

Using Structural Domain and Learner Models to Link Multiple Data Sources for Learning Analytics

Michael D. Kickmeier-Rust, Dietrich Albert

Graz University of Technology, Graz, Austria

michael.kickmeier-rust@tugraz.at, dietrich.albert@tugraz.at

Abstract. The uptake of Learning Analytics is often limited to online courses and other digital environments such as MOOCs and it is sparsely used in schools. The reason is that school-based teaching and learning is still an ‘analogue’ and personal process that is not producing the digital data that are necessary to conduct in-depth learning analytics. The Lea’s Box project is addressing this problem by supporting teachers in their daily practice to collect data as easy and complete as possible to have at least the ‘little data’ required to make their teaching more individual and more formative. In addition, the project attempts to develop competence-oriented techniques for learning analytics on the basis of solid theories that have been developed in the context of intelligent tutorial systems. In this paper we present a summary about the developments and experiences with the tools and techniques in schools.

Keywords: Learning Analytics, Competence-based Knowledge Space Theory, Formal Concept Analysis

1 Introduction

Learning analytics (LA) and educational data mining (EDM) are more than recent buzz words in educational research: they signify one of the most promising developments in improving teaching and learning. While many attempts to enhance learning with mere technology failed in the past, making sense of a large amount of data collected over a long period of time and conveying it to teachers in a suitable form is indeed the area where computers and technology can add value for future classrooms. However, reasoning about data, and in particular learning-related data, is not trivial and requires a robust foundation of well-elaborated psycho-pedagogical theories. The fundamental idea of learning analytics is not new, of course. In essence, the aim is using as much information about learners as possible to understand the meaning of the data in terms of the learners’ strengths and weaknesses, abilities, competences and declarative knowledge, attitudes and social networks, as well as learning progress, with the final goal of providing the best and most appropriate personalized support. Thus, the concept of learning analytics is quite similar to the idea of formative assessment. “Good” teachers of all time have strived to achieve exactly this goal. However, collecting, aggregating, storing and interpreting information about learners that

originates from various sources and over a longer period of time (e.g., a semester, a school year, or even in a lifelong learning sense) requires smart technology. To analyze this vast amount of data, give it educational meaning, visualize the results, represent the learner in a holistic and fair manner, and provide appropriate feedback, teachers need to be equipped with the appropriate technology. With that regard, a substantial body of research work and tools already exist. The European Lea's Box project aims to continue and enrich on-going developments and facilitates the broad use of learning analytics in the "real educational world".

Lea's Box stands for "Learning Analytics Tool Box" and concentrates on a competence-centered, multi-source formative analytics methodology based on sound psycho-pedagogical models, such as the *Competence-based Knowledge Space Theory* (CbKST) and the *Formal Concept Analysis* (FCA) which are used to put structural domain and learner models in the center of LA. The tangible result of Lea's Box manifest in form of a Web platform for teachers and learners provide links to the existing components and interfaces to a broad range of educational data sources. Teachers will be able to link the various tools and methods that they are already using in their daily practice and that provide software APIs (e.g., Moodle courses, electronic tests, Google Docs, etc.) in one central location. More importantly, the platform hosts the newly developed LA/EDM services, empowering educators to conduct competence-based analysis of rich data sets. A key focus of the platform will enable teachers not only to combine existing bits of data but to allow them to "generate" and collect data in very simple forms, not requiring sophisticated hard- or software solutions. Finally, we want to open new ways to display the results of learning analytics - leaving the rather statistical dashboard approach, moving towards structural visualizations and towards opening the internal learner models.

2 Structural Domain and Learner Modelling

The original Knowledge Space Theory (KST), founded by Doignon and Falmagne [1, 2] and extensions such as the CbKST, are coming from the genre of autonomous intelligent and adaptive tutoring systems. The idea was to broaden the ideas of the linear Item Response Theory (IRT) scaling, where a number of items are arranged on a single, linear dimension of "difficulty". In essence, KST provided a basis for structuring a domain of knowledge and for representing the knowledge based on prerequisite relations. More recent advancements of the theory accounted for a probabilistic view of test results and they introduced a separation of observable performance and the actually underlying abilities and knowledge of a person. Such developments lead to a variety of theoretical, competence-based approaches (cf. [3] for an overview). An empirically well-validated approach to CbKST was introduced by [4]; basically, the idea was to assume a finite set of more or less atomic competencies (in the sense of some well-defined, small scale descriptions of some sort of aptitude, ability, knowledge, or skill) and a prerequisite relation between those competences.

