When Less is More: Focused Pruning of Knowledge Bases to Improve Recognition of Student Conversation

Authors names removed for double-blind review

[1] Author identification removed

Abstract. Expert knowledge bases are effective tools for providing a domain model from which intelligent, individualized support can be offered. This is even true for noisy data such as that gathered from activities involving ill-defined domains and collaboration. We attempt to automatically detect the subject of free-text collaborative input by matching students' messages to an expert knowledge base. In particular, we describe experiments that analyze the effect of pruning a knowledge base to the nodes most relevant to current students' tasks, and observe how this change affects the algorithm's ability to correctly identify the content of student chat. We discover a tradeoff. By constraining a knowledge base to its most relevant nodes, the algorithm can detect student chat topics with more confidence, at the expense of its overall accuracy. We suggest that this trade-off be taken into consideration and manipulated to best fit the intended use of the matching scheme in an intelligent tutor.

Keywords: expert knowledge base, ill-defined domains, collaboration, inquiry learning

1 Cognitive Support for Collaborative Inquiry Systems

While great strides have been made in the categorization and improvement of Intelligent Tutoring Systems (ITS) that support work in ill-defined domains [1, 2, 3] inherent challenges exist in working within these loosely structured spaces. One crucial example is the problem of students becoming sidetracked and not focusing on appropriate content [4]. Intelligent tutors in such open-ended environments can support students by returning their focus to the domain topics relevant to their current work when the students stray. Expert knowledge bases can provide a useful domain model by encapsulating the content and semantic relationships necessary for a tutor to understand what section of the subject matter a student is currently focused, and how that portion of the domain is relevant or related to the students' current task. When this relationship is identified, the semantic relationships of the expert knowledge base can be used to guide students toward information most relevant to both their current work and the task at hand.

The first challenge when offering this type of support is to identify the current focus of student work. The introduction of collaboration among students can create an even greater chance that students will become sidetracked, but also provides novel opportunities to automatically recognize the content focus of the students. Previous studies have shown that the topic of student work in such a system can be identified within an expert knowledge base with approximately 70% accuracy [5]. From these recognitions, a system can create awareness for students and teachers about how likely the conversation is on task, and how the ongoing conversation might relate to other portions of the knowledge base. However, with only 70% accuracy on average, the results from such techniques are sometimes erroneous, and so more direct types of feedback and support are risky to attempt, as the system lacks confidence in the recognition.

With this dilemma in mind, our current research focus is to understand how accuracy and confidence can be increased within such a system to allow for more direct student intervention. This paper presents a specific attempt to understand how pruning of an expert knowledge base can affect the match confidence and overall accuracy of content recognition in student discussion in a collaborative inquiry learning system.

For the remainder of this paper, we describe the related research (Section 2) and how our current research relates to it and moves beyond it (Section 3). We present the methodology of the study (Section 4) and the results of the focused pruning of the knowledge base on the match confidence and overall accuracy, most importantly the discovery of a tradeoff between these two values (Section 5). We conclude by describing why this confidence-overall accuracy tradeoff is relevant to the designers of intelligent tutors and suggest methods that take advantage of this tradeoff dependent on the type of awareness or support one might attempt to offer (Section 6).

2 Related Work

Some previous work has focused on utilizing expert knowledge bases to detect patterns in student actions. Rahati and Kabanza describe a system that can successfully detect when student's constrained interactions are useful for learning a topic [6]. Chen and Mostow constructed a model of predictable student responses [7] within a reading tutor and are able to detect on task behavior, but are not able to offer dynamic feedback. This work is based on an assumption that in a relatively constrained system, user actions can be (to a certain extent) predicted ahead of time. There is also work in utilizing simple text matching schemes to grade student answers to open-ended questions automatically without detecting specific content subjects [8]. Most of these attempts either succeed due to the constrained nature of student interactions, or are unable to offer feedback after detection. We build from this work by suggesting that these techniques can be applied, and feedback can be offered in useful ways even in larger domain spaces where student work is less constrained. However, as stated earlier our initial efforts offered accuracy and confidence rates that could limit the use of direct feedback. In our current work, we present methods of pruning our knowledge base that can potentially improve these rates.

