2C4-IOS-3c-6

An estimation method of item difficulty index combined with the particle swarm optimization algorithm for the computerized adaptive testing

Shu-Chen Cheng Guan-Yu Chen I-Chun Pan

Department of Computer Science and Information Engineering, Southern Taiwan University of Science and Technology

The computerized adaptive testing is to provide items that are consistent with the current ability of testee, and to decide the difficulty for the next selected item according to the correctness of the testee's answer. It achieves the goals of adaptive learning through the mechanisms of dynamic adjustment of item difficulty to accelerate the test process or to shorten the number of items in a test. A prerequisite of computerized adaptive testing is to estimate the difficulties of items correctly. In this study, we describe the parameters of items by an adjusted approach. It considers each knowledge block as an independent dimension and gives a value for each dimension of the difficulty. Combined with the particle swarm optimization algorithm, a dynamic item selection strategy is proposed to develop an adaptive testing system. Therefore, it adopts the multiple assessment methods for the abilities by giving a value for each dimension of the ability. By way of the the dynamic item selection in computerized adaptive testing, all the selected items will be highly correlated and more consistent with the current actual abilities of the testees.

1. Introduction

Learning is generally immediately accompanied by testing. Testing is an integral part of the learning process. The results of testing can provide feedbacks to both instructors and learners. An instructor can correct teaching directions based on such feedbacks, and learners can correction their learning from these feedbacks. Due to the popularity of computers and Internet, the teachers or researchers have started to construct the computerized test systems. The computerized adaptive testing (CAT) has been developed to solve the problem that the traditional computerized testing gives the inappropriate test items. It provides the test items with the difficulties which are consistent with the testee's ability. It creates the exclusive content of personal tests by the way of dynamic item selection.

For a big test item bank whit many test items, there are two important challenges to a CAT system. First is how to correctly and quickly estimate the item difficulty index of test item. In this study, the testees' abilities are considered into the estimation process of the item difficulty indices. Those who answered wrongly with higher ability or answered correctly with lower ability are regarded as the answers abnormality. The concept of answers abnormal rate is proposed to develop an estimation method of item difficulty indices. Second is how to quickly locate a suitable test item for a learner's ability. This study adopts the knowledge structure concept for multiple ability evaluation for testees, which is based on the particle swarm optimization (PSO) algorithm, to develop a dynamic item selection strategy.

Through the proposed estimation method based on the answers abnormal rate, the item difficulty indices and the testees' abilities can be estimated mutually. Each test item can also be estimated independently. Therefore, the test item bank can be expanded easily at any time without abundant pre-test samples. And then, the dynamic item selection system adopts the PSO algorithm as a core and integrates with the knowledge structure concept. The quick search advantage in the PSO algorithm and knowledge structure characteristics allow the most suitable test items for a testee's ability to be quickly identified, even in a big test item bank.

2. Literature Reviews

2.1 Computerized Adaptive Testing

The traditional testing is based on classical test theory. It gives all the testees the same test paper with the same test items. However, this is not appropriate for certain types of tests. For the testees with higher or lower abilities, to give them the same test items may be too difficult or too easy. Inappropriate items are not only unable to discriminate the testees' abilities accurately, but even may combat the testees' positive or confidence. Hence, the tests loss the significance (Cheng, Lin, & Huang, 2009; Huang, Lin, & Cheng, 2009)。

In order to improve the lack of traditional testing, the basic concept of computer adaptive testing is to select the test item with the difficulty which is the most consistent with the testee's current ability. When a test item has completed, the test system will assess the testee's ability immediately. And then, the next one test item will be selected according to this ability. In the other words, the testee's answer is correct or not will affects the difficulty of next one test item selected. For the testees with higher abilities, do not have to give them too easy test items; for the testees with lower abilities, do not have to give them too difficult test items. Through this kind of dynamic item selection strategy, the computer adaptive testing can be held according to the different testees' abilities. The adaptive testing is a way of test which is created exclusively and personally. Therefore, the

