Composing Sequential Test Items with Multipart Criteria in Adaptive Testing

Kavitha Rajamani
Department of MCA,
AIMIT, St. Aloysius College (Autonomous)
Mangalore, Karnataka

Vijaya Kathiravan
Department of Computer Science
Government Arts College (Autonomous)
Salem -7, Tamilnadu

ABSTRACT

The traditional learning environment is being rapidly supplemented by an E-Learning environment, particularly Computer Assisted Instruction (CAI). Each learner has different learning status and therefore should use different test items in their evaluation. The Computerized Adaptive Test (CAT) can adjust the degree of difficulty of test items dynamically depending on their ability. A good test will not only help the instructor evaluate the learning status of the students, but also facilitate the diagnosis of the problems embedded in the students' learning process. One of the most important and challenging issues in conducting a good test is the construction of test sheets that can meet various criteria. Therefore, several measures have been proposed to represent the quality of each test item, such as degree of difficulty and discrimination. However, the quality of a test not only depends on the quality of the item bank, but also relates to the way the assessment sheet is constructed. Selection of appropriate test items is important when constructing an assessment sheet that meets multi-criteria assessment requirements, such as expected difficulty degree, expected discrimination degree, number of the test items, estimated testing time and the specified distribution of relevant concept weights. Dynamic question generation is proposed which uses the novel approach of Particle Swarm Optimization. This approach will improve the efficiency of composing near optimal serial test items to meet multiple assessment criteria. The proposed approach can be compared with some existing means in terms of efficiency.

Key Words

Computer adaptive testing, Assessment sheet generation, Intelligent Testing, Intelligent Tutoring.

1. INTRODUCTION

Advancement of information and communication technologies has paved the way for innovation in the education system. By constructive approach of teaching, that is education has to be learner centered and learning occurs in a cognitive manner in learners' mind by means of past experiences gained and active learning that is "learning by doing in nature." Many adaptive learning and intelligent testing systems have been proposed to offer learners the customized courses. Various computer-assisted application platforms have been built, such as intelligent tutoring systems and computerized adaptive testing systems [1]–[3],

Tests are generally the most common and effective way in evaluating a learner's knowledge or ability. Traditional tests cannot always satisfy the need in discriminating the learners' knowledge, and the attributes such as the test completion time and the difficulty degree of a test are hard to be controlled. Computer-based tests have been proven to be more effective and efficient than traditional paper-and-pencil tests due to several reasons:

First, the test sheets can be composed dynamically based on the practical requirements; second, more test items can be presented in multimedia styles; third, the student testing portfolio can be recorded and analyzed to improve their learning performance. With user interactivity and adaptability, computer-based assessment expands testing possibilities beyond the limitations of traditional paper-and pencil tests. Therefore, how to progress an efficient learning process is a critical issue.

A well-scrutinized test is helpful for teachers wanting to verify whether students well digest relevant knowledge and skills and for recognition of students' learning bottlenecks [9]. In a computerized learning environment, this provides students with greater flexibility during the learning process; information concerning the student learning status is even more important [9]. The key to a good test depends not only on the subjective appropriateness of test items, but also on the way the test sheet is constructed.

In modern education, computer-assisted testing systems are promising in generating tests more efficiently and effectively for evaluating a person's skill. Compared to the traditional paper-and-pencil media, computer-assisted testing platforms are more favored by students. In addition, personalized assessments tailored to each student are the developing trends [3]–[5]. Personalized CAT systems select an appropriate question from the question bank based on the examinee's answer to the previous question. Therefore, it is very important to precisely determine the learning status of each student so that proper tutoring strategies can be applied accordingly [5], [6]. A high-quality test is the major criterion for determining the learning status of students.

In [6], researchers propounded a "Knowledge Based Computer Assisted Instruction System", which can change the numeric component of items when the test is in progress, preventing students from memorizing the answers. Another branch of relevant researches is Computerized Adaptive Testing (CAT), which applies various prediction methodologies to shorten the length of the test participation time without loses of precision [7]. However, the quality of a test is not only dependent upon the quality of the item bank, but also the way in which the test sheet has been constructed. Further, it is important to select appropriate test items when constructing a test sheet that meets several assessment requirements, such as average difficulty degree, average discrimination degree, number of test items, and the specified distribution of concept weights.