We can see potential for our theory, and indeed similar outcomes when considering [9]. This project took quite the opposite approach, drastically increasing the size of their knowledge base using an online resource. Using this larger knowledge base to recognize student solutions, they report an increase in recall (number of recognized solutions) along with a decrease in precision (a measure of confidence in the solutions). This mirrors the results presented in Section 5. Finally, we can also look to recent work outside of the ITS community, in the field of machine-learning classification [10] that demonstrates the power of harnessing implicit expert knowledge encoded in the dataset to make informed decisions about pruning, a similar concept to our approach described in Section 4.

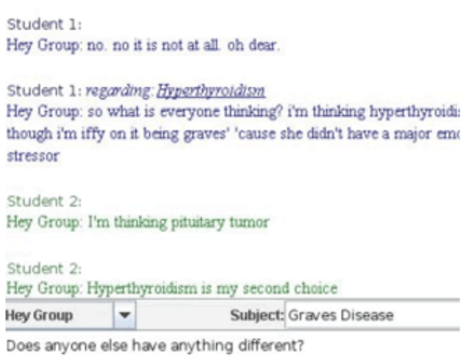
When considering such a task of recognizing textual input in order to support students' learning processes, we must also consider the offerings provided by the field of natural language understanding in this regard. Several researchers offer notable contributions in this manner [11, 12], yet they approach a different problem, and offer a different type of solution. The focus of this prior work is to mine large datasets for valuable information, placing emphasis on the sorting and filtering of unstructured data that might be of use to students. Our work takes the opposite approach, using a smaller, custom-built knowledge base to provide the set of items from which to identify helpful information for student use. In fact, the current work emphasizes this difference by taking this idea one step further, and actually limiting the size of our expert knowledge base further to increase our ability to confidently identify and support student work.

3 An Inquiry Learning System with Collaborative Features

The system on which we have gathered data for this experiment is a collaborative inquiry-learning system that provides the tools and environments necessary for students to consider authentic, real-world problems [4]. Data collection methods (question / answer interface, interactive images, etc.) provide open-ended spaces for student exploration and acquaint students with methods commonly used by professionals to access and organize information. While the system remains domain-independent, our current research relates to challenges involving human biology, where students evaluate ill patients and generate hypotheses about their medical condition.

The system provides several collaborative features to support student's efforts. These features allow students to view and monitor the work of a peer, offer critiques of specific discussable objects, and receive dynamic feedback regarding discussions that might be of interest. These collaborative features have been shown to prompt an increase in hypothesis creation, data collection, and established relationships [13]. The system also provides a chat facility that enables students to have unconstrained discussions with members of their group (Figure 1). Because the chat feature is unconstrained, it is more difficult to understand the topic of student discussion. It is in our interest to detect the content of student chat, in order to provide personalized, contextual feedback and foster collaboration. We attempt to do so by utilizing our system's expert knowledge base.

Our expert knowledge base provides both the enumeration of the individual subjects we seek to identify, as well as the semantics necessary to provide support after identification [4]. The EKB is a directed, acyclic graph of domain concepts connected with supporting and refuting relationships (Figure 2).



Student 1:
Hey Group: no. no it is not at all. oh dear.

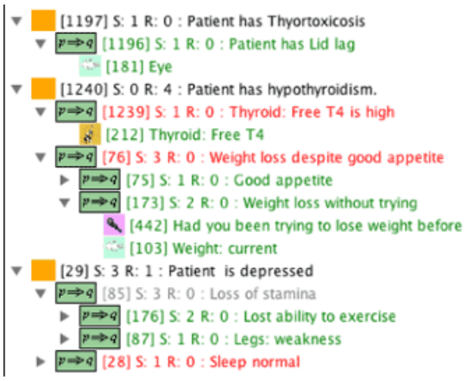
Student 1: *regarding: Hyperthyroidism*
Hey Group: so what is everyone thinking? i'm thinking hyperthyroidi
though i'm iffy on it being graves' 'cause she didn't have a major em
stressor

Student 2:
Hey Group: I'm thinking pituitary tumor

Student 2:
Hey Group: Hyperthyroidism is my second choice

Hey Group ▾ Subject: Graves Disease

Does anyone else have anything different?



