Generative Model of Knowledge Gaps Discovery from Multiple Choice Tests

Dymitr Ruta
EBTIC, Khalifa University,
Abu Dhabi, UAE
dymitr.ruta@ku.ac.ae

Yousef Abosalem
Khalifa University,
Abu Dhabi, UAE
yousef.abosalem@ku.ac.ae

Ling Cen
EBTIC, Khalifa University,
Abu Dhabi, UAE
cen.ling@ku.ac.ae

***Abstract*—This submission represents a proposition of the Full Paper in the Research-to-Practice category in which we present our innovative work on automated student knowledge mapping and gaps discovery.**

Students' ability to learn and retain new knowledge depends on the precise state of their current knowledge distributed along a complex network of related concepts. Uneven distribution of competencies in the required knowledge concepts, or even worse - knowledge gaps - in certain critical concepts, may significantly hinder students' ability to acquire knowledge efficiently and prevent them from maximizing the potential of their talents. Discovery of and bridging these gaps at the individual student and group level is critically important for students to maximize their learning efficiency, for the educators to maximize the impact of the teaching, and for the governments to maximize the effectiveness of their educational programs and policies.

We propose a novel methodology for automated student knowledge mapping and gaps discovery from the results of multiple choice test questions extracted from a single .pdf report through pattern recognition techniques. Automatically lifted test performance data are then decomposed into student competencies in the related concepts required to correctly answer the test questions. Simple constrained linear optimization is deployed to find the optimal student competencies in the related concepts such that their linear combination best reconstructs the achieved test scores. Extracted knowledge maps in the related concepts allowed to discover common concept gaps for individual students and their groups, which has been successfully illustrated in the context of mathematics preparatory course at the university level with hundreds of students and thousands of test questions.

I. Introduction

Learning has always been one of the most important lifelong process of acquiring or discovery of new facts, skills, concepts we did not know before that increase our knowledge and expand our understanding of the world [1]. In an era of data deluge, the amount of information generated every 20 minutes around the globe is equivalent to the knowledge contained in the Library of Congress [2]. In such information-overloaded environment fast, efficient and meaningful learning becomes critically important. With the rapid development of Artificial Intelligent (AI) related emerging technologies like: Machine Learning (ML), Big Data (BD), Predictive Analytics (PR), Internet of Things (IoT), there has been a growing interest in the education community to take advantage of these advanced technologies and pass the benefits in a form of: improved learning performance, enhanced teaching effectiveness, reduced administrative workload to students, teachers and schools respectively [3], [4]. These relatively new technologies have been developed and evolved very quickly and now reached the state of maturity that allows them to efficiently penetrate data-rich education ecosystem to gain a valuable insight into the mechanics of the learning process, and extract solutions and recommendations of how to improve the learning efficiency, performance and knowledge retention that are positively changing and revolutionizing traditional education [4].

According to Wikipedia, knowledge, referring to the theoretical or practical understanding of a subject, is a familiarity, awareness, or understanding of facts, information, descriptions, or skills, which is acquired through experience or education by perceiving, discovering, or learning [5]. Knowledge, in general, can be considered as a complex network structure of hierarchically related concepts that vary in terms of complexity and granularity from the very basic fundamental building blocks up to the large thematic knowledge sections. From the human cognition point of view the concepts complexity and granularity are directly related to the process of knowledge acquisition i.e. learning and depend on the extend, complexity and quality of knowledge concepts already acquired by a specific student. Similar approach to knowledge represented by concepts studied as components of human cognition is adopted within the cognitive science disciplines of linguistics, psychology and philosophy [6], [7]. According to the Ausubel's learning theory, the most important factor influencing learning new knowledge is what students already know, suggesting teaching should be conducted according to prior knowledge of students [8]. It has been shown in the literature that teaching techniques that employ previous knowledge and new knowledge can significantly improve student

978-1-5386-1174-6/18/$31.00 ©2018 IEEE

learning performance [9], [10], [11]. As suggested in concept maps, proposed by Novak and his colleague at Cornell [12] based on Ausubel's learning theory, which use diagrams to organize and represent knowledge, explicit and structured relationships exist among concepts. Students' ability to learn and retain new knowledge depends on the precise state of their current knowledge distributed along a complex network of related concepts.

The reality of a typical school, however, is that the scope and proficiency of knowledge among different students even in the same age and in the same class vary significantly. This is due to many possible factors like different background of education, acquisition ability, learning habits, ability to focus, time management etc., to name the few, which lead to uneven distribution of competencies in the required knowledge concepts, sometimes resulting with deep knowledge gaps in critical enabling concepts. Such gaps may significantly inhibit educational progress, dent student confidence and motivation to learn that in turn prevents students from maximizing the potential of their talents. If not discovered and addressed deep knowledge gaps may significantly restrict students' ability to acquire new knowledge efficiently and as a knock-on effect promote further gaps in a subsequent learning process of related dependent knowledge.