In a first step, CbKST attempts to develop a model of the learning domain, e.g. algebra. Examples for such competencies might be the knowledge what an integer is or

the ability to add two positive integers and so on. The level of granularity to which a domain is broken down depends on the envisaged application and might range from a very course-grained level on the basis of lessons (for example to plan a school term) to a very fine-grained level of atomic entities of knowledge/ability (for example as the basis of an intelligent problem solving support application). In a second step, CbKST looks into a natural course of learning and development and into logical prerequisites between competencies. Usually, learning and the development of new abilities as well as the stabilization of skills occurs along developmental trajectories. On the basis of a set of competencies and a set of prerequisite relationships between them, we can formally derive a collection of so-called competence states (Figure 1). Due to such prerequisite relations between the competencies, not all subsets of competencies (which would result in the power set) are plausible competence states.

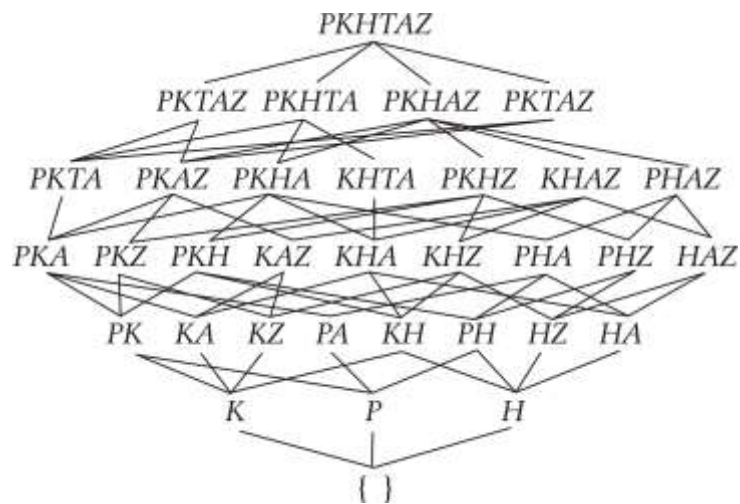


Fig. 1: A prototypical competence space.

So far, the structural model focuses on latent, unobservable competencies; loosely speaking the model makes hypotheses about the brain's black box. By utilizing interpretation and representation functions the latent competencies are mapped to evidence or indicators relevant for a given domain. Such indicators might be test items but might refer to all sorts of performance or behavior (e.g., the concrete steps when working with a spread sheet application). Due to these functions, latent competencies and observable performance can be linked in a broad form. This means that an entire series of indicators can be linked to underlying competencies. CbKST accounts for the fact that indicators such as test items cannot be perfect evidence for the latent knowledge or ability. There is always the possibility that a person makes a lucky guess or exhibits a correct behavior/activity just by chance. In turn, a person might fail in a test item although the necessary knowledge is actually available (e.g., being inattentive or careless). Thus, CbKST considers indicators on a probability-based

level, this means that mastering a test item suggest having the underlying competencies with a certain probability. Conceptually, this constitutes a probability distribution over the competence structure. A further significant advantage of such approach is that learning is not only considered a one dimensional course on a linear trajectory, equal for all learners. Learning rather occur along one of an entire range of possible learning paths.

3 Visualizing Structural Models

As claimed, Hasse diagrams are capable of holding a number of important information for an educator to evaluate the learning progress and also to make recommendations. In this paper we want to highlight such advantages.

3.1 Competence States and Levels

As outlined, a competency space is the collection of meaningful states a learner can be in. Depending on the domain, the amount of possible states might be huge. The big advantage, however, is that depending on the degree of structure in the domain, by far not all possible combinations of competencies are reasonable and thus part of the space. When zooming into the diagram, a teacher can exactly identify the set of competencies that is most likely for the learner, by zooming out color-coding can illustrate the most likely locations of a learner within the space. When looking at the entire space, it is obvious at first site at which completion level a learner is approximately (rather at the beginning or almost finished). These zoom levels are shown in Figure 2. Technically, there is a variety of options to achieve the coding, for example, bolding, greying, or color coding, whereas likely states are displayed more distinctly than such with low probability.