▾ [1197] S: 1 R: 0 : Patient has Thyortoxicosis
  ▾ [1196] S: 1 R: 0 : Patient has Lid lag
    [181] Eye
▾ [1240] S: 0 R: 4 : Patient has hypothyroidism.
  ▾ [1239] S: 1 R: 0 : Thyroid: Free T4 is high
    [212] Thyroid: Free T4
  ▾ [76] S: 3 R: 0 : Weight loss despite good appetite
    ▸ [75] S: 1 R: 0 : Good appetite
    ▾ [173] S: 2 R: 0 : Weight loss without trying
      [442] Had you been trying to lose weight before
      [103] Weight: current
▾ [29] S: 3 R: 1 : Patient is depressed
  ▾ [85] S: 3 R: 0 : Loss of stamina
    ▸ [176] S: 2 R: 0 : Lost ability to exercise
    ▸ [87] S: 1 R: 0 : Legs: weakness
  ▸ [28] S: 1 R: 0 : Sleep normal

Figures 1 (left) and 2 (right): *Students chat with group members to discuss the patient's illness (left). These messages are matched against nodes from the knowledge base (right)*

Prior work has used this knowledge base to offer various types of support when student work is matched to individual nodes [4]. Our current effort lies in how to associate student dialogue to specific knowledge base entries, so that feedback can be generated. This is motivated by the fact that free-text chat, in which student inputs are unconstrained, presents a more difficult subject recognition problem than matching hypotheses, which are generally succinct descriptions of knowledge-base entries.

The system has an established, simple, text-matching algorithm, that matches chat message content to the knowledge base. Previous effort demonstrated an average success rate of 70 % in appropriately matching chat messages to expert knowledge base content [5], demonstrating potential. However, the confidence in any given judgment could be quite low (below 60% in some cases). This discrepancy is due to the fact that many of the inaccuracies were not missed opportunities, but rather cases of incorrect recognition. The low confidence in recognition (see Section 4) makes the prospect of offering direct feedback based on these identifications a questionable tactic. Examples from the chat message stream, along with corresponding expert knowledge base matches, are shown in Figure 3.

Actual Student Message	**Match to Knowledge Base**
Student 1: so I think it is hyperthyroidism	Patient has hyperthyroidism / thyrotoxicosis
Student 2: I concur	No Match
Student 1: web md basically said it was hyperthyroidism	Patient has hyperthyroidism / thyrotoxicosis
Student 3: hello	No Match
Student 3: how's it going?	How is the appetite?
Student 3: are you working on janet stone?	How has Janet been at home?
Student 1: janet stone	How has Janet been at home?
Student 1: because everyone is working on janet stone right?	How has Janet been at home?
Student 2: yes	No Match
Student 3: does she have a caffeine addiction?	Are you addicted to any drugs?
Student 3: Is she currently taking any medications?	Previous Medical Problems

Figure 3: *A sample dialogue and corresponding EKB matches. These students are working on the hyperthyroidism case. Student names have been removed for anonymity.*

As we can see from the conversation, the messages regarding hyperthyroidism and drugs are recognized correctly, while several statements that mention the patient's name (Janet) are recognized incorrectly as being associated with an irrelevant node. Our research focuses on providing methods for increasing the confidence with which these matches can be returned.