Discovering knowledge gaps at the individual student and group level is critically important, which can guide and help educators and students to address these gaps in targeted ways, e.g. personalizing studying review, providing supplementary tutoring, adjusting teaching delivery material, etc. Having bridged knowledge gaps, rather than ignored them, students can more efficiently and confidently acquire new knowledge along any future educational journey, equally delivering efficiencies and relieving educators and educational policy makers from operational stresses and work overload.

Detailed identification of concept gaps among individual students is not an easy task and it is implicitly linked to equally complex quantification of the competencies of various concepts related to the knowledge that is being evaluated. Student themselves may not realize knowledge gaps they are facing. For example, they may be aware of a basic definition of a concept but struggle to grasp its detailed structure, applicability and complex nature of interaction with other concepts that would all impact their ability to consistently answer correctly any questions related to the tested concept. The most common way for teachers to evaluate learning performance of students and identify their knowledge gaps is using classroom assessment techniques (CAT) [13]. Through administering a formal test to students on a specific topic, educators try to find out the extent and structure of the related knowledge concepts [14], although the access to the latter from the test results remains typically vague and uncertain and is often neglected. Employing collaborative learning techniques [15] offers more informal means of identifying students' knowledge proficiencies [14]. As an example, in a so-called think-pair-share process, students are asked to give their thoughts on the tested concepts, share them among their groups, and report the results of discussions and cross-examined interactions on behalf of the whole group, in-line with the assessment guidelines of the evaluated knowledge scope [14].

Whether formal or informal assessment is employed to evaluate student knowledge, the common practice is to follow it with a laborious manual analysis of the tests carried out by the domain expert, the teacher, in order to reliably identify personalized gaps for a specific student. In many cases, since questions can be related to multiple intertwined concepts, additional tests are required to isolate specific misconceptions to give the student correct feedback. This complex manual gaps identification task becomes impossible to implement en masse for hundreds or thousands of students due to immense cost and time requirements and hence in practice is often partially or completely abandoned.

To address this challenge, we propose a novel methodology for automated student knowledge mapping and gaps discovery from the results of multiple choice test questions extracted from a single .pdf test report through a combination of pattern recognition techniques and linear constrained optimization. The major contributions of our work can be concluded as below:

1) An automated push-button method to extract test questions, answers and students' tests performance data from a single bulk .pdf imaged report, developed as a generic and reusable software prototype that employs image patterns recognition and OCR technology.
2) A method for linear decomposition of the knowledge implied in the test question into a set of associated knowledge concepts required to correctly answer the test question
3) A method based on constrained linear optimization uses students' test results and the associated question-to-concepts association map to find the optimal set of student competencies in the related concepts such that their linear combination best reconstructs the achieved test scores.
4) Individual student's and group's common concept gaps are then discovered based on the extracted student knowledge map.

Our proposed methodology has been successfully illustrated in the context of real mathematics preparatory course at the university level with hun-

dreds of students and thousands of test questions grouped into 6 distinct thematic sections.

The remainder of the paper is organized as follows. Conceptual knowledge representation is presented in Section II. In Section III, automated questions and test results extraction from .pdf test report is described. Students' knowledge map extraction is discussed in Section V, followed by knowledge gap discovery in Sections VI. Concluding remarks are given in Section VII.

II. Conceptual knowledge representation

The knowledge in a particular domain is represented as a finite set of distinct and independent quantum concepts $C = c_1, .., c_k$. The knowledge of these concepts can be tested by measuring the accuracy of the answers provided to the questions that cover one or more of such concepts. Formally, we associate the question q, coverage of the knowledge concepts C with the support vector $s = [r_1, .., r_k]$ expressing degrees of association or relationship with the corresponding concepts $c_1, .., c_k$, and assume that $\sum_{i=1}^{k} r_i = 1$, i.e. questions only involving and fully confined to the knowledge space C.

Now let us consider a degree of knowledge of a specific concept c_j that can be attributed to a particular student s_i and associate it with the expectation of correctly answering a question exclusively covering concept c_j: $w(s_i, c_j) = E(test(s_i, c_j))$, where $test_{i,j}$ is a logical test function returning 1 (true) if the i^{th} student answers correctly the question exclusively covering c_j ($r_j = 1$) and E is the expectation operator.

In practice, however, it is unrealistic to expect questions exclusively covering individual concepts. Instead a realistic test question q covers multiple concepts $\mathbf{c}$ with various degrees of relationship $\mathbf{r}$, while students have various degrees of knowledge $\mathbf{w}$ of these concepts. For such general case we now consider the probability of a student s_i with the knowledge $\mathbf{w_i} = w_{i,1..k}$ of concepts $\mathbf{c} = c_1, .., c_k$ to correctly answer the question q that relates to the associated concepts $\mathbf{c}$ with the coverage vector $\mathbf{r_q} = r_{q,1..k}$ as:

$$p(test(s_i, \mathbf{r_q}) = 1) = p_{i,q} = \sum_{j=1}^{k} w_{i,j} r_{q,j} = \mathbf{r}_q \mathbf{w}_i^T \tag{1}$$

Here for simplicity we model the student performance at answering question q as a linear combination function of student competencies in the related concepts necessary to answer this question correctly. Note that although in general knowledge is believed to be structured in a form of concepts hierarchies, such linear representation could be considered sufficient for inferring the knowledge competencies in a small subset narrowed down by the scope of the questions.

