Abstract - We propose an automated construction method of multiple equivalent test forms indicated by test information function based on item response theory, and the method reduces the probabilities of selecting redundant items to the same test form. In previous studies, although their methods minimized the difference between the test information functions of the constructed test forms, they neglected the content similarities of selected items in the same test form. Therefore, the similar content items have probabilities to be selected to the same test form. This affects the test reliability. The main idea of this paper is to reduce the probabilities by applying a latent Dirichlet allocation in the test construction method to detect the content similarities between the selected items and the remaining items in the item banks.

Keywords – Automated Test Construction, e-Testing Forms, Latent Dirichlet Allocation, Multiple Equivalent Test.

I. INTRODUCTION

Educational assessments sometimes need “multiple test forms” in which each form consists of a different set of items but still has qualities that are equivalent (e.g., rate of correct answers, equivalent amounts of test information based on item response theory) to the others. For example, multiple test forms are needed when a testing organization administers a test in different time slots. To achieve this, multiple test forms are constructed in which all forms have equivalent qualities so that examinees, who have taken different test forms, can be objectively evaluated on the same scale.

In order to construct multiple test forms, e-testing, which accomplishes automated test construction, has recently become popular in research areas involving educational measurement (e.g. [1], [2], [3]). The methods in previous studies have been used to construct all forms of a test to satisfy the same test constraints (e.g., the number of test items and the amount of test equivalent qualities. van der Linden and Boekkooi-Timminga [4] proposed a sequential method of constructing test forms using linear programming to minimize the fitting errors to the test constraints. While, Boekkooi-Timminga [5] and Armstrong, et al. [6]
proposed methods that simultaneously constructed all test forms to minimize the differences in the fitting errors on the test forms. The former used linear programming and the latter used network-flow programming. van der Linden and Adema [2] and van der Linden [3] proposed methods that sequentially constructed test forms by minimizing the difference in fitting errors between a currently constructed test form and the remaining set of items in the item bank. Although the reviewed methods minimized the differences among the fitting errors on the test forms, these methods neglected considering the content similarities of the selected items in the same test form. Therefore, in actual constructions of multiple test forms, the similar content items have probabilities to be selected to the same test form. When the similar content items are selected to the same test form, they are redundant and might become hints for the other item(s). This reduces the reliability of the test form.

The main idea of this paper is to solve the problem by applying a natural language processing technique to detect the content similarities of items while constructing test forms. Recently, in natural language processing, one of the famous statistical models, which can be applied to detect the content similarities of items, is the latent Dirichlet allocation (LDA) [7]. Several researches (e.g. [7], [8]) showed LDA provided better performance than the previous statistical models such as Latent Semantic Analysis (LSA) [9] and probabilistic Latent Semantic Analysis (pLSA) [10]. Therefore, we apply LDA to detect the content similarities of items in the multiple test construction.

However, it is well known that the multiple test construction has the trade-off problem between the equivalent of test forms and computational costs. To alleviate the trade-off problem, Songmuang and Ueno [11] proposed a Bees algorithm for multiple test construction (BA) that applies a parallel-computing technique that distributes the computational costs to multiple processors without increasing the differences in fitting errors among test forms.

Therefore, in this approach, we apply LDA in BA to detect the item content similarities, reduce the number of similar content items in the same test form, and construct multiple test forms in a realistic time. The proposed method constructs multiple test forms to minimize the probability of selecting similar content items in the same test form and to minimize the difference of the fitting errors to the test constraints between test forms.

Moreover, we perform an experiment using actual data to show the performance of the proposed method to reduce the number of similar content items in the same test form. The results show that the number of similar content items in the test forms constructed by the proposed method is smaller than that of the test forms constructed by traditional test construction methods.

II. CONSTRUCTION METHOD OF MULTIPLE EQUIVALENT TEST FORMS BASED ON BEES ALGORITHM COMBINING LATENT DIRICHLET ALLOCATION

A. Bees algorithm for multiple test forms construction

In this section, we describe a method of construction test forms based on BA [11] that constructs multiple equivalent test forms by minimizing the difference in fitting errors between test forms. The construction of multiple test forms is classified as an NP-hard problem. To reduce the computational time, we divided the construction of test forms into two steps:

Step A: Construct test forms only to minimize the fitting errors of each form to test constraints without taking into consideration the equivalence of test forms. Here, the constructed test forms are still not equivalent.

Step B: Extract the most equivalent set of test forms from the constructed test forms in Step A that minimizes the difference in fitting errors between test forms.
The proposed construction of multiple test forms based on BA is implemented that includes one server and several workers.

Using a parallel-computing technique, the computational cost of constructing the test forms for each processor core is calculated by dividing the computational cost by the number of processor cores. Therefore, we can decrease the computational time by increasing the total number of processor cores of workers. As a result, we can relax the trade-off by using the proposed method and the parallel computing technique.

However, this method neglects the content similarities of the selected items in the same test form. Next, we introduce LDA and the idea to detect the item content similarity using LDA.

### B. Latent Dirichlet allocation

In natural language processing, many research proposed statistical models (e.g. TF-IDF [12], LSA, pLSA, LDA) for analyzing the similarities among documents using text data. The proposed statistical models were applied in several applications such as separating the group of similar documents, searching for a similar document, analyzing balance of number of test items in an item bank.

Moreover, several researches (e.g. [7], [8]) compared the effectiveness of the proposed statistical models and the results showed that LDA provided better performance to separate the group of similar documents. Therefore, in this paper, we apply LDA to evaluate the similarities of items.

