Machine beats experts: Automatic discovery of skill models for data-driven online course refinement

Noboru Matsuda[1] Tadanobu Furukawa[2] Norman Bier[3] Christos Faloutsos[2]
mazda@cs.cmu.edu tfuru@cs.cmu.edu nbier@cmu.edu christos@cs.cmu.edu
[1]Human-Computer Interaction Institute [2]Computer Science Department [3]Open Learning Initiative
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213, USA

ABSTRACT

How can we automatically determine which skills must be mastered for the successful completion of an online course? Large-scale online courses (e.g., MOOCs) often contain a broad range of contents frequently intended to be a semester's worth of materials; this breadth often makes it difficult to articulate an accurate set of skills and knowledge (i.e., a skill model, or the Q-Matrix). We have developed an innovative method to discover skill models from the data of online courses. Our method assumes that online courses have a pre-defined skill map for which skills are associated with formative assessment items embedded throughout the online course. Our method carefully exploits correlations between various parts of student performance, as well as in the text of assessment items, to build a superior statistical model that even outperforms human experts. To evaluate our method, we compare our method with existing methods (LFA) and human engineered skill models on three Open Learning Initiative (OLI) courses at Carnegie Mellon University. The results show that (1) our method outperforms human-engineered skill models, (2) skill models discovered by our method are interpretable, and (3) our method is remarkably faster than existing methods. These results suggest that our method provides a significant contribution to the evidence-based, iterative refinement of online courses with a promising scalability.

Keywords

Online course refinement, skill model discovery, evidence-base course engineering, MOOC, Q-matrix

1. INTRODUCTION

When designing and implementing large-scale online courses (aka MOOCs), defining a set of skills to be learned and having individual skills associated with particular part of course contents often becomes quite challenging. Making an effective course with explicit associations between a necessary set of skills and course contents requires intensive cognitive task analysis and time-consuming evidence-based iterative engineering [1]. Studies show how important it is to have data-analytics feedback for course improvement and theory development [2-5]. However, cognitive task analysis driven by human experts has an issue in its accuracy and scalability; applying it for a large-scale online course is often impractical.

Research shows the potential for advanced technologies to automatically and semi-automatically discover a set of skills for online courses. Learning Factor Analysis (LFA), for example, semi-automatically refines a given skill set [6]. However, LFA works only when meaningful "features" are given, which (usually) requires cognitive task analysis by subject domain experts. Other studies apply matrix factorization methods for automatic skill set (aka Q-matrix) discovery from students' response data [7, 8]. However, these methods often face the issue of interpretability—i.e., providing meaningful feedback to course designers and developers based on the machine-generated skill set is often troublesome.

We developed an efficient, practical, and scalable method that we call eEPIPHANY, to fully and automatically discover skill sets from online course data, which are the combination of the assessment item text data (i.e., problem and feedback text sentences for assessment items) and student learning interaction data. eEPIPHANY is a collection of data-mining techniques to automatically refine (or rebuild) a human-crafted set of skills, initially given by course designers and developers.

The most important goal of eEPIPHANY is to provide constructive feedback to online course designers and developers for iterative course improvement. We assume that our target online courses have occasional formative assessments to probe students' competency towards learning objectives. We hypothesize that students' response data and assessment item text data both reflect latent skills to be learned, and assessment items can be clustered based on those latent skills. To test these hypotheses, we implemented eEPIPANY as a combination of the matrix factorization to analyze students' response data and bag-of-words techniques to analyze course content data.

The contributions of this work are the following: (1) *A new problem formulation*—We show how to integrate diversified information such as student performance and assessment item text data. (2) *A new algorithm*—Our solution, the eEPIPHANY algorithm, is scalable and effective for practical use for large-scale online course engineering. (3) *Evaluation*—eEPIPHANY outperforms past competitors, including *human experts*, on several, real online course datasets.

The goal of this paper is to introduce the eEPIPHANY method (section 3) and provide empirical evaluation for its effectiveness (section 4). We discuss implications for the application of eEPIPHANY to evidence-based online course refinement (section 5.3). To begin, the next section provides a standard structure of our target online courses and various definitions for later discussions.

2. SKILL MODEL FOR ONLINE COURSES

We assume that our target online courses have occasional low-stake assessments throughout the course—aka formative assessments—to assess students' competency on target *skills*. We assume that each formative assessment has multiple *assessment items* (i.e., problems to answer), each of which is associated with one or more skills.

We assume that online courses have a pre-defined *skill map* (often called Q-matrix [9, 10]) that shows one-to-many mapping between individual skills and one or more assessment items. In this paper a mapping between a single skill and multiple assessment items in the skill map is called a *skill-item association*.

We call a set of skills a *skill model*. The terms "skill model" and "skill map" will be used interchangeably in this paper. The pre-defined skill model is therefore called the "*default*" *skill model*—a human-developed model that is initially guided by authors' intuition in the absence of data, or a human-developed model that has been refined based on student data.

The Open Learning Initiative (OLI) at Carnegie Mellon University [11] is an example of an online course platform that meets the above-mentioned criteria [12]. OLI courses all have a *human-crafted* "default" skill model that is often recognized as semi-optimal, and could always be improved.

To improve skill models to refine online courses, it becomes crucial that the machine-discovered skill models have high interpretability so that course designers and developers can make sense of the proposed skill model improvements. Our proposed method, eEPHIPHANY, discovers accurate and interpretable skill models from learning data and assessment item text data. The next section describes details of the eEPHIPHANY method.

