Using Contrast Data to Determine Student Knowledge Granularity

Communicating Student's Recent Understanding At a Glance

Thomas Li
tzl@andrew.cmu.edu
Carnegie Mellon University School of Computer Science
Pittsburgh, Pennsylvania

Kenneth Holstein[*]
Vincent Aleven[*]
kjholste@cs.cmu.edu
aleven@cs.cmu.edu
Carnegie Mellon University Human-Computer Interaction Institute
Pittsburgh, Pennsylvania

ABSTRACT

Intelligent tutoring systems generate large amounts of data on students' learning and behavior. By leveraging this data, visual learning analytics tools hold the potential to help teachers gain valuable insights into how students learn. This project sought to explore new ways to support teachers in making faster, more accurate inferences about precisely where students are struggling.

By characterizing student performance in terms of more abstract or less abstract canonicalized categories, we are able to create an algorithm that finds non-obvious contrasts to characterize an individual student's current state of learning (i.e., categories that the student has learned and has not yet learned) that the teacher could capitalize on in an individualized session. The resulting outputs were tested with teachers and researchers, to investigate the tool's utility and gain insight into how it might be improved in future work.

CCS CONCEPTS

• **Human-centered computing** → *Usability testing*; HCI theory, concepts and models.

KEYWORDS

educational data mining, canonicalization, student performance visualization

ACM Reference Format:
Thomas Li, Kenneth Holstein, and Vincent Aleven. 2019. Using Contrast Data to Determine Student Knowledge Granularity: Communicating Student's Recent Understanding At a Glance. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 7 (May 2019), 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

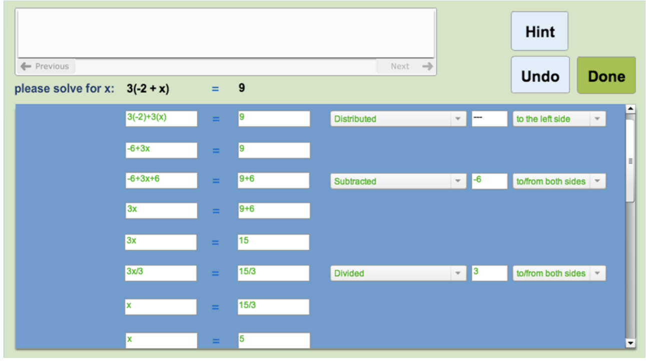
Past research conducted by Holstein and Aleven resulted in the creation of Lynette, the problem solving interface that students use to try and solve single variable equations, and Lumilo, a Microsoft Hololens application for the teacher that receives the data collected and analyzed by Lynette and displays it to quickly give the teacher a better understanding of who is struggling, and on what concepts.

This framework uses a knowledge-component based mastery system, continuously sending students questions on topics which they have not yet mastered until

[*] Both authors contributed equally to this research.

Authors' addresses: Thomas Li, tzl@andrew.cmu.edu, Carnegie Mellon University School of Computer Science, 5000 Forbes Ave, Pittsburgh, Pennsylvania, 15213; Kenneth Holstein, kjholste@cs.cmu.edu; Vincent Aleven, aleven@cs.cmu.edu, Carnegie Mellon University Human-Computer Interaction Institute, 5000 Forbes Ave, Pittsburgh, Pennsylvania, 15213.

1073-0516/2019/5-ART7 $15.00
https://doi.org/10.1145/1122445.1122456



please solve for x: 3(-2 + x) = 9

3(-2)+3(x) = 9 Distributed --- to the left side
-6+3x = 9
-6+3x+6 = 9+6 Subtracted -6 to/from both sides
3x = 9+6
3x = 15
3x/3 = 15/3 Divided 3 to/from both sides
x = 15/3
x = 5

Figure 1: Lynnette's Problem-Solving Interface [Yanjin Long and Aleven 2018]

they get the questions correct enough times to declare a topic mastered. When queried about a specific student, the framework displays both the current question the student is working on, and the most recent mistake made in the concept in which they do the worst.

