Modeling Exercise Relationships in E-Learning: A Unified Approach

Haw-Shiuan Chang, Hwai-Jung Hsu, Kuan-Ta Chen
Institute of Information Science, Academia Sinica, Taipei, Taiwan
{ken77921, hjhsu, swc}@iis.sinica.edu.tw

ABSTRACT

In an e-learning system, relationships between a large amount of exercises are complex and multi-dimensional; measuring the relationships and arranging curriculums accordingly used to be time consuming and costly tasks which require either enormous log collection or large-scale human annotations. Moreover, accurately quantifying the relationships is difficult because there are too many factors which affect our measurement based on the data, such as the ability of exercise takers and the subject bias of annotators. To overcome these challenges, we propose a unified model that extracts information from both human annotations and usage log using regression analysis. The proposed model is applied to quantify the *similarity*, *difficulty*, and *prerequisite* relationships between every two exercises in a curriculum. As a case study, we collaborate with Junyi Academy, a popular e-learning platform similar to Khan Academy, and infer the pairwise relationships of 370 exercises in its mathematics curriculum. We show that the model can predict exercise relationships as well as an expert does with human annotations of a few sample exercise pairs (2% in our experiments). We expect the introduction of the proposed unified model can improve the relationships among exercises and learning pathways of students in other e-learning platforms.

Keywords

Exercise relationships, Prerequisite, Curriculum, Human annotations, Regression Analysis, Khan Academy

1. INTRODUCTION

Estimating relationships between items has a wide range of applications in educational data mining (EDM). For example, curriculum arrangement [2, 5] and adaptive testing [6, 9] are often based on the estimations of difficulty and prerequisite relationships between courses, knowledge components, or exercises. Furthermore, estimating the similarity and prerequisite relationships between exercises can improve the quality of knowledge components [12, 13] and student modeling [3, 1, 4]. In this paper, we focus on studying the relationships of exercises (i.e., complete question units), which can facilitate personalized education in the future.

Meanwhile, in large and dynamic e-learning websites, manually organizing the growing number of exercises becomes more and more difficult. For instance, Junyi Academy[1], an e-learning platform in Taiwan similar to Khan Academy[2]. Junyi Academy provides over 300 interactive exercises for its mathematics curriculum, which is visualized by the knowledge tree as shown in Figure 1. We can see that there have been many complex prerequisite links in the knowledge tree, so it is very time consuming to manually validate how appropriate the prerequisite links are and whether there are better ways to arrange the links of the exercises. Moreover, the instructors need to consider hundreds of exercise candidates when determining the prerequisites for a new exercise.

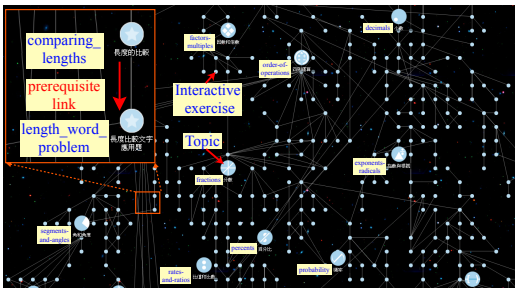

Figure 1: Part of the knowledge map on Junyi Academy. To visualize the prerequisite structure, the knowledge tree is laid out in a 2D plane called knowledge map.

Based on exercise taking log, researchers discover the relationships through item response theory (IRT) [10], inferring Bayesian model of students [3, 12, 1, 4], factor analysis [8], association rule learning [5], assuming a known Q-matrix [13], or assuming students would perform better after they have taken prerequisite or similar exercises [12, 11, 15], etc. Most of the aforementioned data-driven methods develop a specific learning algorithm for estimating a specific relationship between exercises. The learning algorithms usually require a large amount of log data so as to simultaneously infer all latent factors affecting our observation in data, such as relationships of exercises and capability of every student over time. However, data in some e-learning platforms might not be sufficient to accurately profile various behaviors of every student. As a result, the estimation of relationships between exercises might be misleading in a new system with only a small amount of usage log [16, 10].

