Semiautomatic Annotation of Test Materials in an ITS Authoring System

Sunandan Chakraborty[a], Devshri Roy[a], Anupam Basu[a]
[a]*Department of Computer Science & Engineering,*
Indian Institute of Technology, Khargapur, India
{sunandanchakraborty,droy.iit,anupambas}@gmail.com

Abstract: Reducing effort and time consumption has always been the main focus in Intelligent Tutoring System Authoring Systems (ITSAS). In this paper, we propose a methodology to semi-automatically annotate test items (questions) using test item analysis in an ITSAS. We also discuss how this mechanism can reduce the overhead of the teachers and make ITS authoring simpler and less time consuming. It is also explained how question annotation through test item analysis can improve the efficacy of the ITS. Finally, we demonstrate an instance of a test item analysis to describe the annotation of test items and also the benefits received by the authors. It is also shown that semi-automatic annotation can reduce the test material authoring time considerably, through an experiment.

Keywords: Intelligent Tutoring System Authoring System, Test Item analysis

Introduction

In order to add flexibility and versatility in Intelligent Tutoring Systems (ITS) a separate paradigm of systems evolved. They are called Intelligent Tutoring Systems Authoring Systems (ITSAS). Using these authoring systems teachers can configure or modify different features of the ITSs and author different tutors in various domains and for various purposes [1].

A major objective of an ITSAS is to provide a platform for the teachers to create computer-based intelligent tutors using their limited knowledge of computer handling. Thus, an important focus of ITSAS designers is to plan the system in such a way that the effort, time and computer-skill requirement on behalf of the teachers are minimized [1]**.** Simplifying every aspect of ITS authoring can reduce the overall difficulty of the authoring system. One of the most important aspect of an ITSAS is test material authoring or creating questions in an ITS. Simplifying this procedure can greatly contribute in minimizing the effort and time requirement of teachers for developing test materials and apparently help in achieving the most important design goal of an ITSAS.

We built an ITSAS called *Shikshak*, where there are different modules, such as domain model, student model, pedagogical model and facilities for authoring each module [9]. Authoring domain model involves authoring and organizing study and test materials. For customized retrieval both the study and test materials are annotated using informative tags [8][10]. The features annotated with the test materials help in assessment of the student's knowledge state. A detailed analysis on student's test result can detect the flaws in her performance and scopes of improvement in the subsequent lessons. Manual categorization [2] is a plausible method but it may lead to inconsistencies. To avoid such inconsistencies there is a need of automatic categorization of questions.

In this paper, we propose a scheme of semi-automatic annotation of the test materials through test item analysis. Test item analysis would also provide various reports on the test items. This would help the teachers to modify the question base if required and rectify any fallacies made during authoring previously. Hence, we demonstrate how test item analysis can reduce time, effort and computer-skill requirement of the teachers and make ITS authoring in *Shikshak* simpler and effective.

1. Test Item Metadata

The system has a repository, which contains study and test materials. The test material repository currently accommodates multiple-choice type questions. In the rest of the paper, the term question or test item will refer to multiple choice type questions with four alternatives. Each such question is annotated using a set of metadata, which provides a comprehensive description of the questions. The set of metadata used in this purpose is shown in Table 1.

The test material annotation has a twofold advantage.

1. Main focus of any ITS is to provide personalized instructions to the students and to do that the system needs to have a good perception of a student's cognitive ability and her performance level. The test items of the test materials are the only means to evaluate the students' current state of knowledge and they can provide hints on how the system will plan the teaching method for the student subsequently. Therefore, the effectiveness of the ITS and its adaptive teaching depends much on the quality of the test items. Hence, the test items should be designed in such a way that the test results reflect the correct state of the student. For this, a good description of the questions is required. The different important features of a question like *difficulty index, discrimination index, effectiveness of distracter, reliability coefficient* etc. can assess the quality and effectiveness of each question [7] and a better analysis of a student's performance can be done using these different features of a question.

