, an average proficiency value based on the estimated proficiency of other learners in his group in that domain is used. If the information about other learners in his groups is not available, the value of θ_i is set to 0.

At step 3, the strategy used to select the next item consists of choosing the item that will give the maximum information for the learner's current estimated proficiency θ. It would be inefficient to search the entire item pool to compute the information given by each item for a current value of θ and select the most informative one. In practice, an information table is used. This table contains the list of the items ordered by the amount of information provided at various levels of θ [6]. Other constraints relating to a better coverage of the pedagogical contents of the domain and a minimization of some item's exposure are also taken into account. Finally, the episodic memory is checked to prevent the administration of the same questions during different evaluations.

At step 6, the learner proficiency estimate is revised using the adaptive version of MAP.

At step 7, a fixed value of the standard error is used as the stopping criterion. In case of non convergence of θ, the test is stopped after a fixed number of items.

6. Results

Experiments based on real data have been made in order to determine the accuracy of the adaptive testing implemented in PersonFit. We use the data from the "English as second language classification" test administered in Québec's colleges named TCALS. The non adaptive version of TCALS contains 85 questions and is administrated to thousands of examinees. It gives data file containing thousands of answer vectors coming from a real assessment situation. An experiment of an adaptive version of the TCALS using a sample of 515 examinees taken from the data file has been made. First, we would like to check if the initial value of θ (*named θ_i*) could have an impact on its final estimated value. Second, we want to see the number of items needed to reach the convergence of θ. Then, for the experiments, an initial value of the θ_i is set to -3. At each administered item, the average value of θ_i for all the 515 examinees is graphed giving us the dotted curve beginning at -3. The same procedure is repeated with an initial θ_i value of -2 and so on up to 3. The figure 2 shows the results of the experiment. It represents the estimated value of the ability θ for different number of items.

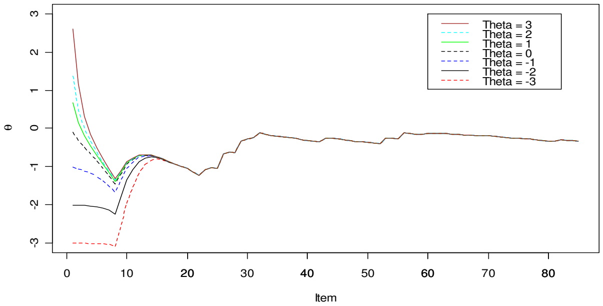


Figure 2. Real data experimentation result

Figure 2 shows that beyond the 15th administrated item, all the curves merged. Then independence from the value of θ_i is obtained. In addition, after the 55th item, there is no significant variation of the value of θ and administration of additional items does not bring us more information. The convergence is then obtained. Then, the adaptive version of TCALS only needs 55 items to estimate the learner ability.

7. Conclusion

Efficient use of IRT combined with Bayesian network allows the implementation of a functional system to administer an adaptive test named PersonFit. The results obtained by using PersonFit's adaptive testing algorithm on real data show how this algorithm is efficient.

References

[1] Baker F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland

[2] Baker. F. B. (2004). *Item response theory – Parameter estimation techniques*. New York: Marcel Dekker.

[3] Carr, B., Goldstein, I., *Overlays: A theory of modelling for computer aided instruction*, Technical Report AI Memo 406, MIT, Cambridge, MA 1977.

[4] Conati, C., Gertner, A., Vanlehn, K. (2002), *Using Bayesian networks to manage uncertainty in student modeling*, Journal of User Modeling and User-Adapted Interaction, volume 12 (4), pages 371-417, 2002.

[5] Hage, H. and Aïmeur, E. *Exam Question Recommender System*. Proceedings AIED (2005)

[6] Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

[7] http://www.imsglobal.org/

[8] http://ltsc.ieee.org/wg12/

[9] http://moodle.org/