Unfortunately, this assessment system is not without concerns. Perhaps the most salient issues raised in regards to the extended use of CAT are item overexposure and face validity [10]. Depending on the item selection algorithm used in CAT application programs, particular items in the item pool may be over-selected. That is, items that provides the most discriminating information to the CAT system about the examinee's ability may be administered to numerous participants and become familiar to test takers prior to testing, thus diminishing test security and reliability. In addition, if items are found to be over-selected and risk exposure, additional item development will be required, in effect increasing costs for CAT maintenance. It is inefficient to require additional development of items while a large proportion of the item pool remains unused. To limit overexposure and its effects, the item selection method needs to select discriminating items while considering pool utilization. Item selection is also confounded by non-statistical issues such as content balancing. By nature of an adaptive test, examinees sitting to take the same test will be administered different items but each must receive the same distribution of items by content area. For example, for a 28 item mathematics test it would not be valid to administer 28 items on arithmetic to one student and 28 items on geometry to another. There must be a balance across content areas or domains measured.

A well-constructed test sheet not only helps evaluation of the learning status of the students, but also facilitates improved diagnosis of any problems within the learning process. In this paper, Dynamic question generation is proposed which uses the novel approach of Adaptive Particle Swarm Optimization. This approach will improve the efficiency of composing near optimal serial test items to meet multiple assessment criteria. The proposed approach is compared with some existing means in terms of efficiency. Experimental results have shown that the approach can achieve better performance than other previously used methods.

2. RELEVANT RESEARCHES

In computer based tests, randomized presentation of items is automatically programmed into testing software to present different items to the test takers. The downside of such randomization is that it prevents planned sequencing of items. Randomizing items does not accommodate a test user or a constructor who wishes to ensure that items progressively become tougher. It may unfairly increase test anxiety for some of the candidates. Increased anxiety at any stage during the test for whatever reason is likely to have a negative effect on that person's performance for the remainder of the test [1]. Instead of giving each examinee the same fixed test, CAT item selection adapts to the ability level of individual examinees. In [5], they proposed an automatic leveling system for e-learning examination pool using entropy measure. The questions were leveled based on the response given by the greater part of learners with similar background. In order to assess the capacity of each question or task to distinguish between those who know and those who do not, the trial group of candidates should possess a range of knowledge from those with good knowledge to those lacking it [6].

Although many computer-assisted testing systems have been proposed, few of them have addressed the problem of systematically composing test sheets for multiple assessment requirements [2], [7]. Most of the existing systems construct a test sheet by manually or randomly selecting test items from their item banks. Such manual or random test item selecting strategies are inefficient and usually are not able to simultaneously meet multiple assessment requirements. Some previous investigations attempted to employ a dynamic programming algorithm to find an optimal composition of the test items [6]. As the time complexity of the dynamic programming algorithm is exponential in terms of the size of input data, the required execution time will become unacceptably long if the number of candidate test items is large. In a testing system, the quality of the test items will significantly affect the accuracy of the test; therefore, several measures have been proposed to represent the quality of each test item, e.g., degree of difficulty and discrimination. These measures can be derived and updated, according to the statistical results of each test. Hwang [9] proposed multiple-criteria where test sheet-composing problem is formulated as a dynamic programming model to minimize the distance between the parameters (e.g., discrimination, difficulty, etc.) of the generated test sheets and the objective values subject to the distribution of concept weights. A critical issue arising from the use of a dynamic programming approach is the exceedingly long execution time required for producing optimal solutions. As the time-complexity of the dynamic programming algorithm is exponential in terms of input data, the execution time will become unacceptably long if the number of candidate test items is large. Consequently, Hwang [9] attempted to solve the test sheet-composing problem by optimizing the discrimination degree of the generated test sheets with a specified range of assessment time and some other multiple constraints.

Particle swarm optimization was proposed by Kennedy and Eberhart in 1995[11]. This algorithm was developed by a simulation of social behavior models. PSO maintains a swarm of particles, such as fish schooling and bird flocking, where each particle represents a potential solution to an optimization problem. The primary stratagem of PSO is that each particle keeps track of its coordinates in an N-dimensional problem space which are related to the optimal solution it has achieved so far.