Given such representation the objective of this study can be formulated as, first, finding the students' competencies in the taught knowledge concepts $\mathbf{w}_i$, to then identifying the knowledge gaps i.e. concepts c_j, for which the student s_i has insufficiently low competencies $w_{i,j}$, subject question-concepts mapping defined by $\mathbf{r}_q$. Note that the actual student performances $p_{i,q}$ in all of the test questions are readily available from the tests results, while the concepts mapping vector $\mathbf{r}_q$ can be defined manually by the related knowledge expert as a one-off task.

Hence, given the test results $p_{i,q}$ and the question-to-concepts mapping $\mathbf{r}_q$ it is possible to extract individual student's competencies or knowledge $w_{i,j}$ in the concepts c_j, as a result of the optimization that finds the optimal $w_{i,j}$ for which the deviation between reconstructed and actual student performance is minimized:

$$\mathbf{w}_i = \underset{w_i}{\operatorname{argmin}} \sum_{q=1}^{Q} (p_{i,q} - \mathbf{r}_q \mathbf{w}_i^T)^2 = \underset{w_i}{\operatorname{argmin}} |\mathbf{p}_i - \mathbf{r}\mathbf{w}_i^T|_2 \tag{2}$$

subject to constraints of $0 \leq w_{i,j} \leq 1$, where $|x|_2$ stands for the $L^2 - norm$ operator.

III. Automated test results extraction from .pdf test reports

Detailed data from test results are rarely available in a well structured and organized electronic format that would allow immediate analysis and knowledge inference. In practice still most of tests are carried out on paper such that an insightful conceptual analysis of errors would only be available via manual human intervention. At the scale of a school or even a class the amount of manual work required to infer detailed feedback from such tests like gap analysis grows prohibitively large such that in practice it is simply limited to plain assessment, nonetheless, absorbing huge amount of expensive skilled assessors' time.

Even if the tests are carried out in a modern computerized environment and the results are stored electronically, the efficiency and data protection constraints very rarely allow to combine together detailed, often graphical information about the test questions and students responses along with their time-ordered characteristics. Instead, typically just the student answers to the specific questions are stored, which allows to generate immediate performance scores after the test. The bigger opportunities to extract detailed structure of the knowledge gaps that prevented the student from correctly answering all the questions are usually missed out.

Often, though, in addition to the electronic answers and scores, every single test is backed up

in an originally presented format i.e. the test paper with student answers and pdf-imaged into a single archive test report. All the information required for detailed gap analysis is included in such report, although in a form that is not immediately consumable for insightful data analytics.

Since exactly such a case was with the Math Preparatory Module and its Assessment run as a part of our university's 1^{st}-year introductory courses, we have decided to develop a software engine for automated extraction of all useful data from the .pdf assessment report and thereby eliminate the time consuming and costly data collection and preparation burden.

Our generic tool was designed to automatically scan through the imaged .pdf report and recognize image patterns corresponding to specific predefined test questions followed with OCR-based extraction of matched students' answers along with other test statistics like test answering delays, durations, attempts, changes, errors etc. The questions detector exploited the fact that electronically edited test questions after rescaling look the same for all students and hence it was relatively easy to match-detect every single question among the total of over 12000 test questions presented to over 430 students in a series of 6 thematic math tests. The detection was observed when the average pixel color difference between the sliding question template and the overlapping region in the report significantly dropped indicating a match.

IV. Decomposition of Questions into Knowledge Concepts

Given the matched predefined test questions, their correct answers and the corresponding student answers in each of the six tests of the Math Preparatory Course, the next task was to decompose the knowledge required to solve each question into a set of underlining elementary math concepts. This activity had to be carried out manually by a domain expert but it was a one-off process that involved simply a teacher solving each question and listing down all distinct mathematical concepts that were required along the way to solve the question correctly. This task involves a bit of an arbitrary decomposition of taught content down to the resolution that the teacher or the course provider is comfortable with and able to address during the course. Equally a standard reference concepts from the domain knowledge could be used and worked against in this case. In our trial, however, we applied knowledge expert based decomposition into a smallest set of concepts, knowledge of which is sufficient to solve correctly all the questions within each of the tested math domains.

Fig. 1 illustrates the questions, all relevant concepts and their associations for each question within arithmetics, college and intermediate algebra, while Fig. 2 presents the same for polynomial & functions, logarithms & exponents and trigonometry, respectively.

The binary question-to-concepts association masks clearly visible in Figs. 1 and 2 simply mean that the concepts (columns) identified by coordinates of the black squares are required to solve correctly corresponding questions (rows). In fact these diagrams are directly translated into the question-to-concepts association maps $\mathbf{r}$ defined in (1) and (2) after simple question-wide normalization to achieve $\sum_{i=1}^{k} r_{q,i} = 1$, i.e. for simplicity assuming that the fully correct solution of question q requires equal contributions of all associated concepts.

