Estimation of prerequisite skills model from large scale assessment data using semantic data mining

Bruno Elias Penteado
ICMC - Institute of Mathematics and Computer Science
University of Sao Paulo
Av. Trabalhador São-Carlense, 400, São Carlos, Brazil
brunopenteado@usp.br

ABSTRACT

Learning sequences are important aspects in learning environments. Students should learn by moving gradually from simpler to more complex concepts, promoting deeper levels of learning. This feature is usually embedded in most intelligent learning environments to guide the student in the study of subject matter. The organization of this knowledge structure is usually an intensive effort of human experts, in creating a logical ordering of what is to be taught - determining the concepts and the prerequisite relations among them. In recent years, some methods have been developed for dealing with this knowledge structuring using data coming from logs of learning environments, applying data mining techniques to discover prerequisite rules and create directed graphs of prerequisites. These methods model both assessment items and skills underlying those items. The automatic methods developed so far present a semantic gap between the probabilistic analysis and the expert knowledge, sometimes causing confusion with the results. This research aims to bridge this gap by adding a minimal layer of semantic information to help in the data mining process. As an application, we intend to analyze large-scale assessment datasets, considering its specificities, and evaluate if those hybrid models can improve the prediction of item success.

Keywords

Skill model, knowledge structure, data mining, semantic data mining.

1. INTRODUCTION

Skills prerequisite structure is an important component in *domain modeling*, used in intelligent learning environments and which serve as a basis for planning learning sequences and adaptive strategies for tutoring systems. Analogously, most intelligent learning environments uses a *student model* for the automatic adaptation of teaching strategies and as a overlay of domain model, influencing how the automatic intervention is carried out. Human experts usually define such prerequisite structure; however, they are rarely validated empirically and improved for better results.

For most of the large scale assessments, the current approach considers all knowledge in a single unidimensional scale, which considers the item difficulty in its ordination. *Computer adaptive tests* tend to use predominantly this ordination for item selection in diagnostic assessments. This approach raises some issues: the *interpretability* of results, since a single value is used to represent a knowledge in a large domain; and the *agreement* about the structure, since most experts cannot see a direct, unidimensional relationship among skills. Given the amplitude of skills, experts seem to agree on other sorts of dependencies, not just the simple ordination for item difficulty. For instance, in the field of Physics, an easy item of spatial movement might not be considered as a prerequisite for a difficult item in geometric optics, since they belong to different branches.

On the other hand, the process of manual creation of these dependencies is highly costly, time-consuming and presents large disagreement among experts modeling the same domain. Pavlik et al. [1] point to 3 other factors: the description of irrelevant skills, redundancy among skills and the ordination of those skills

There seems to be a semantic gap between the automatic extraction from data and the mapping made by human experts. This research aims to explore this gap, trying to bridge it using semantic data mining, and combining the advantages of both approaches.

2. PREVIOUS WORK

The process of prerequisite structure derivation from observable variables (such as assessment items) from data has been investigated by many researchers; yet, the skill modeling is still an open issue, since a student's knowledge is a latent variable, not being observed directly. In [2] it is proposed the POKS (Partial Order Knowledge Structure) algorithm to learn the dependency structure among items, composed only by the observable nodes (answers to the items), outperforming Bayesian networks algorithm, both in predictive performance and computational efficiency. In [1] POKS algorithm is applied to analyze the relations among skills, using observable items and use the result to cluster redundant skills, with a high degree of covariance, simplifying the domain model and determining it structure. In [3] a method is proposed to determine dependency relations among curricular units from student's performance data, using a binomial test for every pair of skills, to evaluate the existence of a prerequisite relationship between them. In [4] a frequent association rules mining method is proposed to discover concept maps, but not considering the uncertainty in the process of knowledge transfer of the student to his performance. In [5] the structure is derived from noisy observations using log likelihood calculated between the precondition model and the model in which the skills are all independent on each pair of skills to estimate which model better fits the student's data. In [6] causal discovery algorithms are used to find a skill prerequisite structure applying statistical tests in the latent variables. In [7] is proposed a probabilistic association rules mining method, having the probabilistic knowledge states estimated by an evidence model, to find a structure from performance data.