Motivation

This project sought to improve upon that framework to increase the amount and quality of data given to the teacher which which to make inferences about student learning. For any given knowledge component, or skill, there are several different levels of granularity involved, depending on how well the students are able to abstract the concept. As an example: in regards to subtracting a term from both sides to maintain equality, and cancel out a term on one side, students may have trouble distinguishing between variables and constants, being perfectly fine with constants, but having trouble with variables, or vice versa. They could have mastered adding both variables and constants, but only when they are non-negative, or only when the term being cancelled out is on the right side of the equation and not the left, and any permutation of the above and more.

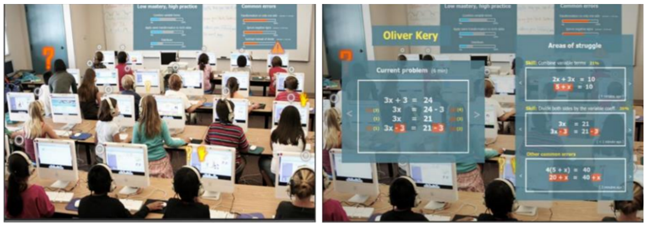


Figure 2: Lumilo's Teacher Interface [Kenneth Holstein and Aleven 2018]

Approach

In order to provide accurate analysis as to at what granularity student knowledge is at, student transaction data from Lynnette is used. Every transaction that corresponds to a student attempting a step in solving a problem is considered, while all other transactions, which include randomly clicking the screen, continuously attempting to skip the problem by pressing "done", or spamming the hints button are ignored. For every valid attempted step, the state of the problem before the attempted step is recorded, and fully canonicalized, for use in the algorithm to determine granularity.

Contribution

In dynamically analyzing past student performance, this project aims to detect the proper level of granularity of the student's current understanding, and reflect this in their output display to the teacher when queried. Naturally, to be able to collect and display this information, it depends almost wholly on the skill granularity component. In addition, records kept of past student performance will further enhance this feedback by allowing the instructor to determine which student(s) had issues with a particular issue, and then managed to overcome it, further allowing them to facilitate group learning, and freeing the teacher for other tasks.

2 DESIGN / APPROACH

Prior work, including the current system in Lumilo, works based off of a traditional Knowledge Component (KC) model. KC models in this field tend to have low granularity, with components such as "add/subtract both sides" serving as large umbrella terms of operation types, instead of specific question format.

The primary hypothesis in this study is that while students are intended to learn these components and be able to apply them for general use; before they learn them, they rely on examples, making their understanding a question-level matter, instead of an operation-level one. As a result, more accurate knowledge of where students are struggling can theoretically be obtained from only the question format itself.

To that end, a tool was developed that fully canonicalizes (tokenizes) all problem steps students are presented with, leaving all of them to fit into the discovered categories. Given that the most specific driving motivation for this was the observation that students tend to have issues performing basic operations when signs change, a "tree" is constructed for every step a student attempts based on the tokenization of the step's initial state. A sample of such a generated tree is displayed below.

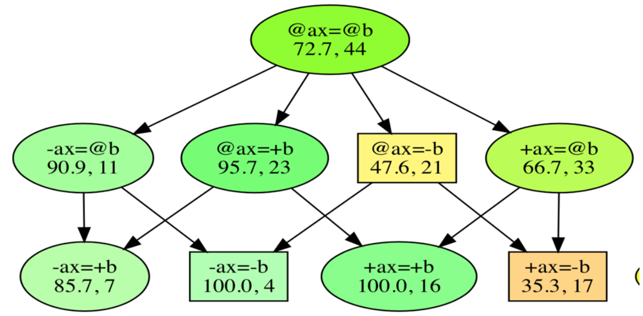


Figure 3: Example of a Canonicalized Tree Found for a Sample Student

Over the course of a student's learning, the number of such observed trees increases pseudo-logarithmically as the pool of question types gets exhausted. However, the total number of such trees can still balloon to a size too large for teachers to process efficiently, as shown below.

