Using Collaborative Filtering in an Intelligent Tutoring System for Legal Argumentation

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

[1] Carnegie Mellon University, HCI Institute, 5000 Forbes Avenue,
Pittsburgh PA 15213, USA
{nielsp,aleven}@cs.cmu.edu
[2] University of Pittsburgh, School of Law, 3900 Forbes Avenue,
Pittsburgh PA 15260, USA
ashley@pitt.edu
[3] University of Pittsburgh, Intelligent Systems Program,
Pittsburgh, PA 15260, USA
collinl@cs.pitt.edu

Abstract. This paper presents a novel way of applying social navigation techniques to provide feedback to students in an intelligent tutoring system in the field of legal argumentation. Using this system, students study transcripts of US Supreme Court oral argument and annotate them by creating a graphical representation of argument flow as the Justices pose hypotheticals in order to challenge tests offered by attorneys. The proposed system is capable of detecting three types of weaknesses in arguments; when it does, it presents the student with a self-explanation prompt. This kind of feedback seems more appropriate than the "strong corrective feedback" typically offered by model-tracing or constraint-based tutors. Structural and context weaknesses in arguments are handled by graph grammars, and the critical problem of detecting and dealing with content weaknesses in student contributions (i.e., the quality of their brief statements of tests and hypotheticals found in the transcript) is addressed through a collaborative filtering approach in which students are asked to evaluate peer solutions to tasks they have done themselves. This avoids the critical problem of natural language processing in legal argumentation. Our group-oriented collaborative evaluation technique is novel in several respects. First, the atomic unit is very fine grained (i.e., small pieces of arguments hyperlinked to textual transcripts), thereby minimizing interruptions caused by reviewing peer documents while working on the same task. Second, while the system does filter student answers for quality, the tool is not primarily designed to show "good" or "matching" answers to users (as most collaborative filtering systems do). Instead, it uses the quality estimations as an input for the intelligent tutoring system, which engages learners in self explanation activities.

1 Introduction

The field of law is an established and interesting application area for AI (e.g. [1, 2]). Argument is central to the practice of law, and therefore training in the skills of argument and advocacy are essential parts of it. Despite the variety of law-related educational systems (e.g. [3]), there are still only few educational technology systems specifically designed for assisting students in the construction of legal arguments. Exceptions include the intelligent tutoring systems CATO [1] and ArguMed [4]. Partially, the small number of computer-based learning environments for legal argumentation can be explained by that fact that legal argumentation is a kind of natural language discourse that focuses on interpreting the meaning of general legal concepts in light of specific facts. The involved texts are rather unstructured and involve a wide range of (legal and world) knowledge. Thus, they are not readily accessible for an ITS without applying natural language processing (NLP) techniques. These, however, would be very error-prone in the interpretive field of legal argumentation. Current NLP technology is not able to automatically determine the meaning of specific statements in the context of the overall argument. In addition, legal argumentation is an ill-structured domain; for most tasks there is no unambiguously defined “correct” solution which could be used as a basis for an ITS. Thus, even if NLP techniques could be applied and resulted in an automatic categorization of arguments along specific dimensions of an argumentation model, this would still not adequately facilitate the assessment of student solutions.

