What Do Argument Diagrams Tell Us About Students' Aptitude Or Experience? A Statistical Analysis In An Ill-Defined Domain*

Collin Lynch[1], Niels Pinkwart[2], Kevin Ashley[1,3] and Vincent Aleven[4]

[1]University of Pittsburgh, Intelligent Systems Program, Pittsburgh, PA, USA
[2]Clausthal University of Technology, Department of Informatics, Germany
[3]University of Pittsburgh, School of Law, Pittsburgh, PA, USA
[4]Carnegie Mellon University, HCI Institute, Pittsburgh, PA, USA

Abstract. In ill-defined domains, argumentation skills are essential in order to define problems and to present, justify, and evaluate solutions. In well-defined domains there exist accepted methods of characterizing student arguments as good or bad. This is not always possible in ill-defined domains, where competing arguments are often acceptable. In this paper, we use a set of statistical analysis methods to investigate whether, despite the lack of an "ideal solution,", student-produced argument diagrams can be diagnostic in that they can be used to reliably classify students into novices and experts or high and low aptitude. Our analysis, based on data collected during three studies with the LARGO ITS, suggests that indeed, argument graphs created by different student populations differ considerably, particularly with respect to the completeness and "connectedness" of graphs, and can thus potentially be used to adapt the system to a particular student's needs.

Keywords: ill-defined domains, assessment, feedback, graph comparison

1 Introduction

Argumentation is a fundamental mode of reasoning and analysis in many domains, such as law, ethics, public policy, and science, that typically involve ill-structured problems.

Ill-structured problems are characterized by an incompletely stated goal, relevant constraints not stated in the problem, and competing, possibly inconsistent, reasonable solutions [19]. By contrast, well-structured problems have clearly stated goals and a strong well-supported domain theory that may be used to validate solutions [19].

In addressing ill-structured problems, argumentation is essential in order to define the problem and to present, justify, and evaluate solutions. The solver must frame the problem, refining the goal and inferring constraints [5]. Solvers may frame

* This research is supported by NSF Award IIS-0412830.

the problem in different ways depending on their knowledge, values, and interests, and often there is less consensus on the "right" way to frame the problem [19]. A solution "usually is justified by verbal argument that indicates why the solution will work, and provides a rebuttal by attacking a particular constraint or barrier to the solution or by attempting to refute an anticipated opposing position." [19].

Argumentation skills are therefore essential for students to learn how to address ill-structured problems not only in domains like law or ethics. Although mathematics and science are typically taught using well-structured problems, when it comes to discovering new knowledge, the problems are ill-structured. Skilled mathematicians and scientists engage in argument schema that are similar to those used by attorneys or ethicists, such as evaluating a proposed definition or decision rule by testing how it works on hypothetical examples as described below [6,7]. Even in solving a well-structured problem, a skillful student can state an argument to justify his solution approach. One expects such arguments to be more uniform, however, because the goal and constraints are stated in the problem and there is more of a consensus about how to proceed.

Increasingly, argument diagrams are used to help students acquire argumentation skills. Argument diagrams are graphical formats that reify aspects of the arguments' structure [1]. They can be a simple tree, whose root contains the argument's conclusion, each child of which can be supported by additional nodes that represents an argument premise. [13]. Often, argument diagrams are based on Toulmin's model of argumentation in which the argument structure represents data moving through a warrant to support a claim [15]. As described in [13] and at http://www.phil.cmu.edu/projects/argument_mapping/, other schemes for diagramming arguments have been developed, some of which have been employed in teaching.

The use of argument diagramming appears to be most widespread in courses on critical thinking and in philosophy [4,16,18]. To a lesser extent, they have been applied in teaching legal and scientific reasoning [2,11,14]. Students use them to represent their own arguments or to reconstruct arguments in some source text [18] such as, in our present research, transcriptions of oral arguments before the U.S. Supreme Court.

Whatever the diagramming scheme, software is often used to assist students in constructing the diagrams [4,13,18]. Initially, the diagrams served primarily as representational tools; the important thing was the diagramming process and its impact on learning, not the details of the resulting diagrams. In a software environment, however, the diagrams can also convey diagnostic information about the student's understanding. A small but increasing number of intelligent tutoring systems (ITS) focus on argumentation skills, employ diagrams, and even evaluate student-generated diagrams in order to provide feedback [11,14].

