Exploring the Relationships between Reading Behavior Patterns and Learning Outcomes Based on Log Data from E-Books: A Human Factor Approach

Chengjiu Yin, Masanori Yamada, Misato Oi, Atsushi Shimada, Fumiya Okubo, Kentaro Kojima & Hiroaki Ogata


To link to this article: https://doi.org/10.1080/10447318.2018.1543077

Published online: 16 Nov 2018.
Exploring the Relationships between Reading Behavior Patterns and Learning Outcomes Based on Log Data from E-Books: A Human Factor Approach

Chengjiu Yin*, Masanori Yamada, Misato Oi, Atsushi Shimada, Fumiya Okubo, Kentaro Kojima, and Hiroaki Ogata
Faculty of Arts and Science, Kyushu University, Japan

ABSTRACT
Online learning environments presently accumulate large amounts of log data. Analysis of learning behaviors from these log data is expected to benefit instructors and learners. This study was intended to identify effective measures from e-book materials used at Kyushu University and to employ these measures for analyzing learning behavioral patterns. In an evaluation, students were grouped into four clusters using k-means clustering, and their learning behavioral patterns were analyzed. We examined whether the learning behavioral patterns exhibited relations with the learning outcomes. The results reveal that the learning behavior of “backtrack” style reading exerts a significant positive influence on learning effectiveness, which can aid students to learn more efficiently.

1. Introduction
With the development of e-publishing technologies and standards, an increasing number of traditional textbooks have been replaced by e-books (Rainie, 2012). Compared with traditional textbooks, e-books provide numerous benefits such as cost-saving, quick accessibility from the Internet, low space consumption, and higher portability (Shepperd, Grace, & Koch, 2008).

Moreover, many countries plan to use e-books in schools (Nakajima, Shinohara, & Tamura, 2013). For example, the Korean Education and Research Information Service announced a digital textbook usage plan in 2007 (Shin, 2012), while the Japanese government is scheduled to modify all its textbooks for elementary, middle, and high schools into digital textbooks by 2020 (Yin et al., 2014).

By using e-books, a significant amount of logged educational data can be created. These log data are a recording of learning practices such as reading, writing, test-taking, and performance of various tasks in actual or virtual environments with peers (Mostow, 2004). To date, research on e-book logs in the classroom has received limited academic attention. Therefore, our research is focused on the log data of student activities.

Learning analyses have been undertaken utilizing log data. Researchers have reported that log data can positively relate to student performance and outcomes. (Iglesias-Pradas, Ruiz-de-Azcárate, & Agudo-Peregrina, 2015; Campbell, DeBlois, & Oblinger, 2007; Macfadyen & Dawson, 2012; Archer, Chetty, & Prinsloo, 2014; Hrastinski, 2005; Zhou, Hu, Zhang, Xu, & Chen, 2013). Several studies have utilized frequency measures of student log data, such as the number of markers, frequency of logins, and time spent reading pages, which are the most typical measures used to explain individual variations in online learning styles (Morris, Finnegan, & Wu, 2005; Yamada et al., 2015; Yin et al., 2015, 2017; You, 2016).

Researchers have indicated that the analysis of user behaviors and experience can facilitate the design of learning systems, materials, or activities (Brajnik & Gabrielli, 2010; Law & Larusdottir, 2015; Sutcliffe & Hart, 2016). For those e-learning behaviors that are clearly evident and can be easily recorded and used for statistical analyses, such as “add marker” and “write memo,” the analysis process is simple and straightforward. We call such behaviors “Observable Learning Behaviors (OLB).” However, certain learning behaviors cannot be easily observed and recorded; therefore, we call them “Hidden Learning Behaviors (HLB).” For example, “preparing for the lessons” and “backtrack reading.” It is difficult to observe these HLB and to determine their frequency of occurrence (Arroyo & Woolf, 2005; Zhou, Hu, Zhang, Xu, & Chen, 2013).

However, learning outcomes are generally affected by not only the visible measures, but also the hidden ones. Therefore, in this article, we propose the use of a large volume of e-book reading log data to visualize and identify effective measures that exhibit strong correlations with learning outcomes. These include HLB such as backtrack reading.

By visualizing the log data, we determined that certain students often return immediately to previous pages to review them or reflect on them. This behavior is referred to as the “backtrack” learning behavior pattern. Immediate reviewing helps students confirm that they understand the information, thereby reducing the time required to relearn it when they review it in the future (Mind Tools Ltd., 2015).

CONTACT Chengjiu Yin (yin.academic@gmail.com) Information Science and Technology Center, Kobe University, 1-1 Rokko-dai, Nada, Kobe Hyogo 657-8501, Japan.
*Present address: Information Science and Technology Center, Kobe University, Japan.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.
© 2018 Taylor & Francis Group, LLC.
To further examine whether backtracking learning behavior patterns relate to learning outcomes, cluster analysis was employed, wherein students were grouped using k-means clustering to analyze their learning behavior patterns.

2. Literature review

2.1. Educational data mining method selection

Educational data mining implies applying data mining techniques to educational data to resolve educational research issues. There are numerous data mining approaches such as Apriori, k-means clustering, and Support Vector Machine (SVM) (Table 1).

