JOURNAL OF LEARNING ANALYTICS

(2016). Adaptive geography practice data set. *Journal of Learning Analytics*, *3*(2), 317–321. http://dx.doi.org/10.18608/jla.2016.32.17

Adaptive Geography Practice Data Set

Jan Papoušek
Masaryk University Brno, Czech Republic

Radek Pelánek
Masaryk University Brno, Czech Republic
xpelanek@fi.muni.cz

Vít Stanislav
Masaryk University Brno, Czech Republic

ABSTRACT: We present a data set on student learning of geography facts in an open online system — slepemapy.cz. The data set has a simple format with intuitive interpretation. At the same time, it offers rich possibilities for modelling and analysis; for example, of prior knowledge, forgetting, or response times.

Keywords: Adaptive practice, geography, learning, student modelling

1 INTRODUCTION

The presented data set comes from an online adaptive system — slepemapy.cz — providing adaptive practice of geography facts (e.g., names and locations of countries, cities, mountains). The system estimates student knowledge and, based on this estimate, it adaptively asks questions of suitable difficulty (Papoušek, Pelánek, & Stanislav, 2014). The system uses a target success rate (typically 75%, which varies in time and for different users) and adaptively constructs questions in such a way that achieved performance is close to this target (Papoušek & Pelánek, 2015). The system uses open questions and multiple-choice questions with 2 to 6 options and 2 different directions ("Where is France?" or "What is the name of the highlighted country?"). Students answer questions using an interactive "outline map." Students can also access a visualization of their knowledge using an open learner model.

The presented data set has several advantages. The data set is based on an open education system — an open source project freely available online — with available description of algorithms used (Papoušek, Pelánek, & Stanislav, 2014). Researchers can thus try the system themselves before using the data set and inspect the details of its realization. This is in contrast with many current education data sets whose origin is not completely clear or easily inspectable (e.g., data sets based on Carnegie Learning systems, which are commercial). In such cases it is hard to understand the data exactly.

The presented data set is also easy to interpret, with simple structures and records with clear intuitive meanings. The data set deals with geographical items that can be easily visualized. This offers

(2016). Adaptive geography practice data set. *Journal of Learning Analytics*, *3*(2), 317–321. http://dx.doi.org/10.18608/jla.2016.32.17

possibilities for quick inspection and analysis of data. The data set contains all the important aspects of asked questions; it does not contain any assumptions or pre-processing steps by authors (e.g., use of predefined knowledge components, as is often done in case of currently available data sets).

The content of the data set — learning of facts in a realistic setting — supplements those currently available. Most fine-grained data sets of learning processes, as available for example in DataShop (Koedinger et al., 2010), focus on the learning of procedural skills (e.g., math). The learning of factual knowledge has been studied thoroughly before, but mostly in laboratory experiments with small groups and artificial facts. The presented data set comes from a realistic, large-scale application used by students in schools or in preparation for exams.

Although the structure of the data set is simple, the recorded student learning behaviour is complex and captures many interesting aspects of learning: widely varied prior knowledge, forgetting and short term memory effects, and the relation between response times and correctness of answers. The potential of the data set is illustrated by previous research (in many cases the research offers just preliminary results, showing a potential for more complex modelling and deeper analysis):

- Nižnan, Pelánek, & Řihák (2015) — modelling prior knowledge (considering only first answers on each item)
- Pelánek (2015) — modelling students' memory (short term memory effects, forgetting) based on repeated answers utilizing time between attempts
- Papoušek, Pelánek, Řihák, & Stanislav (2015) — analysis of response times

2 THE DATA SET

Creator: Adaptive Learning Research Group at Masaryk University Brno.

Access details: The data set is available at http://www.fi.muni.cz/adaptivelearning/data/slepemapy/

The adaptive geography practice data set is made available under the Open Database Licence: http://opendatacommons.org/licenses/odbl/1.0/

Provenance, date, version, maintenance: The data set is based on the online system http://slepemapy.cz. The system is available in Czech, English, and Spanish, with most users from the Czech Republic (78%) and Slovakia (10%). The system uses adaptive algorithms — described in detail in Papoušek, Pelánek, & Stanislav (2014) and Papoušek & Pelánek (2015) — for choosing questions. This first publicly available version of the data set is static and captures student interactions up to 21 May 2015. The basic statistics of the data set are as follows: 91,331 students, 1,459 geographical items, and 10,087,306 answers.

