A Heuristic Method for Large-Scale Cognitive-Diagnostic Computerized Adaptive Testing

Jill-Jênn Vie
RIKEN Center for Advanced Intelligence Project (AIP)
Tokyo, Japan
vie@jill-jenn.net

Fabrice Popineau
Laboratoire de recherche en informatique (LRI)
Orsay, France
fabrice.popineau@lri.fr

Françoise Tort
ENS Paris-Saclay
Cachan, France
francoise.tort@stef.ens-cachan.fr

Benjamin Marteau
French Ministry of Education
Paris, France
benjamin.marteau@education.gouv.fr

Nathalie Denos
Université Grenoble-Alpes
CNRS, LIG
Grenoble, France
nathalie.denos@univ-grenoble-alpes.fr

ABSTRACT

In formative assessments, one wants to provide a useful feedback to the examinee at the end of the test. In order to reduce the number of questions asked in an assessment, adaptive testing models have been developed for cognitive diagnosis, such as the ones encountered in knowledge space theory. However, when the number of skills assessed is very huge, such methods cannot scale. In this paper, we present a new method to provide adaptive tests and useful feedback to the examinee, even with large databases of skills. It will be used in Pix, a platform for certification of digital competencies for every French citizen.

Author Keywords

adaptive testing; cognitive diagnosis; item response theory; knowledge space theory; q-matrices; knowledge components.

INTRODUCTION

In online assessments, it is crucial to uncover the latent knowledge of examinees efficiently, in order to tailor the learning experience to their needs. Therefore, the cost to be minimized is the number of questions asked during an assessment. In summative assessments such as the ones encountered in certifications (GMAT, GRE), several models have been proposed in order to reduce the number of questions asked, using item response theory [5]. In formative assessments though, one also wants to provide a useful feedback at the end of test. To address this outcome, several models based on cognitive-diagnostic computerized adaptive testing (CD-CAT) have been proposed [2]. However, most of them are not suitable for learning at scale, i.e., they cannot handle many knowledge components.

In this paper, we show how to provide a CD-CAT that can handle more knowledge components, that efficiently computes the next question to ask, and that provides useful feedback in few questions.

This paper is organized as follows. We first present the existing models for CD-CAT and their limitations. Then, we present the curriculum we want to assess, and our new heuristic method. Finally, we describe the context in which our method has been used: a certification of digital competencies for all French citizens. Our conclusions follow.

L@S 2017, April 20 - 21, 2017, Cambridge, MA, USA

ACM ISBN 978-1-4503-4450-0/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3051457.3054015

MODELS FOR COGNITIVE-DIAGNOSTIC COMPUTERIZED ADAPTIVE TESTING (CD-CAT)

In order to provide a feedback called cognitive diagnosis, most models rely on a *q-matrix* [6], i.e., a binary matrix that draws a link between questions and knowledge components (KCs) required to solve them. Formally, the (j,k) entry of the q-matrix q_{jk} is 1 if the KC k is involved in the resolution of the question j, 0 otherwise. Using this q-matrix, it is possible, based on the performance of a learner, to tell them their strong or weak points at the end of the test.

An adaptive test can be represented as a tree-shaped automaton called CAT tree, of which the states are the questions asked, and the edges are labeled with 0 or 1 according to a false or true answer from the learner. Thus, an execution of the adaptive test can be seen as a path in the automaton, according to the learner's performance.

We will now present two adaptive testing models based on a q-matrix: the DINA model and the attribute hierarchy model, also related to knowledge space theory.

DINA model

In the DINA model (Deterministic Input, Noisy And), a learner should master every KC involved in the resolution of a certain question in order to solve it. This model is also robust to careless errors from the learners, using *slip* (s_j) and *guess* (g_j) parameters for every question j. The learner has a latent state $c \in \{0,1\}^K$ of which the k-th component c_k represents their mastery or non-mastery of the k-th KC. The probability that a certain learner answers the question j correctly is:

$$\begin{cases} 1-s_j & \text{if the learner masters every KC involved in } j \\ g_j & \text{otherwise.} \end{cases}$$

Formally, a learner of latent state c masters every KC involved in j if for all KC k, $q_{jk} = 1$ implies $c_k = 1$.

Using this model, it is possible to infer the most probable latent state of the learner based on their answers, update the probability distribution of the 2^K possible latent states after each question in a Bayesian way. One can also compute which question is the most informative, i.e., reduces the uncertainty (entropy) over the latent state of the learner the most. For details on potential algorithms for next item selection using the DINA model, see [2].

