Bayesian based adaptive question generation technique

Nabila Khodeir\textsuperscript{a,*}, Nayer Wanas\textsuperscript{a}, Nevin Darwish\textsuperscript{b}, Nadia Hegazy\textsuperscript{a}

\textsuperscript{a} Informatics Department, Electronics Research Institute, Dokki, Giza, Egypt
\textsuperscript{b} Department of Computer Engineering, Cairo University, Giza, Egypt

Available online 15 March 2014

Abstract

In this paper we aim to estimate the student knowledge model in a probabilistic domain using automatic adaptively generated assessment questions. The student answers are used to estimate the actual student model. Updating and verification of the model are conducted based on the matching between the student’s and model answers. Moreover, a comparative study between using the adaptive and random generated questions for updating the student model is investigated. Results suggest that utilizing adapted generated questions increases the approximation accuracy of the student model by 40% in addition to decreasing of the required assessing questions by 35%.

© 2014 Production and hosting by Elsevier B.V. on behalf of Electronics Research Institute (ERI).

Keywords: Intelligent Tutoring System; Student modeling; Abduction; Adaptive assessment; Item response theory

1. Introduction

Asking questions is a tool for many systems to achieve a specific goal including assessment and enhancing learners’ engagement (Graesser et al., 2005) in Intelligent Tutoring Systems (ITSs). In addition, ITSs rely on assessing of the student answers to presented questions to model his/her knowledge. Based on the student model, ITSs have the ability to personalize their support and interactions for each individual student (Brusilovsky, 2003).

Automatic generation of questions supports functionality of ITSs, in addition to dialog systems (Pwek, 2010), and Question Answering (QA) systems (Kalady, 2010). Most question generation techniques revolve around linguistic study including syntactic and semantic analysis for the given document to generate questions (Heilman and Smith, 2009; Becker, 2010). In turn, factual and definitional questions are the common types of generated questions in these approaches (Heiman, 2010; Becker, 2010). However, queries associated with some domains cannot be generated or answered based on linguistic analysis. For example, mathematics and physics tutoring systems need to define
some rules and their applying sequences in solving or generating a specific question. Probabilistic domain represents a difficult problem in this regard. The uncertainty incorporated in the knowledge representation makes generation of questions and answers a difficult task. In such domains, automatic generation of questions and their answers need to be based on a knowledge representation of the domain. In this paper we propose an approach to generate questions and their answers automatically by utilizing Bayesian Network (BN) knowledge representation (Korb and Nicholson, 2011) for probabilistic domains. The purpose of the generated questions is to model the student knowledge within ITS.

Probabilistic domains are domains that consider uncertainty in defining relations between their items. For example, medical domains need to define the probability of association of a symptom to a specific disease. The modeling using BN of such domains is based on defining a set of nodes and a set of directed arcs or links. The nodes in the BN represent a set of random variables $X = x_1, \ldots, x_j, \ldots, x_n$ from the domain. The set of links connects pairs of nodes, $x_i \rightarrow x_j$ representing direct dependencies between variables. The strength of these relations is defined using the Conditional Probability Distribution (CPD) associated with each node. The CPD lists the probability that the child node takes on each of its different values for each combination of values of its parents (Korb and Nicholson, 2011). On the other hand, reasoning in BNs is a process of inferring new information conditioned by observing values of some variables. The process of inference is performed via a flow of information through the network to compute the posterior probability distribution for a set of query nodes given values for some evidence (or observation) nodes.

Probabilistic domains are usually associated by diagnostic questions which require identifying the most probable explanation given a set of evidences. We consider such questions especially in relation to ambiguous cases, where more than one hypothesis that explains the question evidences exist. In such cases the student is asked to provide a ranked list of possible hypotheses for the question evidences. Diagnostic questions for ambiguous cases are chosen since answers for such questions reveal more information about the student knowledge. Consequently fewer questions will be sufficient for the student knowledge modeling.

The generated question takes the following form “If you have a case with evidence_1, evidence_2, evidence_n. Choose and rank from the following hypothesis: Hypothesis_1, Hypothesis_2, Hypothesis_3, \ldots, Hypothesis_n.”

The answer form is a ranked list of likely diagnosis hypotheses. Hypotheses are chosen from the available choices that are associated with the question. The student is asked to type the rank value corresponding to the chosen hypothesis. Fig. 1 is a snapshot of the question and answer form.

