Coping with Poor Advice from Peers in Peer-Based Intelligent Tutoring: The Case of Avoiding Bad Annotations of Learning Objects

John Champaign[1], Jie Zhang[2], and Robin Cohen[1]

[1] 200 University Ave W; Waterloo, Ontario N2L 3G1 Canada
[2] School of Computer Engineering Block N4 #02c-110
Nanyang Avenue Singapore 639798
{jchampai,rcohen}@uwaterloo.ca
zhangj@ntu.edu.sg

Abstract. In this paper, we examine a challenge that arises in the application of peer-based tutoring: coping with inappropriate advice from peers. We examine an environment where students are presented with those learning objects predicted to improve their learning (on the basis of the success of previous, like-minded students) but where peers can additionally inject annotations. To avoid presenting annotations that would detract from student learning (e.g. those found confusing by other students) we integrate trust modeling, to detect over time the reputation of the annotation (as voted by previous students) and the reputability of the annotator. We empirically demonstrate, through simulation, that even when the environment is populated with a large number of poor annotations, our algorithm for directing the learning of the students is effective, confirming the value of our proposed approach for student modeling. In addition, the research introduces a valuable integration of trust modeling into educational applications.

1 Introduction

In this paper we explore a challenge that arises when peers are involved, in the environment of intelligent tutoring systems: coping with advice that may detract from a student's learning. Our approach is situated in a scenario where the learning objects[1] presented to a student are, first of all, determined on the basis of the benefits in learning derived by similar students (involving a process of pre- and post-tests to perform assessments). In addition, however, we allow students to leave annotations of those learning objects. Our challenge then becomes to determine which annotations to present to each new student and in particular to be able to cope when there are a large number of annotations which are, in fact, best not to show, to ensure effective student learning.

[1] A learning object can be a video, chapter from a book, quiz or anything else a student could interact with on a computer and possibly learn from as described in [1].

Joseph A. Konstan et al. (Eds.): UMAP 2011, LNCS 6787, pp. 38–49, 2011.

Our work is thus situated in the user modeling application area of intelligent e-learning and, in particular, in the context of peer-based intelligent tutoring. We seek to enhance student learning as the primary focus of the user modeling that we perform. Our user modeling in fact integrates a) a modeling of the learning achieved by the students, their current level of knowledge and their similarity to other students and b) a modeling of the trustworthiness of the students, as annotators.

The decision of which annotations to ultimately show to each new student is derived, in part, on the basis of votes for and against, registered with each annotation, by previous students. In this respect our research relates as well to the general topic of recommender systems (in a style of collaborative filtering). In the Discussion section we reflect briefly on how our work compares to that specific user modeling subtopic.

We ground the presentation of our research and our results very specifically in the context of coping with possible "bad" advice from peers. And we maintain a specific focus on the setting of annotated learning objects. From here, we reflect more generally on advice for the design of peer-based intelligent tutoring systems, in comparison with other researchers in the field, emphasizing the kind of student modeling that is valuable to be performing. We also conclude with a view towards future research. Included in our final discussion is also a reflection on the trust modeling that we perform for our particular application and suggestions for future adjustments. As such, we present as well a few observations on the value of trust modeling for peer-based educational applications.

2 Overview of Model Directing Student Learning

In this section, we present an overview of our current model for reasoning about which learning objects and which annotations to present to a new student, based on a) the previous learning of similar students b) the votes for annotations offered by students with a similar rating behaviour c) a modeling of the annotation's reputation, based, in part, on a modeling of the overall reputation of the annotator. The user modeling that is involved in this model is therefore a combination of student modeling (to enable effective student learning), similarity of peers (but grounded, in part, in their educational similarity) and trust modeling of students as annotators.