Automatically evaluating argument diagrams for providing feedback necessarily presents a challenge, especially when problems are ill-structured. For reasons discussed above, arguments supporting correct solutions to a well-structured problem are unlikely to diverge widely. The problem constraints are clearly specified, and the solution is deterministic and may be derived by more or less algorithmic means. As a result, such arguments can take relatively few plausible paths and one would expect to see detailed similarities across the alternatives (e.g., invoking the same physics

formulas or geometry theorems). Thus, one may readily construct criteria for assessing argument quality. For example, ActiveMath [9] can check the structure of mathematical arguments (proofs) on a fairly detailed level. By contrast, with ill-structured problems, given the need to frame the problems and the divergence of plausible frames, it is far less likely that all good arguments will look alike, at least in detail of the framings and justifications. Good arguments about ill-structured problems may have similarities, indeed they probably do, but the similarities may be manifest *only* at a more abstract level in the form of more generalized patterns in the argument diagrams.

Some ITS developers have manually developed rules for identifying patterns in the diagrammed arguments that are the occasions for meaningful feedback [11,14]. These patterns range from avoiding loops in the arguments to checking whether the graph reflects the important moves in the argument text being reconstructed to identifying portions of the argument diagram that are complete enough for the student to reflect on their significance.

As databases of argument diagrams accumulate, the question arises whether they might disclose other diagnostic patterns. Given the relative novelty of computer-supported argument diagramming and the difficulty of manually extracting the patterns, pedagogically valuable patterns may not yet have been discerned and the criteria for assessing argument diagrams in terms of such patterns may not yet be known. The problem-solving tasks involved in ill-structured problems likely give rise to distinctive patterns that reflect students' understanding of the tasks. For example, given the various argument schemes for evaluating a proposed solution, such as posing a hypothetical to critically evaluate a proposed decision rule, one would expect to see patterns involving proposed rules, hypotheticals, responses, and the links among them. The existence and nature of such patterns and criteria is a matter of research.

Manually analyzing the diagrams for such patterns is tedious. As noted above, we have developed automated techniques so that an ITS can identify and analyze known patterns. In this paper, we focus on how much information may be obtained from diagrams using standard statistical measures. In the context of our research teaching law students about hypothetical reasoning, we operationalize the question as follows. When students reconstruct a real-life argument graphically, are these graphs diagnostic of (a) general aptitude for legal reasoning (as measured by the Law School Admission Test (LSAT) a preparatory exam used by law schools not unlike the GRE) and (b) experience as law students (as measured by years in law school), and are these graphs predictive of (c) the gains in argumentation skill / knowledge that result from making the graphs (as measured by performance on the LARGO study pre- and post-tests)? Similarly, we want to understand better the differences in graphs that correlate with post-test differences relating to hypothetical reasoning skills. One's belief that argument diagramming activity reflects something real about legal argument skills would be bolstered to the degree that more advanced students create graphs that are recognizably different from those created by beginning students; and likewise, though perhaps not as strongly, to the degree that high LSAT students create recognizably different graphs from those of low LSAT students. In this paper, we investigate the research questions using statistical analysis techniques in order to detect significant

differences in the sets of argument diagrams created by the participants in our studies with LARGO.

ITSs have long used methods to infer, from student-system interactions, characteristics of students' current knowledge (e.g., through knowledge tracing), and these methods have been applied successfully to predict post-test scores based on how well students were able to solve problems within the tutor [3]. The current effort is focused differently, as a consequence of the ill-definedness of our domain: the focus is on discovering and validating measures by which to assess student argument diagrams.

2 Study Context

The LARGO ITS [11] for legal argumentation supports students in the process of analyzing oral argument transcripts taken from the U.S. Supreme Court. These are complex, real-world examples of argumentation of the kind in which professors seek to engage students in class. Since U.S. Supreme Court oral arguments tend to be more complicated than classroom arguments, students probably need support in order to understand and reflect on them.

Students annotate oral arguments using a graphical markup language with Test, Hypothetical and Fact nodes and relations between them. The test and hypothetical nodes represent assertions of tests and hypotheticals in the arguments. They may be linked to relevant portions of the transcript and contain a student-authored summary of the assertion.