When the number of KCs is high, say 30, the number of possible states is big (2^{30}), and so is the support of the probability distribution. Therefore, the next question to ask cannot be computed efficiently.

Attribute Hierarchy Model and Knowledge Space Theory

Prerequisite graphs have been suggested in order to reduce the complexity of CD-CAT: $G = (V,E)$ where the nodes V are the KCs and $(u,v) \in E$ whenever u should be mastered before v. Therefore, the possible latent states should verify $\forall (u,v) \in E, c_v = 1 \Rightarrow c_u = 1$. Thus the support of the probability distribution is reduced and the next question to ask can be computed efficiently. Such an approach has been referred to as the *attribute hierarchy model* [7].

Knowledge space theory follows the same principle, but the underlying models are not robust to careless errors [4]. They have been tested to provide an adaptive test at the beginning of a MOOC of mathematics [8]. See [12] for a review of models used for adaptive assessment.

If the prerequisite graph is large though, or composed of many connected components, the number of states is still intractable, therefore other approaches should be used.

OUR NEW HEURISTIC METHOD

We now present our method, based on a q-matrix together with a prerequisite graph. An extra requirement is that every KC should have an intrinsic value of difficulty level, which is a positive real number. Questions assess one main KC, and in what follows, we will call *difficulty* of question j the difficulty level of the main KC validated by j.

Curriculum Data

We had 48 knowledge components being assessed by 48 questions. Each question assesses the mastery of a single KC. Each of our KCs has a difficulty level comprised between 1 and 6 that will help choose the next question. The prerequisite graph is composed of 10 connected components that are simple paths, such as $web4 \rightarrow web5$ or $url1 \rightarrow url2 \rightarrow url4 \rightarrow url5 \rightarrow url6$, see Figure 1. Thus, validating a KC propagates the validation to its parents, while invalidating a KC propagates the non-validation to its children.

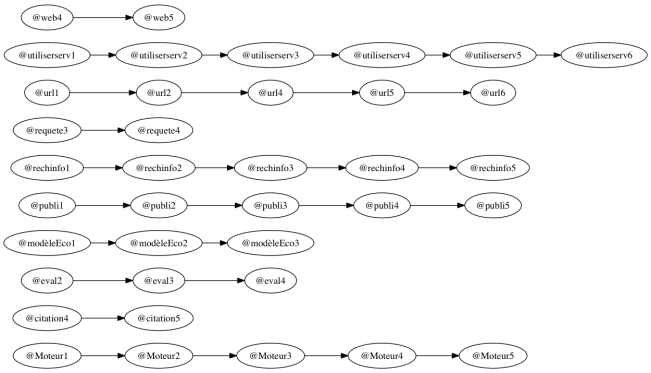

Figure 1. The prerequisite graph for our adaptive test in Pix.

In our case, even with our prerequisite graph, the number of possible latent states is too big (15M), because the numerous connected components lead to a combinatorial explosion. Seemingly, edges are missing, but we did not want to force such extra connections before we could examine learner data. Based on this usage data, some extra prerequisites could be added, such as $url2 \rightarrow web4$, that could reduce the number of possible latent states.

Algorithm

Instead of maintaining a probability distribution over potentially millions of latent states, in order to build the CAT tree, we maintain two sets called `acquired` and `not_acquired` that collect the KCs seemingly mastered or non-mastered by a learner throughout the test. Those sets do not necessarily reflect the final diagnosis provided to the learner but will allow to choose efficiently the next question to ask. At the end of the test, one can compute a diagnosis of the learner based on the collected information, for example using the general diagnostic model [3], or models based on slip and guess parameters.

When a learner solves correctly a question that requires a certain KC, they also validate the parents of this KC in the prerequisite graph, all of them being added to the `acquired` set. When they provide an incorrect answer to a question validating a KC, they also invalidate the children of this KC in the graph, being added to the `not_acquired` set. For example, if the prerequisite graph contains the path $url1 \rightarrow url2 \rightarrow url4$, then solving correctly a question that requires $url2$ will add both $url1$ and $url2$ to the `acquired` set, while solving incorrectly such a question will add both $url2$ and $url4$ to the `not_acquired` set. It is thus possible to compute for each KC the number of KCs acquired $N_{\texttt{acquired}}$ or not acquired $N_{\texttt{not\_acquired}}$, using a simple depth-first search. Those sets do not reflect the true knowledge of the learner, but allow choosing early questions of various difficulty levels.