4 Research Design

Led by intuition, our initial investigations found that student chat tends to focus on the most relevant aspects of the case. Thus, we decided to analyze the change in our matching algorithm's efficacy after pruning the knowledge base by removing the hypotheses least relevant to the case. Thus, we first prune only hypotheses of least relevance. We defined relevance computationally using the connectedness of particular hypothesis within the knowledge graph. Thus, the relevance of a hypothesis H is simply:

$$Relevance(H) = |\, inEdges(H) \,| + |\, outEdges(H) \,|$$

The attributes of our knowledge base allow us to prune data using a more direct method. Data nodes are annotated with a Boolean value that defines whether the data is case-specific or generic knowledge. Using this value, we determined which data nodes are relevant to the case, and which are not. The algorithm was executed using the original knowledge base, and then three successive levels of knowledge base pruning. Each of these conditions was repeated over the chat messages from two distinct human biology cases. The first case was an anemia case, while the second was a hyperthyroidism case. The four levels are as follows:

- ***All:*** *Full Knowledge Base*
- ***Min Hypo Importance > 2:*** *The minimum hypothesis importance must be greater than or equal to 2 to be included in the search.*
- ***Min Hypo Importance > 5:*** *The minimum hypothesis importance must be greater than or equal to 5 to be included in the search.*
- ***Min Hypo Importance > 5 + Restricted Data:*** *Same as above, with the additional condition that only case-specific data nodes are included in the search.*

For each of these conditions, the algorithm outputs the chat message itself (as the student wrote it), and what node (if recognized) is the subject of that message. If no match is found, then the algorithm assumes the message not related to domain content (e.g. not related to human biology) and outputs a message simply saying "No Match".

A human judge examined the algorithm's output (from each condition) and placed each line of output into one of four categories according to whether the message subject was: correctly matched (+); correctly ignored / not matched (+); incorrectly matched (-); or not matched / ignored even though an appropriate match existed in the knowledge base (-). Once this was completed, we analyzed the results to determine how the algorithm was affected by the pruning of the expert knowledge base. We considered two related statistics.

Match Confidence: A measure of how likely the average match given is correct. The percentage of correct identification when a match is returned.

Confidence = Correct Matches / [Correct Matches + Incorrect Matches]

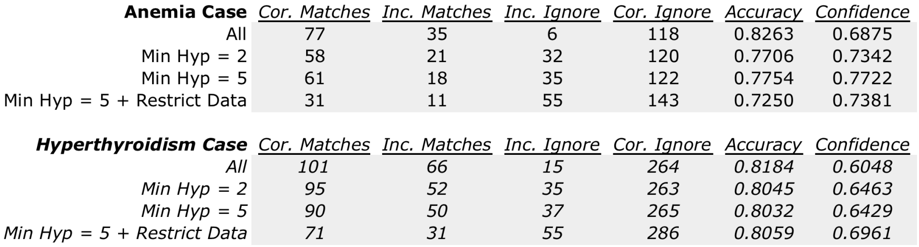
Overall Accuracy: A measure of the total efficacy of the algorithm. The percentage of total chat messages that are matched or left unmatched correctly.

Accuracy = [Correct Matches + Correct Non-Matches] / Total Messages

The data collection spanned multiple dialogues produced by students of a variety of ages (middle-school, high-school, college), within multiple environments, and after varying amounts of work time (from 45 minutes to 2 hours). Additionally, bias in the human judge's responses was eliminated by categorizing the algorithm's matches from each condition and shuffling them randomly. The human judge made judgments without knowledge of which matches belonged to which condition, after which the dialogue was resorted to reveal the results.

5 Results

Tables 1 and 2 below show the raw data results for the experiment. The data for each case are analyzed separately because the knowledge base is customized and enhanced for each case, meaning that the size and structure of the knowledge base varies, as do the resulting alterations from pruning the knowledge base.



Anemia Case	*Cor. Matches*	*Inc. Matches*	*Inc. Ignore*	*Cor. Ignore*	*Accuracy*	*Confidence*
All	77	35	6	118	0.8263	0.6875
Min Hyp = 2	58	21	32	120	0.7706	0.7342
Min Hyp = 5	61	18	35	122	0.7754	0.7722
Min Hyp = 5 + Restrict Data	31	11	55	143	0.7250	0.7381

Hyperthyroidism Case	*Cor. Matches*	*Inc. Matches*	*Inc. Ignore*	*Cor. Ignore*	*Accuracy*	*Confidence*
All	*101*	*66*	*15*	*264*	*0.8184*	*0.6048*
Min Hyp = 2	*95*	*52*	*35*	*263*	*0.8045*	*0.6463*
Min Hyp = 5	*90*	*50*	*37*	*265*	*0.8032*	*0.6429*
Min Hyp = 5 + Restrict Data	*71*	*31*	*55*	*286*	*0.8059*	*0.6961*

Tables 1 – 2: *Raw data results for the 2 cases, under the 4 conditions per case.*

Figure 4 shows these results in line graph form. We see that as we restrict the knowledge base, the number of matched chats decreases, while the number of unmatched chats increases.