In semantic technologies, ontologies are explicit specifications of conceptualization and a formal way to define the semantics of knowledge and data. Dou et al. [8] surveys this semantic data mining in multiple domains - formal ontologies have been introduced to semantic data mining to: i) bridge the semantic gap between data, data mining algorithms and results; ii) provide data mining algorithms with a priori knowledge, guiding the mining process or reducing the search space; iii) provide a formal way for representing the data mining flow, from data preprocessing to mining results. Bellandi et al. [9] presented an ontology-based association rule mining method, using the ontology to filter instances in the process, constraining the search space of itemsets, excluding items and characterizing others according to an abstraction level, enabling generalization of an item to a concept of the ontology. Marinica and Guillet [10] presented a post-processing method for the results of the association mining, pruning invalid or inconsistent association rules with the help of the ontology.

Large scale assessments present some specificities: they are very strict in their skill model, with reference matrices specifying what is expected in the test; they are periodic, meaning that they are applied, in some cases, in an annual basis, with no single item in common between applications; the test items are organized in blocks (incomplete balanced blocks) and the test is comprised of a few blocks with a fixed number of items, so that many versions of the test are available at a time; the items are all pre-tested before the actual application, to estimate psychometric parameters (following Item Response Theory principles) being equalized into the same scale. A challenge for this research is to work with datasets from multiple years (i.e., no common items), balanced in blocks trying to discover generalizations in the underlying skill model.

3. METHOD AND MATERIAL

In this work, we will work with microdata from ENEM – an annual Brazilian exam for high school students, used as a classification ranking for admission in many public federal universities in Brazil. This exam is composed by 4 knowledge areas (Mathematics, Natural Sciences, Human Sciences and Languages), each composed by 30 skills in the reference matrix specified for this exam. Each item is mapped to a single skill and a score is given for each of these knowledge areas. The test is composed by 45 multiple-choice items for each knowledge area, along with an essay, in a 2-day time span. Different tests are organized in an *incomplete balanced blocks* design. In this approach, each test is composed by multiple blocks of items, with fixed ordination and in increasing order of difficulty. The blocks are arranged in different tests so to alleviate possible biases like the position of an item and a fatigue factor for items in the end of the test.

The datasets contain every alternative selected by every student whom participated in the exam. We plan to conduct this study using the Mathematics dataset, from 2009 to 2014, in a sum of 270 items answered my tens of millions of students.

Working along with Math experts, we will try to create simple ontologies, just with constraints of what should or not be considered in the final model, to prune some of the spurious results.

This research will adopt a quantitative approach and use data mining techniques as a method to construct the mapping of the prerequisite structure which, from the items mapped to their respective skills and the performance data (correct and incorrect answers) for every respondent, is able to extract relations among the skills, generalized by different observations in different items.

The evaluation of the method will be based on the capacity of prediction of success on the items individually, assessing the goodness of fit against the human experts mapping. The method will be compared to state-of-the-art algorithms such as POKS, probabilistic association rules mining and with some expert mapping.

4. PRELIMINARY WORK

This is a research project in its earlier stages, narrowing the research questions to be pursued. As an initial effort, I found that more simplistic approaches tend to model just the difficulty of items in the creation of a prerequisite structure, i.e., an easier item is a prerequisite for a more difficult item, disregarding contextual information on the respective topics.

Early examples for ENEM using data from Mathematics test applied in 2014 are depicted in Figure 1 (skill prerequisites). They were generated by the author using the POKS algorithm, with source code available in [11] and show the algorithm results.

In Figure 1, the previous items were mapped to their respective skills and the algorithm was run. Skills are numbered according to the official codes available at ENEM website. We can see that some skills are more fundamental, specially numbers 1, 3, 4 and 17. Skills 12, 15 and 22 were not assessed in this test.

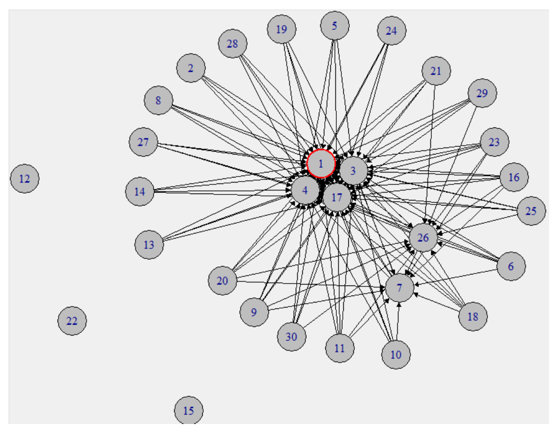

Figure 1. Prerequisite skills derived from Math assessment.