2. These features can help the teachers to evaluate the standard of the questions. Some of the features like *discrimination index, effectiveness of distracter, reliability coefficient* etc can give a quantitative analysis on the quality of the test items. This analysis can be interpreted by the teachers and the test items can be modified if flaws are found in them. This would make the test material repository more robust and would reduce the number of inconsistent questions in it.

Table 1
A set of metadata associated with test items

S. No.	**Name**	**Purpose**	**Range of Values**	**Type of Annotation**
1	Topics	The list of topics this question focuses	Name of the topics	Manual
2	Difficulty Index	Indicates the difficulty level of the test item	[0,1]	Automatic
3	Discrimination Index	Indicates the effectiveness of the test item in discriminating between low and high scorers on the whole question set	[-1,1]	Automatic
4	Effectiveness of Distracters	Analyses the effect of the distracters on the students (applicable only for multiple choice type questions)	[-1,1]	Automatic

5	Focus[1]	It means whether the question focuses the comprehension ability (C) or the problem solving skills (P) of a student.	*C/P* or *both*	Manual
6	Time Limit	Time limit to solve the whole test	In minutes	Manual
7	Relevance	Relevance of a question with respect to the topic	[0,1]	Manual
8	Reliability Coefficient[2]	Measure of internal consistency reliability i.e. how well the test set measures a single cognitive factor	[0,1]	Automatic

2. Test Material Authoring

Presently, our ITSAS supports authoring of only multiple-choice type questions. There is an interface called Question Editor, which allows the teachers to create questions and annotate them simultaneously. Each question is annotated with 8 features as shown in Table 1. The metadata annotation is done in a semi-automatic manner, where some of the features like *topic, focus, relevance* and *time limit* are provided by the teachers while preparing the question. Values of other features like *difficulty index, discrimination index, Reliability Index and effectiveness of the distracters of 4 options in a multiple choice question* are automatically calculated by the system through test item analysis. Screenshot of the Question Editor is shown in Figure 9.

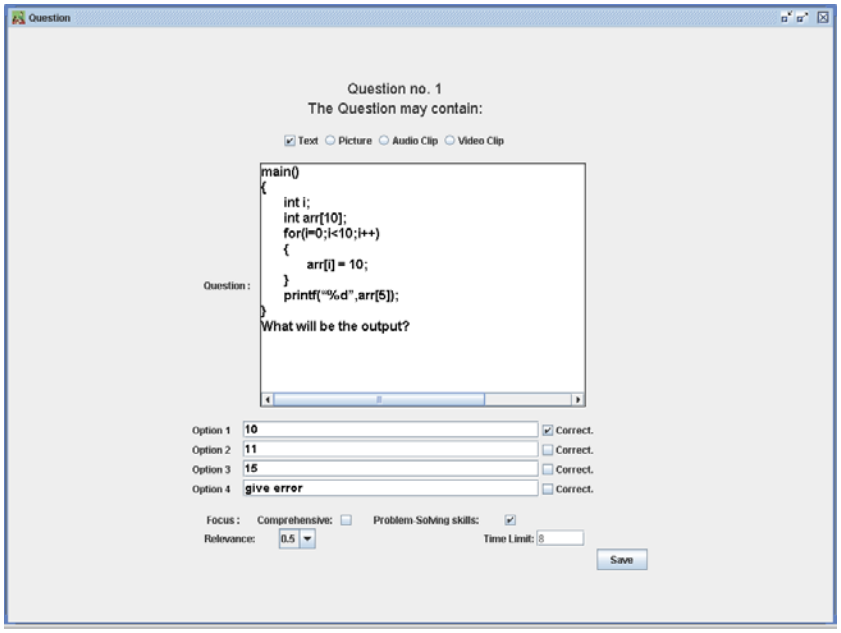

Figure 9. Screenshot of the Question Editor interface with a sample question

All the questions created through Question Editor are stored in the test material repository. The repository contains questions belonging to different topics. The teachers can select any number of questions to form test sets in any topic. The metadata *topic* associated

[1] Focus of a question means, whether the question focuses the comprehension ability(C) or the problem-solving skills(P) of a student. A question can focus both these attributes also. These attributes are a part of the Student Model. Explanations of these terms are beyond the scope of this paper.