Initially, PSO generates a swarm of random particles and then searches for the optimal solution by updating each iteration. In every generation, each particle updated its location according to the velocity function. The velocity function follows two values. The first one is the personal best experience (fitness value) of each particle in the past iterations. This value is called PBest. The other one is the global optimal solution of total particles in the past iterations. This value is called GBest. When the termination criteria or maximum number of iterations has been attained, the PSO process would terminate.

From the literature, it is very well seen that, there is a need of adaptive assessment with intelligence to satisfy compound criteria through some new enrichments.

3. METHODOLOGY

In this paper, we propose a model that formulates the dynamic question generation problem under different assessment criteria. With regard to each test item, this model maintains three assessment considerations which are the difficulty level of each test item, discrimination level, the relevance association between each question and each topic, the proportion of the concepts in the test and the exposure frequency of each question. Assume the item bank consists of N test items. When i questions are selected to test learners from the item bank, these questions will be a subset of N test items. Also, assume that the test aims at M concepts.

Following are the some of the attributes needed for composing test.

3.1 Item Attributes in a Test

A test of *n* questions should be generated. Each question has several attributes, as a unique item id, difficulty degree, discrimination degree and the weight of concept(s) that the question involves.

Item Difficulty

Item difficulty is used to find out how each item affects a student's overall success throughout the test in terms of difficulty. Because it is being tried to group questions according to their difficulty level. Normally, item difficulty is scaled in a range from 0.00 to 1.00. Actually, it is inversely proportional to the number of correct answers of each question. This means that if any question has the least amount of correct answer is the hardest question in test. Hence, Item Difficulty can be calculated as,

$$ID = MSCA / SCAE$$

where,
ID is Item Difficulty,
MSCA is Minimum Sum of Correct Answers, SCAE is Sum of Correct Answers of Each Question.

Normally, items having difficulty values in two extremes of range will be pruned from further analysis, because this will not give proper inference on the ability of the learner.

Item Discrimination

Item discrimination degree indicates a question's ability to discriminate between the students who know the knowledge and those who do not. Generally, it is computed by ranking the students according to the total score. The value of a discrimination degree ranges in [−1.00, 1.00].

$$\text{Item Discrimination} = (Up / U) – (Lp / L)$$

Where,
Up= Number of high performers with question right
Lp = Number of low performers with question right
U = Number of high performers
L = Number of Low performers

The higher the discrimination degree, the better the question does in evaluating the students' knowledge. A discrimination degree that is no smaller than 0.3 is usually regarded as acceptable. If the discrimination degree is smaller than zero, the question is not suitable for the test and should be deleted.

Item – Concept Weightage

As questions are used for assessing whether the student has grasped the concept(s), each question is related with one or more concept(s). Suppose *M* concepts are checked in the test. Using 0 to 4 representation scheme, the relations between concepts and questions can be formulated as,
0: Test item has no relationship with that concept
1: Test item has weak relationship with that concept
2: Test item is related to that concept
3: Test item has high relationship with that concept
4: Test item has very high relationship with that concept.
Assume that an examination aims at M topics which consist of i questions, therefore each question selected from the item bank should relevant to one or more of these topics, say rj, $1 \leq j \leq m$

Test – Concept Weightage

The test should be administered in such a way that all the relevant concepts are covered that too with proper weightage. Hence, the instructor has to provide the proportion of the concepts for the test in terms of a vector. Each topic has a different weight wj, $1 \leq j \leq m$, which is assigned by the instructor. For example, to test the data structure knowledge of learners which consists of ''Stack", ''Queue", and ''Array", the teacher can assign different weights to these topics, such as w1 = 0.3 (weight of Stack), w2 = 0.3 (weight of Queue), and w3 = 0.4 (weight of Array).

Initial ability of the learner

Normally, in the beginning of the test, a moderate test item is posed. Based on the response given to that item, successive item is posed. Instead of following this strategy, a multistage testing is adopted. A set of moderate items are posed and the ability of the learner is estimated. This ability is taken as the start for the real test.

Exposure Frequency of the Test Items

Additionally, the system records the exposure frequency of the N test items that were selected in the past examinations, n1, . . .,nN. The maximum exposure frequency of N test items is called max(n1, . . .,nN). Assume a question k has been selected nk times in the past tests, called $0 \leq nk \leq$ max(n1, . . .,nN). If a question's exposure frequency is higher than others, this question's answer will possibly be remembered by the students. Therefore, the dynamic question generation system would control the exposure frequency of each test item.