On the other hand, the collected data are often noisy [1] and have different statistical characteristics in different systems, which might violate the assumptions made by a data-driven model. For example, many e-learning websites, such as Khan Academy and Junyi Academy, allow learners to browse any exercise without actually answering them. In fact, around 70% of the first answers are correct for the first problem of each mathematical exercise on Junyi Academy, which shows that learners tend to skip exercises they cannot answer. The freedom of selecting exercises would degrade the performances of purely data-driven approaches on more difficult exercises with less responses [16], and also cause challenges to identify the difficulty and prerequisite

[1]Junyi Academy (`http://www.junyiacademy.org/`) is established in 2012 on the basis of the open-source code released by Khan Academy.
[2]`https://www.khanacademy.org/`

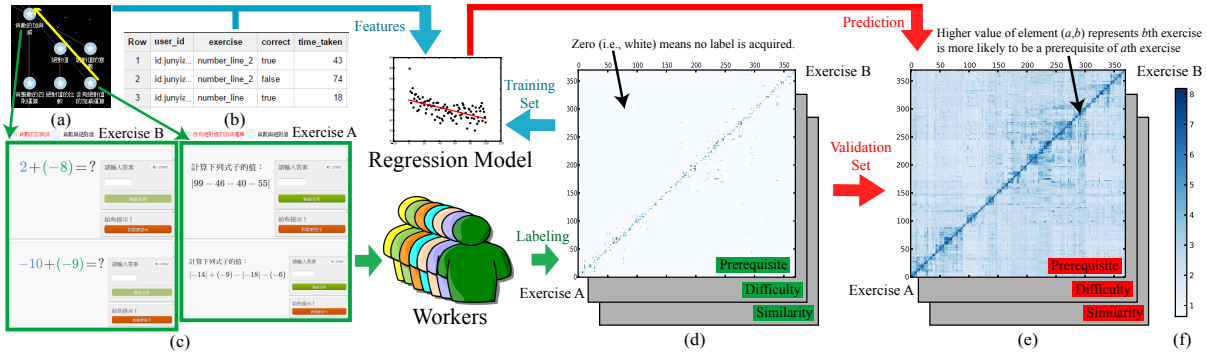

Figure 2: The proposed work flow. (a) A screen shot of the local knowledge map, (b) examples of usage log, (c) a example of exercise pairs, (d) the sparse similarity matrices labeled by workers, (e) the dense similarity matrices predicted by a regression model, and (f) the color code of (d) and (e).

relationships between exercises (See details in Sec. 2.3).

To solve the challenges, we advocate a hybrid method which integrates the power of crowdsourcing and machine learning as [14] did for finding prerequisite relationships among documents. As illustrated in Figure 2, we first quantify the *similarity*, *difficulty*, and *prerequisite* relationships of mathematical exercise pairs using crowd wisdom. Then, we characterize each exercise pair by various types of features extracted from the user practice log and website contents. Given labels and features, a regression model can be trained to predict relationships of every exercise pair. Finally, collected labels can be used to quantitatively evaluate both the prediction of machines and humans. Our experiments show that predictions generated by the proposed models are closer to the crowd consensus (i.e., average opinions of workers) than most of individuals' ratings.

2. RELATIONSHIP DISCOVERY

2.1 Label Collection

As previously discussed, the exercise relationships are hard to define objectively from usage log. Recently, Wauters et al. [16] pointed out that as more annotators judge difficulty of each exercise, their average score converges to a more steady value, which is highly correlated with the difficulty inferred by IRT model. Therefore, if we collect more subjective labels with high quality, their average responses are more representative (i.e., more likely to be agreed by most learners and instructors) and less sensitive to subject bias.

To collect high-quality labels from wide range of people, we divide the task of comparing exercise relationships into several questionnaires and apply several quality control methods. The method includes mathematical ability qualification, malicious workers detection by checking the elapsed time and the variances of their responses in each questionnaire, and outlier filtering using crowd consensus as [7] did.