V. Students' Knowledge Map Extraction

Given student test results and the established associations between the knowledge required to solve a given question and the related set of concepts we can proceed to the extraction of student competencies in each of the related concepts.

This process in line with (2) is defined as an optimization process that tries to find the i^{th} student knowledge vector $\mathbf{w_i}$ in the related knowledge concepts such that the average squared error between the actual student result for this question $p_{i,q}$ and the expected result based on his/her knowledge of the underlining associated concepts $\mathbf{r}_q\mathbf{w}_i^T$ is minimized across all questions inline with (2), subject to the constraints of $\forall w, 0 \leq w \leq 1$ i.e. that the student competency in each concept scales between a complete ignorance and a full knowledge, respectively. Note that in such generative student performance model the student is expected to answer:

- all related questions correctly, only if all concept competencies are equal to 1
- all related questions incorrectly, only if all concept competencies are equal to 0
- p fraction of all related questions correctly, if all concepts competencies are equal to p

Figs. 3(a) -3(f) illustrate the knowledge concepts' distribution among students based on multiple choice tests in arithmetics, college algebra, intermediate algebra, polynomials and functions, logarithms and exponents, and trigonometry. In these figures, the x-axis represents the elementary math concepts that are sorted from the most difficult to the easiest, and the y-axis depicts the numbers of students that are sorted by students' overall test average scores from the lowest to highest. The color changing from read to green illustrates the concept competencies of students, in which pure red corresponds to full ignorance and pure green corresponds to full competency.

$\left(\frac{3}{2}+1\frac{2}{3}\right)\div\frac{4}{5}=$

a) $\frac{95}{24}$
b) $\frac{55}{24}$
c) $\frac{31}{24}$
d) $\frac{29}{24}$

$\frac{1}{2}\times3\frac{2}{3}+\frac{7}{4}=$

a) $\frac{16}{3}$
b) $\frac{31}{12}$
c) $\frac{49}{12}$
d) $\frac{43}{12}$

$2.8\times2.3-3.6=$

a) 2.84
b) 1.84
c) 4
d) 0.84

$12+16\div4\times2=$

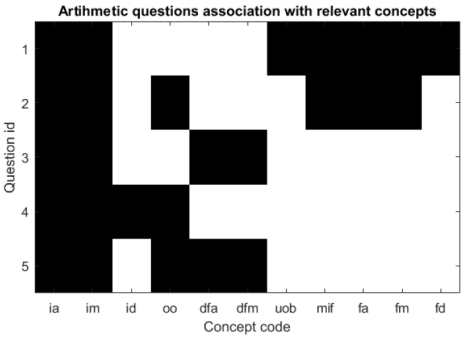
a) 2
b) 20
c) 12
d) 14

$4.4+1.9\times1.7=$

a) 8.84
b) 3.93
c) 4.03
d) 7.63

ARITMETIC CONCEPTS DERIVED FROM TEST QUESTIONS

code	name	type
ia	integer addition	arithmetic
im	integer multiplication	arithmetic
id	integer division	arithmetic
oo	operations order	arithmetic
dfa	decimal fractions addition	arithmetic
dfm	decimal fractions multiplication	arithmetic
uob	use of brackets	arithmetic
mif	mixed to improper faction conversion	arithmetic
fa	fractions addition	arithmetic
fm	fractions multiplication	arithmetic
fd	fractions division	arithmetic



(a) Arithmetics test questions and associated concepts

The product of $(x-2)(3x^2+4x-1)$?

a) $3x^3-6x^2+9x+2$
b) b) $3x^3-6x^2+9x+1$
c) c) $3x^3+2x^2+9x-2$
d) d) $3x^3-2x^2-9x+2$

$-6x+4=-2(5-3x)+2x$

a) 1
b) 9
c) -4
d) -7

The solution of the inequality $-5(-1+4x)\geq-23-6x$ is:

a) $x\leq2$
b) $x\leq-8$
c) $x\geq-23$
d) $x\geq-8$

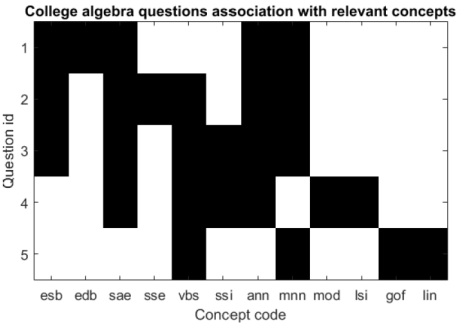
Solve the inequality $|x+3|>2$.