Step 1: Selecting a learning object

We begin with a repository of learning objects that have previously been assembled to deliver educational value to students. From here, we attach over time the experiences of peers in order to select the appropriate learning object for each new student. This process respects what McCalla has referred to as the "ecological approach" to e-learning [1]. The learning object selected for a student is the one with the highest predicted benefit, where each learning object l's benefit to active student s is calculated as [2]:

$$p[s,l] = \kappa \sum_{j=1}^{n} w(s,j)v(j,l) \quad (1)$$

where v is the value of l to any student j previously exposed to it (which we measure by mapping onto a scale from 0 to 1 the increases or decreases in letter grade post-test assessment compared to pre-test assessment), w is the similarity between active student s and previous student j (measured by comparing current letter grade assessments of achievement levels) and κ is a normalizing factor currently set to $\frac{1}{n}$

Step 2: Allow annotations of learning objects and votes on those annotations

As students are presented with learning objects, each is allowed to optionally attach an annotation which may be shown to a new student. Once annotations are shown to students, they register a thumbs up or a thumbs down rating.

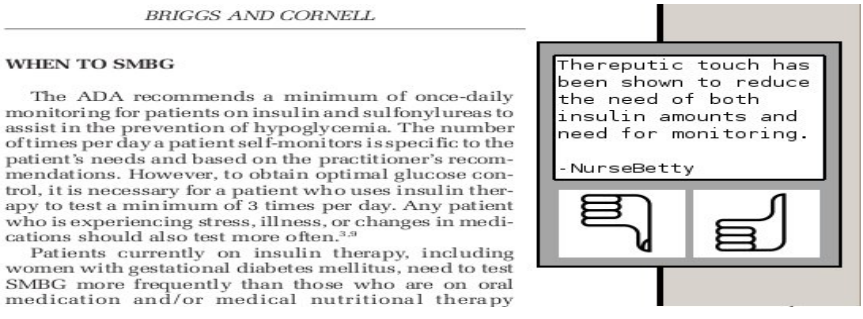

Fig. 1. Example of a low-quality annotation, adapted from [3]

The learning object presented in Figure 1 is for tutoring caregivers in the context of home healthcare (an application area in which we currently projecting our research [4]). The specific topic featured in this example is the management of insulin for patients with diabetes. This annotation recommends therapeutic touch (a holistic treatment that has been scientifically debunked, but remains popular with nurse practitioners). It would detract from learning if presented and should be shown to as few students as possible.

Consider now that: *In an ITS devoted to training homecare and hospice nurses, one section of the material discusses diet and how it is important to maintain proper nutrition, even for terminal patients who often have cravings for foods that will do them harm. One nurse, Alex, posts an annotation saying how in his experience often compassion takes higher precedence than strictly prolonging every minute of the patient's life, and provides details about how he has discussed this with the patients, their families and his supervisor.*

This annotation may receive many thumbs up ratings from caregivers who can relate to its advice. Since it is a real world example of how the material was applied, and it introduces higher reasoning beyond the standard instruction, that turns out to be a very worthwhile annotation to show to other students.

Some annotations may be effective for certain students but not for others. Consider now: *A section on techniques for use with patients recovering from eye surgery in a home healthcare environment has some specific, step-by-step techniques for tasks such as washing out the eye with disinfected water. A nurse, Riley, posts an advanced, detailed comment about the anatomy of the eye, the parts that are commonly damaged, a link to a medical textbook providing additional details and how this information is often of interest to recovering patients. The remedial students struggling with the basic materials find this annotation overwhelming and consistently give the annotation bad ratings, while advanced students find this an engaging comment that enhances the material for them and give it a good rating.*

Since our approach reasons about the similarity of students, over time, this annotation will be shown to advanced students, but not to students struggling with the material.

Some annotations might appear to be undesirable but in fact do lead to educational benefit and should therefore be shown. We present an example below. *An annotation is left in a basic science section of the material arguing against an assertion in the text about temperatures saying that in some conditions boiling water freezes faster than cooler water. This immediately prompts negative ratings and follow-up annotations denouncing the original annotator to be advocating pseudo-science. In fact, this is upheld in science (referred to as the Mpemba effect). A student adds an annotation urging others to follow a link to additional information and follow-up annotations confirm that the value of the original comment that was attached..*

While, at first glance, the original annotation appeared to be detracting, in fact it embodied and led to a deeper, more sophisticated understanding of the material. Our approach focuses on the value to learning derived from annotations and thus supports the presentation of this annotation.