Contact: Shu-Chen Cheng, Department of Computer Science and Information Engineering, Southern Taiwan University of Science and Technology, No.1, Nantai St., Yongkang Dist., Tainan City 710, Taiwan (R.O.C.), +886-6-2533131#3228, kittyc@mail.stust.edu.tw

adaptive testing is widely used in different areas (Anatchkova, Saris-Baglama, Kosinski, & Bjorner, 2009; El-Alfy & Abdel-Aal, 2008; Badaracco & Martínez, 2013). Because of the feature of dynamic item selection strategy according to the testees' abilities, to implement the computer adaptive testing can not only shorten the number of test items, but also can assess the testees' abilities accurately. It archives the goal of individualized learning (Cheng, Lin, & Huang, 2009; Huang, Lin, & Cheng, 2009).

2.2 Particle Swarm Optimization

The basic concept of particle swarm optimization (PSO) algorithm is derived from a social group behavior simulation, first used by Eberhart & Kennedy (1995) to develop an optimization method based on characteristics of foraging behavior in fish shoals and bird flocks. This method assumes that a flock of birds forages for food in an area. There is only one place where the food is. The birds do not know the position of food, but they know how far they are from the food. Thereby, the simplest or most effective strategy to find the food is to search in the adjacent areas to be closest to the food.

Since the PSO algorithm was formally proposed, it has been widely applied in many applications due its many advantages, including its simple structure, few parameters, fast convergence, and applicability to dynamic environments and almost optimization problems. Many relevant studies have even applied the PSO algorithm to digital learning and used it in adaptive testing systems for conducting item searches (Cheng, Lin, & Huang, 2009; Huang, Lin, & Cheng, 2009), applied it to online learning systems for teaching (Huang, Huang, & Cheng, 2008), applied it to automatic learning partner recommendation (Lin, Huang, & Cheng, 2010), or applied it to blogs for searching recommended posts (Huang, Cheng, & Huang, 2009).

There are two important functions of PSO algorithm: the fitness function and the velocity function (Musii, Daolio, & Cagnoni, 2011). The fitness value, calculated by the fitness function, determines whether the position where the particle falls is good or bad. The velocity function is been used to determine the particle's velocity as (1) and to determine the new particle's falling position as (2).

$$\begin{aligned} &v_{id} \leftarrow w \times v_{id} + C_1 \times R_1 \times (Pbest_{id} - x_{id}) \\ &\square\square\square\square\square\square\boxplus C_2 \times R_2 \times (Gbest_d - x_{id}) \end{aligned} \tag{1}$$

Where v_{id} is the velocity of *i*-th particle in *d*-th dimension; w is the inertia weight; C_1 and C_2 are the acceleration functions, which are usually 2; R_1 and R_2 are the random values between 0 and 1; $Pbest_{id}$ is the particle's optimal solution, which is the position of optimal solution of *i*-th particle in *d*-th dimension; $Gbeest_d$ is the global optimal solution, which is the position of current optimal solution among all particles in *d*-th dimension.

$$X_{id} \leftarrow X_{id} + V_{id} \tag{2}$$

Where x_{id} is the position of *i*-th particle in *d*-th dimension.

In its initial state, the PSO algorithm randomly generates particles in a search space, where each particle has a different velocity. After using the fitness function to obtain an adaptive value for the current position, the algorithm determines whether the current position is good or bad. Each particle can memorize its own optimal fitness value, which named the particle's optimal solution (*Pbest*). It then passes a message to identify the position with an optimal fitness value among the positions passed by all particles, which named the global optimal solution (*Gbest*). It then uses (1) to compute a new velocity for each particle, (2) to determine a new position for the particle, and update *Pbest* and *Gbest* with an iterative approach until the optimal solution is found. Fig. 1 shows a flowchart for the PSO algorithm.