3. eEPIPHANY

eEPIPHANY is a collection of data mining techniques for automatic discovery of skill models from online course data. The primary input to eEPIPHANY is a matrix representing a chronological record of students' responses to assessment items, called an A-matrix (Figure 6-a). The A-matrix is a three-dimensional matrix showing a history of attempts on individual assessment items made by individual students. Each attempt is a vector of binary values representing the correctness of a student's response—0 indicates incorrect and 1 indicates correct. The A-matrix contains at most one correct response per student per assessment item.

The goal of eEPIPHANY is to find a skill model (Q-matrix) that produces the best prediction of the A-matrix. The predictive power is measured by cross-validation. eEPIPHANY can either find a Q-matrix by itself or refine a given Q-matrix by the following steps: (1) clustering assessment items with latent features that would best characterize the similarity in the difficulties of assessment items (section 3.1), (2) proposing a new skill model by assuming that the above-mentioned cluster of assessment items provides a hint for new skills (section 3.2), and (3) searching for the best skill model by comparing multiple skill model candidates (section 3.3).

3.1 Feature Extraction

We have developed two latent-feature extraction strategies: (1) the Matrix Factorization (MF) strategy, and (2) the Bag-of-Words (BoW) strategy. The goal of feature extraction, regardless of the strategy difference, is to generate a two-dimensional matrix, the P-Matrix, showing a mapping between assessment items and "skill candidates" (Figure 6-d. Also see below).

3.1.1 Matrix factorization (MF) strategy

For the MF strategy, the A-matrix is first transformed into the difficulty matrix (D-matrix), which is a two-dimensional matrix representing an individual student's difficulty for each assessment item. We hypothesize that the record of individual students' performance on assessment items reflect their "difficulties" in answering assessment items, and that those students who show a similar distribution pattern of difficulties share a similar competency on latent skills.

The item difficulty *id*, by definition, is computed as $id = 1 - 1/d$ where d is the number of attempts made on an assessment item. We only include attempts until the first correct attempt is made, i.e., *id* is the length of the vector of attempts in the A-matrix (Figure 6-a). We hypothesize that students would more likely skip items that look too easy for them hence no difficulties at all. Therefore, we defined *id* as 0 for missing data in the A-Matrix (i.e., skipped items).

The D-matrix is then factorized into U and V matrices (i.e., $D = U \times V$) by the Non-Negative Matrix Factorization method [13]. The V-matrix is a two-dimensional (assessment item by latent feature) matrix. It is therefore a collection of *feature vectors*, each corresponding to an assessment item (Figure 6-b).

Assessment items in the V-matrix are then clustered by the k-means method [14], resulting in an F-matrix (Figure 6-c). We hypothesize that each cluster in the F-matrix represents a "skill candidate" that can be used to construct the P-Matrix (Figure 6-d).

The P-Matrix is a two-dimensional binary matrix showing which assessment item belongs to which skill candidate. The P-matrix represents the association of each assessment item to a skill candidate. By its nature, in the current eEPIPHANY algorithm, each assessment item has an association to at most one skill candidate (if any).

3.1.2 Bag-of-words (BoW) strategy

The BoW strategy creates the F-matrix directly from a collection of *item stems* (i.e., assessment item text data showing problem and feedback texts) for assessment items. That is, the assessment items are clustered by the bag-of-words method using item stems.

We first transform each assessment item into a set of component words from a collection of item stems using a part-of-speech tagger, TreeTagger[1]. We then apply the Latent Dirichlet Allocation model (LDA) [15] to cluster assessment items. Assessment items are clustered based on the probability of topic distribution—i.e., individual assessment items are assigned to the topic with the highest topic probability, which results into the F-Matrix from which the P-Matrix is generated as mentioned above.

[1] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

3.2 Skill Model Construction

eEPIPHANY refines a given "default" skill model by either modifying it or replacing it with a new skill model. In either case, eEPIPHANY first proposes candidates for new skills, and then finds the best way to refine the default skill model in terms of the accuracy of the data fit. This subsection describes the former step, whereas the latter step is described in section 3.3.

Given a P-matrix, there are three strategies to refine the "default" skill model: (1) Replacing the entire "default" skill model with an entirely new skill model, (2) appending new skill-item associations to the "default" skill model, (3) splitting given a skill-item association(s) in the "default" skill model into multiple skill-item associations.

3.2.1 Replace Strategy

To replace the default skill model with an entirely new skill model, the P-matrix is straightforwardly converted into the Q-matrix. Namely, each skill candidate becomes a new skill. Assessment items that are associated with the skill candidate become members of the skill-item association for the newly defined skill.

3.2.2 Append Strategy

The *append* strategy adds more skill-item associations to the default skill model, while the original skill-item associations in the default skill model remain intact. Skill-item associations that are being newly added are the same set of skill-item associations proposed by the *replace* strategy. The following example illustrates this process (Figure 1):

Assume that there is a skill-item association a_i for a skill s_i with assessment items $q^i_1 \ldots q^i_5$ in the default skill model. Also, assume that there is a skill candidate c_1 and c_2 in the P-matrix where c_1 has a skill-item association with assessment items q^i_1, q^i_2, and q^i_3; and c_2 has a skill-item association with assessment items q^i_4 and q^i_5. The *append* strategy enters two new skill-item associations, one for c_1 and another one for c_2 into the default skill model. As a consequence, the assessment item q^i_1, for example, is now associated with two skills, s_i and c_1.