[2] Reliability Coefficient is a feature of a test set, others are features of a single question

with the test item helps in selecting the questions from the repository to prepare the test set for a particular topic. The metadata *Reliability Coefficient* gives the inter-correlations among the test items**.** The *Reliability Coefficient* of the test set is low if the content of the test set are more diverse from the subject matter [6]. This metric helps the teachers to understand the quality of the tests authored by them and indicates if any modifications are required in test set.

3. Test Item Analysis and Automatic annotation

In this section we discuss how test item analysis is used in annotating the test questions.

3.1 Test Item Analysis (TIA)

Test item analysis provides a method to evaluate the test items or the questions quantitatively. Through TIA individual test question can be assessed as well overall quality of the test can be judged.

Values of different features of a test item can be obtained from test item analysis. These features can be used to annotate the test items. Thus test item analysis also provides a way of automatic annotation of the test materials. In an authoring system automatic annotation can reduce time and efforts required by the teachers and simplify the authoring process. Moreover, it can reduce errors generated due to manual annotation.

Test item analysis is done for each test item. Each student has to undergo a test after studying a particular topic. All the answers provided by the students in those tests are stored in the system. After every 40 responses of each question the *Test Item Analyzing Agent (TIAA)* retrieves all the stored answers (given by the students) of that question and performs some operations to calculate the values of some features, describing the question. Such repetitive analysis and revision of the feature values gives an updated and more accurate description of the questions. The different features used in this analysis and the methods used are discussed below.

Difficulty Index (DfI): DfI of a question is the proportion of respondent selecting the right answer to that question [5] and it reflects the difficulty of a question. DfI is calculated as [7],

$$DfI = \frac{No.\ of\ \ students\ answered\ the\ item\ correctly}{Total\ \ no.\ of\ students} \tag{1}$$

Its values lies in the range [0,1], where 0 represent the hardest question and 1 the easiest question

Discrimination Index (DI): This quantity gives a measure about how a question can discriminate between high-achieving and low-achieving students [7]. Its value lies in the range [-1,1]. High DI valued question implies that the question discriminates between the two groups of the students very well and vice versa. A negative value of DI of a question infers that more number of students from the low-achieving group answered the question correctly than the high-achieving one. Hence, the question might be defective and not achieving its purpose. DI is calculated as [7],

$$DI = \frac{U - L}{n} \tag{2}$$

where, U is the number of students that correctly answered the question from the upper group; L is the similar count from the lower group and n is the number of students in each group. To construct the groups, top 27% and bottom 27% of the students are considered, and then n becomes 27% of the total students [4][7].

Effectiveness of Distracters (ED): In a multiple choice type questions usually there is only one correct option and other options are distracters. The distracters are effective when they are plausible alternatives and lower ability students select it more in number than the higher ability ones. Its value lies in the range [-1,1]. ED of 1 represent no higher ability student (top 27%) selected it but all lower ability students (bottom 27%) selected the option [7]. Effective distracters should have positive values. However, the correct alternative of the question will have a negative ED value, but it is not a distracter. Effectiveness of Distracter of an alternative **a** is calculated as,

$$ED_a = \frac{L_a - U_a}{L_a + U_a} \tag{3}$$

where, L_a is the no. of students from lower groups who selected the option **a**, and U_a is the no. from the upper group.

Reliability Coefficient (RC): It is a measure of internal consistency reliability of the test set. High reliability means that the questions of a test set have stronger relationship among themselves. Students who answered a given question correctly were more likely to answer other questions correctly. Low reliability means that the questions tended to be unrelated to each other. Kuder-Richardson-20 (KR-20) formula gives the appropriate reliability coefficient for multiple choice test papers. KR 20 is defined as [11],

$$KR = \frac{n}{n-1}\left(\frac{v - \sum_{i}^{n} p_i q_i}{v}\right) \tag{4}$$

where, n is the no. of questions in the test, v is the variance of all the scores, p_i is the proportion of correct answer of question i in the test, and q_i is similarly the proportion of incorrect answers. The value lies between 0 (no reliability) and 1 (perfect reliability). In practice, their approximate range is from .50 to .90 [3][6].