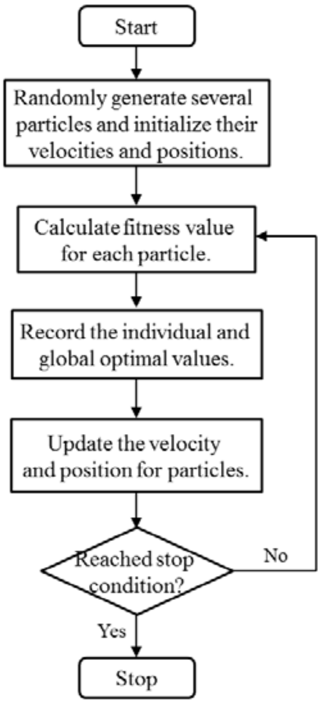
Start
Randomly generate several particles and initialize their velocities and positions.
Calculate fitness value for each particle.
Record the individual and global optimal values.
Update the velocity and position for particles.
Reached stop condition?
No
Yes
Stop

Fig. 1 PSO algorithm flowchart.

2.3 Item Difficulty Index

There are usually two methods to estimate the item difficulty indices. First, the item difficulty index of a test item is represented by the percentage of correct answers. It is shown as (3).

$$P = \frac{R}{N} \times 100\% \tag{3}$$

Where *P* is the item difficulty index; *N* is the number of all the testees; *R* is the number of testees who answered correctly.

There is another method to estimate the item difficulty index. First, the testees are sorted by their scores. Then, the groups of highest scores and lowest scores are designated as the higher score group and the lower score group. To compute the percentage of correct answers for these two groups. Finally, the average of their percentage is taken as the item difficulty index. It is shown as (4).

$$P = \frac{P_H + P_L}{2} \tag{4}$$

Where P_H is the percentage of correct answers in the higher score group and P_L is the percentage of correct answers in the lower score group.

Typically, these two extreme groups are token 25%, 27%, or 33%. (Haladyna, 1999; Suen, 1990).

3. Methods

3.1 Item Difficulty Index Estimation

In this study, an estimation method of item difficulty indices based on the answers abnormal rate is proposed to refer to the item response theory (IRT) model. The testee's abilities are considered into the estimation process of item difficulty indices.

The answers abnormal rate of one test item for the testees with the abilities greater than the item difficulty index is represented by the wrong answer rate. It is shown as (5).

$$hAAR_i = \frac{hW_i}{hN_i} \tag{5}$$

Where $hAAR_i$ is the answers abnormal rate of higher ability group for one test item if its difficulty is the i-th level; hW_i is the number of wrong answers of higher ability group for one test item if its difficulty is the i-th level; hN_i is the number of all the testees of higher ability group for one test item if its difficulty is the i-th level.

The answers abnormal rate of one test item for the testees with the abilities smaller than the item difficulty index is represented by the correct answer rate. It is shown as (6).

$$lAAR_i = \frac{lR_i}{lN_i} \tag{6}$$

Where $lAAR_i$ is the answers abnormal rate of lower ability group for one test item if its difficulty is the i-th level; lR_i is the number of correct answers of lower ability group for one test item if its difficulty is the i-th level; lN_i is the number of all the testees of lower ability group for one test item if its difficulty is the i-th level.

The answers abnormal rate of one test item for the testees with the abilities equal to the item difficulty index is represented by the absolute value of the difference between the correct answer rate and 0.5. It is shown as (7).

$$eAAR_i = \left| \frac{eR_i}{eN_i} - 0.5 \right| \tag{7}$$

Where $eAAR_i$ is the answers abnormal rate of equal ability group for one test item if its difficulty is the i-th level; eR_i is the number of correct answers of equal ability group for one test item if its difficulty is the i-th level; eN_i is the number of all the testees of equal ability group for one test item if its difficulty is the i-th level.

To add the three parts (5), (6), and (7) together is the answers abnormal rate of one test item. It is shown as (8).

$$AAR_i = hAAR_i + eAAR_i + lAAR_i \tag{8}$$

Where AAR_i is the answers abnormal rate for one test item if its difficulty is the i-th level.

To take the level of minimum answers abnormal rate as the item difficulty index of one test item. It is shown as (9).

$$D = \arg\min(AAR_i) \tag{9}$$

Where D is the item difficulty index of one test item.