a) $x>-1$
b) $x>1$
c) $x>-1, or\ x<-5$
d) $x<5, or\ x>-1$

Which of the following equations has a graph perpendicular to the graph of $3x+4y=4$

a) $y=\frac{3}{4}x+1$
b) $y=\frac{-3}{4}x+1$
c) $y=\frac{4}{3}x+1$
d) $y=\frac{-4}{3}x+1$

code	name	type
esb	expanding single bracket	college algebra
edb	expanding double brackets	college algebra
sae	simplifying algebraic expressions	college algebra
sse	solving simple equation	college algebra
vbs	reorganizing variables both sides	college algebra
ssi	solving simple inequality	college algebra
ann	adding negative numbers	college algebra
mnn	multiplying negative numbers	college algebra
mod	expanding modulus	college algebra
lsi	solving system of inequalities	college algebra
gof	graph of a function	college algebra
lin	line function properties	college algebra



(b) College algebra test questions and associated concepts

$\frac{2y}{6x^2}-\frac{3x}{4y^2}=$

a) $\frac{4y^3-9x^3}{12x^2y^2}$
b) $\frac{2y-3x}{6x^2-4y^2}$
c) $\frac{-xy}{2x^2-y^2}$
d) $\frac{8y^3-10x^3}{24x^2y^2}$

Solve for x: $\frac{3}{x}+\frac{1}{4}=1$

a) $x=3$
b) $x=4$
c) $x=\frac{1}{4}$
d) $x=0$

$\frac{6x^2+7x+2}{6x^2-x-2}=$

a) $\frac{2x+1}{2x-1}$
b) $\frac{2x+1}{3x-2}$
c) $-\frac{7x+2}{x-2}$
d) $\frac{3x+2}{3x-2}$

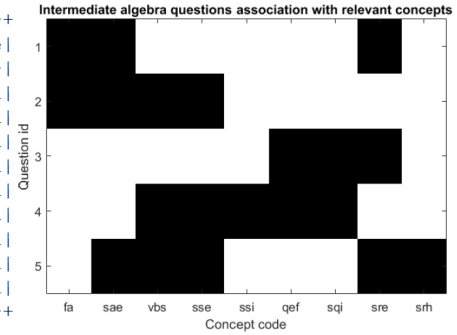
Solve for x: $3x^2-5x>2$

a) $x>2$ or $x<\frac{-1}{3}$
b) $\frac{-1}{3}<x<2$
c) $x>\frac{-1}{3}$ or $x>2$
d) $-2<x<\frac{1}{3}$

Make t the subject of the equation $2\sqrt{\frac{v+t}{t}}=\frac{1}{v}$

a) $t=\frac{4v^2}{1-4v^2}$
b) $t=\frac{v^2}{v^2-1}$
c) $t=\frac{4v^3}{1-4v^2}$
d) $t=\frac{3v^3}{1-2v^2}$

code	name	type
fa	fractions addition	intalg algebra
sae	simplifying algebraic expressions	intalg algebra
vbs	reorganizing variables on both sides	intalg algebra
sse	solving simple equation	intalg algebra
ssi	solving simple inequality	intalg algebra
qef	quadratic equation factorization	intalg algebra
sqi	solving quadratic inequality	intalg algebra
sre	simplifying reciprocal expressions	intalg algebra
srh	square root handling	intalg algebra



(c) Intermediate algebra test questions and associated concepts

Figure 1. Concepts and their associations with test questions from: basic arithmetics, college and intermediate algebra.

The domain of $f(x)=\sqrt{x-7}+2$ is:

a) $x \geq 2$
b) $x > 2$
c) $x > 7$
d) $x \geq 7$

For $f(x)=x^2+x-9$, $f(a+2)=$

a) a^2+a-7
b) a^2+2a-5
c) a^2+a-3
d) a^2+5a-3

If $8x^3+2x+1$ *is divided by* $2x+1$, then the remainder is:

a) 1
b) 2
c) - 1
d) - 2

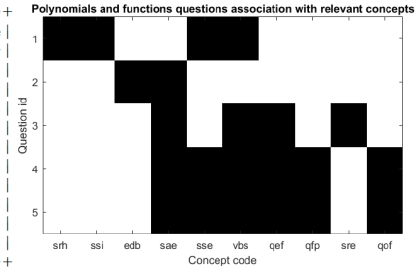
The points of intersection of the curves $y=x^2-3x$ *and* $y=2x-6$ are:

a) (2,3),(- 2,0)
b) (2,0),(- 2,3)
c) (0,2),(- 2,3)
d) (2,- 2),(3,0)

The value of a such that the function $f(x)=x^2-2ax+1$ has a minimum at $x=1$ is:

a) $a=1$
b) $a=-1$
c) $a=2$
d) $a=-2$

code	name	type
srh	square root handling	polynomials functions
ssi	solving simple inequalities	polynomials functions
edb	expanding double brackets	polynomials functions
sae	simplifying algebraic expressions	polynomials functions
sse	solving simple equation	polynomials functions
vbs	reorganizing variables both sides	polynomials functions
qef	quadratic equation factorization	polynomials functions
qfp	quadratic function properties	polynomials functions
sre	simplifying reciprocal expressions	polynomials functions
qof	graph of a function	polynomials functions