It is worth noting that the skill model produced by the *replace* strategy is the proper subset of the skill model produced by the *append* strategy. The number of skills in the skill model produced by the *append* strategy is the sum of the number of skills in the default skill model and the number of skills in the skill model produced by the *replace* strategy.

3.2.3 Split Strategy

The *split* strategy refines the default skill model by individually splitting skill-item associations into multiple new skill-item associations. These splits are based on skill-item associations in

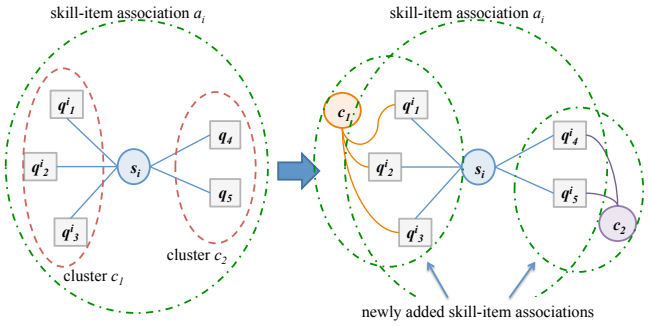

Figure 1. The *append* strategy appends new skill-item associations to the default skill model

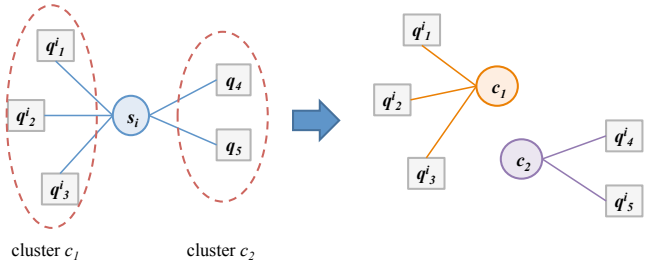

Figure 2. The *split* strategy breaks given skill-item associations into new ones with newly discovered skills

the P-Matrix. The following example illustrates this process (Figure 2):

Assume the same situation as mentioned above for the *append* strategy. That is, there is a skill-item association a_i for a skill s_i with assessment items $q^i_1 \ldots q^i_5$ in the default skill model. Also, assume that there is a skill candidate c_1 and c_2 in the P-matrix where c_1 has a skill-item association with q^i_1, q^i_2, and q^i_3; and c_2 has a skill-item association with q^i_4 and q^i_5. The *split* strategy then replaces the original skill-item association a_i with two new skill associations a_{i-1} and a_{i-2}, where a_{i-1} has c_1 as a skill and q^i_1, q^i_2, and q^i_3 as associated assessment-items, while a_{i-2} has c_2 as a skill and q^i_4 and q^i_5 as associated assessment-items.

3.3 Model Search

We hypothesize that two different types of feature-extraction strategies (section 3.1) present pros and cons for our purposes. For example, the item stem (i.e., problem sentences and feedback messages) might reflect skills necessary to answer the assessment item correctly. On the other hand, the student response data might reflect skills that students have actually acquired. The BoW strategy might provide better interpretability, but the student response data might provide more accurate skill models. The BoW strategy can be applied even before the course has been used (i.e., before student data is available).

With the lack of a predictive theory of parameter selection to compute the best skill model, eEPIPHANY exhaustively searches for the best skill model by comparing all possible skill models with different combinations of the following four parameters. The comparison is done by the model fit using the Bayesian Knowledge Tracing as a predictor:

(1) The number of components used for the Matrix Factorization (N_C)—This determines a dimension of the V-matrix. N_C reflects the variance in the pattern of student competency over the latent features. Although, the greater N_C value would result in the smaller reconstruction error (i.e., ||D-U*V||), it might also result in the over fit to the data (which is penalized in the AIC and BIC scores). *Nc* varies from 10 to the number of students, increased by 10 during the model search.

(2) The number of clusters in k-means (N_k)—We hypothesize that each feature is shared by at least five assessment items. Therefore, N_k varies from 25 to $N_Q/5$ where N_Q is the number of assessment items; increased by 25 during the model search.

(3) The number of topics used for LDA (section 3.1.2) to compute the bag-of-words clustering (N_T)—Here again, applying the same hypotheses as for N_k. N_T varies from 25 to $N_Q/5$, increased by 25 during the model search.

(4) The threshold used for the split strategy (β)—Assume that skill s is associated with n assessment items, $q_i,..,q_n$. Also assume

that in the P-matrix, these n assessment items are associated with k skill candidates, C = $<c_1,\ldots,c_k>$. The skill-item association for s will be split into new skill-item associations with skill candidate c in C, if the number of assessment items associate with the skill candidate c is greater than $n \times \beta$. β is set to 0.05, 0.25, and 0.5 in this order during the model search.

3.4 Model Interpretation: The DoE Analysis

The most important goal of the skill-model discovery and refinement proposed in the current paper is to improve online courses. Providing ***interpretable*** feedback based on a machine-discovered skill model and model refinement is therefore crucial. We hypothesize that to achieve this goal, two subgoals must be met: (1) to identify what part of the default skill model has been improved the most, and (2) to understand the improvement from a domain perspective.