3.2 Fitness Function of PSO

The fitness function of the PSO algorithm determines the strengths and weaknesses of a particle based on its position. In this study, the three evaluation criteria are the test item difficulty, test item relevance degree and knowledge block, and selected number of test items. These items are used as evaluation criteria to locate the most suitable test item for the learner's ability using the relevant parameters.

The main purpose of (10) is to evaluate the gap between the testee's current ability and the difficulty of selected item. A smaller gap gives a smaller value, when $0 \le DL_k \le 1$. The optimal situation is that there is no gap between the selected item and the testee's ability, and its value is 0.

$$DL_k = \frac{\sum_{j=1}^{m} \left| d_{kj} - D_j \right| r_{kj}}{q_k} \tag{10}$$

Where D_j is the testee's current knowledge block ability, where $0<D_j<1$; d_{kj} is the current knowledge block difficulty of current selected item, where $0<d_{kj}<1$; q_k is the number of relevant knowledge blocks for the current selected item; m is the number of relevant knowledge blocks; r_{kj} is the relevance of the k-th test item and the j-th relevant knowledge block, which its value is 1 if it has relevance and 0 otherwise.

Equation (11) locates the relevance degree between the test item and weight value of knowledge block set by the testee, where $0 \le RD_k \le 1$. A smaller value indicates greater relevance.

$$RD_k = 1 - \frac{\sum_{j=1}^{m} (w_j - U_j / T) r_{kj}}{q_k} \tag{11}$$

Where U_j is the number of items currently selected from knowledge block j and T is the total number of test items expected to be given in the test; w_j is a different weight value that can be set for each relevant knowledge block, where $0 \le w_j \le 1$.

Equation (12) is the exposure control factor. It balances the number of selected times for all test items and adopts an item exposure rate control as its objective. They thus take the most frequently selected item in the item bank and the current selected item as the major evaluation terms. To maintain the main purpose of adaptive testing, it selects the items that meet the learner's ability and relevant knowledge. To improve the item selection accuracy, these functions use RD_k as a constraint condition. The smaller values indicate that are has been previously selected fewer times, where $0 \le ECF_k \le 1$.

$$ECF_k = (1 - RD_k) \cdot \frac{n_k}{Max(n_1, \ldots, n_k, \ldots, n_N)} \tag{12}$$

Where n_k is the number of times that the currently selected test item has been selected, where $0 \le n_k$; Max($n_1,\ldots,n_k,\ldots,n_N$) is the number of times that the most frequently selected item in the item bank has been selected, where $0 \le$ Max($n_1,\ldots,n_k,\ldots,n_N$); N is the number of test items in the test item bank.

Adding (10), (11), and (12) allows the fitness function (13) in the PSO algorithm.

$$\text{Minimum}\square Z(X_k) = DL_k + RD_k + ECF_k \tag{13}$$

Where Z is the fitness value and X_k={DL_k, RD_k, ECF_k} is the particle's position vector.

In this algorithm, a smaller fitness value denotes greater fitness between item difficulty, relevant knowledge, and testee's current ability; in other words, it indicates that the test item is more suitable for the testee.

3.3 Velocity Function of PSO

After defining the fitness function, we come to another important function in the PSO algorithm: the velocity function. Velocity affects the direction and distance of a particle's movement in the search space. It is shown as (14).

$$\begin{aligned} V_{t+1} &= W \times V_t + C_1 \times R_1 \times (X_p - X_t) \\ &\square\square\square\square\square\square\boxed{+}C_2 \times R_2 \times (X_g - X_t) \end{aligned} \tag{14}$$

Where V_t indicates the velocity of a particle in the t-th iteration; V_{t+1} indicates the velocity of a particle in the (t+1)-th iteration; C_1 and C_2 are learning factors that influence the individual and global optimal solutions for each iteration; R_1, R_2, and W all prevent falling into the local optimal solution in the search process; X_p is the particle's individual optimal solution; X_g is the global optimal solution; X_t is the position of a particle in the t-th iteration.