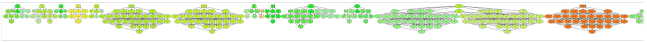


Figure 4: Part of the Full Set of Trees Generated for a Sample Student

The tree in **Figure 3** displays an example of the type of patterns in student performance we are looking for in the square-shaped nodes. The nodes are colored according to how well the student understands how to solve problems of that type, with green indicating high understanding, and red indicating low understanding. The square nodes indicate that swapping the sign of the variable term in questions of the form $?x =?$ produces a tangible difference in student performance. The algorithm developed is tailored to look for distinctions such as these based on a student's transaction data, as recorded by Lynnette. In semantic terms, the algorithm searches for the starkest, most general form of specific contrasts of this type.

3 EVALUATION

Multiple teachers of middle-school math were asked to rate the usefulness of the provided examples on a 1-5 scale. 36 teachers responded - centered around the Pittsburgh area, but with some from remote locations. All polled teachers had experience teaching middle school math, though their overall experience ranged anywhere from 1 to over 10 years.

Teachers were asked to rate the usefulness of the following outputted data formats:

- `ALGORITHMIC (ALG)`: An example of a mistake, and an example of a correct response determined by the algorithm. The examples are chosen from the graph nodes corresponding to the identified split, and are intended to bound the specific mistake the student is making to the best degree possible.
- `SEMI-RANDOM (RAND)`: The mistake from above, along with a randomly-selected correct response. The correct response is randomly picked from all canonicalized categories, albeit weighted towards recency. The semi-random contrast is here as a weak control condition, in order to determine how much extra information the algorithmic contrast encodes.
- `SINGLETON (SING)`: Just the mistake from above. The singleton data format is analogous to the way Lumilo currently chooses which examples to show Teachers, and functions as a strong control condition.

From a Datashop repository of collected transaction data taken from Lynnette, the transaction data of 4 students were randomly selected. For these four students, a sample error message was created for each of the above data formats. For each data format-student combination, polled teachers were asked a series of 4 questions:

(1) In a short sentence or two: What do you think this student is struggling with?
(2) How useful were the excerpts above in figuring out what this student is struggling with? (On a scale from 1 to 5)
(3) In a short sentence or two: How do you think you would help this student, if you were sitting next to them right now?
(4) How useful were the excerpts above in deciding how you would help this student, if you were sitting next to them right now? (On a scale from 1 to 5)

In addition: teachers were asked how many years' experience they had teaching algebra for administrative purposes. Due to the somewhat repetitive nature of the data formats, the exact order of the questions was randomized, with emphasis placed on similar questions not appearing adjacent to each other.

4 RESULTS

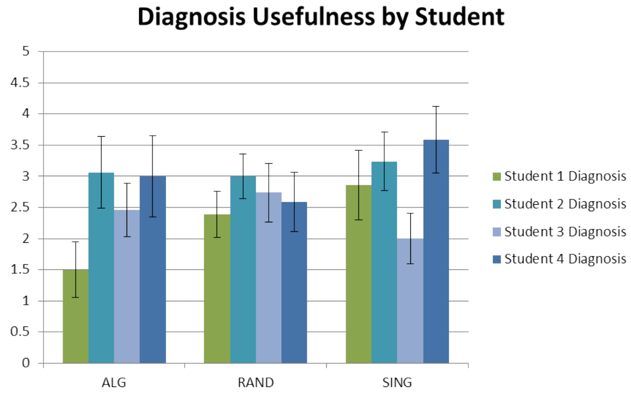


Figure 5: Usefulness for Problem Diagnosis per Output Format by Student

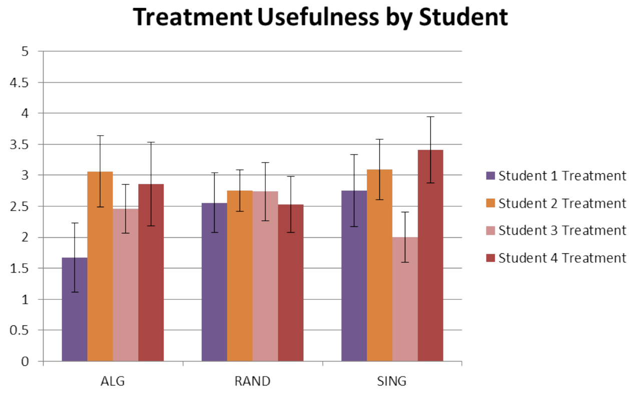


Figure 6: Usefulness for Problem Treatment per Output Format by Student

Surprises

Large variation was found between students, as shown above, so a Two-Factor ANOVA with Replication was conducted to see if there were significant interactions. The results of the ANOVA are displayed below in **Figure 8**.