LARGO provides support by analyzing student diagrams for "characteristics". These are associated with phases (1=orientation, 2=transcript markup, 3=diagram creation, 4=analysis, and 5=reflection). Characteristics in phases 1-3 can be thought of as diagram "weaknesses" (i.e., areas of potential problems or errors), while characteristics in phases 4 and 5 are opportunities for reflection. The system provides feedback in the form of self-explanation prompts which (in the later phases) encourage reflection about the diagram and the argument or (in the earlier phases) inform the student about misconceptions. Failing to link a test or hypothetical node to the transcript triggers the UNLINKED_TEST or UNLINKED_HYPO characteristics, both phase 1, and advice suggesting that the student link them. TEST_REVISION_SUGGESTED, phase 5, is triggered when the student has rated other students' test formulations using collaborative filtering and his own formulation was rated poorly indicating that a change might be needed. For more information on the characteristics see [10,11].

We conducted three studies with LARGO. In the Fall of 2006 we tested it with 28 paid volunteers from the first year Legal Process course at the University of Pittsburgh School of Law. Students were randomly assigned to analyze a pair of cases using LARGO or a text-based note-taking tool without feedback. We found no overriding differences in terms of post-test scores or system interactions between the conditions. However, lower aptitude students, as measured by LSAT score, showed higher learning gains than their low-LSAT text peers. Also, the use of the help was strongly correlated with learning [11].

Since participation was voluntary, the students self-selected for their interest in the curriculum, system, and pay. A second study was necessary to further examine and substantiate the findings with non-volunteers. We developed a LARGO curriculum covering three personal jurisdiction cases integrated into one section (85 students) of the 2007 first-year Legal Process course at the University of Pittsburgh School of Law. Participation was mandatory. The students were not paid but were given coffee gift cards as a token of appreciation. The curriculum was structured as preparation for a graded writing assignment on personal jurisdiction, worth 10% of their grade. As in 2006, students were randomly assigned to LARGO and text conditions, balanced by LSAT scores. The curriculum consisted of six weekly two-hour sessions, one more than in the first study. In this study, we found no significant differences between conditions [12]. Post-hoc analysis revealed that students in the first study made far more use of the advice functions than students in the second study, which may explain the difference between the study outcomes. We are presently conducting a third (2008) study with LARGO involving experienced (third-year) law students at the University of Pittsburgh. Students in this study are assigned to mark up the same set of cases used in the fall 2007 study. At the time of this paper writing, a total of 17 third-year students have completed this study. Their data is used below.

Analysis of student's pre- and post-test scores in the three studies shows that the experienced students performed significantly higher than the novices in terms of post-test score ($t(5.03)=48.91, p < 0.001$). There was no pre-test score difference between the groups ($t(1.65)=34.00, p < 0.1$).

Figure 1 shows two example graphs created by a novice (left) and experienced student (right). These graphs describe the same oral argument transcript, but clearly differ from another. The novice student has produced a linear series of notes with few structural interrelationships (indicated by arcs) nor any link to the transcript. When the student generates a test or hypothetical node with no link to the transcript the box is labeled with a hand symbol  . When a link is made the symbol changes to a highlight  . The experienced student, by contrast, has produced a linked graph representing the relationships among the elements with commentary on the arcs. He has also linked the tests and hypotheticals to the transcript. He has similarly made use of additional And-Clause and Outcome fields in the test and hypo nodes to produce more structured representations. These differences may just be due to the fact that in the ill-defined domain of legal argumentation there are no "ideal" diagrams – thus all diagrams will likely be different. On the other hand, there may well be typical diagram aspects that are characteristic of specific student groups. We now analyze the extent to which certain differences between diagrams are diagnostic and thus allow for a prediction about the student who created them.