To identify the part of the skill model that has been improved most, we analyze the *degree of enhancement* (DoE) of the proposed change in skill models. We hypothesize that the DoE would be maximized among a skill(s) for which the accuracy of students' performance prediction improved the most [16]. The accuracy of student performance prediction is operationalized as the root mean squared error (RMSE) in cross-validation for the model-fit evaluation.

Based on this hypothesis, we identify skills with the most DoE in the default skill model M_D relative to a refined (i.e., machine-discovered) skill model M_R as follows:

(1) For each skill s_i in the default skill model M_D, let I_D^i be a set of assessment items associated with s_i.

(2) Find all skills c_i^j (j=1,…,n_i) in the refined skill model M_R that are associated with any assessment items in I_D^i.

(3) Compute xI_D^i, the extended version of I_D^i, by adding all assessment items associated with any of c_i^j to I_D^i.

(4) Compute RMSEs_i that is an RMSE in predicting student performance on assessment items in xI_{Di} using corresponding s_i in M_D as the predictor.

(5) Compute RMSEc_i that is an RMSE in predicting student performance on assessment items in xI_{Di} using corresponding c_i^j in M_R as a predictor.

(6) Let d_i = RMSEs_i – RMSEc_i be the *DoE score* of skill s_i relative to c_i^j.

(7) Find a skill s in M_D with the largest DoE score. The skill s has the largest error reduction from M_D to M_R.

Once the skill with the largest error reduction is found, the next step is to understand what the improvement is about, that is, to interpret the machine-discovered model refinement with the focus on the skill with the largest error reduction.

To interpret the proposed model refinement, we use the bag-of-words analysis in combination with manual inspection of the assessment item text. For each skill-item association in the refined skill model, a set of keywords is extracted from the item stem (i.e., the combination of text sentences for the items and their feedback messages). The χ^2 value is computed for individual word w appearing in the item stem for a skill-item association k as follows [17]: $\chi^2(k, w) = (\text{aic}(k, w) - \text{aict}(k, w))^2 / \text{aict}(k, w)$ where $\text{aic}(k, w)$ is the number of assessment items that contains w in k, and $\text{aict}(k, w)$ is a theoretical implication for $\text{aic}(k, w)$, i.e., $\text{aict}(k, w) = \text{aic}(k, *) \times \text{aic}(*, w) / \text{aic}(*,*)$. The word w is considered as a keyword only when $\text{aict}(k, w) < \text{aic}(k, w)$.

Table 1. Three OLI datasets used for the evaluation

	Statistics	Biology	C@CM
#Students	1,013	481	100
#Transactions	538,062	418,344	94,612
#Unique Items	1,791	916	912

4. EVALUATION

To evaluate the efficiency and effectiveness of the eEPIPHANY method, we applied it to actual online course data.

4.1 Data

Three OLI courses—Computing@CarnegieMellon (C@CM), Biology, and Statistics—were used for evaluation. All three courses are actively used at Carnegie Mellon University and other educational institutions for registered, academic students and in open sections for independent learners. Table 1 shows the number of students, transactions (i.e., students' responses to assessment items), and unique items; these datasets represent use in academic contexts. All these OLI data are available on DataShop [18]. It turned out that the C@CM data only contains randomly selected students' data from a larger pool of the OLI data that contains more than 1300 academic students enrolled.

4.2 Method

For each of the three OLI datasets, we applied eEPIPHANY and had it search the best skill model by finding the optimal clustering parameters (section 3.3). During the search we recorded the model-fit for three feature-extraction strategies (matrix factorization, bag-of-words, and their combination as described in section 3.1) crossed over three skill-model construction strategies (*split, add*, and *replace* as in section 3.2). The model-fit was computing by cross-validation using the Bayesian Knowledge Tracing technique.

4.3 Results

4.3.1 Comparison of feature extraction and refinement strategies

Table 2 shows the best skill models, annotated with the strategies and parameters used to discover them. As the table shows, *the matrix factorization (MF) strategy always outperformed the BoW strategy for the three datasets used in the study*. When the MF strategy is used, *replacing the default skill model with a completely new skill model discovered by eEPIPHANY yielded the best skill model* for all dataset.

To understand how the size of cluster impacts the quality of the resultant skill model, we compared different skill models with different sizes measured as the number of skills. Figure 3 plots the

Table 2. ePIPHANY always found better skill model than experts. FS: Feature Extraction Strategy, SC: Skill Construction Strategy, #S: Number of items

FS	SC	#S	AIC	BIC	RMSE
Statistics					
MF	Replace	63	307730	310731	0.447
BoW	Append	143	317808	323802	0.456
Biology					
MF	Replace	86	224944	228514	0.389
BoW	Split	187	228597	236360	0.393
C@CM					
MF	Replace	41	59497	60998	0.364
BoW	Split	137	61648	66661	0.371