At each section of questionnaires, we consecutively compare an exercise A with 1–7 other exercises which might be related to A. Note that potentially related exercises are paired according to student modeling and knowledge tree in Figure 1, and the order of comparisons is randomly determined. An example of comparison could be seen in Figure 2(c).

Any target relationship of exercise pairs could be quantified by a specific question. In this work, we ask the workers to choose the 1–9 score for the following questions, which query about *similarity*, *difficulty*, and *prerequisite* relationships of each exercise pair (A and B), respectively.

- How similar is the knowledge required for answering these two exercises?
- How much more difficult is exercise B compared to exercise A, where a higher score means B is more difficult than A and a score of 5 indicates that they have the same difficulty?
- After students learned to correctly answer exercise B, how appropriate is utilizing exercise A to deepen the students' knowledge on the topic step by step?

2.2 Feature Extraction

To automatically predict the relationships, we extract the usage log from Oct. 2012 to July 2014 on Junyi Academy, which contains over 10 million answering records from over 100 thousand users. When describing relationships between exercise A and exercise B, we extract the potentially helpful features from usage log and cluster them into 6 categories:
(i) ***Student Modeling (4 features)*** is extracted based on the practice history of each student. To be more specific, the student is modeled by applying random forest regressor to predict his/her accuracy on every exercise which has not been done by the student. Then, we compute original and normalized feature importance of log data in B for predicting students' accuracy in answering A, and the corresponding importance of A for the prediction of B.
(ii) ***Answering Time Duration (6 features)*** includes the difference between the average answering time duration of A and that of B (i.e., $(time\ for\ A)-(time\ for\ B)$), the logarithm difference of their average answering time duration (i.e., $\log(time\ for\ A)-\log(time\ for\ B)$), the difference and the logarithm difference of their answering time duration on the average of users' correct answers, and on the average of users' first correct answers of the exercises.
(iii) ***#Problems Taken in Exercises (4 features)*** (# means *the number of*) includes the difference and the logarithm difference between #total problems taken in A and B, the difference and the logarithm difference of #problems which are answered correctly in A and B.
(iv) ***Answering Accuracy (6 features)*** includes the difference and the logarithm difference between accuracy of A and that of B on the average of users' first, last, and all answers in the exercises, where the accuracy is defined by $\frac{\#correct\ answers}{\#total\ answers}$. Note that we only count the first answer of each learner in the same problem.
(v) ***#User Taking Exercises (3 features)*** includes the difference and the logarithm difference between #users taking A and that of B, and the Euclidean distance between #users vectors of A and that of B. The ith element in the #users vector of A records the #users who have done exercise i correctly before A.
(vi) ***User Answering Orders (6 features)*** include #users who practice A before B (denoted as $\#U[A \rightarrow B]$), #users who do B before A ($\#U[B \rightarrow A]$), $\frac{\#U[A \rightarrow B]}{\#U[A \rightarrow B]+\#U[B \rightarrow A]}$, #correct answers for A before answering B ($\#C[A \rightarrow B]$), the corresponding #answers for B before A ($\#C[B \rightarrow A]$), and $\frac{\#C[A \rightarrow B]}{\#C[A \rightarrow B]+\#C[B \rightarrow A]}$.

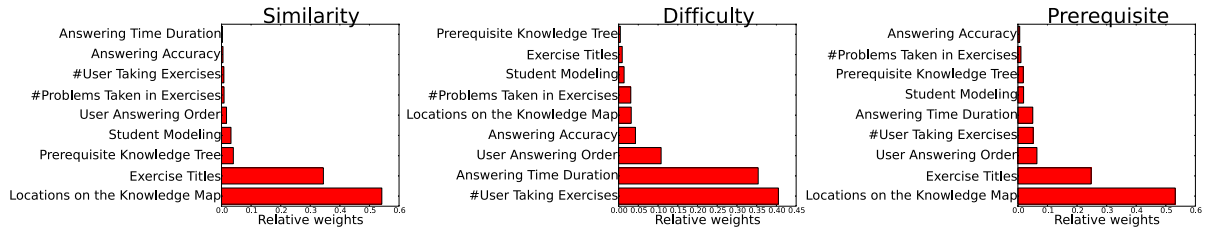

Figure 3: The feature importance for predicting relationships on Junyi Academy. The red bar of each category means the summation of all the feature importance in the category, and symbol # represents *the number of*.