After each particle obtains its velocity according to (14), (15) is been used to update the particle's new position.

$$X_{t+1} = X_t + V_{t+1} \tag{15}$$

Where X_{t+1} updates parameters of the selected test item and determines the particle's (t+1)-th item position.

3.4 Dynamic Item Selection Strategy

The CAT primarily seeks to allow a system to provide the test items in line with the testee's ability when the learner takes an online computerized test and to determine the difficulty of the next item according to each answer. The system uses this mechanism to achieve computerized adaptive learning objectives and accurately estimate the user's ability. This study realizes the CAT with the knowledge structure concept to interpret the relationship between test item knowledge and the testee's ability. It adopts the PSO algorithm as the core to find the most suitable test item for the testee's current ability from a big test item bank. Fig. 2 shows the flowchart for PSO adaptive testing.

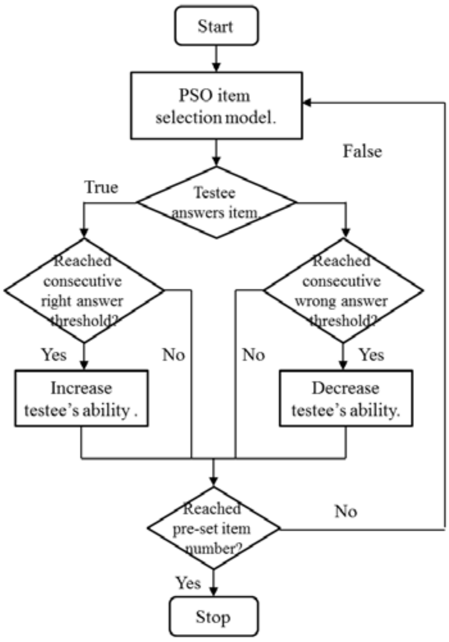
Start
PSO item selection model.
False
True
Testee answers item.
Reached consecutive right answer threshold?
Reached consecutive wrong answer threshold?
Yes
No
No
Yes
Increase testee's ability .
Decrease testee's ability.
Reached pre-set item number?
No
Yes
Stop

Fig. 2 PSO adaptive testing flowchart.

4. Experiments and Results

4.1 Searching Speed

This experiment aims to observe the search time for the computerized adaptive dynamic item selection model proposed in this study over different sizes of test item banks with different parameter settings, and to compare them with the search time of a sequential search. The implementation method for this experiment is to observe the results of giving different numbers of particles and iterations to the computerized adaptive dynamic item selection model. There are 10 and 20 particles; 5, 10, 15 and 20 iterations; and 10 item selection processes are performed.

Fig. 3 and Fig. 4 compare the PSO and sequential search times. These two comparison figures show that, when the number of test items is below 1000, the average PSO and sequential search times are not significantly different; when the number of test items is over than 1000 and approaches 5000, the PSO search speed is significantly faster than sequential search. It proves that PSO search is effective for item selection in a big test item bank.

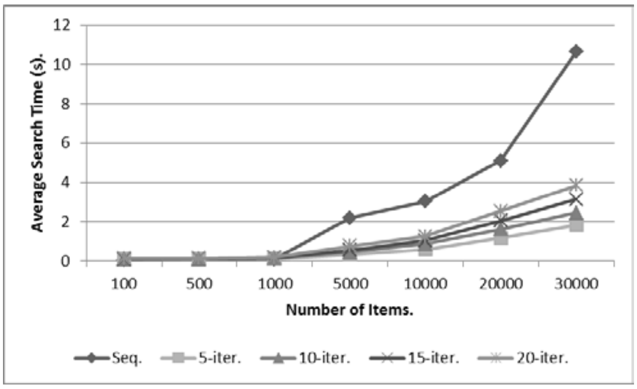

Fig. 3 Search time comparison of sequential search and PSO search (10 particles with different iterations).