The rest of this paper is organized as follows. Section 2 presents the proposed question and answer generation techniques. Thereafter, we explore the experimental results that illustrate achieving of generated questions to their purpose of modeling the student knowledge in Section 3. Section 4 presents a discussion and conclusion of the paper.
2. Question and answer generation

The target of question generation process is to model the student knowledge and misconceptions. The question generation process utilizes the student model which is initialized by the domain knowledge. Domain knowledge represents number of hypotheses that are selected based on their overlapping evidences to allow creating questions with ambiguous answer (answer contains more than one possible hypothesis). Each time the student provides answer to a generated question, the model is updated according to his/her answer. The updating process is illustrated by Khodeir et al. (2010). Using the updated student model in generating consequent questions grantees tracking of the student misconceptions that accordingly appear in the updated student model. Moreover, the question generation process accommodates the student knowledge by generating questions with different difficulty levels and evidences scope. The question generation process is based on six components, as shown Fig. 2, namely the (i) student knowledge estimator, (ii) evidences selector, (iii) question evaluation module, (iv) question difficulty level estimator, and (v) question selector.

Student knowledge estimator estimates the student knowledge using all responses provided by the student. Due to its simplicity, the one parameters model (Lord, 1980) is selected from item response theory (IRT) models to estimate the student knowledge (ability) which takes the form shown in Eq. (1).

\[
P(\text{correct}|\theta) = \frac{1}{1 + \exp^{-(\theta - \delta)}}
\]  

(1)

Where \(\theta\) is the student knowledge and \(\delta\) is the question difficulty level. Based on the IRT, maximum likelihood procedure (Lord, 1980) is used to estimate the student’s knowledge \(\hat{\theta}\). The maximum likelihood procedure is an iterative process, which begins with some a priori value for the knowledge of the student and the known values of the questions or item parameters (the parameters of the question which is generated based on the initial student knowledge). These are used to compute the probability of correct response to each item for that student. Then an adjustment to the knowledge estimate is obtained that improves the agreement of the computed probabilities with the students’ items responses.
vector. The process is repeated until the change in the estimated knowledge is negligible. Eq. (2) is used to iteratively estimate the student knowledge $\hat{\theta}_{S+1}$.

$$
\hat{\theta}_{S+1} = \hat{\theta}_S + \frac{\sum_{i=1}^{N} [u_i - P_i(\hat{\theta}_S)]}{\sum_{i=1}^{N} P_i(\hat{\theta}_S)(1 - P_i(\hat{\theta}_S))} 
$$

(2)

where $\hat{\theta}_{S+1}$ is the estimated knowledge of the student in iteration $S$, $u_i$ is the binary value representing the response provided by the student to item $i$, where $u_i = 1$ for a correct response and $u_i = 0$ for an incorrect response. $P_i(\hat{\theta}_S)$ is the probability of correct response to item $i$ at knowledge level $\hat{\theta}$ within iteration $S$.

Evidences selector selects the number of question evidences which is proportional to the estimated student knowledge. This stems from the fact that when the number of evidences increases the question will require more analysis for the relations between these evidences and the connected hypotheses. Moreover, more inference is required to compare between the possible hypotheses that explain the question evidences and rank them according to their explanation power. It is worth noting that the number of evidences is selected within a range depending on the domain BN. The range is selected to avoid both the generation of a very ambiguous question, which may be difficult to answer due to large number of possible explanation hypotheses, as well as questions that give one explanation hypothesis as an answer that will not reveal sufficient information about the student knowledge.

Question evaluation module generates the student model answer to the presented diagnostic question based on the student model. That is achieved by applying an inference mechanism on the student model BN. To match the type of questions generated the abduction inference mechanism suggested by Nilsson (1998) is used.

This mechanism is an efficient algorithm for finding the most probable hypotheses in probabilistic expert systems that explain a given evidences set. The evidence set represents the values of variables that are presented in the diagnostic questions. The probable hypotheses generated are associated by their probabilities that express their explanation power for the given evidences set and are used in ranking. That is to say, the verification module generates the student model answer in the form of a ranked list of hypotheses that explains the evidences set mentioned in the question by utilizing of the student model BN.