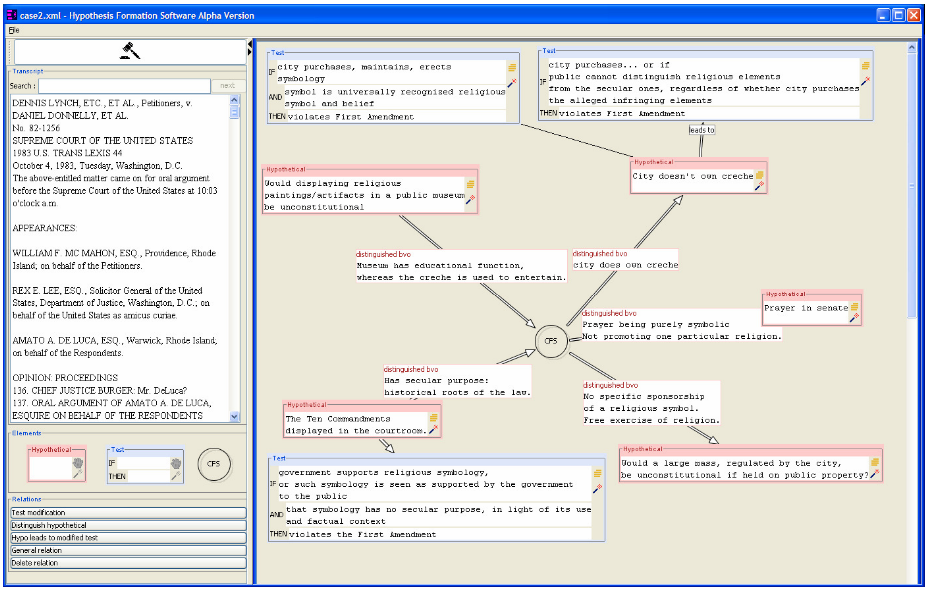
This is where social navigation principles come into play in our approach: we make use of peer students working on the same task, and let students rank peer solutions as an integrated part of their own learning activity. By active and passive evaluations, the system is able to build a heuristic measurement of the quality of a student’s answers, and is able to react to poor argument descriptions without having to parse the content of the student’s answers. Our application of the group-oriented collaborative evaluation technique is novel in several respects. First, the atomic unit of evaluation is very fine-grained (i.e., small pieces of arguments hyperlinked to textual transcripts). This minimizes interruptions caused by reviewing peer solutions while working on the same task. Second, while the system does filter student answers for quality, the tool is not primarily designed to show “good” or “matching” answers to users (as most collaborative filtering systems do). Instead, it uses the quality estimations as an input for the intelligent tutoring system. Specifically, pieces in student answers which are of low quality (as measured by the system heuristics as a result of the collaborative filtering process) are used as self explanation prompts, engaging learners to re-think the presumably weak parts of their work.

In the following sections of this paper, we first describe the underlying task of analyzing and graphically annotating transcripts of US Supreme Court oral arguments. The collaborative filtering approach, which is based on the argument graphs created by the students, is presented subsequently.

2 Annotating US Supreme Court Transcripts to Visualize Argument as Hypothesis Testing

In US Supreme Court oral arguments, contending attorneys each formulate a hypothesis about how the problem at hand should be decided with respect to a set of issues. They may propose a test and identify key points of the facts at hand on which the issue should turn. The Justices test those hypotheses by posing hypothetical scenarios. These scenarios are designed to challenge the hypotheses' consistency with past decisions and with the purposes and principles underlying the relevant legal rules. These oral arguments provide interesting material for legal educators. They are concentrated examples of many conceptual and reasoning tasks that occur in Socratic law school classrooms. As discussed in [2], the oral arguments illustrate important processes of concept formation and testing in the legal domain. As such, studying the transcripts of these arguments can be an educationally valuable task for law students. However, this task is quite difficult for beginning law students due to the complexity of the argument. As discussed above, the construction of an intelligent tutoring system based on the textual information is also difficult.

One idea to overcome these problems is to augment the textual documents with structured graphical representations that express the argument structure explicitly, thereby providing data usable by an underlying intelligent support system. The use of graphical representations for legal argumentation is not a new approach. Carr [5] has used Toulmin schemas for collaborative legal argumentation, and the Araucaria system [6] makes use of premise/conclusion visual argument structures. ArguMed [4] provides intelligent feedback through an argumentation "assistant" that analyzes structural relations between contributions in diagrams. Out of the three, only Carr conducted an empirical evaluation. Yet, he does not report on a significant learning gain caused by his system. In summary, though a lot of promising general approaches for graphically supporting argumentation exist, current literature does not show much evidence for the educational effectiveness in the domain of legal argumentation.