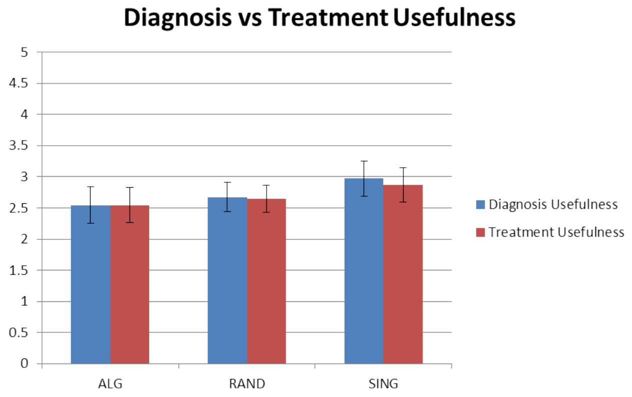
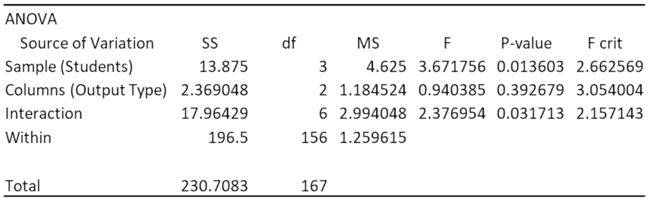


Figure 7: Average Usefulness for Diagnosis vs Average Usefulness for Treatment over Presentation Formats



ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample (Students)	13.875	3	4.625	3.671756	0.013603	2.662569
Columns (Output Type)	2.369048	2	1.184524	0.940385	0.392679	3.054004
Interaction	17.96429	6	2.994048	2.376954	0.031713	2.157143
Within	196.5	156	1.259615			
Total	230.7083	167				

Figure 8: 2-Factor ANOVA with Replication Results

Significant variation detected, centered around the students the examples were selected from being different. This implies that the differences inherent between students are by far the greatest determinator in the usefulness of provided feedback

The reasons for the interaction, and variation amongst students was likely due to both 1) the small number of students sampled from, and 2) the specific errors that were polled. The following examples were part of Students 1 and 3, the students whose results were farthest from the norm.

- Student 1: $2x = -4 \implies 2x/-4 = 4 - 2$
- Student 3: $3 = x \implies 3x = 0$

Teachers polled were, in general, very confused about what the students were doing in these cases, which likely greatly affected the relative usefulness ratings. Sample comments regarding their answer to the question "In a short sentence or two: What do you think this student is struggling with?" are as follows:

- "I can't tell. There are many errors in this one step"
- "I have never seen this error before."

Teacher comments, with a few singular exceptions, focused only on the error example provided, ignoring any provided contrasts. The presence of a contrasting example where the student solved the problem correctly was mentioned rarely, and almost exclusively by teachers with > 5 years of experience.