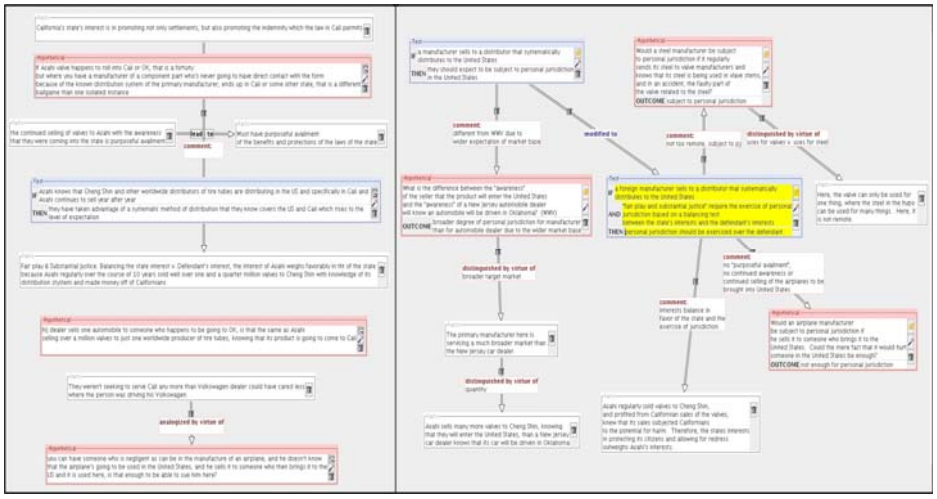

Figure1: Selected novice (left) and experienced (right) student graphs.

3 Results

Our first analysis of the diagrams was based on simple statistical measures such as the number of contained nodes, or relations, or the ratio of relations to nodes. The results can be summarized as follows:

- In none of our studies, did we observe a statistically significant correlation between the number of diagram elements and either the LSAT score, post-test score, or pre/post test gain.
- In 2007, the number of graph relations correlated positively with students' LSAT scores ($r=.32$, $p<.05$), yet the other studies do not support this (2006: $r=-.21$, $p>.4$; 2008: $r=.02$, $p>.9$).
- In 2007, the ratio of relations per nodes correlated positively with students' LSAT scores ($r=.32$, $p<.05$). A similar trend could be observed in 2006 ($r=.37$, $p<.2$), but not with the experienced students in 2008 ($r=.1$, $p>.7$).
- An ANOVA of the number of elements (nodes and relations) revealed significant differences between the three studies ($(F2,72)=9.3$,$p<.001$ for nodes, $(F2,72)=23.3$,$p<.001$ for relations). Also the ratio "relations per nodes" was different in the three studies ($F(2,72)=21.2$, $p<.001$). A post-hoc Tukey test revealed that experienced students produce significantly ($p<.05$) more relations ($m=12.3$) than the volunteer novice students ($m=7.9$), who produced significantly more than the non-voluntary novices ($m=5.2$). For the elements in graphs, there is a significant difference ($p<.05$) between both experienced students ($m=10.5$) and the volunteer novices ($m=9.6$) as compared to the non-volunteers ($m=7.5$); the first two did not differ significantly. All novices (volunteer or not) had a significantly ($p<.05$) smaller link-to-node ratio than experienced students. The averages were 1.14 for the experienced students,

and .82 and .67 for the two groups of first semester students.

This indicates that some (even very simple) measures are characteristic of specific student groups or aptitudes. However they are not sufficient as diagnostic tools that could distinguish the work of good students from that of poor students: None of the simple statistics discussed above is sufficient to accurately predict post-test scores or pre/post gains.

We therefore conducted a second set of analyses on the characteristic patterns that LARGO detects in student graphs and uses to give feedback. We analyzed the graphs created by students in all three studies for the occurrence of characteristics as detected by LARGO's graph grammar engine. A first analysis showed that the groups did not differ in terms of the **total** number of characteristics triggered by the graphs ($F(2,65)=1.2$, $p>.3$): on average, a student graph had between 14.3 (2008 study) and 18.4 (2006 study) characteristics. Thus LARGO can give as much feedback to an advanced student as to a first semester student.

We then investigated if the **types** of characteristics in the diagrams varied between the studies. A plausible heuristic here is that better (or more advanced) students will produce graphs with fewer weaknesses (i.e., characteristics of phase 1 – 3), and more opportunities for reflection (i.e., characteristics of phase 4 and 5). Figure 2 shows the relative frequencies of detected characteristics by phase. As the figure illustrates in the 2008 study, almost 50 percent of all detected characteristics were of the "phase 5" (reflection phase) type. For novice students, these frequencies are only 30 and 41 percent, respectively. This difference between the average amount of graph characteristics that indicate "opportunities for reflection" is not statistically significant, though ($F(2,65)=2.4$, $p=.09$). For none of the other phases (1-4), is the difference significant either.