DENNIS LYNCH, ETC., ET AL., Petitioners, v.
DANIEL DONNELLY, ET AL.
No. 82-1256
SUPREME COURT OF THE UNITED STATES
1983 U.S. TRANS LEXIS 44
October 4, 1983, Tuesday, Washington, D.C.
The above-entitled matter came on for oral argument before the Supreme Court of the United States at 10:03 o'clock a.m.

APPEARANCES:

WILLIAM F. MC MAHON, ESQ., Providence, Rhode Island; on behalf of the Petitioners.

REX E. LEE, ESQ., Solicitor General of the United States, Department of Justice, Washington, D.C.; on behalf of the United States as amicus curiae.

AMATO A. DE LUCA, ESQ., Warwick, Rhode Island; on behalf of the Respondents.

OPINION: PROCEEDINGS
136. CHIEF JUSTICE BURGER: Mr. DeLuca?
137. ORAL ARGUMENT OF AMATO A. DE LUCA, ESQUIRE ON BEHALF OF THE RESPONDENTS

Test
IF city purchases, maintains, erects symbology
AND symbol is universally recognized religious symbol and belief
THEN violates First Amendment

Test
IF city purchases... or if public cannot distinguish religious elements from the secular ones, regardless of whether city purchases the alleged infringing elements
THEN violates First Amendment

Hypothetical
Would displaying religious paintings/artifacts in a public museum be unconstitutional

Hypothetical
City doesn't own creche

distinguished bvo
Museum has educational function, whereas the creche is used to entertain.

distinguished bvo
city does own creche

Hypothetical
Prayer in senate

distinguished bvo
Prayer being purely symbolic
Not promoting one particular religion.

distinguished bvo
Has secular purpose:
historical roots of the law.

distinguished bvo
No specific sponsorship
of a religious symbol.
Free exercise of religion.

Hypothetical
The Ten Commandments displayed in the courtroom.

Hypothetical
Would a large mass, regulated by the city, be unconstitutional if held on public property?

Test
IF government supports religious symbology, or such symbology is seen as supported by the government to the public
AND that symbology has no secular purpose, in light of its use and factual context
THEN violates the First Amendment

Fig. 1 Graphical argument model example

In contrast to the systems referred to before, we recommend a special-purpose argument representation geared toward a particular kind of argumentation process in which a normative rule (or "test") is proposed, tested, and "debugged," primarily by means of hypotheticals. Similar to the approach adapted in Araucaria, we allow the student to explicitly relate tests and hypotheticals to the transcript of the oral argument using simple markup techniques. This design enables students to substantiate their solutions using the authentic material. There is empirical evidence to believe that students indeed make use of such markup functions if the system makes it easy to do so [7].

Figure 1 shows the result of one use of the system in an exploratory study. The left side of the figure contains the transcript of the oral argument in a case called Lynch vs. Donnelly, 465 U.S. 668 (1983). At the bottom left, there is a palette with the elements (tests, hypotheticals, current fact situation) and relations (test modification, distinction of hypothetical, hypothetical leading to test change, general relation) that the user can apply to construct a graphical representation of the argumentation in the transcript. The workspace on the right side of the figure contains the argument representation. The diagram records five hypothetical cases presented by the Justices and also contains the attorney's responses to these hypotheticals, in which he distinguished them from the facts of the case or formulated new tests.

Our approach gives feedback to the students about their argument representation, including all aspects of the diagram – structure, links to the transcript, and content of the diagram elements. As argued however, rules which are guaranteed to detect errors in the student's argument graphs are virtually impossible, as there are no "ideal solutions" in the ill-structured domain of legal argumentation. As such, more heuristic methods are needed for determining *on what* to give feedback, and for *how* to give feedback in the absence of a clear-cut domain model. For the latter, our approach uses self explanation prompts as is described in more detail in [8]. In this paper, we focus on the former question: even if a precise notion of errors cannot be exactly defined, the student's conception of the argument may have *weaknesses* (in the sense of indicators for *potential* problems) that can be classified into several types. The most challenging part of this approach – the detection of weaknesses in the textual parts of the diagrams – uses peers' activities with the system to estimate quality based on a collaborative evaluation approach. In order to enable this approach, some basic argument graph pre-checks are required in order to guarantee some minimal relations between the graph and specific important parts of the transcript. These checks are briefly described in the next section.