Step 3: Determine which annotations to show a new student

Which annotations are shown to a student is decided in our model by a process incorporating trust modeling, inspired by the model of Zhang [5] which determines trustworthiness based on a combination of private and public knowledge (with the latter determined on the basis of peers). Our process integrates i) a restriction on the maximum number of annotations shown per learning object ii) modeling the reputation of each annotation iii) using a threshold to set how valuable any annotation must be before it is shown iv) considering the similarity of the rating behaviour of students and v) showing the annotations with the highest predicted benefit.

Let A represent the unbounded set of all annotations attached to the learning object in focus. Let r_j^a = [-1, 1] represent the jth rating that was left on

annotation a (1 for thumbs up, -1 for thumbs down and 0 when not yet rated). The matrix R has R^a representing the set of all ratings on a particular annotation, a, which also represents selecting a column from the matrix. To predict the benefit of an annotation for a student s we consider as Local information the set of ratings given by other students to the annotation. Let the similarity[2] between s and $rater$ be $S(s, rater)$. Global information contains all students' opinions about the author of the annotation. Given a set of annotations $A_q = \{a_1, a_2, ..., a_n\}$ left by an annotator (author) q we first calculate the average interest level of an annotation a_i provided by the author, given the set of ratings R^{a_i} to the a_i, as follows:

$$V^{a_i} = \frac{\sum_{j=1}^{|R^{a_i}|} r_j^{a_i}}{|R^{a_i}|} \tag{2}$$

The reputation of the annotator q is then:

$$T_q = \frac{\sum_{i=1}^{|A_q|} V^{a_i}}{|A_q|} \tag{3}$$

which is used as the Global interest level of the annotation.

A combination of Global and Local reputation leads to the predicted benefit of that annotation for the current student. To date, we have used a Cauchy CDF[3] to integrate these two elements into a value from 0 to 1 (where higher values represent higher predicted benefit) as follows:

$$\text{pred-ben}[a, current] = \frac{1}{\pi}\arctan(\frac{(vF^a - vA^a) + T_q}{\gamma}) + \frac{1}{2} \tag{4}$$

where T_q is the initial reputation of the annotation (set to be the current reputation of the annotator q, whose reputation adjusts over time, as his annotations are liked or disliked by students); vF is the number of thumbs up ratings, vA is the number of thumbs down ratings, with each vote scaled according to the similarity of the rater with the current student, according to Eq. 5. γ is a factor which, when set higher, makes the function less responsive to the vF and vA values.

$$v = v + (1 * S(current, rater)) \tag{5}$$

Annotations with the highest predicted benefit (reflecting the annotation's overall reputation) are shown (up to the maximum number of annotations to show, where each must have at least the threshold value of reputation).

[2] The function that we used to determine the similarity of two students in their rating behaviour examined annotations that both students had rated and scored the similarity based on how many ratings were the same (both thumbs up or both thumbs down). The overall similarity score ranged from -1 to 1. Other similarity measures that could be explored are raised in the Discussion section.

[3] This distribution has a number of attractive properties: a larger number of votes is given a greater weight than a smaller number (that is, 70 out of 100 votes has more impact than 7 out of 10 votes) and the probability approaches but never reaches 0 and 1 (i.e. there is always a chance an annotation may be shown).

There is real merit in exploring how best to set various parameters in order to enable students to achieve effective learning through exposure to appropriate annotations (and avoidance of annotations which may detract from their learning). In the following section, we present our experimental setting for validating the above framework, focusing on the challenge of "bad" annotations.

3 Experimental Setup

In order to verify the value of our proposed model, we design a simulation of student learning. This is achieved by modeling each student in terms of knowledge levels (their understanding of different concepts in the course of study) where each learning object has a target level of knowledge and an impact [2] that increases when the student's knowledge level is closer to the target. We construct algorithms to deliver learning objects to students in order to maximize the mean average knowledge of the entire group of students (i.e. over all students, the highest average knowledge level of each student, considering the different kinds of knowledge that arise within the domain of application).