As pointed out in [6, 5, 10, 8, 13], different types of tags on exercises or courses labeled by experts are useful information for determining their relationships. Therefore, we additionally extract exercise-related information from website contents on Junyi Academy, which can be grouped into following 3 categories:
(i) ***Prerequisite Knowledge Tree (5 features)*** includes whether B is a parent of A in the knowledge tree (i.e., the directed acyclic graph), whether B is a sibling of A, distance between A and B in the directed acyclic graph, and the corresponding distances after reversing and removing the direction of every edge in the graph.
(ii) ***Locations on the Knowledge Map (3 features)*** include Euclidean distance between A and B on the knowledge map, and coordinate difference between A and B on x-axis and y-axis in the knowledge map (e.g., the length and the coordinate vector of the yellow arrow in Figure 2(a)).
(iii) ***Exercise Titles (3 features)*** include edit distances of Chinese and English titles between A and B, and summation of the minimal edit distances among English words in their titles.

2.3 Relationship Prediction

Given the features and relationship labels, we formulate the relationship prediction task as a regression analysis. In Sec. 3, we use the collected labels to experiment on the effects of using different regression algorithms. To know the effectiveness of our 40 dimension features, we show the importance of feature categories which are determined by random forest regressor in Figure 3.

Compared with *Answering Accuracy*, *#User Taking Exercises* is a much better type of features for predicting the difficulty difference of exercises, because learners tend to skip exercises they cannot answer as we mentioned in Sec. 1. For the similarity and prerequisite relationships, the *Locations on the Knowledge Map* are the strongest type of features for the tasks, while the *Prerequisite Knowledge Tree* surprisingly has relatively low feature importance. An explanation for the observation is that instructors usually maintain similar exercises in close distance on the knowledge map, which are often good prerequisite candidates for each other. However, when they manually assign the prerequisite links in the knowledge tree, the graph needs to be kept sparse to ensure the clarity and simplicity of its layout.

Figure 3 also illustrates that the information contained in the *Exercise Titles* is much more correlated with the prerequisite relationships on Junyi Academy than features based on *Student Modeling* and *Answering Accuracy*, of which the analysis is extensively studied by many previous works such as [3, 12, 1]. Therefore, it would be interesting to investigate whether the observation is still valid in other platforms which probably have different rules of naming titles or of recommending exercises to learners.

3. EXPERIMENTS

Our proposed method is evaluated in the exercise system of Junyi Academy. To prevent scarce usage log skewing the statistical distribution of our features, we exclude the exercises which are answered by less than 100 users. The remaining 370 exercises of interest are randomly divided into two sets: the training set containing 240 exercises, and the testing set with 130 exercises. On average, each exercise of interest in training set is paired with 4.7 other exercises where around 10% of exercises are randomly selected, and each one in testing set is paired with 6.3 other exercises where the percentage of randomly selected exercises reaches around 30% to verify our generalization capability.

To evaluate how good humans and machines perform, one of metrics we adopt is relative squared error (RSE), which is defined as $\frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$, where $\hat{y}_i$ and y_i are our prediction and the ground truth for a relationship of exercise pair i, respectively, and $\bar{y}$ is the mean of y_i over all i. In addition, we transform every score of exercise relationships into its rank, and compare the similarity between the ranks from the predicted scores and the ranks from the ground truth scores. Then, we evaluate the predicted rank by Spearman's ρ and Kendall τ rank correlation coefficients.

3.1 Performance of Workers

After excluding malicious and unqualified workers, we hire 3 teachers, 8 online workers, and 43 people to work in the lab. All workers in the lab are at least graduated from senior high school, and most of them have a college degree. Each exercise pair in the training set are labeled 6.6 times on average by total 51 normal workers, and teachers are asked to score all the exercise pairs in the testing set. For the interest of the consistency between judgements from crowd consensus (i.e., the average scores from all workers) and that from experts, we also ask 2 among 3 teachers to label every pair in the training set. The total costs of collecting above labels are around 1,000 USD.