Equal to individual states, Hasse diagrams can represent group distributions. Defined by a certain confidence interval of probabilities those states and areas can be made more salient that hold the highest percentage of learners of a group. By this means, specific areas in the competency space become apparent within which the most learners are and, in contrast also positive or negative outliers pop out the diagram. A different method was suggested by [4], who altered the size of the nodes to represent the groups' sizes; the larger a node the more learners hold a particular state.

3.2 Learning Paths

In addition to having insight into groups' and individuals' current states of learning, the learning history, the so-called learning paths, are of interested for educators; on the one hand for planning future activities, on the other hand, for negotiation and documenting the achievements of a learning episode (e.g., a semester). Learning paths can be simply displayed by highlighting the edges between the most likely state(s) over time. As for the states, various probable paths can be realized by making more

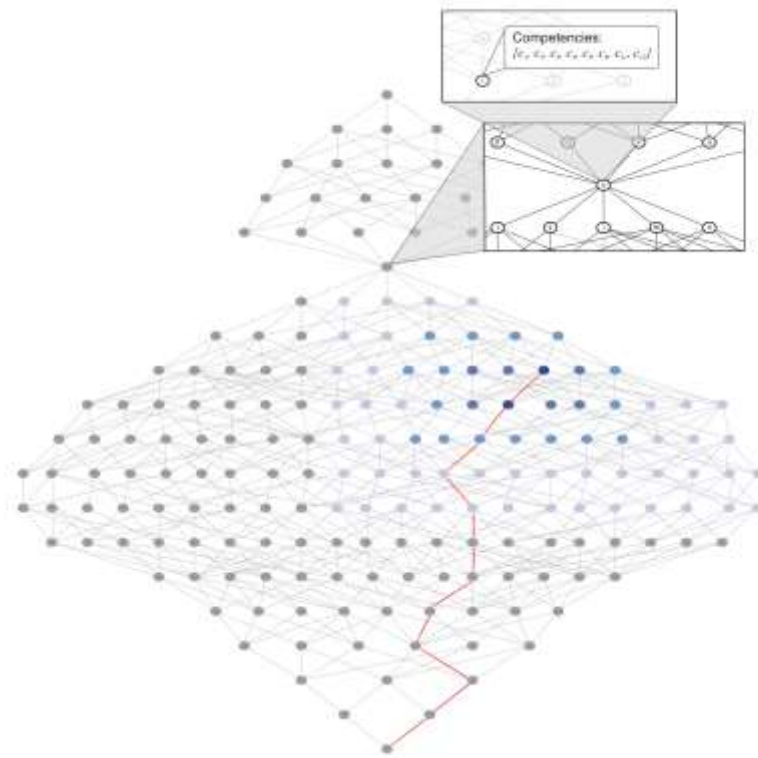


Fig. 2: Hasse diagram illustrating the probability distribution over a competence space on three zoom levels.

likely paths more intensive (by color coding or line thickness). Figure 2 shows a simple example (red line). A key strength of presenting learning paths, as indicated, is opening up the learner model to the learners (perhaps parents) themselves [4] – to explain where they started at the beginning of a course and how they proceeded during the course and which competencies they hold today. This perhaps can be complemented with comparisons to others or groups. Not least, learning paths can unveil information about the effectiveness and impact of certain learning activities, materials, or the teacher herself.

4 Multi-Modal, Multi-Source Data

The features of CbKST-ish modelling bear clear advantages for LA. However, the key question is, where do they data for computing the probabilities of competence states come from. In typical digital learning scenarios such as e-Learning lessons,

MOOCs, or Webinars the data come rather naturally. Students are permanently acting in one closed digital system and consequently rich and clean is the basis of data. In most other educational settings (e.g., in typical K18 classroom scenarios) students do not act in or in front of electronic devices. Most activities occur in an analogues way and digital activities are done with a vast range of different devices, apps, and software tools – from Google Docs to Facebook, from a multimedia app to a Moodle quiz. This means that data is recorded sparsely, the sources are manifold, the data sets are extremely heterogeneous in nature and so is the explanatory power of data.