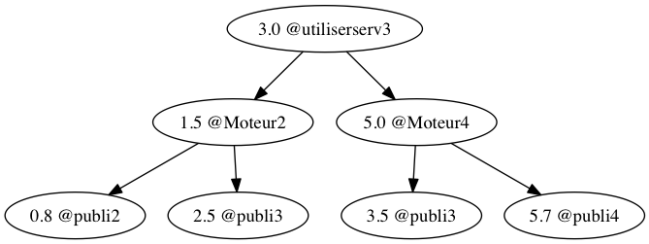

Figure 2. An example of CAT tree provided by our method.

The learner is modelled by a single parameter θ representing their proficiency. We denote R_j the outcome of the learner over the question j, either right (1) or wrong (0). This outcome verifies:

$$p_j(\theta) \triangleq \Pr(R_j = 1) = \Phi(\theta - d_j)$$

where d_j is the difficulty of question j and $\Phi : x \mapsto 1/(1+e^{-x})$ is the logistic function. This is the 1-parameter logistic model of item response theory.

At each step, we ask the question which achieves the highest expected number of KCs added either to the `acquired` set or the `not_acquired` set. This way, we can compute a value of the information collected at each question, based on the current estimated level of the learner:

$$score(j) = \Pr(R_j = 1) \cdot N_{\texttt{acquired}} + \Pr(R_j = 0) \cdot N_{\texttt{not\_acquired}}.$$

After some questions have been asked, parameter estimation of the learner's level θ is performed by maximizing the likelihood:

$$L(\theta) = \prod p_j(\theta)^{a_j} (1 - p_j(\theta))^{1-a_j}.$$

In other words, we just need to find the zeroes of:

$$\frac{\partial \log L}{\partial \theta} = \sum_{j=1}^{N} a_j (1 - p_j(\theta)) - (1 - a_j) p_j(\theta) \qquad (1)$$

which is usually performed using the Brent algorithm. When the samples are all-right or all-wrong, we add either a positive outcome for a question of difficulty 0 or a negative outcome for a question of maximal difficulty (in our case, 7), in order for the maximum-likelihood estimator to exist. Such an update of the parameter θ allows asking more difficult questions if the learner is performing well.

Computing the zeroes of Equation 1 is fast, therefore a 20-question-depth CAT tree (containing one million nodes) can be computed in 1 minute, using a Python implementation of our method. The CAT tree obtained this way is shown in Figure 2. An edge towards the right indicates a correct answer while an edge towards the left indicates a wrong answer. Within each node are described the estimated level of the learner at this node, together with the main KC of the question that will be asked to them next. One can see on this figure that the questions are increasingly difficult whenever the learner performs well, and vice versa. Also, consecutive questions deal with different components of the graph, because it is more informative to test various subjects, according to the objective of maximizing the KCs added or removed at each step.

The test is available online at `http://pix.jiji.cat`.

CONTEXT: PIX, CERTIFICATION OF DIGITAL COMPETENCIES FOR FRENCH CITIZENS

Goal and curriculum

This method will be applied to Pix[1], an online platform for assessment and certification of digital skills for every French citizen. It is managed by the French Ministry of Education, in close relationship with public and private stakeholders. It aims at revealing and stimulating the training needs necessary to face the digital transformation of our societies, by measuring, promoting and developing digital competencies. It is built upon DigComp 2.0, the European Digital Competence Framework for Citizens [10] composed of five areas:

1. information and data literacy;
2. communication and collaboration;
3. digital content creation;
4. safety;
5. problem solving (in a digital environment).

The main goal of Pix is to provide a free assessment to any French citizen (scholar, student, professional, retired, etc.) that can assess their digital skills, and put a name on what they do not know (e.g., "most wiki-like websites have a publicly available history"). At the end of the test, they can receive a diagnosis, summarizing their strong and weak points, and possibly do the test again at will. Therefore, Pix provides a formative assessment, and people can learn more by sitting for the test again.

Pix allows citizens to monitor their progress using an account. After each test administration, they will be acknowledged with points on a 1024 pix scale, together with a competency profile. Progress will be encouraged with targeted recommendations of learning resources. Within a test, a level of proficiency between 1 and 8 is computed, together with the acquisition or non-acquisition of knowledge components called *acquix*, which are learning outcomes.