3.2 Fuzzification of the Features and Report Generation

As explained earlier, the metadata used in test item annotation are used to update the student model and obtain information about a student's performance. Apart from this the updated values of the features after test item analysis produces important information about the questions. This information can be perceived by the teachers and can be modified if necessary. After each phase of TIA the authoring tool within the system produces a report about all the questions and a facility to edit them. The report helps the teachers to analyze the nature of the questions previously authored by them and if required they can modify the questions after viewing the report. For example, if a question has less than the desirable value of DI, the teacher can modify it and try to improve its efficiency. Again, if in a question one of the alternatives shows negative or very low value of ED, the teacher may replace that alternative with a more plausible one. This reduces the flaws in the authored questions and might lead to better performance of the ITS.

In order to make this part of the authoring task simpler the values of the features are fuzzified. This enables the display of the values of these features in linguistic terms. The fuzzified terms are more natural and easier to understand for the teachers, than crisp numerical values [13]. For example, instead of showing the numerical values of difficulty index, the difficulty value of a question is expressed as *hard, medium,* and *easy*. This lessens the difficulty of using the system and makes authoring simpler. Different fuzzy sets used to represent the features and their fuzzy membership functions [6][12] are shown in Table 2.

In Figure 10 the interface is shown where the reports of a question is shown along with editing facilities.

Table 2
Different features and fuzzy sets to describe them with membership functions (μ)
v denotes the crisp values

Difficulty Index	**Discrimination Index**
$\mu_{hard}(v) = 1 \text{ if } v \leq 0.5$ $= 0 \text{ otherwise}$ $\mu_{medium}(v) = 1 \text{ if } 0.5 < v < 0.8$ $= 0 \text{ otherwise}$ $\mu_{easy}(v) = 1 \text{ if } v \geq 0.8$ $= 0 \text{ otherwise}$	$\mu_{good}(v) = 1 \text{ if } v > 0.4$ $= 0 \text{ otherwise}$ $\mu_{fair}(v) = 1 \text{ if } 0.1 < v \leq 0.4$ $= 0 \text{ otherwise}$ $\mu_{poor}(v) = 1 \text{ if } v \leq 0.1$ $= 0 \text{ otherwise}$
Effect of Distracters	**Reliability Coefficient**
$\mu_{high}(v) = v \text{ if } v > 0$ $= 0 \text{ otherwise}$ $\mu_{moderate}(v) = 1\text{-}v \text{ if } v > 0$ $= 0 \text{ otherwise}$ $\mu_{low}(v) = 1 \text{ if } v \leq 0$ $= 0 \text{ otherwise}$	$\mu_{excellent}(v) = 1 \text{ if } 0.8 \leq v \leq 1$ $= 0 \text{ otherwise}$ $\mu_{good}(v) = 2v \text{ if } 0 \leq v < 0.5$ $= (2\text{-}2v) \text{ if } 0.5 \leq v \leq 1$ $\mu_{low}(v) = 1 \text{ if } 0 \leq v \leq 0.2$ $= 0 \text{ otherwise}$