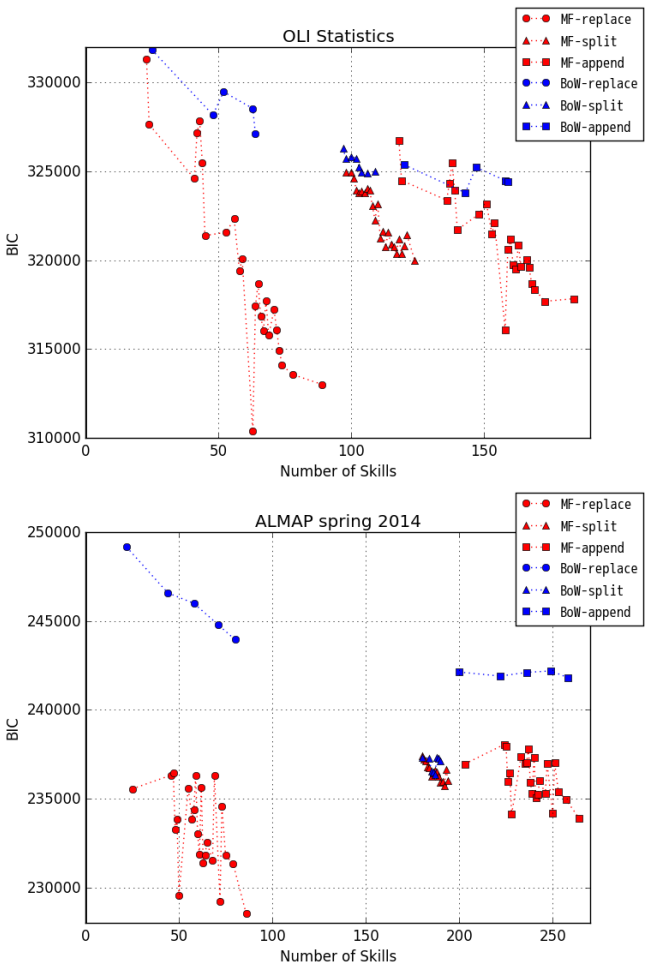

Figure 3. MF-replace wins or ties with MF-append: Comparison of skill models with different size. OLI Statistics (top) and Biology (bottom)

BIC (Y-axis) against a number of skills (X-axis). In the figure, two feature extraction strategies—MF and BoW—are crossed three skill-model construction strategies—replace, split, and append.

As the figure shows, it turned out that *for any strategy combination, the bigger the size of the model (i.e., the number of the clusters) the better the model*. It can be also seen that the *replace strategy is relatively better than other two skill-model construction strategies* (as depicted by more dots towards the bottom).

4.3.2 Comparison with other methods

Table 3 shows the comparison of the model-fit between skill models discovered by LFA, an OLI course designer (OLI), and eEPIPHANY (eEPI) on the OLI Statistics course. In DataShop, skill models discovered by LFA and human expert only contain data from Unit 1. Therefore, for this analysis, we applied eEPIPHANY only to the OLI data from Unit 1.

The table shows the number of skills (#S) and the number of assessment items (Obs.). The model fit was evaluated by AIC, BIC, and RMSE scores computed by using Additive Factor Model (AFM) [19] and Bayesian Knowledge Tracing (BKT). As shown in the table, eEPHIPHANY outperformed human expert (OLI),

Table 3. eEPIPHANY beats human expert on OLI Statistics. The analysis contains data only from Unit 1.

Method	#S	Obs.	AIC	BIC	RMSE
AFM					
eEPI	22	75955	72125	80901	0.412
LFA	28	75955	69108	77984	0.404
OLI	19	75955	74787	83507	0.418
BKT					
eEPI	22	75955	74560	75373	0.407
LFA	28	75955	74343	75378	0.404
OLI	19	75955	77405	78107	0.414

Table 4. Assessment items involved in the most beneficial skill model refinement

ID(Skill)	Assessment item (item stem)
Q881(c31)	The ability or tendency of organisms and cells to maintain stable internal conditions is called homeostasis (value:A) metabolism (value:B) evolution (value:C) emergent property (value:D)
Q885(c31)	Why do organisms maintain fairly steady conditions within their cells and bodies? They need to keep conditions stable so that they can obtain food. (value: A) Organisms just change along with whatever is happening in the outside world, which is usually quite steady. (value: B) They must maintain stable conditions to keep their enzymes working and generally to enable the chemical reactions of life. (value: C) Unstable conditions will destroy the DNA in cells; this is the most important risk for a cell facing physical or chemical stress. (value: D)
Q901(c31)	An organism or cell exhibits _____ when it maintains steady internal conditions despite changes in the outer environment. homeostasis (value: A) evolution (value: B) natural selection (value: C) balance (value: D)
Q717(c3)	Humans maintain a blood pH between 7.35 and 7.45. In order to maintain homeostasis, how will your body respond if your blood pH drops to 7.0? If your blood pH is 7.0, your body will raise your pH. (value: A) If your blood pH is 7.0, your body will lower your pH. (value: B) A blood pH of 7.0 is close enough to 7.35. Your body won't do anything. (value: C)

and arguably tied with LFA. We will further discuss this result in section 5.3.

4.3.3 Model interpretation

Figure 5 shows the skill *k153* with the largest DoE score (section 3.4) in the OLI Biology course. In the figure, the skill *k153* in the default skill model was associated with four assessment items. In the discovered skill model, these 4 assessment items are associated with two skills—*c31* and *c3*. The newly constructed skills *c31* and *c3* have 16 and 19 assessment items associated respectively. The RMSE is computed for those 35 steps using skills in the default skill model. The RMSE is then re-computed using *c31* and *c3*. According the DoE analysis, splitting skill *k153* into two skills *c3* and *c31* yields the biggest DoE score. This addressed the first subgoal of the model interpretation.