The algorithm, over both cases, achieved overall accuracies between 72 and 82 percent. In addition, the confidence of returned matches ranged from 60 percent to 77 percent.

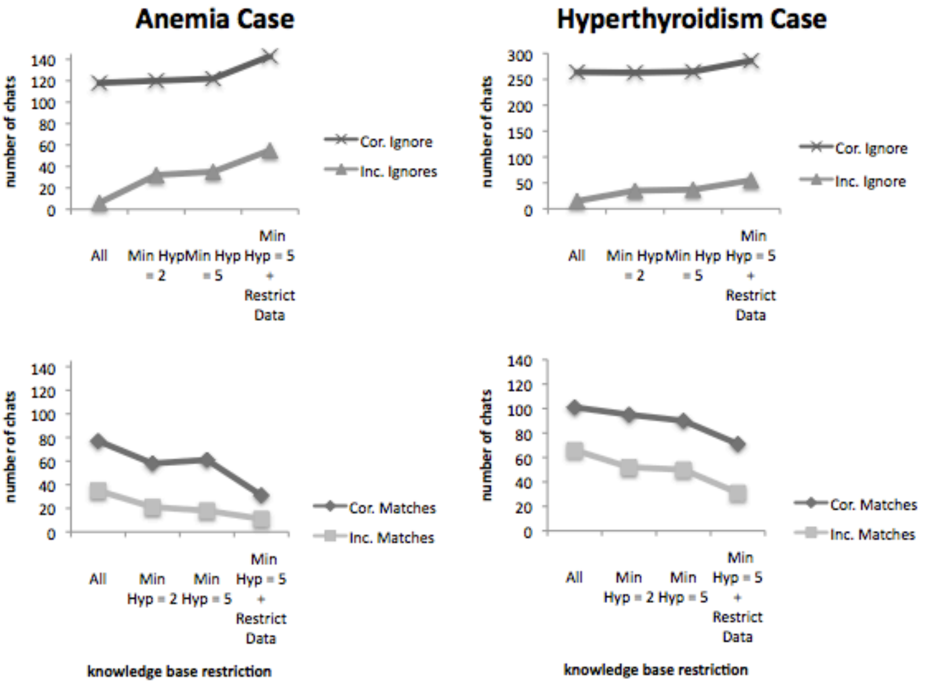
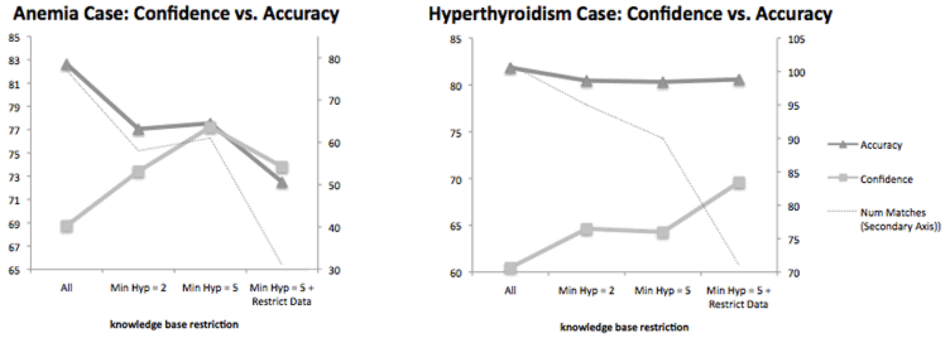


Figure 4: Raw data results for both cases

Figures 5 and 6 show the relationship between the overall accuracy of the algorithm, and the match confidence when restricting the expert knowledge base. We see that the overall accuracy tends to decrease, while the percentage of correct matches tends to increase.