Lessons Learned

- It was surprising that teachers found no significant difference in usefulness between output formats. That teachers found no significant difference between the usefulness in diagnosing a student's current problems and the usefulness in figuring out how to treat a student's current problems is interesting, but not surprising. It is, in fact, encouraging, as it points to the teachers surveyed tying the specific advice they would give to struggling students with the students' observed problems, as they should.
- Regarding the differing formats, the two contrasting cases used both provide a positive, and negative example. The rationale behind using contrasts was to provide bounds on the specific contrast that was algorithmically identified as an interesting point of student struggle. The algorithmically-defined contrasts thus, theoretically encoded more information into the contrast, than just randomly selecting a positive and negative sample, but whether or not teachers could find and extract a noticeable difference was unknown.
- That no significant difference was found between algorithmically finding the specific canonicalized point of intrigue, and semi-randomly selecting the criterion can be partially explained through the actual selection criterion of the semi-random pair. The negative portion of the pair was the same as the one taken from the algorithmic contrast, for control purposes. For the positive half, the selection process was weighted towards recency. In the normal process of learning, where students make mistakes, and then learn from the mistakes, the highest points of intrigue, where they tend to make the most mistakes, tend to be the topic they have most recently tackled. As a result, more recent positive examples tend to be examples of problems similar to the chosen negative example, meaning the semi-random contrast actually may encode more information than expected.
- That no significant difference was found between teachers being shown a contrasting pair of positive and negative examples, and teachers being shown a sole single negative example was surprising. It indicates that teachers tend to focus on what students did wrong, as opposed to what students did right. A minor observation was that there seemed to be a general trend that more experienced teachers tended to reference the provided examples, both positive and negative, more in the text comments in addition to ratings that may have been higher. Due to the (relatively) small number of participants, however, there was not enough data to conclusively determine anything, an area for future work.
- A prototype version of the tool should have been rushed into testing earlier, because it would have caught the general lack of interest earlier, at which point efforts could have been diverted into determining why that is the case.

5 CONCLUSIONS

- The developed tool is able to find contrasts in student learning, and provide appropriate examples of errors to aid teachers in diagnosing the students' current problems.
- However, providing these contrasts is not more useful than single examples of mistakes, and the degree of specificity encoded into these contrasts is of approximately the same use to teachers as a semi-randomly-selected contrast.
- Student variability in both ability and problem types attempted produces much more noise than expected - enough that it can statistically account for virtually all of the variation in the usefulness of the examples selected for teachers.

6 FUTURE WORK

At this point, whether the developed tool provides an explicit improvement over the existing system is unknown. What is known is that teachers find no perceived improvement over the Lumilo system. Precisely why this is provides intriguing topics for future research.

- Are accurate contrasts actually more useful, but the teachers are somehow unable to extract the information?
  If they are, then how can the increased usefulness be communicated to teachers?
  If not, then why? Do teachers draw on other knowledge (e.g. knowledge of students, of curriculum, typical challenges / order of learning, etc.) to make the same inferences the contrasts provide?
- Are the contrasts shown too complicated?
  The so-called "tyranny of choice" is a well-documented phenomenon where having more information available to you induces choice overload instead of better informing your decisions. Is a similar effect at work here?
  If not, then do they hold less information than expected because of their complexity? Does the very specificity that was sought transform the available information into a form that is indecipherable?

ACKNOWLEDGMENTS

To Peter Schaldenbrand, for providing code that parsed raw transaction data, as well as a rudimentary tokenization system which was incorporated into the final tool.

REFERENCES

Brittney Johnson Been Kim, Elena Glassman and Julie Shah. 2015. iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction. MIT-CSAIL-TR-2015-010.

Rishabh Singh Philip J. Guo Elena L. Glassman, Jeremy Scott and Robert C. Miller. 2015. Overcode: Visualizing variation in student solutions to programming problems at scale. *ACM* 22, 2, Article 5 (March 2015), 35 pages. https://doi.org/10.1145/2699751

Bruce M. McLaren Kenneth Holstein and Vincent Aleven. 2017. Intelligent Tutors as Teachers' Aides: Exploring Teacher Needs for Real-time Analytics in Blended Classrooms. *LAK* 54, 2, Article 5 (March 2017), 50 pages. https://doi.org/10.1145/3027385.3027451

Bruce M. McLaren Kenneth Holstein and Vincent Aleven. 2018. Student Learning Benefits of a Mixed-reality Teacher Awareness Tool in AI-enhanced Classrooms. Carnegie Mellon University.

Kenneth Holstein Yanjin Long and Vincent Aleven. 2018. What Exactly Do Students Learn When They Practice Equation Solving? Refining Knowledge Components with the Additive Factors Model. *LAK* 54, 2, Article 5 (March 2018), 50 pages. https://doi.org/10.1145/3170358.3170411