(a) Polynomials & function test questions and associated concepts

$\log_a b = 4$ means:

a) $b^4 = a$
b) b) $a^4 = b$
c) c) $4^a = b$
d) d) $4^b = a$

Solve $\frac{9^{x+1}}{3^{x-1}} = 27$

a) $x = 0$
b) $x = 1$
c) $x = -1$
d) $x = 2$

If $f(x)=\frac{1}{1-e^x}$, $f^{-1}(x)=$, *for* $x>1$.

a) $1-e^x$
b) $\frac{-1}{1-e^x}$
c) $\frac{\ln(x-1)}{x}$
d) $\ln(\frac{x-1}{x})$

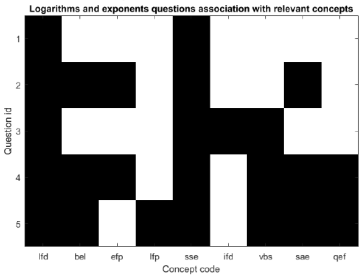
Solve $3^{2x+1}+3^{x+2}-\log_3(27)=3^x$

a) $x=\{0,1\}$
b) $x=\{1\}$
c) $x=\{1\}$
d) $x=\{1,1\}$

Solve $\log_2(x^2)-\log_x(2)=-1$

a) $x=\{2^{\frac{1}{3}}, 2^{-1}\}$
b) $x=\{2, 2^{-1}\}$
c) $x=\{2,4\}$
d) $x=\{2^2,-2\}$

code	name	type
lfd	logarithm function definition	logarithms and exponents
bel	basic exponents and logarithms	logarithms and exponents
efp	exponential function properties	logarithms and exponents
lfp	logarithm function properties	logarithms and exponents
sse	solving simple equation	logarithms and exponents
ifd	inverse function derivation	logarithms and exponents
vbs	reorganizing variables both side	logarithms and exponents
sae	simplifying algebraic expressions	logarithms and exponents
qef	quadratic equation factorization	logarithms and exponents



(b) Logaithms & exponents test questions and associated concepts

In the right triangle ABC shown below, the length of AB is 6 cm and the measure of angle ACB is x°. What is the length of AC?

a) $\frac{6}{\cos(x)}$
b) b) $6\cos(x)$
c) c) $\frac{6}{\sin(x)}$
d) d) $6\sin(x)$

If $\sin(x+y)=\sin x\cos y+\cos x\sin y$, *then* $\sin\left(\frac{\pi}{2}+\alpha\right)=$

a) $\sin\alpha$
b) - $\sin\alpha$
c) - $\cos\alpha$
d) $\cos\alpha$

For all real numbers x, $\cos^2(99x)+\sin^2(99x)=$

a) *0*
b) *b) 1*
c) *c) cos(198x)*
d) *d) sin(198x)*

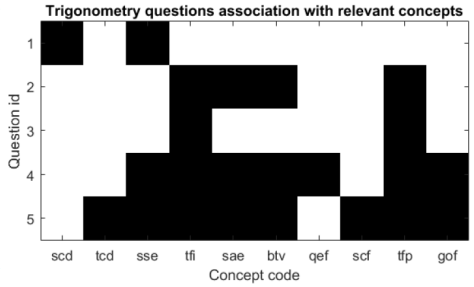
Solve $2\cos^2(\frac{x}{2})-\sin(\frac{x}{2})-1=0$, for $0° \leq x \leq 360°$