LDA is a generative model in natural language processing. The basic idea of this model is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The graphical model of LDA is shown in Fig. 1, where

- $\alpha$: the parameter of the Dirichlet prior on the per-document topic distributions
- $\beta$: the parameter of the Dirichlet prior on the per-topic word distribution
- $\theta$: a topic distribution for each document
- $\phi$: a word distribution for topic
- $z$: a topic for a word in a document
- $K$: a total number of topics
- $w$: a specific word
- $N$: a total number of specific words
- $M$: a total number of documents

The specific word is only one observable variable, while the other variables are latent variables that are estimated and are inferred from the specific word.

One of the popular techniques for parameter estimation and inference are Gibbs sampling method [13] and we use this method in this approach.

After we estimate the variables of LDA, we infer a topic of a document using the topic distribution of the document. The similar documents, which have similar topic distributions, belong to the same topic. However, the documents in the same topic are not always similar.

As a result, it is difficult to indicate the similarities of documents using only the inferred topics of the documents. Therefore, the idea to apply LDA to evaluate the similarities of items is that we compare the topic distributions of items instead of the inferred topics of items. Next, section we describe how to apply LDA to reduce the number the similar content items in the same test form in BA.

### C. Bees algorithm with latent Dirichlet allocation for constructing multiple test forms

In this approach, we apply LDA in BA to minimize the probability of selecting similar content items of each test form in multiple test constructions by install LDA in step A of BA. For more details, in step A of BA, BA
generates the selection probabilities of items inversely proportional to the content similarities of the selected items in the test form and the remaining items in the test form and the remaining items in an item bank. The original selection probability, \( p \), of BA for item \( I \) after \( t \) items are selected to the constructing test form is as follows:

\[
p_t = \frac{d_j}{q_{ij}} \sum_{i \in A} q_{ij}^{-1},
\]

where \( q \) is an item-selection coefficient which increases with the fitting errors between the test constraints and a constructing test from and \( d \) is a binary variable that equal zero if item \( i \) does not satisfy any test constraint or is one otherwise. Here, \( A \) is the set of indexes of remaining items in the item bank.

Here, we combine LDA with the function (1) of BA to minimize the probability of selecting similar content items to the same test from. The item-selection probability for item \( I \), after \( t \) items are selected to the constructing test form, is as follows:

\[
p_t^{(BA)} = \frac{d_j}{q_{ij}} \sum_{i \in L_t} \left( \frac{\sum_{i \in L_t} \sum_{i \in L_t} |\theta_{ik} - \theta_{ik}|}{\sum_{i \in L_t} d_i / q_{ij} \sum_{i \in L_t} \sum_{i \in L_t} q_{ij} - \theta_{jk}} \right)
\]

where \( L_t \) is the set of items in the constructing test form, \( \theta_{ik} \) is the value of the topic distribution of item on the topic \( k \). The expression in (2) means the difference between topic distributions of the items in the constructing test form \( L \) and the topic distribution of the remaining item \( i \). When the difference is large, the content of the remaining item is different from the contents of the selected items in the constructing test form. On the other hand, when the difference is small, the content of the remaining item is similar to the contents of the selected items in the constructing test form. Therefore, the item-selection probability increases with the difference between the topic distributions of the items in the constructing test form and the topic distribution of the remaining item.

Next section, we perform an experiment to evaluate the performance of the proposed method to detecting similar content items while constructing multiple test forms.

**III. EXPERIMENT**

In this experiment, we used actual data to evaluate the performance to detect similar content items of the traditional test construction methods and the proposed method. The item bank related to Theory of Distance Education had 72 items, each item had 10-25 words, and corpus had 344 words. The rates of correct answers of items were random between 0.25 and 0.92.

The multiple test construction methods were as follows:
1. Bees algorithm,
2. Bees algorithm with TF-IDF,
3. Bees algorithm with LDA.

We used the three construction methods to construct ten multiple test forms with five item for each test form.

Here, the Bees algorithm with TF-IDF was a multiple test construction based on BA and moreover, we installed TF-IDF to detect the similarities of item contents. TF-IDF is a weight used to evaluate how important a word is to a document in a corpus, which proportionally increases with the number of time a word appears in the documents but is offset by the frequency of the word in the corpus. We used the weight distributions of documents to indicate the similarities of item contents instead of the difference between topic distributions of the items in function (2).

Before constructing test forms, we defined the test constraints as follows:
1. Each test has 10 items,
2. Minimum and maximum rates of correct answers of test are between 0.45 and 0.65.

The minimum and maximum rates of correct answers of test is a test constraint that indicates the test quality. In this experiment, if the test forms satisfied all test constraints,
we assumed that the test forms were equivalent. After we constructed the multiple test forms to satisfy the test constraints using the three test construction methods, we compared the similar content items in the same test forms. Here, a test content expert defined the sets of similar content items. The experiment was repeated 100 times.

Fig. 2 shows the average numbers of similar content items in the same test forms that are constructed by the three test construction methods. The results show that the number of similar content items in the test forms constructed by the proposed method is smaller than that of the test forms constructed by the traditional methods. This indicates that the proposed method provides better performance to detect the content similarities than the traditional methods.

IV. CONCLUSIONS

We proposed a construction method for multiple equivalent test forms based on Bees Algorithm with latent Dirichlet allocation in which this method detects the item content similarities, reduce the number of similar content items in the same test form, and construct multiple test forms in a realistic time. Moreover, we performed the experiment using actual data to compare the performances of the proposed method and the traditional methods to reduce the number of similar content items in the same test form. The results showed the number of similar items in the constructed test by the proposed method was smaller than that of the traditional methods. This means the proposed method provides better performance to detect the content similarities than the traditional methods. However, we note that the performance to reduce computational costs of the proposed method was not evaluated since it was difficult to show the significant difference of the computational costs using the actual data in this paper.

REFERENCES