Question difficulty level estimator is used to characterize each question by its difficulty level. According to the IRT, the difficulty level is estimated based on item characteristic curve which is fitted to the probability of correct response $P_i(\hat{\theta}_S)$ at each knowledge level. This means that each question or item has to be answered by different students with different abilities to obtain the item characteristic curve. The question difficulty level is estimated by identifying the knowledge value that will make certain that at least half the students presented by the question deliver a correct response.

The idea of estimation of the question difficulty level from previous students’ responses which needs a significant amount of data, in this paper we suggest a method to calculate the question parameters by utilizing the structure of the student knowledge model and the hypotheses in the predicted student answer. The question difficulty level is estimated through calculating the number of relations between the question evidences and the possible hypotheses in the predicted student answer. These relations are considered the question coverage relations $Q_{CR}$. This value is normalized by dividing it on the total number of relations between the question evidences and all possible hypotheses which are considered the question evidence relations $Q_{ER}$. In addition, the inverse of the standard deviation $\sigma_i$ between the probabilities of different hypotheses in the answer is added to the question difficulty. This follows from the fact that if the difference between the different hypotheses rank probability is small then the question will be more difficult. The estimation of the difficulty level of the question, $\delta_i$ is given by Eq. (3).

$$
\delta_i = \frac{Q_{CR}}{Q_{ER}} + \frac{1}{\sigma_i}
$$

(3)

This expression is assumed to represent the question difficulty level because when the number of question coverage relations is large, the student needs to know and utilize more relations. It is worth mentioning that the question difficulty level value is normalized to be within the range of the values acceptable in IRT.

Question selector uses the Maximum Information (Weiss, 1982) to select one question gives a difficulty level and an estimate of covering of the student model. This entails selecting the item that maximizes the item information for
the previous student knowledge level estimated until that moment, where the Information Function $I_i(\hat{\theta})$ for the \textit{i}th item given the current estimation of the student knowledge level $\hat{\theta}$ is given by Eq. (4).

$$I_i(\hat{\theta}) = \frac{P_i(\hat{\theta})^2}{P_i(\hat{\theta}) - [1 - P_i(\hat{\theta})]}$$  

(4)

Finally, the selected question takes the final form using a template in addition to the selected list of evidences.

\section*{2.1. Question generation process}

The question generation process aims to generate a question that accommodates the student knowledge in addition to tracking of the student misconceptions. The generation process proceeds as follows

1. According to the student knowledge the evidence selector identifies the suitable number of evidences. Moreover, the evidence selector considers the updated student knowledge model to select the evidences that have more than one common hypothesis. This selection guarantees that the generated question has more than one explanation.

2. The question evaluation module verifies the generated question by applying the abduction algorithm using the question evidence and the student BN. The output of applying the algorithm represents the student model answer to the presented question which is the predicted student answer. The answer is re-checked to ensure that it contains more than one explanation. If the number of explanations is greater than one, the question generation process proceeds, otherwise step 1 is repeated while excluding the previously selected evidence.

3. The question difficulty level is estimated using the predicted student response in addition to the student knowledge. That is possible by determining the relations which are covered by the generated question, the relations between the question evidence and the predicted student answer hypotheses, and the inverse of the standard deviation between hypotheses.

4. The question generation process is repeated to generate all possible questions for the available uncovered relations. In addition, the difficulty levels of all generated questions are estimated.

5. The question selector selects the most informative question among all generated questions. Then final question form is produced for the selected question and presented to the student.

\section*{3. Evaluation}

In this paper we presented an evaluation of using the questions generation algorithm in building of the student model. To evaluate the modeling performance, we need to run the experiment on students that we know prior information about their knowledge. This allows comparison between the resulted student knowledge model and the actual student knowledge. Therefore we use a simulated student approach (Millan and Perez-De-La-Cruz, 2002) in the evaluation process. The proposed mechanism to generate the simulated students is based on an existing domain BN, simulated students are randomly modified BNs that represent the students knowledge. The simulated student response is assessed by processing the generated BN to produce the target student answer to the posted question. It is worth mentioning that 50 simulated students are randomly generated and used in the evaluation process.