3.2 Test Sheet Composing

In an item bank, a subset of n candidate test items Q1, Q2, . . .,Qn will be selected for composing a test sheet. The model proposed here considers different compound assessment requirements. Assume there are 'n' items in the item pool and 'm' concepts to be dealt with. The measures which are in need of assessment are as follows:

(1) Decision variables xi : $1 \leq i \leq n$, xi is 1 if test item i is selected; 0, otherwise.
(2) Degree of Difficulty of an item di : $1 \leq i \leq n$
(3) Degree of discrimination of item ei : $1 \leq i \leq n$
(4) Concept involved cj : $1 \leq j \leq m$
(5) Degree of association between an item Qi and concept Cj r_{ij} : $1 \leq i \leq n$ and $1 \leq j \leq m$
(6) Lower Bound on Expected Concept Relevance CLj : $1 \leq j \leq m$
(7) Upper Bound on Expected Concept Relevance CHj : $1 \leq j \leq m$
(8) Exposure Frequency : nk, $0 \leq nk \leq$max(n1,.,nN).

To generate test sheets for multiple requirements, this paper proposes a new approach. Without loss of generality, one can assume each test item contains the following information:
• a measure of its difficulty level
• a measure of its discrimination value
•a weight to represent the relationship between each test item and each concept
• an exposure frequency.
Based on the assumptions, the compound criteria test-generating problem can be described as follows:

1. The instructor sets the relevant parameters of a test, which include the initial ability of the learner, difficulty degree of test items, relevant knowledge for this examination, and the weight of each topic.

2. The dynamic question generation model applies the PSO algorithm that can select tailored test items automatically for each learner according to the multiple criteria. Since a test item is selected in each stage, the difficulty level, the discrimination level, and the distribution of concept weights are recomputed, and the new values are used as the starting values of the next stage.

3. The students then answer the questions based on their capability. If a leaner answers a question that turns out to be too easy, then the next test item to be posed must be more difficult, and vice versa.
4. This step is to judge whether this procedure can be terminated, and if not then it goes back to the second step and selects the best question for each student. The final difficulty level of the test sheet should be closer to the desired values under the constraint, so that the weights of the concepts satisfy the specified distribution.

The formal definition of the dynamic question generation model is described as follows:

Minimize Z = f + C1 + C2

The above formula is the fitness function of the dynamic question generation model, and it consists of three constraints which are described as follows:

f = = $|d_k - D_i|$, 1 ≤ k ≤ n

f indicates the difference between the degree of difficulty of selected test items and the target difficulty level.

$C_1 = 1 - \sum_{j=1}^{M} w_j r_j$, 1 ≤ j ≤ m

C1 represents the degree of relevance between the selected questions and particular topics.

$C_2 = \frac{n_k}{Max(Max(n_1,\ldots,n_k,\ldots,n_N),a)}$, 1 ≤ k ≤ n

C2 indicates the exposure frequency of selected test item.

Z(Ik) is a fitness function. Through the computation and iteration each test item has obtained a fitness value from the fitness function. If a test item Ik contains minimal fitness value, it will be selected by the dynamic question generation system.

The above approach tries to generate a test sheet which is having difficulty degree closer to the expected difficult level. Also, only the items having maximum discrimination values are taken. Concurrently the degree of association between the item and the concept is checked with the given range of association. Also, exposure frequencies of the items are controlled very much. Hence forth, this approach provides an intelligent test sheet with more quality.

4. EVALUATION AND DISCUSSION

The performance of the proposed approach has been evaluated through an experiment completed for 5 cases which specify various degrees of difficulty and discrimination with different similarity thresholds. The subject Data Mining is taken for the study test. All the cases were administered with 25 test items. The proposed approach is compared with the Random Item Selection method and with the objective requirements.

The random selection program generates the test sheet by selecting test items randomly to meet the constraints of number of test items.