As mentioned, one concern is to avoid annotations which may detract from student learning. As will be seen in Figure 2, in environments where many poor quality annotations may be left, if annotations are simply randomly selected, the knowledge levels achieved by students, overall, will decline. This is demonstrated in our experiments by comparing against a Greedy God approach which operates with perfect knowledge of student learning gains after an annotation is shown, to then step back to select appropriate annotations for a student. The y-axis in our graphs shows the mean, over all students, of the average knowledge level attained by a student (so, averaged over the different knowledges being modeled in the domain).

As well as generating a random set of target levels for each learning object, we also generated a random length of completion (ranging from 30 to 480 minutes) so that we are sensitive to the total minutes required for instruction. The x-axis in each graph maps how student learning adjusts, over time. We used 20 students, 100 learning objects and 20 iterations, repeating the trials and averaging the results. For these experiments we ran what is referred to as the raw ecological approach [2] for selecting the appropriate learning object for each new student; this has each student matched with the learning object best predicted to benefit her knowledge, based on the past benefits in learning achieved by students at a similar level of knowledge as in Step 1 of Section 2. Ratings left by students were simulated by having each student exposed to an annotation providing a score of -1 or 1; we simulated this on the basis of "perfect knowledge": when the annotation increased the student learning a rating of 1 was left[4].

[4] This perfect knowledge was obtained by running the simulated learning twice, once with the annotation and learning object, and once with just the learning object. A student gave a positive rating if they learned more with the annotation and a negative rating if they learned more without.

The standard set-up for all the experiments described below used a maximum of 3 for the number of annotations attached to a learning object that might be shown to a student; a threshold of 0.4 for the minimum reputability of an annotation before it will be shown; a value of 0.5 as the initial reputation of each student; and a value of 20% for the probability that a student will elect to leave an annotation on a learning object. While learning objects are created by expert educators, annotations created by peers may serve to undermine student learning and thus need to be identified and avoided.

3.1 Quality of Annotations

We performed experiments where the quality of annotations from the group of simulated students varied. For each student we randomly assigned an "authorship" characteristic which provided a probability that they would leave a good annotation (defined as an annotation whose average impact was greater than 0). A student with an authorship of 10% would leave good annotations 10% of the time and bad annotations 90% of the time, while a student with an authorship of 75% would leave good annotations $\frac{3}{4}$ of the time and bad annotations $\frac{1}{4}$ of the time. In each condition, we defined a maximum authorship for the students and authorships were randomly assigned, evenly distributed between 0.0 and the maximum authorship. Maximum authorships of 1.0 (the baseline), 0.75, and 0.25 were used. For these set of experiments, we elected to focus solely on Local information to predict the benefit of annotations, i.e. on the votes for and against the annotations presented by peers (but still adjusted according to rater similarity as in Eq. 5).

The graphs in Figure 2 indicate that our approach for selecting annotations to show to students (referred to as the Cauchy), in general does well to begin to achieve the learning gains (mean average knowledge) attained by the Greedy God algorithm. The random selection of annotations is not as compromised when there is a greater chance for students to leave good annotations (100% authorship) but degrades as a greater proportion of bad annotations are introduced (and does quite poorly when left to operate in the 25% authorship scenario). This reinforces the need for methods such as ours.

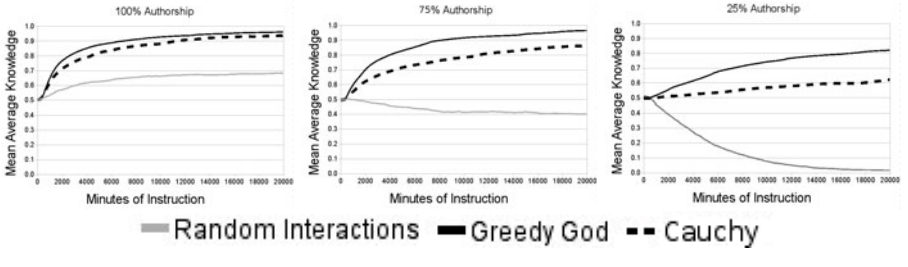

Fig. 2. Comparison of Various Distributions of Bad Annotations

3.2 Cutoff Threshold

One approach to removing annotations or annotators from the system is to define a minimum reputation level, below which the annotation is no longer shown to students (or new annotations by an annotator are no longer accepted). A trade-off exists: if the threshold is set too low, bad annotations can be shown to students, if the threshold is set too high, good annotations can be stigmatized.