In addition, Figures 5 and 6 show the effect of pruning the expert knowledge base on the number of chats recognized. As we prune the knowledge base, we cannot recognize as many total individual pieces of dialogue, but can recognize that smaller portion with higher confidence.



Figures 5 and 6: *Confidence vs. Accuracy for the each case.*

Lastly, we performed several chi-square tests to ensure that the differences between the conditions were significant. Table 3 shows the results of these tests under both of the cases we explored. We find a statistical significance between all of our conditions, with two exceptions. The changes, under both problem cases, were not significant when changing from condition two (Min Hyp = 2) to condition three (Min Hyp = 5).

Anemia Case	Min Hyp = 2	Min Hyp = 5	Min = 5 + Data
All	*1.75E-26	*1.03E-32	*4.39E-97
Min Hyp = 2		0.826	*2.48E-08
Min Hyp = 5			*4.07E-07

Hyperthyroidism Case	Min Hyp = 2	Min Hyp = 5	Min = 5 + Data
All	*1.38E-06	*3.89E-08	*2.79E-29
Min Hyp = 2		0.925	*3.66E-06
Min Hyp = 5			*7.71E-05

Table 3: *Chi-square statistics, significant changes in algorithmic behavior were found (denoted by *).*

6 Conclusions and Future Work

We find that simple keyword matching to an expert knowledge base holds serious potential for identifying the content of student conversation within noisy environments. We find that the breadth of a knowledge base has a direct effect on the quality of subject recognition. In particular, the entries of a knowledge base can be evaluated for relevance to any particular student task. If nodes are pruned to the most relevant, then subject recognition can be done with a significant increase in confidence, at the cost of the breadth of student input that can be identified.

We believe that this tradeoff is a useful observation for designers of Intelligent Tutors who utilize expert knowledge bases. By allowing students the ability to interact in loosely controlled activities such as chat discussions while engaging in learning activities, a system can offer students the ability to more freely interact while at the same time gaining additional information about student focus that would have been otherwise unavailable. Normally the freedom of interaction would come at a decrease in the system's ability to understand and provide feedback. However, by intelligently pruning and searching a knowledge base to suit the application's needs, a tutor can maximize its ability to understand and aid a student.

Figure 7 shows the nature of the observed tradeoff at work, using the same dialogue from Figure 3 with the addition of our algorithms responses using the most restricted knowledge base. In the middle column, we see that the overall accuracy of the algorithm is strong, but there are some notable mistakes in subject recognition. When the knowledge base is pruned down to its most relevant nodes for the same conversation, we see that although the number of matches found has decreased, and some messages are incorrectly not matched, the subjects the algorithm does recognize are increasingly confident.

Actual Student Message	**Full Knowledge Base**	**Most Pruned Knowledge Base**
Student 1: so I think it is hyperthyroidism	Patient has hyperthyroidism / thyrotoxicosis	Patient has hyperthyroidism / thyrotoxicosis
Student 2: I concur	No Match	No Match
Student 1: web md basically said it was hyperthyroidism	Patient has hyperthyroidism / thyrotoxicosis	Patient has hyperthyroidism / thyrotoxicosis
Student 3: hello	No Match	No Match
Student 3: how's it going?	How is your appetite?	How is your appetite?
Student 3: are you working on janet stone?	How has Janet been at home?	No Match
Student 1: janet stone	How has Janet been at home?	No Match
Student 1: because everyone is working on janet stone right?	How has Janet been at home?	No Match
Student 2: yes	No Match	No Match
Student 3: does she have a caffeine addiction?	Are you addicted to any drugs?	Do you drink a lot of caffeine in coffee or sodas?
Student 3: Is she currently taking any medications?	Previous Medical Problems	No Match

Figure 7: *An example dialogue, and the algorithm's matches for the least and most restricted expert knowledge base.*