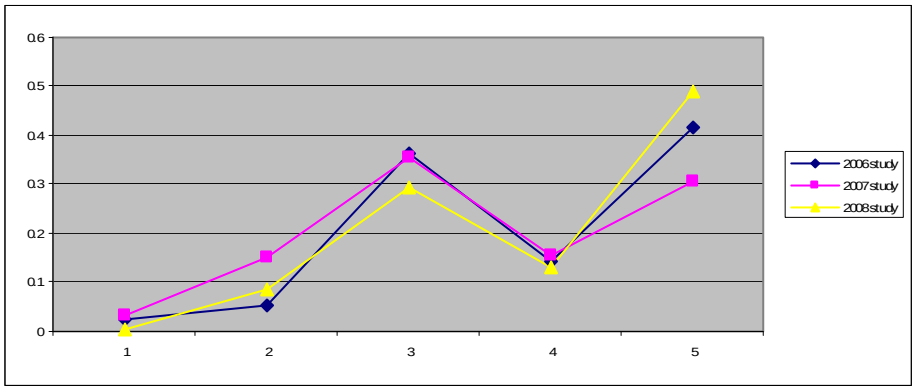

Figure 2: Relative frequencies of characteristics in student graphs by phase

Next, we analyzed whether the characteristics detected in students' graphs correlate with their pre/post gains or LSAT scores: does, for instance, a larger number of "weaknesses" mean that the student will perform poorer on the post-test? We compared the average number of phase 1-3 and phase 4+5 characteristics to the student's post-test score, the student's pre/post gain score, and the student's LSAT score. In the 2006 and 2008 studies, we observed no significant correlations between these variables. In the 2007 study, the number of "reflection characteristics" was

negatively correlated to the student's gain scores ($r=-.35$, $p<.05$) – i.e., students who have learned more during tool usage tend to produce final graphs that contain fewer opportunities for reflection.

A third analysis investigated if any of the individual diagram characteristics are by themselves diagnostic – out of the many types of characteristics that LARGO can detect, which ones are most predictive? We conducted a chi^2 analysis on the sets of characteristics (as detected by LARGO) contained in the final student-created diagrams. Due to training differences between the 2006 and 2007+8 studies we used only the 2007 novice and expert data. Our analysis demonstrated that three of the individual characteristics (the three described above) can be used to classify students into "above-median" and "below-median" in terms of their post-test score [8]. These characteristics are "UNLINKED_HYPO" ($c^2(10.40,N=51)=1.00$, $p<0.01$, precision=47/51), "UNLINKED_TEST" ($c^2(10.88,N=51)=1.00$, $p<0.001$, precision=38/51) and "TEST_REVISION_SUGGESTED" ($c^2(7.04,N=51)=1.00$, $p<0.01$, precision=35/51).

We further examined to what extent the three study populations differed. Our findings show that the following characteristics can be used to predict the group membership: NO_FACTS: ($c^2(8.61,N=51)=1.00,p < 0.01$, precision=32/51), UNLINKED_TEST: ($c^2(4.46,N=51)=1.00,p < 0.1$, precision=32/51), TEST_REVISION_SUGGESTED: ($c^2(12.40,N=51)=1.00,p < 0.001$, precision=41/51) and TEST_FACTS_RELATION_SPECIFIC ($c^2(7.44,N=51)=1.00,p < 0.01$, precision=39/51). The novice subjects exhibited more occurrences of NO_FACTS and UNLINKED_TEST than the expert subjects, while the expert subjects exhibited more instances of TEST_REVISION_SUGGESTED and TEST_FACTS_RELATION_SPECIFIC.

4 Discussion

Overall, the analysis results confirm our hypothesis. While – due to the ill-defined nature of the domain of legal argumentation – the diagrams created by students vary considerably and do permit a direct classification of "good" and "poor" diagrams, the differences between diagrams are not random. Even with rather simple statistical measures, it is possible to detect systematic differences between novice-produced and expert-produced diagrams, between diagrams of students with different aptitudes, and between those of students who use the system voluntarily or on a mandatory basis. Also, our results show that there are some aspects in diagrams that are suitable as diagnostic measures for learning gains.