a) $x = 60°, 150°$
b) $x = 60°, 300°$
c) $x = 30°, 150°$
d) $x = 30°, 60°$

If $\sin\theta=\frac{5}{13}$ *and* $\tan\theta<0$ *then* $\sec\theta=$

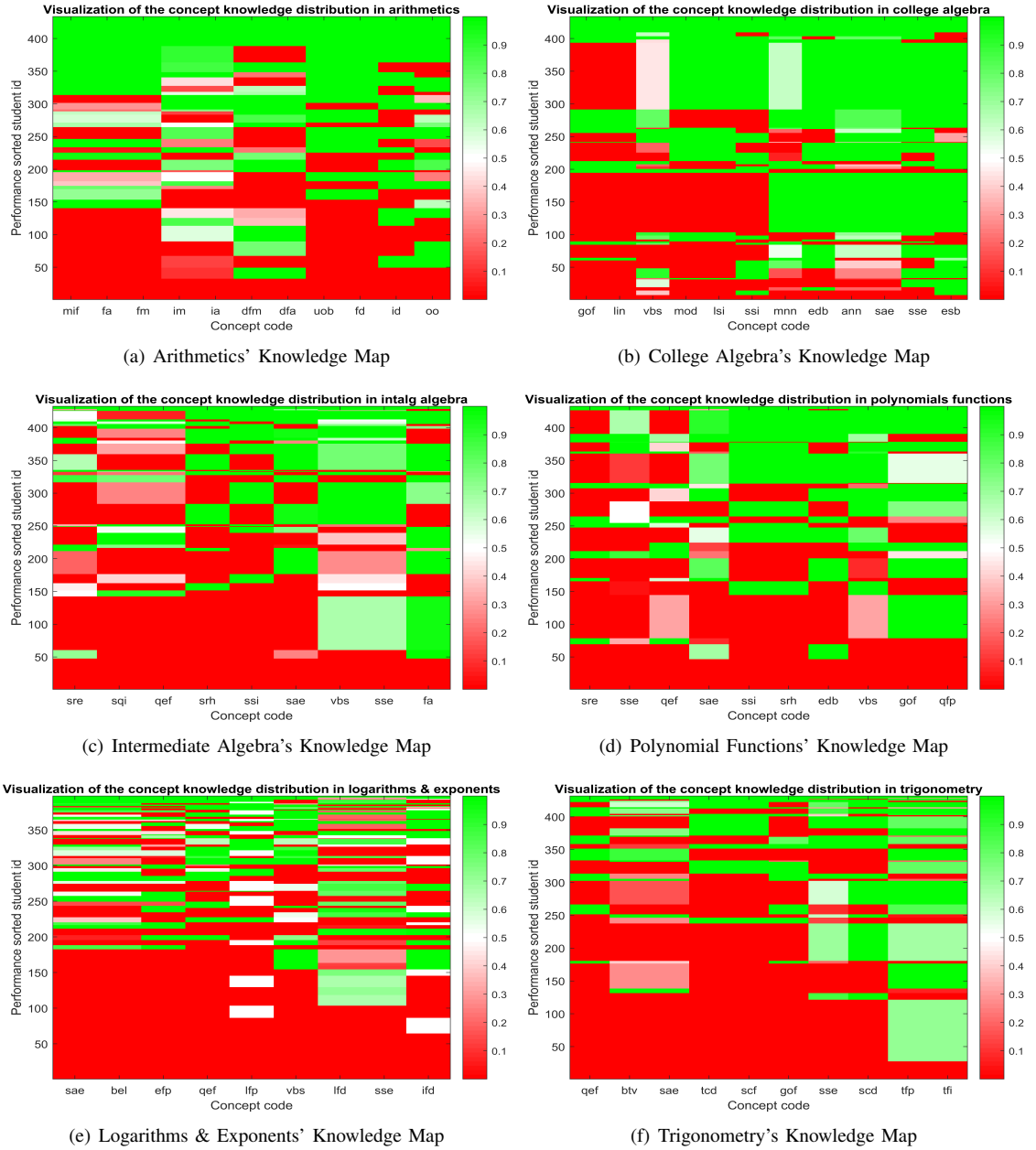
a) $\frac{13}{5}$
b) $\frac{-12}{13}$
c) $\frac{-13}{12}$
d) $\frac{-13}{5}$

code	name	type
scd	sin cos definition	trigonometry
tcd	tan cot definition	trigonometry
sse	solving simple equation	trigonometry
tfi	trigonometric function identities	trigonometry
sae	simplifying algebraic expressions	trigonometry
btv	basic trigonometric values	trigonometry
qef	quadratic equation factorization	trigonometry
scf	sec, cosec function definitions	trigonometry
tfp	trigonometric function properties	trigonometry
gof	graph of function	trigonometry



(c) Trigonometry test questions and associated concepts

Figure 2. Concepts and their associations with test questions from: polynomials & functions, logarithms & exponents, trigonometry.



(a) Arithmetics' Knowledge Map

(b) College Algebra's Knowledge Map

(c) Intermediate Algebra's Knowledge Map

(d) Polynomial Functions' Knowledge Map

(e) Logarithms & Exponents' Knowledge Map

(f) Trigonometry's Knowledge Map

Figure 3. Visualization of the knowledge concepts' distribution among students based on multiple choice tests in (a) arithmetics, (b) college algebra, (c) intermediate algebra, (d) polynomials and functions, (e) logarithms and exponents, (f) trigonometry.

VI. Knowledge Gap Discovery

Students' ability to learn and retain new knowledge depends on the precise state of their current knowledge. Uneven distribution of competencies in the required knowledge concepts, especially knowledge gaps in core concepts, may significantly hinder students' ability to acquire knowledge efficiently and prevent them from maximizing the potential of their talents. The knowledge concepts' distribution extracted from multiple choice tests shown in Section V can be very effective for discovery of knowledge gaps at both individual and group levels. Bridging these gaps are critically important for students to maximize their learning efficiency, for the educators to maximize the impact of teaching, and for the governments to maximize the effectiveness of educational programs.

It can be seen from the arithmetics' concept distribution shown in Fig. 3(a) that around 140 lower-performing students lack the competence in: *mif* (mixed to improper faction conversion), *fa* (fractions addition), *fm* (fractions multiplication), and *fd* (fractions devision) concepts. These 4 concepts are all related to fraction operations that appear to significantly hinder students' performance in this entry test, with all the misconceptions likely to involve various rules related to these operations, like converting to the common denominator, frac-

tion conversions, numerator/denominators engagement in multiplication and division etc. Shallow knowledge in this space might have been a result of disconnected memorization of rules without understanding their intrinsic meaning, and perhaps lack of adequate practice to boost their retention. On the other hand, the knowledge competence in fraction operations among higher-performing students are very high (nearly the maximum 1 for highest-performing students). It appears, therefore, that deep understanding of fractions' operations emerges as the critical enabler in the arithmetics test i.e. its deep knowledge almost guaranties high overall performance, while its ignorance consistently leaves student with deep gaps that significantly affect the arithmetics test performance. Immediate and perhaps most efficient feedback that the teacher might provide is additional supplementary tuition on fundamental concepts in fraction operations. Intrinsic relationship between the fractions operations and the other arithmetic concepts covered in the test implies that the corrective action in this space alone would most likely maximally lift the overall performance of students in the test.