For future work, we intend to adjust the knowledge base size used by our matching algorithm according to the needs of any particular module. For example, our system contains an intelligent coach that monitors student work and provides active feedback, presenting the student with pop-up dialog boxes that prompt the student to consider changes to their focus that appear to be to be inconsistent with expectations derived from the expert knowledge base. This coach would benefit from high confidence while analyzing student chat, and thus would utilize a highly pruned knowledge base. Other modules may use the full knowledge base to increase overall accuracy at the cost of confidence. For example, the chat interface is outfitted with a "suggested links" interface, which is located off to the side of students' main work area and offers links to different information that might be interesting according to their conversation. As this feature like this is intended to be passive, providing more general and easily-ignored feedback, the ability to offer current and varied suggestions is preferable to few, more confident suggestions. As another example, teacher tools could take advantage of this tradeoff. The area of a teacher tool offering an overview of all student work, discussion, etc. may want to offer the most information possible, even if introducing a slightly higher rate of noise. However, if a teacher requests to see specific details on an individual student for assessment reasons, then more confidence is necessary at the expense of total recognition. Once the recognition capabilities of these tools are maximized, studies can be done to measure any additional benefits (learning, collaborative activity, etc.) that result.

Acknowledgements

This research was funded by an award from the [name removed]. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Lynch, C., Ashley, K., Aleven, V., & Pinkwart, N. (2006). "*Defining Ill-Defined Domains; A literature survey*." In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.), Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems (p. 1-10). Jhongli (Taiwan), National Central University.
2. Lynch, C. Ashley, K. D. Pinkwart, N., Aleven, V. (2009) "*Concepts, Structures, and Goals: Redefining Ill-Definedness*" In Aleven, V., Lynch, C. Pinkwart, N., Ashley, K. (Eds). *International Journal of AI in Education; Special Issue on Ill-Defined Domains*. Volume 19. Number 3. pp. 253-266
3. Fournier-Viger, P., Nkambou, R. & Mephu Nguifo, E. (2010). *"Building Intelligent Tutoring Systems for Ill-Defined Domains."* In Nkambou, R., Mizoguchi, R. & Bourdeau, J. (Eds.). Advances in Intelligent Tutoring Systems, Springer, p.81-101.
4. Citation removed for double-blind review.
5. Citation removed for double-blind review.
6. Rahati, Amin; Kabanza, Froduald; "*Persuasive Dialogues in an Intelligent Tutoring System for Medical Diagnosis*" In Proceedings of the 10th International Conference on Intelligent Tutoring Systems; Pittsburgh, PA; 2010
7. Chen, Wei; Mostow, Jack; Aist, Gregory "*Exploiting Predictable Response Training to Improve Automatic Recognition of Children's Spoken Responses*" In Proceedings of the 10th International Conference on Intelligent Tutoring Systems; Pittsburgh, PA; 2010
8. Williams, Claire; D'Mello, Sidney; "*Predicting Student Knowledge Level from Domain-Independent Function and Content Words*" In Proceedings of the 10th International Conference on Intelligent Tutoring Systems; Pittsburgh, PA; 2010
9. Kazi, H., Haddawy, P., & Suebnukarn, S. (2009). *"Expanding the Space of Plausible Solutions in a Medical Tutoring System for Problem-Based Learning."* International Journal of Artificial Intelligence in Education, 19 (3), 309-334.
10. Ali Mirza Mahmood, Mrithyumjaya Rao Kuppa: *"A novel pruning approach using expert knowledge for data-specific pruning."* Eng. Comput. (Lond.) 28(1): 21-30 (2012)
11. Ravi, S., Kim, J., Shaw, E. *"Mining On-line Discussions: Assessing, Technical Quality for Student Scaffolding and Classifying Messages for Participation Profiling."* Educational Data Mining Workshop for the Conference of Artificial Intelligence in Education. 70-79. Marina del Rey, CA. USA. July (2007)
12. Bernhard, D., Gurevych, I. *"Answering learners' questions by retrieving question paraphrases from social Q&A sites"*. Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications for the Association for Computational Linguistics Columbus, Ohio: 44-52 (2008)
13. Citation removed for double-blind review.