Our first set of analyses, focusing on the number of elements and relations in student graphs, showed that graphs created by non-volunteers have fewer elements and relations than graphs created by volunteers. This may be caused by motivational factors: volunteers might have been more willing to use the system, resulting in more complex diagrams. A different analysis [12] provides further support for this. Also, experts create more relations and elements than novices, and produce more links per node than the novices resulting in more "connected" diagrams. "Connectedness" is important: if multiple parts of a larger argument graph are connected to each other

more frequently, this may be an indication of deeper reflection on the argument transcript that the graph represents, since the "relations" in LARGO diagrams can only be drawn reasonably if one thinks about the argument and understands the argument model. Connectedness can also be used to distinguish between novices with higher and lower LSAT scores. This further supports the hypothesis that "highly connected" graphs are an indication of more advanced or talented students.

Our second set of analyses indicates that even though the total number of diagram characteristics (as detected by LARGO) does not differ greatly between student groups, advanced students tend to produce diagrams that have more "high level" characteristics which indicate opportunities for reflection. This supports the classification of characteristics as implemented in LARGO. A surprising finding is that for first-semester students who used the system on a mandatory basis, the *fewer* opportunities for reflection in the final student-created diagrams LARGO detects, the *higher* the student learning gains. As such, a diagram with a small number of these high-level diagram characteristics would predict a high learning gain. This finding was not reproduced in the other two studies, though. A possible explanation for this effect may be that the diagrams we analyzed are the result of students' one-hour sessions with the tool. Many of the students who used the system in 2007 on a mandatory basis appeared not to be highly motivated to do so [12]. The few who did use the system intensively have received more feedback on their diagram than their peers, resulting in diagram changes (responding to feedback) that reduce the number of detected characteristics in the final diagrams. In that sense, fewer diagram characteristics in the final diagrams can be an indication of more activity with the system, which in turn leads to higher learning gains. An in depth-investigation of these relationships still remains to be done. An analysis of the log files of the student's activities with the system over the whole usage time (as opposed to analysis of students' final argument diagram, as we report in the current paper) will enable us to investigate, for example, whether relations between detected diagram characteristics and learning differed depending on the amount of help use.

The third set of analyses we conducted focused on individual characteristics (of the type that LARGO can detect using its built-in graph grammar engine). This analysis revealed that apparently, the linking behavior of students (i.e., whether they connect their argument diagram elements to the argument text or not) can be used to distinguish students by aptitude, as measured by LSAT score: better students tend to link their elements to the text more consistently. Also, the results of this analysis confirm the utility of LARGO's algorithm for peer-review of test formulations: a standing recommendation that the students change their test formulation correlates with the aptitudes of the students. This indicates that the students made use of the peer-review process but, paradoxically, that they did not always reformulate their test in response. The results of this analysis also substantiated that that the graphs created by experts are different from those created by novices, and gave a further and more specific means of distinction (in addition to the general "graph connectivity" aspect discussed above). In other words, some "key characteristics" are suitable as heuristic diagnostic tools – their presence (or absence) indicates that the diagram was created by a more (or less) advanced student. Consistent with our expectations, the characteristics typical of experts belong to "higher level" feedback messages, while

the typical "lower aptitude student" characteristics correspond to "beginner's mistakes".

5 Conclusion

No "ideal argument diagram" can be defined for the task of legal argumentation that LARGO is designed to teach. Consequently, there is a rich variety of diagrams that would be called of good (or poor) quality, but that are substantially dissimilar. Still the question, whether some properties of an argument diagram created by a student are diagnostic of that student's skills, is interesting: is the variation between argument diagrams created by different students purely random, and a result of the ill-definedness of the domain, or are there properties of diagrams that are characteristic of specific types of students?

We applied statistical analysis techniques to analyze argument diagrams created by students in three studies with LARGO. The subject populations in these studies differed in terms of their experience (beginning law students vs. advanced), aptitude (as measured by LSAT score), mode of participation in the studies (voluntary vs. mandatory), and learning gains. We found clear evidence that the graphs created by these different populations differ from each other. For some distinctions, rather simple statistical measures (such as the number of relations in a graph) are sufficient. Others require more advanced analysis methods, such as counting the different "graph characteristics" as detected by the feedback mechanism in LARGO.