Table 1. Comparison of degree of difficultness

Case No.	Expected Difficulty	Random Selection	Intelligent approach
1	65	75.5	64
2	78	67	80
3	91	87	88
4	63	68.25	64
5	40	31	44

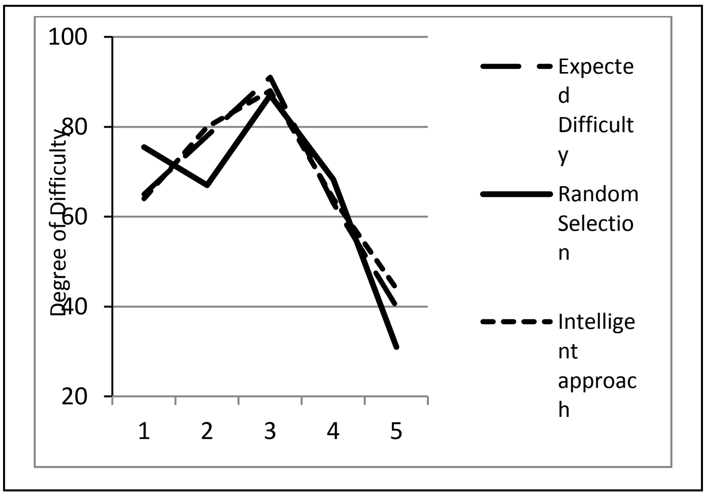

Fig 1: Degree of Difficulty of different approaches

From the above table and figure, it can be apparently seen that assessment sheets with near expected difficulty degrees can be obtained than by the random selection approach.

5. CONCLUSION

The approach of using computer-assisted testing systems to release teachers from the burden of composing tests and improve the assessment quality of tests is significant and promising in modern education. The multiple criteria test-sheet-generating problem is formulated, and an intelligent approach is proposed to generate test sheets that meet multiple assessment requirements. The question attributes in a question bank are adaptively adjusted, always reflecting students' learning states. From some experimental results, the approach achieves desirable performance under considerations of difficulty.

Several other AI or optimization based technologies and heuristic algorithms could be exercised to develop more efficient test sheet generating approaches for very large item banks. The combination of intelligence and personalization is the future direction, which will be addressed in the forthcoming work.

REFERENCES

[1] C. Chou, "Constructing a computer-assisted testing and evaluation system on the world wide web-the CATES

experience," *IEEE Trans. Educ.*, vol. 43, no. 3, pp. 266–272, Aug. 2000.

[2] S. Piramuthu, "Knowledge-based web-enabled agents and intelligent tutoring systems," *IEEE Trans. Educ.*, vol. 48, no. 4, pp. 750–756, Nov. 2005.

[3] E. Guzm´an and R. Conejo, "Self-assessment in a feasible, adaptive webbased testing system," *IEEE Trans. Educ.*, vol. 48, no. 4, pp. 688–695, Nov. 2005.

[4] A. Kavˇciˇc, "Fuzzy user modeling for adaptation in educational hypermedia," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 34, no. 4, pp. 439–449, Nov. 2004.

[5] G.-J. Hwang, B. M. T. Lin, H.-H. Tseng, and T.-L. Lin, "On the development of a computer-assisted testing system with genetic test sheet generating approach," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*,vol. 35, no. 4, pp. 590–594, Nov. 2005.

[6] Xiao-Min Hu, Jun Zhang, "An Intelligent Testing System Embedded with an Ant-Colony optimization based Test composition Method", *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 6, pp. 659–669, Nov. 2009.

[7] Marks, A. M., & Cronje, j. C. "Randomised items in computer-based tests: Russian roulette in assessment?". Journal of *educationaltechnology & society*, *11* (4), 2008.

[8] Betül Erdoğdu, "Computer based testing evaluation of question classification for Computer Adaptive testing", A Master Thesis (2009).

[9] Yin, P.-Y., Chang, K.-C., Hwang, G.-J., Hwang, G.-H., & Chan, Y., "A Particle Swarm Optimization Approach to Composing Serial Test Sheets", Journal of *Educational Technology & Society*, 9 (3), 3-15.

[10] Chi-Keung Leung, Hua-Hua Chang, and Kit-Tai Hau, "Computerized Adaptive Testing: A Comparison of Three Content Balancing Methods", The Journal of Technology, Learning, and Assessment, Volume 2, Number 5, December 2003.

[11] Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. In Proceedings of the IEEE international conference on neural networks 4, pp. 1942–1948.