With multi-modal data, that is, data that is collected by a multitude of sensory devices such as gaze trackers, smart watches, wearables, etc., the problem gets worse and more complicated. The delicate task is to interpret heart rate information in terms of engagement or learning performance. Building one solid picture of a learner on this basis demands an underlying anchor – in form of a robust learner and domain model. In this paper we argued that the CbKST modelling offers a range of options to bring data of multiple sources and even of multiple devices together. Heart rates or gaze paths can serve as probabilistic indicators for strong or weak learning processes, for engagement, or motivation.

The approach of separating latent competencies (which more or less develop and exist in the black box named ‘human brain’) from the performance (that what we can observe), bears particular advantages. On the one hand, performance, e.g. test scores, classroom participation, homework, etc., is not only determined by competencies or aptitude; there is a variety of aspects contributing to a certain performance, e.g., motivation, daily constitution, tiredness, external distractors, nutrition, health status, etc. On the other hand, CbKST-ish structures are rather stable, once set up and validated properly. The advantage lays in the fact that performance such as test results, behaviors, achievements, etc. is considered as probability-based indicators for certain competencies. Mathematically this relationship is established in form of interpretation and representation functions [5], which links an arbitrary set of performances/behaviors to one or more competencies, either in an increasing or in a decreasing sense. This, in the end, allows linking all available and perhaps changing data sources to one and the same competence space. It’s not about a single test, it’s about all available information we can gather, even it is considered being of little importance, all sorts of information may contribute to strengthen the model, the view of the learner. In case the amount or quality of data is weak, CbKST allows conservative interpretations, based on the arising probability distributions, in case there is a richer data basis, the probability distributions are more reliable, valid, and robust. For the educator, and this is important, the uncertainty is mirrored in the degree of likelihood. On a weak data basis, the probabilities of competence states differ substantially less than on the basis of richer data. Such information, however, can change the educator’s view and evaluation of a student’s achievements. In the end, this approach supports a fairer and more substantiated approach to grading or providing formatively inspired feedback.

The described approach is realized in the Lea’s Box project in form of an easy to use web platform. This platform is designed for teachers and provides a number of internal tools for recording and analyzing data. The platform provides also APIs (e.g., xAPI) to link various external data sources to the central domain and learner models.

Teachers and students can retrieve information about the learning processes, the competency probabilities, and learning trajectories in manifold ways and with a variety of visualizations. These range from simple cartoonish visualizations to complex directed graphs (cf. Figure 1). Our experiences show that the approach provides practical LA solutions for practitioners and interesting new insights into learning processes. Moreover, the structural models offer promising beacons to bring the data of diverse data sources together in a meaningful way.

5 Acknowledgements

This work described in this paper is based on LEA's BOX project, supported by the European Commission contracted under number 619762, of the 7th Framework Programme. This document does not represent the opinion of the EC and the EC is not responsible for any use that might be made of its content.

References

1. Doignon, J.-P., & Falmagne, J-C. Knowledge spaces. Berlin: Springer (1999).
2. Falmagne, J-C., & Doignon, J.-P. Learning Spaces. Berlin: Springer (2011).
3. Albert, D., & Lukas, J. Knowledge spaces: Theories, empirical research, and applications. Mahwah, NJ: Lawrence Erlbaum Associates (1999).
4. Nakamura, Y., Tsuji, H., Seta, K., Hashimoto, K., and Albert, D. Visualization of Learner's State and Learning Paths with Knowledge Structures. In A. König et al. (Eds.), KES 2011, Part IV. Lecture Notes in Artificial Intelligence 6884, pp. 261-270. Berlin: Springer (2011).
5. Korossy, K. Modelling knowledge as competence and performance. In D. Albert & J. Lukas (Eds.), Knowledge Spaces: Theories, empirical research, and applications (pp. 103–132). Mahwah, NJ: Lawrence Erlbaum Associates (1999).