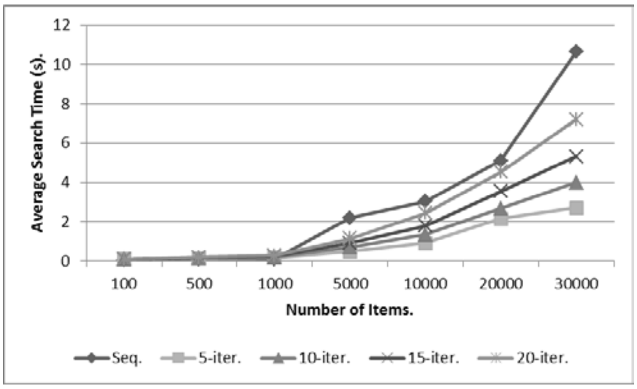

Fig. 4 Search time comparison of sequential search and PSO search (20 particles with different iterations).

4.2 Searching Accuracy

This experiment uses three search methods, PSO, random, and sequential searches, to perform item selection over item banks with different sizes and further compared the fitness values of the searched items. In this experiment, three different search methods are adopted to conduct 10 item selection actions over 7 different sizes of test item banks. The parameters for this experiment are w_1=0.5, w_2=0.3, w_3=0.2; the testee's abilities in knowledge blocks are all set to 0.5; the optimal fitness value is 0 and the worst fitness value is 2.

Fig. 5 shows that all fitness values selected by PSO searches are close to the optimal solution, except for the 5-particles and 5-iterations condition. Although a sequential search ensures that the optimal solution can be always found, the search speed experiments suggest that a sequential search increases the time cost and that the PSO search speed is significantly better than the sequential search speed, even in a big test item bank.

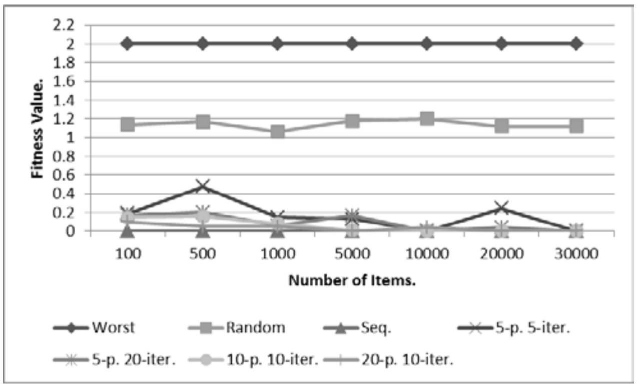

Fig. 5 Comparison of search accuracy rate. p=particles.

Fig. 6 shows that, along with the increased number of particles, less iteration is required to locate test items with the optimal fitness values. Exploring the reasons reveals that the particle distribution range expands when the number of particles increases, so the probability of an optimal solution located around the particles is greater and the opportunity to locate an optimal solution also increases. During the initial setup, the number of iterations can be correspondingly reduced; i.e., an optimal solution can be found without too much iteration. More can be found from the experiment: the search stability is more stable when applying a search scheme with 10 particles and iterations.

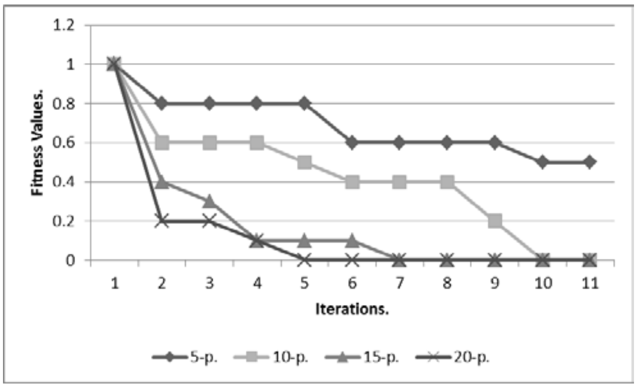

Fig. 6 Changes of fitness values of different number of particles in iterations.

4.3 Item Difficulty Index Estimation

The CAT system used in this study is developed in an online English learning system. That system provides the learning materials for technology English. And then, the CAT is used assessing the learners' outcomes (Cheng, Lin, & Huang, 2009). The item difficulty indices for the item bank in the test system is estimated by the method based on the answers abnormal rate proposed in this study. Combining the PSO dynamic item selection strategy in that system (Huang, Lin, & Cheng, 2009), a complete and robust CAT system is constructed.