Impact

Within the next months, 4000 people will try the adaptive test. The next year, every student from grade 8 to 12 will try the platform: 3.5M students, potentially half of all French higher education students (1.25M), together with employment integration organizations. The source code of the platform is freely available on GitHub[2], under the license AGPLv3.

[1] `https://pix.beta.gouv.fr`
[2] `https://github.com/sgmap/pix`

Problem statements

Problem statements are built in a way similar to evidence-centered design [9], because in order to solve them, people have to bring to the system the proof that they managed to perform the requested task. For example, if the short-answer question is: "In the city Montrésor, what street leads to *Rue des Perrières*?" The answer is *Rue de la Couteauderie*, and the most straightforward way for anyone to find it — except if they unfortunately know Montrésor by heart — is to use a mapping service. No matter whether they used Google Maps or OpenStreetMap, if the answer is correct, the learner will prove they master the corresponding knowledge component *@utiliserserv3*, which means they can "find and use a service to get an answer, without a hint." The problems are thus challenge-based, and fun to solve. To date, there are 697 items in the database, designed by a team including teachers and researchers from French universities and educational institutions.

Research

Data collected by the platform will be sanitized (e.g., removing personal information in user input) and made publicly available for research purposes.

Problem statements will be continually improved according to usage data. For example, some of the questions expect a short answer, therefore new correct solutions may be added to the system, using approaches such as the Divide and Correct framework developed in [1].

CONCLUSION

In this work-in-progress paper, we showed how it was possible to provide a CD-CAT when the number of latent states is potentially very large, and few prerequisites over the KCs are known. Our method consists of a combination of the Rasch model from item response theory and existing techniques based on knowledge components. We applied this technique to Pix, the French platform of certification of digital skills.

After a first administration of this adaptive test to thousands of students, we will be able to calibrate other models that need existing data in order to be trained, such as adaptive testing models based on the general diagnostic model [11]. Such models could help suggest new links between KCs, or could express the fact that a single question could require two KC with different weights. We leave this for further work.

ACKNOWLEDGEMENTS

This project was funded by the French Ministry of Education. Pix has been designed since March 2016 and developed as a State Startup since June 2016 within the incubator of the SGMAP[3] (General Secretariat for the Modernization of Public Action), similar to task forces such as 18F[4] or the United States Digital Service[5].

[3] https://beta.gouv.fr
[4] https://18f.gsa.gov
[5] https://www.usds.gov

REFERENCES

1. Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and correct: using clusters to grade short answers at scale. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 89–98.
2. Ying Cheng. 2009. When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 4 (2009), 619–632.
3. Matthias Davier. 2005. A general diagnostic model applied to language testing data. *ETS Research Report Series* 2005, 2 (2005), i–35.
4. Jean-Claude Falmagne, Eric Cosyn, Jean-Paul Doignon, and Nicolas Thiéry. 2006. The assessment of knowledge, in theory and in practice. In *Formal concept analysis*. Springer, 61–79.
5. Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of item response theory*. Sage.
6. Alan Huebner. 2010. An Overview of Recent Developments in Cognitive Diagnostic Computer Adaptive Assessments. *Practical Assessment, Research & Evaluation* 15, 3 (2010), 7.
7. Jacqueline P. Leighton, Mark J. Gierl, and Stephen M. Hunka. 2004. The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement* 41, 3 (2004), 205–237.
8. Danny Lynch and Colm P. Howlin. 2014. Real world usage of an adaptive testing algorithm to uncover latent knowledge. (2014).
9. Robert J. Mislevy, John T. Behrens, Kristen E. Dicerbo, and Roy Levy. 2012. Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *JEDM-Journal of Educational Data Mining* 4, 1 (2012), 11–48.
10. Vuorikari Riina, Yves Punie, Stephanie Carretero Gomez, and Godelieve Van Den Brande. 2016. *DigComp 2.0: The Digital Competence Framework for Citizens. Update Phase 1: the Conceptual Reference Model*. Technical Report. Institute for Prospective Technological Studies, Joint Research Centre.
11. Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard. 2016a. Adaptive Testing Using a General Diagnostic Model. In *European Conference on Technology Enhanced Learning*. Springer, 331–339.
12. Jill-Jênn Vie, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard. 2016b. A review of recent advances in adaptive assessment. In *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art*. Springer, in press.