In order to determine an appropriate level in the context of a simulation, we examined cut-off thresholds for annotations first of 0.2 and then of 0.4. We considered the combination of Local and Global information in the determination of which annotations should be shown (as outlined in Step 3 of Section 2). In conjunction with this, we adjusted the initial reputation of all students to be 0.7. Students were randomly assigned an authorship quality (as described in Section 3.1) evenly distributed between 0.0 and 1.0.

The results in Figure 3 indicate that our algorithm, both in the case of a 0.4 threshold and that of a 0.2 threshold (together with a generous initial reputation rating of 0.7 for annotator reputation), is still able to propose annotations that result in strong learning gains (avoiding the bad annotations that cause the random assignment to operate less favourably).

3.3 Explore vs. Exploit

Even for the worst annotators, there is a chance that they will leave an occasional good comment (which should be promoted), or improve the quality of their commentary (in which case they should have a chance to be redeemed). For this

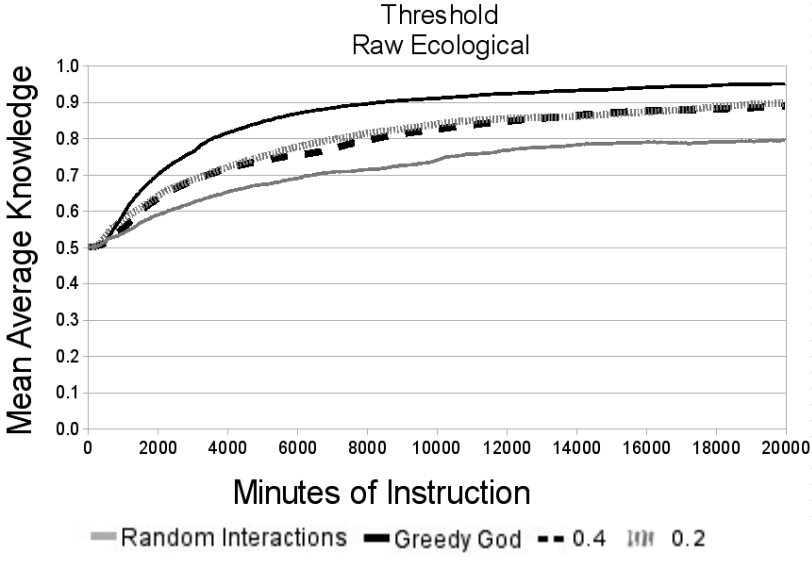

Fig. 3. Comparison of Various Thresholds for Removing Annotations

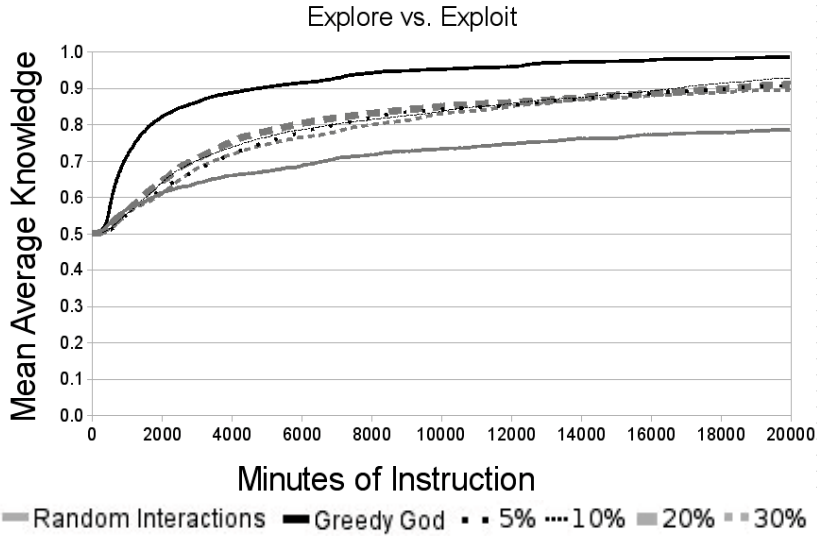

Fig. 4. Explore vs Exploit

experiment, we considered allowing an occasional, random display of annotations to the students in order to give poorly rated annotations and annotators a second chance and to enhance the exploration element of our work. We continued with the experimental setting of Section 3.2, where both Local and Global reputations of annotations were considered. We used two baselines (random and Greedy God again) and considered 4 experimental approaches. The first used our approach as outlined above, the standard authorship of 100%, a cut-off threshold of 0.4 and a 5% chance of randomly assigning annotations. The second used an exploration value of 10%, which meant that we used our approach described above 90% of the time, and 10% of the time we randomly assigned up to 3 annotations from learning objects.We also considered conditions where annotations were randomly assigned 20% and 30% of the time.

Allowing a phase of exploration to accept annotations from students who had previously been considered as poor annotators turns out to still enable effective student learning gains, in all cases. Our algorithms are able to tolerate some random selection of annotations, to allow the case where annotators who would have otherwise been cut off from consideration have their annotations shared and thus their reputation possibly increased beyond the threshold (if they offer an annotation of value), allowing future annotations from these students to also be presented.

4 Discussion

We first note that there is value of being aware of possible bad advice from peers and avoiding it – not just for our context but for peer-based intelligent tutoring