To interpret model improvement, we investigated four assessment items associated with *k153* in the default skill model to see why they were split into two groups. Table 4 shows four assessment

Table 5. Bag of words for a skill (k153) split into two new skills (c31 and c3)

Skill	Bag of Words
k153	homeostasis range internal maintain steady condition narrow tendency metabolism raise optimal entitiy exhibit sensitive balance chemistry drop world despite happening
c31	steady homeostasis evolutionary stress valid theme progress favor module tree ancestor selection adapt internal evolution ancestry natural conclusion environmental whale
c3	hazy fundamentally matter space play concept structo yet mass nutrient exchange determine sometimes dramatically biology rule ability quite period peanut

items and their skill association in the refined skill model. Table 5 shows the bag-of-words associated with each skill cluster.

In the default skill model, the skill *k153* is to "Define homeostasis and explain its role in maintaining life." All four assessment-items related to *k153* in the default skill model mention "homeostasis" and "sustainable life." However, a closer look shows that this skill is most appropriate for the three out of four assessment items—Q881, Q885, and Q901. In the refined skill model, these three assessment items are correctly tagged as one skill *c31*.

Although the fourth assessment item Q717 relates to homeostasis, a closer look shows that learners are being asked to engage in a more sophisticated task—i.e., determine (or predict) necessary action to achieve homeostasis, which results in a separate association with skill *c3*.

For those four rows, the machine-generated split is very coherent from a subject-matter expert's perspective. This satisfies the second subgoal of the model interpretation.

4.3.4 Efficiency

One of the notable strengths of the eEPIPHANY method is its efficiency. As described in section 3.3, eEPIPHANY searches the best skill model by a brute-force search by merely changing the number of clusters, which takes linear time O(*n*). This linear computation must be repeated nine times for three different feature-extraction strategies crossed with three different skill-model construction strategies, which still takes O(n).

The Learning Factor Analysis (LFA) method [6] requires an intensive search for each skill (*s*) over multiple difficulty factors (*d*) that takes O(s^d).

During the evaluation study that used three real OLI course data, eEPIPHANY found the best model in 2 to 3 hours per dataset running on a single-core personal computer, showing its practical potential for actual application to large-scale online course improvement.

5. DISCUSSION

5.1 Strategy comparison

Our study showed that using student response data (i.e., the number of attempts made on assessment items before a student finally made their first correct response) always yields a better skill model than using the bag-of-words with item stems. We also found that *even only using the bag-of-words, eEPIPHANY always yields a better skill model than the default skill model that is hand-crafted by human experts*.

As for the skill-model construction strategy, the *replace* strategy always discovers the best skill model in our study, suggesting that *the Matrix Factorization strategy efficiently discovers a latent skill model from the student learning data*. On the other hand, the *split* strategy always resulted in producing an inferior skill model in our study; suggesting that *the* split *strategy hardly improves on the human-crafted skill*.

The above observation also implies that *eEPIPHANY can actually find a better skill model completely automatically without human interaction (which is what the replace strategy does) from real online course data*.

5.2 Interpretability

To interpret skill models proposed by the Matrix Factorization (MF) strategy is to interpret clusters of assessment items, which is often quite challenging. For the purpose of course refinement however, interpretability becomes crucial.

To overcome this issue, while still taking the advantage of the MF strategy to produce high-quality skill models, we applied the degree of enhancement (DoE) analysis to identify the instance of refinement that received the most benefit—i.e., identifying the skill that received the largest benefit from skill decomposition. We also combined the bag-of-words technique with manual inspection. Our study demonstrated that *this hybrid technique allows course designers to make meaningful interpretations of the proposed refinements of the skill model*.

Yet the obvious limitation of the current technique is its dependence on manual inspection. We hypothesize that one idea to overcome this issue is to combine MF and BoW, namely, to expand the V-matrix (Figure 6-b) by adding the bag-of-words keyword information as a latent feature, and then applying k-mean clustering. The resulting clusters (i.e., the skill candidates) would have better interpretability supported by the bag-of-words keyword information. Testing this hypothesis is an important future study.

5.3 Implication for evidence-based online course refinement

Our study demonstrated that eEPIPHANY discovers skill models that reflect student learning more accurately than human-crafted skill models on all three OLI course data. Even though eEPIPHANY requires human labor to interpret the discovered skill models (with the aid of DoE), it is arguably still less time consuming than creating skill models by hand. Figure 4 depicts this argument as a two-dimensional plot.

We also argue that eEPIPHANY is less labor intensive than LFA, because LFA requires human experts to generate the P-Matrix, which usually requires time-consuming cognitive task analysis. The high demand on human labor might not practical and hence might not scale up to apply to large online courses such as OLI. In fact, as far as we know, there has been no actual application of LFA with human-crafted P-Matrix to OLI courses. In the comparison in Table 3, the data for LFA is taken from DataShop [18], but LFA for these skill models used other existing skill models as P-Matrix (personal communication), therefore, it is not actually a fair comparison—LFA shows in this paper does not use the P-Matrix created by human experts. On the other hand, eEPHIPHANY automatically discover the P-Matrix from data.

Nonetheless, as our study has shown, eEPIPHANY and LFA discovered equally accurate skill models. We also found that different evaluation criteria (i.e., AFM vs. BKT in Table 3) show different favors on different search algorithm. LFA uses AFM and ePIPHANY uses BKT as a search bias, and that might have affected the results. We have yet to investigate this issue.