These findings are an important step in our research with LARGO, and they have implications for other researchers in the field of ITSs in ill-defined domains. Even in domains where it is impossible to make sharp distinctions between "good" and "bad" solutions due to the lack of ideal solutions or a domain theory, the solution differences are meaningful. The diagrams created in LARGO contain diagnostic information about the student's understanding – i.e., whether he is likely to be of lower aptitude, or if he is more likely to be an advanced student or a beginner. This knowledge can potentially be used to adapt the system to the particular needs of a student.

An important related question is how to induce the distinguishing characteristics of good argument diagrams, that is, how to identify new diagnostic patterns. We pursue this line of inquiry in a different paper using machine learning techniques [8].

References

1. Buckingham Shum, S. 2003. The Roots of Computer Supported Argument Visualization. In Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making. Springer, London, UK.

2. Carr, C. 2003. Using Computer Supported Argument Visualization to Teach Legal Argumentation. In Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, 75–96. Springer, London, UK.

3. Corbett, A.T., Anderson, J.R. (1995) Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.

4. Harrell, M. 2007. Using Argument Diagramming Software to Teach Critical Thinking Skills. In Proceedings of the 5th International Conference on Education and Information Systems, Technologies and Applications.

5. Goldin, I., Ashley, K., & Pinkus, R. 2006. Teaching Case Analysis through Framing: Prospects for an ITS in an ill-defined domain. In Proceedings of the ITS 2006 Workshop on ITS for Ill-Defined Domains. Jhongli, Taiwan.

6. Hurley S. L. 1990. Coherence, Hypothetical Cases, and Precedent. 10 Oxford J. of Legal Studies 221, 230-234.

7. Lakatos, I. 1976. Proofs and Refutations. London: Cambridge University Press.

8. Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. 2008. Argument graph classification with Genetic Programming and C4.5. To appear in Proceedings of 1st International Conference on Educational Data Mining.

9. Melis, E., & Siekmann, J. 2004. ActiveMath: An Intelligent Tutoring System for Mathematics. In Proceedings of Seventh International Conference 'Artificial Intelligence and Soft Computing (ICAISC), 91-101. Springer.

10. Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. 2006. Using Collaborative Filtering in an Intelligent Tutoring System for Legal Argumentation. In Proceedings of Workshops held at the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. Lecture Notes in Learning and Teaching, 542-551. Dublin (Ireland), National College of Ireland.

11. Pinkwart, N., Aleven, V., Ashley, K., & Lynch, C. 2007. Evaluating Legal Argument Instruction with Graphical Representations using LARGO. In Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED2007), 101-108. Amsterdam, IOS Press.

12. Pinkwart, N., Lynch, C., Ashley, K., & Aleven, V. 2008. Re-evaluating LARGO in the Classroom: Are Diagrams Better than Text for Teaching Argumentation Skills? To appear in Proceedings of the 9th International Conference on Intelligent Tutoring Systems.

13. Reed, C., & Rowe, G. 2007. A pluralist approach to argument diagramming. Law, Probability and Risk 6, 59−85.

14. Suthers, D., Weiner, A., Connelly, J., & Paolucci, M. 1995. Belvedere: Engaging students in critical discussion of science and public policy issues. In Proceedings of the 7th World Conference on Artificial Intelligence in Education, 266 -273.

15. Toulmin, S. E. 1958. The Uses of Argument. Cambridge: Cambridge University Press.

16. Twardy, C. 2004. Argument maps improve critical thinking. Teaching Philosophy 27, 95–116.

17. van den Braak, S. W., van Oostendorp, H., Prakken, H., & Vreeswijk, G. A. W. 2006. A critical review of argument visualization tools: do users become better reasoners? In Working Notes of the 6th Workshop on Computational Models of Natural Argument (CMNA2006).

18. van Gelder, T. 2007. The Rationale for Rational™. In Law, Probability and Risk 6, 23-42.

19. Voss, J. (2006) “Toulmin's Model and the Solving of Ill-Structured Problems” in Hitchcock D. and Verheij, B. (eds.) Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation. Springer.