in general. In [6] the authors deal with the situation of providing incentives to encourage student participation in learning communities. They use a variable incentive model, based on classroom marks, to encourage behaviours helpful to the community of students. For example, if a student shares a small number of good resources, they will be given a greater incentive to contribute more. In the case of students who contribute a reasonable quantity of low-quality resources, the incentive to contribute is lowered, and the user is prompted with a personalized message to try to have them contribute less and to improve their quality. These incentives do not, however, eliminate scenarios where bad annotations may be left. Our work investigates this consideration. In addition, our approach does not focus on adjusting the contribution frequency of various students, but instead looks to preferentially recommend the more worthwhile contributions.

We contrast with researchers in peer-based intelligent tutoring who are more focused on assembling social networks for ongoing real-time advice [7,6], as we are reasoning about the past experiences of peers. Some suggestions for how to bring similar students together for information sharing from [8] may be valuable to explore as an extension to our current work.

Our research also serves to emphasize the potential value of trust modeling for educational applications (and not just for our particular environment of education based on the selection of learning objects that have brought benefit to similar peers, in the past). As discussed, we are motivated by the trust modeling approach of Zhang [5]. Future work would consider integrating additional variations of Zhang's original model within our overall framework. For example, we could start to flexibly adjust the weight of Local and Global reputation incorporated in the reasoning about which annotation to show to a student, using methods which learn, over time, an appropriate weighting (as in [5]) based on when sufficient Local information is available and can be valued more highly. In addition, while trust modeling would typically have each user reasoning about the reliability of each other user in providing information, we could have each student maintain a local view of each other student's skill in annotation (though this is somewhat more challenging for educational applications where a student might learn and then improve their skill over time and where students may leave good annotations at times, despite occasionally leaving poor ones as well). In general, studying the appropriate role of the Global reputation of annotations, especially in quite heterogeneous environments, presents interesting avenues for future research (since currently this value is not in fact personalized for different users).

Collaborative filtering recommender systems [9,10,11] are also relevant related work. However, intelligent tutoring systems have an additional motivation when selecting appropriate peer advice, namely to enable student learning. Thus, in contrast to positioning a user within a cluster of similar users, we would like to ideally model a continually evolving community of peers where students at a lower level are removed and more advanced students are added as a student works through the curriculum. This is another direction for future research. Some research on collaborative filtering recommender systems that may be of value for us to explore in the future includes that of Herlocker et al. [11] which explores

what not to recommend (i.e. removing irrelevant items) and that of Labeke et al. [12] which is directly applied to educational applications and suggests a kind of string-based coding of the learning achieved by students, to pattern match with similar students in order to suggest appropriate avenues for educating these new students.