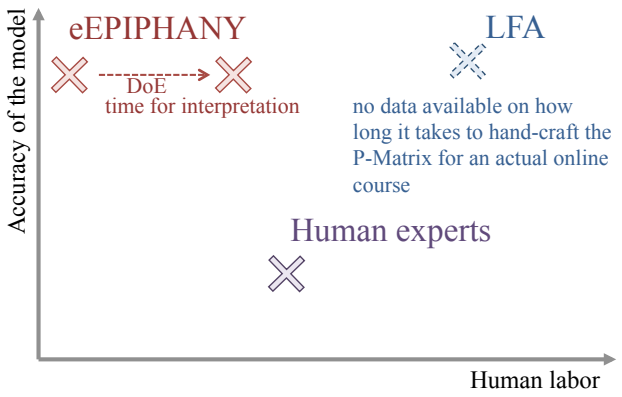

Figure 4. eEPIPHANY discovers skill models better than human experts and quicker than LFA

For our core goal—to provide evidence-based feedback for online course refinement—our study also suggests that eEPIPHANY can be used for a dual purposes with regard to skill model improvement: (1) When the online course is initially implemented, we should apply eEPIPHANY with the bag-of-words strategy. (2) When the online course is actually used and student learning data are collected, then we should apply eEPIPHANY with the student data to further improve the course.

The above observations further suggest that *authors of online courses would not need to create a default skill model at all—eEPIPHANY can find the default model by itself using the bag-of-words method*. This rather strong argument must be investigated as future research.

6. CONCLUSION

We found that eEPIPHANY is an efficient, practical, and quick method to automatically discover skill models from online course data without human interaction. Our empirical study showed that eEPIPHANY always finds skill models that are better than human-crafted skill models used in actual online courses. We also demonstrated that eEPIPHANY-crafted skill models have reasonable interpretability with the added help of the text analysis technique.

Creating effective online courses often requires intensive, iterative system engineering. Studying techniques for automatic skill model refinement and its application for evidence-based course refinement therefore is a critical research agenda for the successful future of online education.

ACKNOWLEDGEMENT

The research reported here was supported by National Science Foundation Awards No.1418244.

7. REFERENCES

1. Fishman, B., et al., *Creating a Framework for Research on Systemic Technology Innovations.* The Journal of the Learning Sciences, 2004. **13**(1): p. 43-76.
2. Stamper, J.C. and K.R. Koedinger, *Human-machine student model discovery and improvement using data*, in *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, S.B. G. Biswas, J. Kay, & A. Mitrovic, Editor. 2011, Springer: Berlin. p. 353-360.
3. Koedinger, K.R., et al., *Using Data-Driven Discovery of Better Student Models to Improve Student Learning*, in *Proceedings of the International Conference on Artificial Intelligence in Education*, H.C. Lane, et al., Editors. 2013, Springer: Memphis, TN. p. 421-430.
4. Velmahos, G.C., et al., *Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory.* Am. J. Surg, 2004. **18**: p. 114–119.
5. Koedinger, K.R. and E.A. McLaughlin, *Seeing language learning inside the math: Cognitive analysis yields transfer*, in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, S. Ohlsson and R. Catrambone, Editors. 2010, Cognitive Science Society: Austin, TX.
6. Cen, H., K. Koedinger, and B. Junker, *Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement*, in *Intelligent Tutoring Systems*, M. Ikeda, K. Ashley, and T.-W. Chan, Editors. 2006, Springer Berlin Heidelberg. p. 164-175.
7. Desmarais, M.C., *Mapping Question Items to Skills with Non-negative Matrix Factorization.* SIGKDD Explor. Newsl., 2012. **13**(2): p. 30--36.
8. Sun, Y., et al., *Alternating Recursive Method for Q-matrix Learning*, in *Proceedigns of the International Conference on Educational Data Mining*, J. Stamper, et al., Editors. 2014. p. 14-20.
9. Barnes, T., *The Q-matrix Method: Mining Student Response Data for Knowledge*, in *Proceedings of AAAI 2005 Educational Data Mining Workshop*. 2005.
10. Tatsuoka, C., et al., *Developing Workable Attributes for Psychometric Models Based on the Q-Matrix.* Journal for Research in Mathematics Education, accepted.
11. Lovett, M., O. Meyer, and C. Thille, *The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning.* Journal of Interactive Media in Education, 2008.
12. Bier, N., R. Strader, and D. Zimmaro, *An Approach to Skill Mapping in Online Courses*, in *Learning with MOOCs*2014: Cambridge, MA.
13. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**.
14. MacQueen, J., *Some methods for classification and analysis of multivariate observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. 1967.
15. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet Allocation.* The Journal of Machine Learning Research, 2003. **3**.
16. Koedinger, K.R., E.A. McLaughlin, and J.C. Stamper, *Automated student model improvement*, in *Proceedings of the 5th International Conference on Educational Data Mining*, K. Yacef, et al., Editors. 2012. p. 17-24.
17. Manning, C.D., P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. 2008, New York, NY: Cambridge University Press.
18. Koedinger, K.R., et al., *A Data Repository for the EDM community: The PSLC DataShop*, in *Handbook of Educational Data Mining*, C. Romero, et al., Editors. 2010, CRC Press: Boca Raton, FL.
19. Cen, H., K.R. Koedinger, and B. Junker, *Is over practice necessary? – improving learning efficiency with the Cognitive Tutor through educational data mining*, in *Proceedings of 13th International Conference on Artificial Intelligence in Education*, R. Luckin, K.R. Koedinger, and J. Greer, Editors. 2007, IOS Press: Amsterdam. p. 511-- - 518.