Manually quantifying the relationships between mathematical exercises is a demanding cognitive task, which requires a certain level of skills in abstract reasoning. Using the average of ratings from all workers (including teachers) as our ground truth, we first evaluate the performances of recruited workers and whether teachers (i.e., experts) perform better in the tasks. The results in the training set are presented in Table 1. Note that smaller RSE and larger rank coefficients indicate better performances. From Table 1, it is clear that the performance of workers (including experts) measured by RSE is significantly lower than the ones measured by rank coefficients compared with the performances of machines. The results illustrate that workers' annotations often contain systematic subject bias (i.e., workers tend to rate every query higher or lower than most of other people), so averaging scores rated by multiple workers is an effective way to improve the labeling quality for the task.

Table 1: Performance comparisons of different methods in the training set of Junyi Academy using cross validation, and best performances among regressors are highlighted in bold font.

Methods			Similarity			Difficulty			Prerequisite		
			RSE	Spearman's ρ	Kendall's τ	RSE	Spearman's ρ	Kendall's τ	RSE	Spearman's ρ	Kendall's τ
Humans	An Normal Worker	Range	0.193–1.124	0.188–0.854	0.208–0.750	0.492–3.235	0.063–0.820	0.050–0.747	0.316–2.381	0.000–0.813	-0.007–0.725
		Mean	0.574	0.598	0.524	1.096	0.516	0.439	0.986	0.458	0.387
	A Teacher	Range	0.493–0.543	0.648–0.718	0.560–0.625	0.619–0.741	0.625–0.634	0.539–0.540	0.858–1.054	0.571–0.684	0.504–0.594
		Mean	0.518	0.683	0.593	0.680	0.630	0.539	0.956	0.638	0.549
Regressors	Linear Regression		0.370	0.658	0.567	0.470	0.593	0.504	0.424	**0.624**	**0.541**
	nu-SVR		0.349	**0.683**	**0.594**	0.483	**0.611**	**0.526**	0.402	0.611	0.520
	Random Forest Regression		0.320	0.662	0.575	0.493	0.576	0.493	0.376	0.608	0.516
	GBR		**0.288**	0.680	0.590	**0.453**	0.610	0.521	**0.346**	0.600	0.514
Features (GBR)	w/o KT and KM		0.311	0.681	0.589	0.474	0.626	0.532	0.378	0.607	0.515
	w/o KT, KM, and ET		0.433	0.607	0.521	0.472	0.642	0.546	0.472	0.567	0.478
	w/ SM, AA, UN, and PT		0.548	0.516	0.438	0.502	0.610	0.516	0.595	0.463	0.377
	w/ SM and AA		0.598	0.524	0.446	0.632	0.486	0.400	0.666	0.417	0.346
	w/ KT		0.674	0.463	0.418	0.869	0.448	0.382	0.717	0.360	0.318

Table 2: Performance comparisons of different methods in the testing set of Junyi Academy. Note that the meaning of all abbreviations is the same as Table 1.

Methods			Similarity			Difficulty			Prerequisite		
			RSE	Spearman's ρ	Kendall's τ	RSE	Spearman's ρ	Kendall's τ	RSE	Spearman's ρ	Kendall's τ
Humans	A Teacher	Range	0.200–0.300	0.764–0.848	0.656–0.757	0.398–0.474	0.732–0.791	0.629–0.696	0.322–0.467	0.696–0.764	0.583–0.665
		Mean	0.235	0.814	0.719	0.427	0.762	0.661	0.406	0.721	0.617
Regressors	GBR		0.269	0.786	0.678	0.553	0.580	0.476	0.311	0.771	0.660

3.2 Prediction Accuracy

For the training set, we evaluate our prediction by 5-fold cross validation, and Table 1 compares the resulting outputs generated by different regression models and different subsets of features. The table summarizes the results of five regression algorithms including linear regression, nu support vector regression (nu-SVR), random forest regression, and gradient boosting regression (GBR). Compared with teachers' ratings in the training set, our approach can generate competitive performances measured by rank coefficients while having lower RSE, especially for more complex regressors such as the random forest or gradient boosting algorithms. This means that after being trained by collected labels, machines could predict exercise relationships closer to crowd consensus than most of the individuals. Note that to make the comparison fair, we round all of the scores predicted by machines into integers between 1–9.