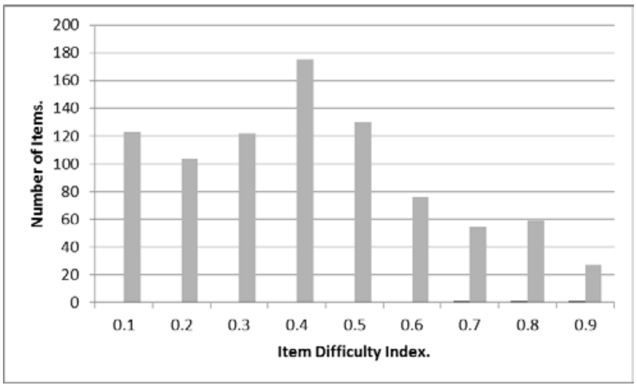

Fig. 7 Distribution of Item Difficulty Index.

This experiment in this study is online tests. The item difficulty indices are 9 levels ranged from 0.1 to 0.9. The initial values of item difficulty indices are all set to 0.5. The participants are the students, who elect the course named "Technology English", in the departments related with computer and information in a university in southern Taiwan. The students' abilities are also divided into 9 levels ranged from 0.1 to 0.9. The initial values of student's abilities are all set to 0.2. The experimental period is 6 weeks, and the way for the tests is that the students exercise in the after-school time freely. Then, the item difficulty indices will automatically be estimated every week according to the results of tests. Fig. 7 shows the distribution of the item difficulty indices in the test item bank after this experiment.

Although there are not enough test data, it needs more data for corroboration, it can be seen that if the selected times of a test item reaches a certain number of candidates, the results of estimation will be stable. Fig. 8 shows the number of adjusted test items in each time of estimation during this experiment. It can be seen that the number of adjusted test items is decreasing quickly. Fig. 9 shows the average adjusted levels of item difficulty indices in each time of estimation. Their values fall between 0.1 and 0.2. It represents that the average gap of item difficulty indices in each time of estimation is not too large.

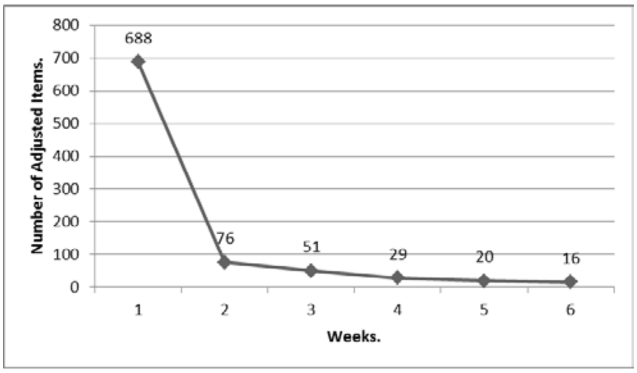

Fig. 8 Number of Adjusted Items for Each Week..

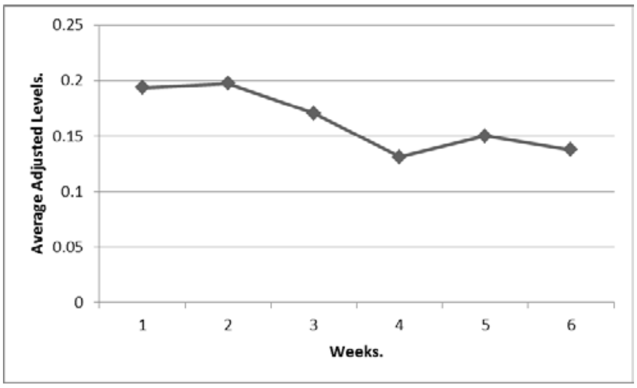

Fig. 9 Average Adjusted Difficulty Levels for Each Week.