3 Basic Argument Graph Checks

The task of annotating the transcript with argument diagrams leaves a lot of freedom to the student. This is consistent with the openness of the task and the ill-defined nature of the domain: in our pilot studies, students have created a number of appropriate and qualitatively good visual representations of arguments – however, these diagrams were far from being identical in structure much less in textual content. Our collaborative filtering approach relies on two preconditions: (1) the system must be able to determine which part of the transcript a piece of the diagram is related to, and (2) the most important parts of the transcript, containing the essential parts of the oral argument, must be represented in the diagram.

While condition 1 is guaranteed through the mechanism of a student's highlighting transcript parts and constructing hyperlinks from the diagram to these, the second one is encouraged by means of feedback messages (self explanation prompts that invite students to re-read these passages). Our system has knowledge about the central passages in the text and encourages students to re-read these parts if no diagram element refers to them. This system-side knowledge is comparable to a "lightweight expert solution", as it encodes certain properties of a good solution. This may often be possible even in ill-defined domains – e.g., if there is a central test formulation in the transcript, this should somehow be reflected in the diagram. Detecting argument features much beyond this (towards a "full expert solution" which specifies in more detail a correct solution) is not possible in our target domain, due to the variety of possible good diagrams for a specific text.

We use a graph-grammar-based engine to detect which central parts of the transcript are represented in the argument graph, and if any irrelevant text passages have been marked up. XML files of the following style are used to parameterize the graph grammar library – these files can easily be changed in order to use the tool with other transcripts that have different "central passages". In the example, the location of one important test and four hypotheticals are specified, in addition to one irrelevant part. These passages (counted in text characters) are kept relatively large in order to accommodate the fact that it is often not possible to provide a very precise definition of where, for instance, a test has been formulated. Files such as that shown in figure 2 can easily be created with our tool by simply marking up the essential passages and saving to a special format. As described in more detail in [8], we are using the graph grammar not only to check the relations between the diagram and the transcript, but also in order to detect structural problems in graphs. For instance, isolated elements trigger weakness detections (since typically statements in the oral argument are not isolated). Uncommon relations between element types, like "a test that is distinguished from the facts" (which would not make legal sense), indicate weaknesses that are then used to generate self explanation prompts.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Locations>
    <Test begin="7452" end="8262"></Test>
    <Hypo begin="9860" end="10863"></Hypo>
    <Hypo begin="14454" end="15381"></Hypo>
    <Hypo begin="16118" end="16235"></Hypo>
    <Hypo begin="22407" end="23098"></Hypo>
    <Irrelevant begin="1" end="760"></Irrelevant>
</Locations>
```

Fig. 2 XML file for specifying central parts of a transcript

4 Using Collaborative Evaluation Techniques to Estimate Content Weaknesses

The system feedback based on graph grammar weakness detections is intended to help students create good argument structures that are related to the transcript in a reasonable way. In addition, the tutoring system explicitly encourages students to consider the most important parts of the transcript and include references to them (i.e., diagram elements that paraphrase this passage) in their diagrams.

Yet, students may have difficulties in understanding, e.g., the essence of a proposed test, as evidenced by a poor paraphrase in the corresponding test node they add to the graph. Obviously, this type of weakness is harder to detect than the structural weakness outlined above, since it involves interpretation of legal argument in textual form. For instance, in figure 1, one of the test versions the students noted is

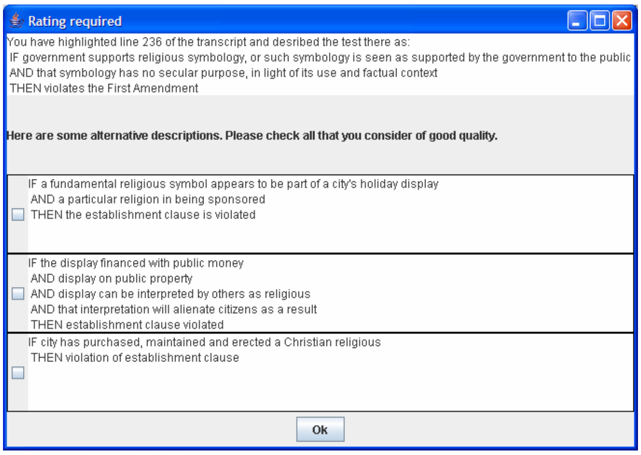
```
"IF government supports religious symbology, or such symbology is
seen as supported by the government to the public AND that symbology
has no secular purpose, in light of its use and factual context THEN