To explore the potential of the proposed adaptive question generation technique in increasing the efficiency of the modeling algorithm, two main experiments are conducted (i) non-adaptive experiment based on randomly generated questions in evaluation of the modeling algorithm, and (ii) adapted experiment which relies on student knowledge adapted generated questions. The comparison between the different approaches is based on: (i) the verification performance, (ii) the output differences between the resulted student model BN and the simulated student BN ($BN_D$), and (iii) the number of questions required ($N_q$). The first two factors illustrate the impact of using the adapted questions on the accuracy of the modeling algorithm, while the latter indicates the performance of using adapted questions. It is worth mentioning that we estimate the differences between two BNs by comparing the summation of the relations weights in each BN. In addition, different granularity levels are evaluated by changing the value of the updating step and the output of this parametric study is indicated.

\textbf{Table 1} indicates comparison between difference in student model and the simulated student ($BN_D$), and the number of questions required ($N_q$) using modeling and adapted modeling techniques. \textbf{Fig. 3} shows the comparison between
Table 1  
Comparison between difference in student model and the simulated student (BN_D), and the number of questions required (N_q) using modeling and adapted modeling techniques.

<table>
<thead>
<tr>
<th>Step</th>
<th>Modeling (BN_D)</th>
<th>Adaptive modeling (N_q)</th>
<th>Adaptive modeling (BN_D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>9.50 ± 2.60</td>
<td>13.12 ± 1.60</td>
<td>5.73 ± 1.27</td>
</tr>
<tr>
<td>0.1</td>
<td>9.53 ± 2.30</td>
<td>13.36 ± 1.41</td>
<td>6.23 ± 1.52</td>
</tr>
<tr>
<td>0.2</td>
<td>10.35 ± 2.71</td>
<td>13.26 ± 1.63</td>
<td>7.18 ± 1.55</td>
</tr>
<tr>
<td>0.3</td>
<td>10.85 ± 2.87</td>
<td>13.58 ± 1.49</td>
<td>8.58 ± 2.07</td>
</tr>
<tr>
<td>0.4</td>
<td>11.57 ± 2.91</td>
<td>13.40 ± 1.68</td>
<td>9.88 ± 2.57</td>
</tr>
<tr>
<td>0.5</td>
<td>12.67 ± 3.60</td>
<td>13.12 ± 1.86</td>
<td>9.54 ± 3.04</td>
</tr>
<tr>
<td>0.6</td>
<td>11.98 ± 3.17</td>
<td>13.16 ± 1.74</td>
<td>9.48 ± 2.99</td>
</tr>
<tr>
<td>0.7</td>
<td>12.35 ± 2.95</td>
<td>13.34 ± 1.72</td>
<td>9.97 ± 3.21</td>
</tr>
<tr>
<td>0.8</td>
<td>12.36 ± 2.85</td>
<td>13.68 ± 1.52</td>
<td>9.64 ± 3.01</td>
</tr>
<tr>
<td>0.9</td>
<td>13.41 ± 3.31</td>
<td>12.18 ± 2.88</td>
<td>9.82 ± 3.13</td>
</tr>
<tr>
<td>1.0</td>
<td>13.23 ± 3.30</td>
<td>12.32 ± 2.93</td>
<td>9.78 ± 3.12</td>
</tr>
</tbody>
</table>

Fig. 3. Comparison between verification performance of modeling and adaptive modeling techniques.

verification of modeling and adaptive modeling techniques. The modeling and updating process are illustrated in Khodeir et al. (2010).

As shown in figure there is a significant difference distinguishing the adaptive modeling technique over the non-adaptive modeling technique especially in lower step values. In addition, Table 1 indicates a comparison between the two techniques based on the number of questions (N_q) and in the differences between the generated student model BN and the simulated student BN (BN_D). The table illustrates considerable differences that prefer the adaptive technique over the non-adaptive corresponding technique. The difference between the two BNs decreased by 40% using the adaptive technique in addition to decreasing the number of questions by a factor up to 35%.

4. Discussion and conclusion

In this paper, we presented an algorithm to generate questions that tailor the student knowledge and misconceptions. Student knowledge is considered while controlling the difficulty level of the generated questions. Student misconceptions are taken into account by relying on the updated student model in the question generation process. The student answers to student knowledge adapted generated questions are used to estimate the actual student model. Updating and verification of the model are conducted based on the matching between the student’s and model answers.
Two different approaches to updating are used, namely coarse, and refined. Moreover, comparison between using the adapted questions and random questions is investigated. The results proved that, all updating techniques achieve a considerable improved performance, even when we use random questions. Results suggest that utilizing adapted generated questions increases the approximation accuracy of the student model by 40% in addition to decreasing the number required assessing questions by 35%.

References