5. Conclusions and Future Works

In the method of item difficulty index estimation based on the answers abnormal rate proposed in this study, the testee's abilities are considered into the process of estimation. The item difficulty indices and the testees' abilities can be estimated mutually at the same time. It can accelerate the process of item difficulty index adjustment estimation to be stable. Every test item is deemed to be independent, so its item difficulty index can be estimated independently. The item bank can be expanded easily at any time. New test items and old existing ones work together in the system. Their item difficulty indices can be estimated quickly and reasonably.

The dynamic item selection system adopts the PSO algorithm as a core and integrates with the knowledge structure concept. The quick search advantage of the PSO algorithm and the characteristics of knowledge structures allow the most suitable test items for a testee's ability to be selected quickly, even in a big test item bank.

In this study, we only discuss with the estimation of item difficulty indices. However, with respect to the mathematical model of IRT, it will be more suitable to discuss with the item discrimination indices and the item guess indices for the type of choice items. The description of item parameters can be more complete by them. Therefore, the item discrimination indices and the item guess indices will be discussed and researched based on the answers abnormal rate in the future. The participants of this study are university students in the departments related with computer and information. Thereby, the results of experiments are case dependent. In the future, the participants of experiments can involve different field of departments to obtain more general results.

Acknowledgements

This research was partially supported by the National Science Council, Taiwan, ROC, under Contract No.: NSC 100-2511-S-218-008-MY3 and NSC 101-2511-S-218-004-MY2.

References

[Anatchkova 2009] Anatchkova, M. D., Saris-Baglama, R. N., Kosinski, M., & Bjorner, J. B., Development and Preliminary Testing of a Computerized Adaptive Assessment of Chronic Pain, The Journal of Pain, 10(9), pp. 932-943, 2009.

[Badaracco 2013] Badaracco, M. & Martínez, L., A fuzzy linguistic algorithm for adaptive test in Intelligent Tutoring System based on competences, Expert Systems with Applications, 40(8), pp. 3073-3086, 2013.

[Cheng 2009] Cheng, S.-C., Lin, Y.-T., & Huang, Y.-M., Dynamic question generation system for web-based testing using particle swarm optimization, Expert Systems with Applications, 36(1), pp. 616-624, 2009.

[El-Alfy 2008] El-Alfy, E.-S. M. & Abdel-Aal, R. E., Construction and analysis of educational tests using abductive machine learning, Computers & Education, 51(1), pp. 1-16, 2008.

[Haladyna 1999] Haladyna, T. M., Developing and validating multiple-choice exam items (2 ed.), Mahwah, NJ: Lawrence Erlbaum Associates, 1999.

[Huang 2009] Huang, T. C., Cheng, S. C., & Huang, Y. M., A Blog Article Recommendation Generating Mechanism Using an SBACPSO Algorithm, Expert Systems with Applications, 36(7), pp. 10388-10396, 2009.

[Huang 2008] Huang, T. C., Huang, Y. M., Cheng, S. C., Automatic and Interactive e-Learning Auxiliary Material Generation Utilizing Particle Swarm Optimization, Expert Systems with Applications, 35(4), pp. 2113-2122, 2008.

[Huang 2009] Huang, Y.-M., Lin, Y.-T., & Cheng, S.-C., An adaptive testing system for supporting versatile educational assessment, Computers & Education, 52(1), pp. 53-67, 2009.

[Kennedy 1995] Kennedy, J. & Eberhart, R. C., Particle Swarm Optimization, Proceedings of the IEEE International Conference on Neural Networks, 4, pp. 1942-1948, 1995.

[Lin 2010] Lin, Y. T., Huang, Y. M., & Cheng, S. C., An Automatic Group Composition System for Composing Collaborative Learning Groups Using Enhanced Particle Swarm Optimization, Computers & Education, 55(4), pp. 1483-1493, 2010.

[Musii 2011] Musii, L., Daolio, F., & Cagnoni, S., Evaluation of parallel particle swarm optimization algorithms within the CUDA architecture, Information Sciences, 181(5), pp. 4642-4657, 2011.

[Suen 1990] Suen, H. K., Principles of exam theories, Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.