We hope, by the end of this research, discover possible prerequisite relations among skills that constitute the ENEM exam, complementing the traditional model of ordination by item difficulty in the IRT model, by creating a generalized graph of dependencies among skills, estimated from empirical data of application and combined with ontology constraints.

From this mapping, it should be possible to build an intelligent learning environment that might diagnose in which point of the graph the student is and the possible sequences he can choose to study. Another practical implication may be the interpretation of results and extension to practices in public policies. As this sort of exam is applied in different moments in K-12, the model could generalize and describe how learning happens in public education system, since literacy through high school.

5. ADVICES SOUGHT

For this doctoral consortium, advice is sought regarding some concerns:

a) *What data mining methods should be used to model these prerequisite skills?* At first, POKS was used but other methods could also be evaluated, like LFA, Rule Space and BKT. As this is a high stake exam, the skills are wider, different from other more granular skill models from ITS domains. An example (skill 17, a basic skill from Figure 1): "analyze information involving variations in quantity as a resource for argument construction". In addition, the same skill can vary a lot depending on the items being assessed. Second, items being that different and having different difficulty parameter,

b) *Should difficulty be embedded in the model?* so that different items of a same skill can influence differently in the model.

c) *Should these information be included in the model?* which may result in different graphs for different populations. Besides the standard item accuracy prediction. This dataset has no other interaction data, as in ITS systems, but has contextual data about the respondents, with high impact features in performance, like geographic region and socioeconomic status.

d) *Is it valid to measure a interrater agreement metric (like Kappa) to compare the generated model with those from experts?* as a means of comparing how close the model fit the expert modeling.

6. REFERENCES

[1] Pavlik Jr., P.I., Cen, H., Wu, L., Koedinger, K.R.: Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor. In Proceedings of the 1st International Conference on Educational Data Mining, Montreal, Canada, 77-86, 2008.

[2] Desmarais, M.C., Meshkinfam, P., Gagnon, M.: Learned Student Models with Item to Item Knowledge Structures. User Modeling and User-adapted Interaction, 16(5), 403-434, 2006.

[3] Vuong, A., Nixon, T., Towle, B.: A Method for Finding

Vuong, A., Nixon, T., Towle, B.: A Method for Finding

Prerequisites within a Curriculum. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands, 211-216, 2011.

[4] Tseng, S.S., Sue, P.C., Su, J.M., Weng, J.F., Tsai, W.N.: A New Approach for Constructing the Concept Map. Computers & Education, 49(3), 691-707, 2007.

[5] Brunskill, E.: Estimating Prerequisite Structure from Noisy Data. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands, 217-222, 2011.

[6] Scheines, R., Silver, E., Goldin, I.: Discovering Prerequisite Relationships among Knowledge Components. In Prerequisites within a Curriculum. In Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, Netherlands, 211-216, 2011.

[7] Chen, Y., Wuillemin, P. H., Labat, J. M. Discovering prerequisite structure of skills through probabilistic association rules mining. In Proceedings of the 8th International Conference on Educational Data Mining, Montreal, Canada, 77-86, 2015.

[8] Dou, D., Wang, H., Liu, H. Semantic Data Mining: a survey of ontology-based approaches. In Proceedings of the 9th International Conference on Semantic Computing, Anaheim, USA, 244-251, 2015. DOI: 10.1109/ICOSC.2015.7050814.

[9] Bellandi, A., Furletti, B., Grossi, V., Romei, A. Ontology-driven association rule extraction: A case study. Contexts and Ontologies Representation and Reasoning, page 10, 2007.

[10] Marinica, C., Guillet, F. Knowledge-based interactive postmining of association rules using ontologies. Knowledge and Data Engineering, IEEE Transactions on, 22(6):784–797, 2010

[11] Desmarais, M., Bhatnager, S. Prerequisite skill structures in ASSISTments. Available in: https://github.com/sameerbhatnagar/POKS-skills. Accessed: March, 20th, 2016.