violates the First Amendment"
```

It is hard to tell for a human if this is an adequate summary of the test as formulated by the attorney during the argument or not. For a computer program it is certainly not easier. The structure offered through the graph and its links to the transcript, together with peers working on the same task either individually or in small groups (which is not an unrealistic assumption in educational scenarios) can help here, since it enables a quality heuristic for single argument components (such as the test description shown above) based on collaborative filtering [9].

In our variant of the collaborative filtering method, students are asked to rate samples of other's work. For selected important parts of the transcript (a subset of the ones the student is prompted to look at if he does not consider them in his diagram), after a student has created a corresponding element in the graph, he is presented with a small number of alternative answers (given by peers) and asked to select all those he considers *of good quality*.

Based on the evaluations a student makes, a first heuristic of the quality of the student's own answer can be calculated. One may assume that recognizing good answers is an indication of having understood the argument component, which in turn is a prerequisite for having created a good quality contribution oneself. We call this first heuristic measure the *base rating*. If a student had n answers to choose from, and the ones he evaluated positively had a quality measure q_1, …, q_k (0 for very bad, 1 for very good, see below for the calculation of quality measures for peer answers), while those he evaluated negatively had quality measures q_{k+1}, …, q_n, then the base rating b is calculated as

$$b = \frac{1}{n}\left(\sum_{i=1}^{k} q_i + \sum_{i=k+1}^{n}(1-q_i)\right)$$



Rating required

You have highlighted line 236 of the transcript and desribed the test there as:
IF government supports religious symbology, or such symbology is seen as supported by the government to the public
AND that symbology has no secular purpose, in light of its use and factual context
THEN violates the First Amendment

Here are some alternative descriptions. Please check all that you consider of good quality.

□ IF a fundamental religious symbol appears to be part of a city's holiday display
AND a particular religion in being sponsored
THEN the establishment clause is violated

□ IF the display financed with public money
AND display on public property
AND display can be interpreted by others as religious
AND that interpretation will alienate citizens as a result
THEN establishment clause violated

□ IF city has purchased, maintained and erected a Christian religious
THEN violation of establishment clause

Fig. 3 Example rating dialog

Figure 3 shows an example. Three answers are available for evaluation by the student. Let us assume quality ratings of 1, 0.8, and 0.3 (i.e., two good ones and two bad ones) for the peer answers. If now the student selects the first two, then the base rating for the student's answer is b = 0.33*(1+0.8+0.7) = 0.83.

The base rating of an answer is immediately available after the student has done his evaluations. It measures in how far a student can recognize good answers and thus serves as a heuristic of his own answer's quality, but does not rate the answer the student has actually typed in. Following the collaborative filtering idea, this can be measured by the positive and negative evaluations that a student's answer receives. We call this the *evaluation rating* e. If j students evaluated a specific answer and p students with own ratings q_1, …, q_p (on this answer) have given positive evaluations (by selecting this answer as being "of good quality") while the other j-p students gave negative evaluations, then e is calculated as:

$$e = \frac{1}{\sum_{i=1}^{j} q_i} \left(\sum_{i=1}^{p} q_i \right)$$

Here, the evaluations given by peers with higher quality ratings receive a higher weight. If we continue our example and assume that three other students had the chance to evaluate the test description, and one of them (with own rating 0.4 on this argument segment) gave a positive evaluation, whereas three others (with own ratings 0.2, 0.8 and 0.7) did not consider the descriptions as being of high quality, then the evaluation rating is e = 0.4 / (0.4 + 0.8 + 0.7 + 0.2) = 0.19. This rather low score results from the mostly negative evaluations by the peers.

Finally, an overall *quality rating* q of a student answer can be calculated as the weighted average of the base and evaluation ratings. It is reasonable to make the impact of the of the evaluation rating on the overall quality rating dependent on the amount of evaluations; if a student answer was subject to evaluation by p peers, and c is the number of answers that was presented to the student for evaluation himself, q can be calculated as:

$$q = \frac{c}{p+c}b + \frac{p}{p+c}e$$

The design of the formula ensures a normalization of q in [0,1] and takes into account the importance of peer's opinions (with large p, the base rating gets less important) while at the same time eliminating the cold start problem through the inclusion of the base rating. In our example, we have p=4 and c=3, and thus an average q = 3/7 * 0.83 + 4/7 * 0.19 = 0.46. This medium rating takes into account both the high base rating and the low evaluation rating. Since the evaluation rating is based on statements of four peer students only, it determines only 4/7 (~57%) of the overall score, the remaining 43% come from the base rating. The definition of the quality rating formula gradually fades out the impact of the base rating once a larger number of peer evaluations are available. While this approach works fine for most of the students in the group, the first and last students who work on a specific part of the transcript (and thus are the first and last to comment on it and subsequently evaluate other answers) need special attention. For the *first* students that annotate a specific passage of the text, peer answers are of course not available yet. Here, we use system provided answers of known quality (some bad, some good) in order to deal with the cold start problem. Specifically, we are using material from previous studies [10] that was graded by legal writing experts. These expert grades ensure a good initial quality heuristic in the system. This is of critical importance in our approach. If the system has a poor heuristic of the first answers initially shown to students in evaluation dialogs, this increases the number of needed evaluations in order to stabilize the quality of the overall quality heuristic. For relatively small user groups in educational settings, the time available to the system might then not be sufficient to produce good quality ratings. The answers of the *last* students that work on a specific part of the text will not be evaluated by peers. Therefore, the quality rating of their answers is equal to the base rating of their answers, which again stresses the importance of the latter (cf. next section).

If the quality measure for a student's answer is below a certain minimal threshold, this indicates a content weakness for the corresponding answer, and the system presents the student with a self explanation prompt that asks him to review and reflect upon the corresponding part of the transcript. This way, system feedback can be adaptive with respect to the student's argument graph and based on the evaluations given in the system, and natural language parsing of the diagram contents is avoided.

Our approach is similar to the reciprocal review system of SWoRD [11], but differs in three respects. First, no textual reviews are required and only quick yes/no decisions are employed within the evaluation questions. While qualitative comments might be helpful for learners in order to help them improve their answer (as done in peer review systems such as SWoRD), our approach is geared towards not distracting the learner from his main activity and includes the evaluation of peer answers as a "side activity". Another difference to SWoRD and other classical peer review systems is that that a rating has immediate implications for the system heuristic about both the *rated* text and also the *rater's* own text. For the rated text, the evaluation feeds into the evaluation rating part of the quality heuristics, and for the rater's text, the evaluation constitutes the base rating. Finally, a difference to SWoRD is that the object of rating is of finer granularity – while SWoRD uses larger samples of student writing, our approach is based on very small annotations of a specific part of a learning resource (i.e., the argument transcript). This probably helps integrating student's own learning activity with the evaluation activity, since the thematic proximity of student's own work and the statements to be evaluated is likely to be very close. Compared to other recommender systems, which essentially rely on large user group sizes, our system is designed also to work with fewer numbers (through the inclusion of the base ratings). The following self explanation prompt is an example of ITS feedback that could be generated if the quality rating of a test formulation given by a student is below a specific intervention threshold (a suitable number for such a borderline is still to be determined in pilot tests).