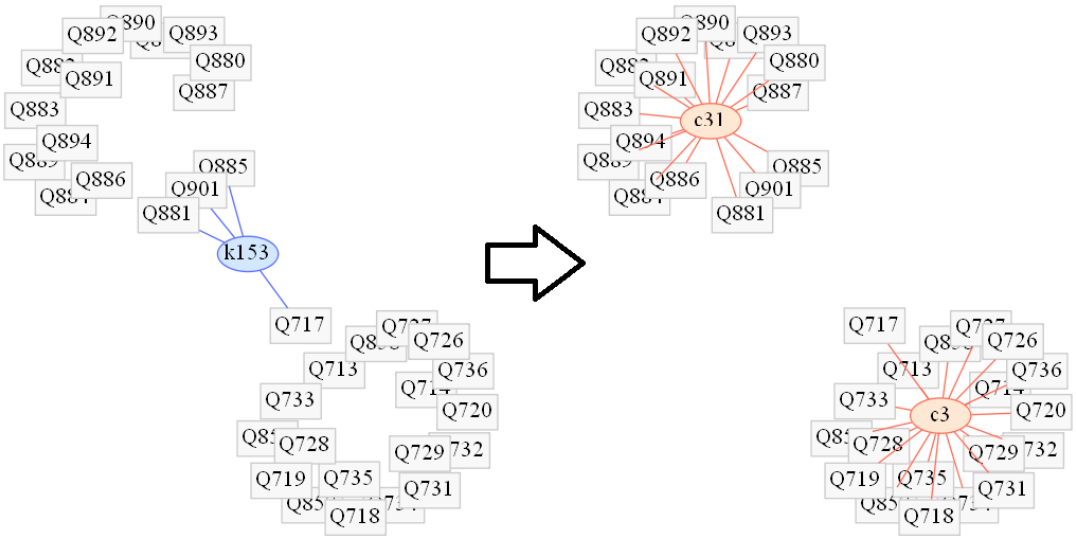

Figure 5. eEPIPHANY agrees with intuition: Assessment items are plotted in a skill-item association. (a) In the default skill model (left), skill *k153* are associated with assessment items Q881, Q885, and Q901 in the default skill model). (b) In the refined skill model (right), these three assessment items are associated with two skills (*c3* and *c31*) among others. In the figure, those other skills plotted in the "default" skill model are the ones contained in xI_D' (section 3.4).

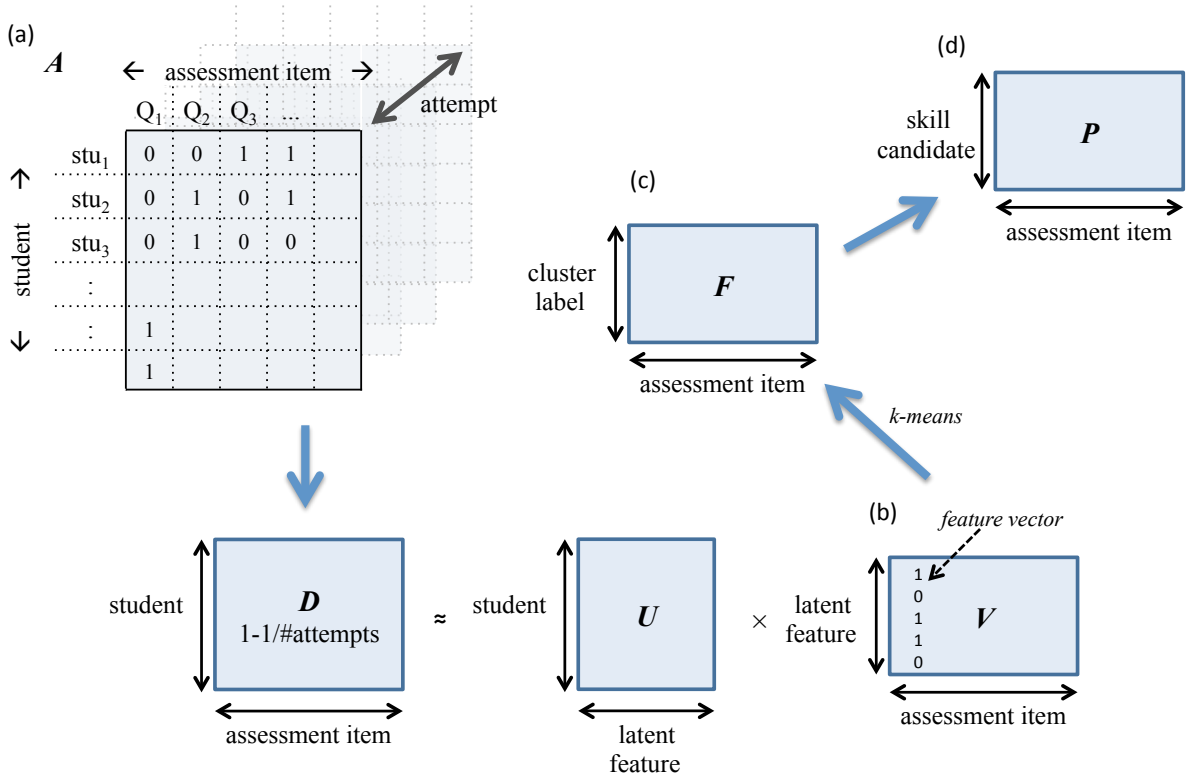

Figure 6. Overview of eEPIPHANY