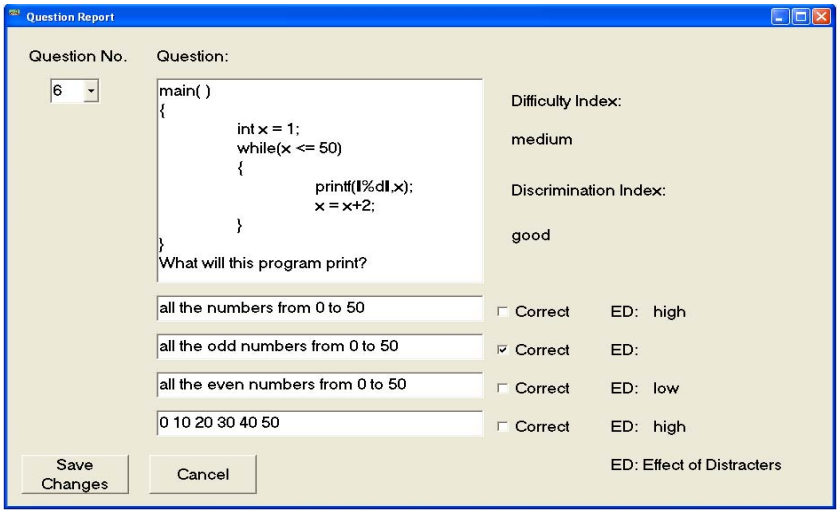

Figure 10. Screenshot showinga report of a question and facility to edit it

4. Results

In order to demonstrate the test item analysis of the system we present the results of an instance of test item analysis, where responses of 10 questions were considered. The test item analyzing agent in the system gathered the responses stored after 40 students answered the questions. Using the equations 1 and 2 it computed the DI, DFI of the questions. The values computed by the system are shown in Table 3. In the table along with crisp values the resulting fuzzy values are also shown. It might be relevant to mention that the crisp values are used to update these features (DI, DfI) of these questions automatically. The crisp values are also used by the system to update the student model and analyze the students' performance. This helps the ITS to adapt itself to the student and provide customized education. However, the fuzzy values will be presented to the teachers in form of reports. Analyzing them the teachers can modify the questions and remove any discrepancies, if found. We also present the ED of the 4 options of two questions obtained from this analysis. The results are summarized in Table 4.

Table 3
Difficulty index and Discrimination Index
of 10 questions computed by the system

Question no	Difficulty Index		Discrimination Index	
	Crisp	Fuzzy	Crisp	Fuzzy
1	0.49	hard	0.52	good
2	0.58	medium	0.4	fair
3	0.67	medium	0.32	fair
4	0.40	hard	0.56	good
5	0.96	easy	0.08	poor
6	0.61	medium	0.64	good
7	0.69	medium	-0.12	poor
8	1	easy	0	poor
9	0.81	easy	0.4	fair
10	0.96	easy	-0.08	poor

Table 4
Effectiveness of Distracters (ED) of each alternative computed
from two questions with detailed response (the figures show no. of responses)
correct options are marked with '*' (they are not distracters, ED values not shown)

Question 1 (correct option: 4)					
		OPT 1	**OPT 2**	**OPT 3**	**OPT4***
Lower 27%		5	4	7	3
Upper 27%		4	0	2	13
ED	**Crisp**	0.11	1	0.556	-0.63
	Fuzzy	moderate	high	high	-
Question 2 (correct option: 2)					
		OPT 1	**OPT 2***	**OPT 3**	**OPT 4**
Lower 27%		6	5	5	3
Upper 27%		1	12	5	1
ED	**Crisp**	0.71	-0.41	0	0.5
	Fuzzy	high	-	low	high

Finally we, show the reliability coefficient for the question sets or tests authored in the system. This is also computed by the TIAA and presented to the teachers as a part of the report. Teachers can make modifications in the tests if the value of the reliability coefficient is poor. Results found in 3 tests are shown in Table 5.

Table 5
Reliability coefficient of 3 tests

Test No.	Reliability Coefficient	
	Crisp	Fuzzy
1	0.86	excellent
2	0.80	excellent
3	0.71	good

As an example, it can be said that after viewing a report on these questions, a teacher might be interested to modify question no. 7, 8 and 10 (Table 3), for showing 'poor' DI values or change the option 3 of question 2 (Table 4) for showing 'low' ED value.

We also performed an experiment to demonstrate the influence of TIA in test material authoring. Initially 4 teachers authored 50 questions and annotated them totally manually; the feature of TIA was disabled. Then the same teachers authored another set of 50 similar questions and TIA was active and annotation was semi-automatic. The average time taken by each of the teacher was noted in both the cases. Their values are shown in Figure 11 and

compared. Second authoring phase shows reduced time consumption by 7.2%. In an opinion survey, all the 4 teachers agreed that using TIA authoring becomes simpler.