> "Evaluations given by your peer learners indicate that possibly your test formulation XYZ does not adequately describe the meaning of what the attorney proposed in the argument section you refer to with your diagram element. Please re-think this section of the argument and try to find a better formulation for the test."

Since we need only a rough heuristic of the quality of student's answers in order to decide whether to present such prompts or not, a less precise but quickly available approximation of quality seems the better option.

5 Conclusion and Outlook

The approach as presented in this paper is designed to support first-year law students in learning legal argumentation skills. The ITS used to generate this feedback is based on two formalisms, which enable a check of student answers for different types of weaknesses: a graph grammar formalism and a collaborative filtering technique.

The latter uses evaluations of peer solution components on a micro level (i.e., single argument elements) as a resource to build a quality heuristic of a student's answers which is based on both active evaluation acts (selecting good / poor answers) and passive evaluations (being evaluated positively or negatively by peers). This constitutes a novel use of collaborative filtering techniques and offers an alternative to the use of natural language processing techniques, which would be error-prone in the interpretive field of legal argumentation.

Based on first pilot studies we conducted, which essentially confirmed the suitability of the ontological categories and the graphical representation format, a currently ongoing second series of pilot studies tests the evaluation interface and some ITS feedback in form of self explanation prompts. Feedback based on the collaborative filtering is currently being implemented (it cannot be pilot tested in studies with single users since it requires at least small groups). We are planning to test this part of the system functionality in some further pilot studies. Further research will then try to find empirical evidence for the effectiveness of the presented tutoring approach, both compared to control groups that make use of the diagram tool without feedback, and also to groups that work traditionally with text resources. Also, we seek research results in how far the claimed correlation between the base rating and the evaluation rating (i.e., the relation between "recognizing good answers" and "providing a good answer") actually holds in practice. As emphasized in the previous section, this correlation is of particular importance for the last students in a learning group.

Acknowledgements

This research is sponsored by NSF Award IIS-0412830. The contents of the paper are solely the responsibility of the authors and do not necessarily represent the official views of the NSF.

References

1. Aleven, V. 2003. Using Background Knowledge in Case-Based Legal Reasoning: A Computational Model and an Intelligent Learning Environment. *Artificial Intelligence* 150:183-238.
2. Ashley, K. 1990. *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. Cambridge MA, MIT Press/Bradford Books.
3. Muntjewerff, J., and Breuker, J. 2001. Evaluating PROSA, a system to train solving legal cases. In Proceedings of the 10th International Conference on Artificial Intelligence in Education, 278–285. Amsterdam, IOS Press.
4. Verheij, B. 2003. Artificial argument assistants for defeasible argumentation. *Artificial Intelligence* 150:291-324.
5. Carr, C. 2003. Using Computer Supported Argument Visualization to Teach Legal Argumentation. In *Visualizing Argumentation*, 75-96. London, Springer.
6. Reed, C., and Rowe, G. 2004 Araucaria: Software for Argument Analysis, Diagramming and Representation. *International Journal of AI Tools* 14:961-980.
7. Farzan, R., and Brusilovsky, P. 2005. Social Navigation Support through Annotation-Based Group Modeling. In Proc. of UM, 387-391. Berlin, Springer.
8. Pinkwart, N., Aleven, V., Ashley, C., and Lynch, C. 2006. Toward Legal Argument Instruction with Graph Grammars and Collaborative Filtering Techniques. To appear in *Proceedings of Intelligent Tutoring Systems*.
9. Konstan, J., and Riedl, J. 2002. Collaborative Filtering: Supporting social navigation in large, crowded infospaces. In *Designing Information Spaces: The Social Navigation Approach*, 43-81. Berlin: Springer.
10. Aleven, V., Ashley, K., and Lynch, C. 2005. Helping Law Students to Understand US Supreme Court Oral Arguments: An Experiment in Progress. In Proc. of the 10th International Conference on AI and Law, 55-59. New York, ACM Press.
11. Cho, K., and Schunn, C. in press. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*.