(2016). Adaptive geography practice data set. *Journal of Learning Analytics*, *3*(2), 317–321. http://dx.doi.org/10.18608/jla.2016.32.17

Ethical and privacy considerations: The educational system is used mainly by students in schools or by students preparing for exams. Nevertheless, it is an open online system, which can be used by anybody, and details about individual users are not available. Users are identified only by their anonymous ID. Users can log into the system using their Google or Facebook account. This login is used only for identifying the user within the system and is not included in the data set. Unlogged users are tracked using web browser cookies. The system also logs the IP address from which users access the system; the IP address is included in the data set in anonymized form. We separately encode the country of origin, which can be useful for analysis, and its inclusion is not a privacy concern. The rest of the IP address is replaced by a meaningless identifier to preserve privacy.

3 FORMAT

The data set is available in the standard CSV format. The core of the data set is the answer.csv file with detailed information about student answers (see Table 1). Some of the presented attributes have been implemented recently, so they may contain undefined value for some answers. A supplementary file describes the used geographical items (file place.csv with columns: ID, code, name of place, type of place, list of maps on which the place occurs, file place_type.csv with description of types of places).

Table 1: Columns of CSV file with answers

Column	Description
id	Answer identifier
user	User's identifier
place_asked	Identifier of the asked place
place_answered	Identifier of the answered place; empty if the user answered "I do not know"
type	Type of question: 1) find the given place on the map; 2) pick the name for the highlighted place
options	List of identifiers of options (including the asked place)
inserted	Time (yyyy-mm-dd hh:mm:ss) when the answer was inserted in the system
response_time	Time spent by the user in answering (measured in milliseconds)
place_map	Identifier of the map used in the question (e.g., a question about France may be asked in the context of the map of Europe or the map of world)

(2016). Adaptive geography practice data set. *Journal of Learning Analytics*, *3*(2), 317–321. http://dx.doi.org/10.18608/jla.2016.32.17

Column	Description
ip_country	Country retrieved from the user's IP address
ip_id	Meaningless identifier of the user's IP address
language	Language version of the system: 0) Czech, 1) English, 2) Spanish

4 LIMITATIONS

The data set has very limited information about students. The system can be used by anybody and no demographic data on users is available (beyond anonymized IP address and the associated information about user location). Moreover, logging into the system is voluntary and tracking of unlogged users is done only using web browser cookies, i.e., it may happen that the same person has multiple identifiers in the data set (this is, however, a feature of many similar systems).

The data set is not based on a randomized experiment, but on an adaptive system that uses a student model to choose questions (Papoušek, Pelánek, & Stanislav, 2014). On one hand, this is a strong point as the data correspond to real-life usage of an educational system. On the other hand, this aspect complicates the interpretation of data, e.g., care must be taken with "success rate," since the system actively tries to achieve a predetermined success rate (by choice of suitable questions).

REFERENCES

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, (p. 43). Boca Raton, FL: CRC Press.

Nižnan, J., Pelánek, R., & Řihák, J. (2015). Student models for prior knowledge estimation. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM 2015), (pp. 109–116). Retrieved from http://files.eric.ed.gov/fulltext/ED560514.pdf

Papoušek, J., & Pelánek, R. (2015). Impact of adaptive educational system behaviour on student motivation. In C. Conati, N. Heffernan, A. Mitrovic, M.F. Verdejo (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (AIED '15), (pp. 348–357). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-19773-9_35

Papoušek, J., Pelánek, R., Řihák, J. & Stanislav, V. (2015). An analysis of response times in adaptive practice of geography facts. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining*

(2016). Adaptive geography practice data set. *Journal of Learning Analytics*, *3*(2), 317–321. http://dx.doi.org/10.18608/jla.2016.32.17

(EDM 2015), (pp. 562–563).

Papoušek, J., Pelánek, R. & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In J. Stamper, S. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM 2014), (pp. 6–13). Retrieved from http://educationaldatamining.org/EDM2014/uploads/procs2014/long papers/6_EDM-2014-Full.pdf

Pelánek, R. (2015). Modeling students' memory for application in adaptive educational systems. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM 2015), (pp. 480–483). Retrieved from http://eric.ed.gov/?id=ED560907