In the college algebra's extracted knowledge illustrated in Fig. 3(b), on the one hand the competence in *lin* (linear function properties) and *gof* (graph of a function) emerged as very poor performers, however both were tested in only one question, hence they need further verification in subsequent tests. Moreover, the competencies in the concepts of *lsi* (solving system of inequalities), *ssi* (solving simple inequalities) were poor among not only lower-performing students but also some higher-performing students, especially if modulus (*mod*) was involved. The inequalities related concepts are often misunderstood or treated as equations with the classic sign changes errors when applying transformations with negative numbers. Interestingly, it can be further observed from the intermediate algebra tests in Fig. 3(c) that both concepts related to inequalities, i.e. ssi (solving simple inequality) and sqi (solving quadratic inequality), are also quite low, combined with deep gaps in *sre* (simplifying reciprocal expressions) and *qef* (quadratic equation factorization) that are rather routine technical skills one would expect to be well covered. Also subsequent tests in polynomial functions, logarithms and exponents, and trigonometry, displayed in Figs. 3(d), 3(e), and 3(f), respectively, confirmed *qef* and *sre* as consistently gapping concepts with poor performance across the board. These further tests rejected also the earlier hypothesis that function graphing skills (*gof*) is a consistently poor performer. Further more detailed analysis of concept gaps or misconceptions is possible and can be combined with a measured dissonance with the overall questions difficulty and the factors in concept dependencies to provide more and more accurate feedback and effective intervention to the student or a group of students.

VII. Conclusions

We presented a novel methodology for automated student knowledge mapping and gaps discovery from the results of multiple choice test questions extracted from a single pdf-imaged report through pattern recognition techniques. The automatically lifted test performance data were decomposed into student competencies in the related concepts required to correctly answer the test questions. Simple constrained linear optimization was deployed to find the set of student competencies in the related concepts such that their linear combination best reconstructs the achieved test scores. The extracted student knowledge map allowed to discover individual student's and group's common concept gaps, which has been successfully illustrated in the context of mathematics preparatory course at the university level with hundreds of students and thousands of test questions.

Discovery of and bridging these gaps at the individual student and group level is critically important for students to maximize their learning efficiency, for the educators to maximize the impact of teaching, and for the governments to maximize the effectiveness of their educational programs.

References

[1] I. King, "Big education in the era of big data," *Federated Conference on Computer Science and Information Systems*, 2014.

[2] R. Smolan and J. Erwitt, The human face of big data, 2012.

[3] B. Schmarzo, "What universities can learn from big data higher education analytics," https://infocus.emc.com, 2014.

[4] L. Cen, R. Dymitr, and J. Ng, "Big Education: Opportunities for Big Data Analytics," *Proc. IEEE International Conference on Digital Signal Processing*, Singapore, 2015.

[5] knowledge: definition of knowledge in Oxford dictionary (American English) (US), *oxforddictionaries.com*, Archived from the original on 2010-07-14.

[6] E. Margolis and S. Laurence, Concepts: Core Readings, Chapter 1, MIT Press, 1999.

[7] S. Carey, "Knowledge Acquisition:Enrichment or Conceptual Change?" *S. Carey and R. Gelman (Eds.), The Epigenesis of Mind: Essays on Biology and Cognition*, pp. 257-291, Hillsdale, NJ: Lawrence Erlbaum Ass., 1991.

[8] D. Ausubel, "Educational Psychology: A Cognitive View," Holt, Rinehart and Winston, New York, 1968.

[9] A. Chickering and Z. Gamson, "Seven principles for good practice in undergraduate education," *AAHE Bulletin*, vol. 40, no. 7, pp. 3-7, 1987.

[10] D. Fink, Creating Significant Learning Experiences, 2003.

[11] G. Kuh, "The national survey of student engagement: Conceptual and empirical foundations," *New Directions for Institutional Research*, vol. 141, pp. 5-20, 2009.

[12] J. Novak, "Institute for Human and Machine Cognition (IHMC)," Retrieved 2008-04-06.

[13] T. Angelo and K. Cross, Classroom Assessment Techniques, San Francisco: Jossey-Bass, 1993.

[14] S. Ginsberg, ""Mind the Gap" in the Classroom," *The Journal of Effective Teaching* 10(2) : 74-80, 2010.

[15] E. Barkley, K. Cross, and C. Major, Collaborative learning techniques, 2005. San Francisco: Jossey-Bass.