In Table 1, we also provide control experiments on different types of features using gradient boosting regression. There might not be the knowledge tree (KT) and the knowledge map (KM) in other interactive learning environments, so we first present the performance without related categories of features. The results show that removing KT and KM can still produces competitive performances, but the performance would decrease by a margin if we further remove more features such as *Exercise Titles* (ET), *User Answering Orders*, *Answering Time Duration*, *User Numbers* (UN), and *#Problems Taken in Exercises* (PT), *Student Modeling* (SM), and *Answering Accuracy* (AA).

In order to verify our generalization ability across different types of annotators, we train the regression models on the training set (mostly labeled by normal workers) and evaluate their performance on testing set (labeled by teachers). As shown in Table 2, the performances of regression models are still very promising. Note that the exercise pairs in the testing set are only rated by 3 teachers whose labels have larger impact on ground truth, so the real performances of experts might be worse than this estimation.

4. CONCLUSIONS

The relationships of exercises are important for curriculum arrangement of e-learning platforms. In this work, we demonstrate that the relationships can be quantified by subjective labeling and predicted by regression models. The experiments on Junyi Academy show that the quality of predicted relationships are competitive against teachers' labels.

References

[1] E. Brunskill. Estimating prerequisite structure from noisy data. In *EDM*, 2011.

[2] E. Brunskill and S. J. Russell. RAPID: A reachable anytime planner for imprecisely-sensed domains. In *UAI*, 2010.

[3] C. Carmona, E. MillÃąn, J.-L. P. de-la Cruz, M. Trella, and R. Conejo. Introducing prerequisite relations in a multi-layered bayesian student model. In *User Modeling*, 2005.

[4] M. C. Desmarais. Performance comparison of item-to-item skills models with the IRT single latent trait model. In *UMAP*, 2011.

[5] T.-C. Hsieh and T.-I. Wang. A mining-based approach on discovering courses pattern for constructing suitable learning path. *Expert Syst. Appl.*, 2010.

[6] C. Koutsojannis, G. Beligiannis, I. Hatzilygeroudis, and C. Papavlasopoulos. Using a hybrid AI approach for exercise difficulty level adaptation. *IJCEELL*, 2007.

[7] B. Lakshminarayanan and Y. W. Teh. Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. *CoRR*, 2013.

[8] A. S. Lan, C. Studer, A. E. Waters, and R. G. Baraniuk. Tag-aware ordinal sparse factor analysis for learning and content analytics. In *EDM*, 2013.

[9] D. Lynch and C. P. Howlin. Real world usage of an adaptive testing algorithm to uncover latent knowledge. In *ICERI*, 2014.

[10] M. L. Nguyen, S. C. Hui, and A. C. M. Fong. Content-based collaborative filtering for question difficulty calibration. In *PRICAI*, 2012.

[11] Z. A. Pardos and N. T. Heffernan. Determining the significance of item order in randomized problem sets. In *EDM*, 2009.

[12] P. I. Pavlik, H. Cen, L. Wu, and K. R. Koedinger. Using item-type performance covariance to improve the skill model of an existing tutor. In *EDM*, 2008.

[13] R. Scheines, E. Silver, and I. Goldin. Discovering prerequisite relationships among knowledge components. In *EDM*, 2014.

[14] P. P. Talukdar and W. W. Cohen. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 2012.

[15] A. Vuong, T. Nixon, and B. Towle. A method for finding prerequisites within a curriculum. In *EDM*, 2011.

[16] K. Wauters, P. Desmet, and W. van den Noortgate. Acquiring item difficulty estimates: a collaborative effort of data and judgment. In *EDM*, 2011.