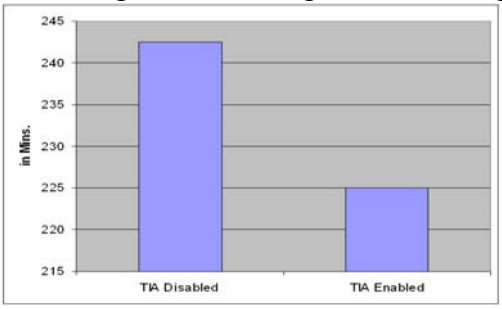

Figure 11. Comparison of Test Material Authoring with and without TIA

4. Conclusion

This paper discussed about test item analysis and how this technique can contribute to automatic annotation of some of the features of a question. It is also demonstrated and discussed how test item analysis and automatic annotation can reduce the cost of authoring test items in terms of time and effort and also helps in creating strong and effective repository of test materials, as manual annotation is prone to human-generated errors. Generation of reports marks out the fallacies in the question base and the facility provided by the authoring system to modify them helps in building a robust test material repository.

Accurate test material annotation should also facilitate improved performance in the ITS. Since, this should help in better estimation of the students' performance and improved scopes of adapting to their requirements. However, such a claim is not yet been evaluated and no such evidence is presented in the current paper.

A major drawback of this approach is that the nature of students may vary from group to groups. Hence, from this analysis the values of the features obtained are not universal. Thus, the values of the different features obtained after test item analysis are only applicable to students of similar groups on whose performance the analysis was done. Otherwise the feature values of the test items may be inconsistent.

Acknowledgement

This project is funded by Media Lab Asia, New Delhi

References

[1] Murray, T.(2003). An overview of intelligent tutoring system authoring tools: *Updated analysis of the state of the art. Authoring Tools for Adv. Tech. Learning Env*, Ainsworth and Blessing, Eds. Kluwer Academic Publishers. Printed in the Netherlands, 2003, pp. 493–546.

[2] Weon, S., Kim J. (2001). "Learning achievement evaluation strategy using fuzzy membership function" i*n Proceedings of the 31st ASEE/IEEE Frontiers in Education Conference,* Reno, NV, 2001, pp. 10–13.

[3] Gronlund, NX. (1981) Measurement and Evaluation in Teaching, New York : MacMillan Publishing Co.,

[4] Kelly, T.L. (1939) The Selection of Upper and Lower Groups for the Validation of Test Items, Journal of Educational Psychology, vol. 30,17-24 .

[5] Anastasi, A. (1968) Psychological Testing, 3rd edition, Macmillan Publishing Co ., Inc.

[6] J. C. Nunnally, Psychometric Theory. New York: McGraw-Hill, 1967, pp. 172-235

[7] Ebel, R.L., & Frisbie, D.A.(1991). Essentials of Educational Measurement. PHI, Eastern Economy Edition, 1991; pp 228-231

[8] Roy, D., Sarkar, S., Ghose, S. (2007). "Learning material annotation for flexible tutoring system". *Journal of Intelligent System,* Vol. 16, No. 4, 2007, pp. 293-305

[9] Chakraborty, S., Bhattachary, T., Bhowmick, P., Basu, A. and Sarkar, S. (2007). Shikshak: An Intelligent Tutoring System Authoring Tool for Rural Education. in the Proceedings of International Conference on Information and Communication Technologies and Development (ICTD2007). Bangalore, India

[10] Verhaart, M., and Kinshuk. (2006). An annotation framework for a virtual Learning portfolio. In the Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT'06)

[11] http://www.asu.edu/uts/pdf/InterpIAS.pdf

[12] Ross, T. J. (1997). *Fuzzy Logic with Engineering Applications*. McGraw-Hill, 1997, pp. 134-146.

[13] Nedic, Z., Nedic, V., and Machotka, J. (2002). Intelligent Tutoring System for teaching 1st year engineering. World Transactions on Engineering and Technology Education, Vol.1, No.2, 2002