Several directions for future work with the model and the simulation would also be valuable to explore. As mentioned previously, we simulated students as accurately rating (thumbs up or thumbs down) annotations based on whether the annotation had helped them learn. It would be interesting to provide for a richer student modeling where each student has a certain degree of "insight", leading to a greater or lesser ability to rate annotations. If this were incorporated, each student might then elect to be modeling the rating ability of the other peers and this can then be an influence in deciding whether a particular annotation should be shown. It might also be useful to model additional student characteristics such as learning style, educational background, affect, motivation, language, etc. The similarity calculation would need to be updated for such enhancements; similarity should then ideally be modeled as a multi-dimensional measure where an appropriate weighting of factors would need to be considered. Similarity measures such as Pearson coefficients or cosine similarity may then be appropriate to examine.

Other variations for our simulations are also being explored. Included here is the introduction of a variation of our algorithm for selecting the learning objects for each student based on simulated annealing (with a view to then continue this simulated annealing approach in the selection of annotations as well). This variation is set up so that during the first 1/2 of the trials there is an inverse chance, based on the progress of the trials, that each student would be randomly associated with a lesson; otherwise the raw ecological approach was applied. We expect this to pose greater challenges to student learning in the initial stages but to perhaps result in even greater educational gains at later stages of the simulation.

We note as well that simulations of learning are not a replacement for experiments with human students; however, the techniques explored in this work are useful for early development where trials with human students may not be feasible and future work could look to integrate human subjects as well; we are currently in discussion with possible users in the home healthcare field. While our current use of simulations is to validate our model, we may gain additional insights from the work of researchers such as [13] where simulations help to predict how humans will perform.

In conclusion, we offer an approach for coping with bad advice from peers in the context of peer-based intelligent tutoring, employing a repository of learning objects that have annotations attached by students. Our experimental results confirm that there is value to student learning when poor annotations are detected and avoided and we have demonstrated this value through a series of variations of our experimental conditions. Our general message is that there indeed is value to modeling peer trust in educational settings.

Acknowledgements. Financial support was received from NSERC's Strategic Research Networks project hSITE.

References

1. McCalla, G.: The ecological approach to the design of e-learning environments: Purpose-based capture and use of information about learners. Journal of Interactive Media in Education 7, 1–23 (2004)
2. Champaign, J., Cohen, R.: A model for content sequencing in intelligent tutoring systems based on the ecological approach and its validation through simulated students. In: Proceedings of FLAIRS-23, Daytona Beach, Florida (2010)
3. Briggs, A.L., Cornell, S.: Self-monitoring Blood Glucose (SMBG): Now and the Future. Journal of Pharmacy Practice 17(1), 29–38 (2004)
4. Plant, D.: hSITE: healthcare support through information technology enhancements. NSERC Strategic Research Network Proposal (2008)
5. Zhang, J., Cohen, R.: Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. Electronic Commerce Research and Applications 7(3), 330–340 (2008)
6. Cheng, R., Vassileva, J.: Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. User Model. User-Adapt. Interact. 16(3-4), 321–348 (2006)
7. Read, T., Barros, B., Bárcena, E., Pancorbo, J.: Coalescing individual and collaborative learning to model user linguistic competences. User Modeling and User-Adapted Interaction 16(3-4), 349–376 (2006)
8. Brooks, C.A., Panesar, R., Greer, J.E.: Awareness and collaboration in the ihelp courses content management system. In: EC-TEL, pp. 34–44 (2006)
9. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734–749 (2005)
10. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering, pp. 43–52. Morgan Kaufmann, San Francisco (1998)
11. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22(1), 5–53 (2004)
12. van Labeke, N., Poulovassilis, A., Magoulas, G.D.: Using similarity metrics for matching lifelong learners. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 142–151. Springer, Heidelberg (2008)
13. VanLehn, K., Ohlsson, S., Nason, R.: Applications of simulated students: An exploration. Journal of Artificial Intelligence in Education 5, 135–175 (1996)