The Apriori problem: Apriori is a rule-based machine learning method for discovering association relationships among different variables (Piatetsky-Shapiro & Frawley, 1991). The Apriori algorithm is used to determine all frequent item sets. The challenge with Apriori is the generation of large candidate sets.

The k-means clustering problem: this involves grouping similar records together in a large multidimensional data set (Aggarwal & Yu, 1999). The records within a group exhibit high similarity to each other; however, they are dissimilar to records in other groups. Dissimilarity and similarity measures are based on the attribute values that describe the records and generally involve distance metrics (Han, Kamber, & Pei, 2011).

The SVM problem: SVM is a supervised method of assigning data items in a data set to target classes. The goal of classification is to predict the target class for each case in the data. For example, Huang, Chu, and Guan (2017) used SVM to predict learners’ performance. The SVM problem is closely related to the k-means clustering problem. The former is a supervised learning problem, while the latter is an unsupervised learning problem (Aggarwal & Yu, 1999).

The purpose of the present research is to classify students into various groups and then to identify the features of the groups. The students in each group exhibit strong similarities with each other. However, Apriori is used to identify frequent item sets among the different variables; therefore, it is not suitable for the present research. SVM can only classify students into two groups and is therefore not suitable for the analysis of the diversity of students in our sample. Therefore, we selected k-means clustering in this research.

Utilizing k-means clustering, Cheng and Tsai (2014) conducted a study to discover the behavior patterns of children and their parents when child–parent pairs collaboratively learned in the context of reading a picture e-book. They reported their success in conducting a k-means clustering analysis to identify the features of reading behaviors. Our research purpose was also to identify the features of reading behaviors. Therefore, k-means clustering analysis was conducted in the current work.

2.2. Previous studies of data collection

Collecting data is the first step in learning behavior analysis (Yin et al., 2013b, 2013a). In May 2015, we therefore performed a review of previous research to survey the categories that can be classified in terms of data collection (Yin et al., 2015). Based on the data source, previous studies on data collection could be classified into three categories (Table 2):

Questionnaire-based Data Collection (QDC). In this category, data are collected using a predesigned questionnaire. For example, Ho, Hung, and Chen (2013) used a questionnaire to investigate the teacher behavior of adopting mobile phone messages as a parent–teacher communication medium.

Manual Data Collection (MDC). In this category, after assessing the criticality of data, users select and save the valuable data manually. They generally employ a data collection system and consciously provide data on their learning behaviors. For example, Cheng and Tsai (2014) analyzed book reading patterns using an augmented-reality book, whereby the users’ learning behaviors during the reading process were videotaped manually.

Automatic Data Collection (ADC). In this category, learning behavior log data are automatically recorded while reading digital learning material. Goda et al. (2015) analyzed undergraduates’ learning behavior patterns in e-learning (Moodle, https://moodle.org/) and investigated their relationship to learning outcomes.

For categories QDC and MDC, the data are consciously collected. Therefore, the data are affected by the users’ own subjective factors. For category ADC, the data are objectively collected, thereby eliminating the subjective factors that affect data authenticity. The present work falls under category ADC. An e-book system was used for this research. In addition, students’ learning behavior log data will continue to be recorded while they read lecture materials.

Table 1. Data mining method comparison.

<table>
<thead>
<tr>
<th>Features Finding</th>
<th>Apriori</th>
<th>k-Means</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association rules</td>
<td>Many groups</td>
<td>Two groups</td>
<td>Exhibit similarity</td>
</tr>
<tr>
<td>Frequent item sets</td>
<td>High similarity</td>
<td></td>
<td>Supervised learning</td>
</tr>
<tr>
<td>Machine learning method</td>
<td>Unsupervised learning</td>
<td>Unsupervised learning</td>
<td>Supervised learning</td>
</tr>
</tbody>
</table>

Table 2. Comparisons with the previous studies.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.3. Previous studies of learning analytics

Learning analytics is an emerging discipline that is concerned with developing methods for exploring the unique types of data that are derived from educational settings and using those methods to more effectively understand student learning (Baker & Yacef, 2009; Yin et al., 2013b). Long and Siemens (2011) demonstrated that learning analytics aids in the analysis and reporting of data on learners and their contexts. This information can foster a more effective understanding of learning processes and optimize learning and its environments. The primary aim of learning analytics is to enhance learning outcomes and the overall learning process in computer-supported education. Learning analytics can aid teachers and learners to search for unobserved patterns and underlying information in learning processes (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014). Analyzing learning behavior patterns is a critical topic in learning analysis.

A few attempts have been made to identify learning behavioral patterns by HLB and to examine the relationship between these learning behavior patterns and the learning outcomes. For example, a few researchers analyzed HLB that were manually video-recorded (Cheng & Tsai, 2014); however, it is challenging to record students’ HLB over a long period of time without prompting their awareness of the recording.

Three previous studies are presented in Table 2. All of these used a marginal amount of data, while a large amount of data was used in the present research (approximately 400,000 records). Ho et al. (2013) collected data by using a questionnaire, and used statistical methods to analyze the data. Cheng and Tsai (2014) video-recorded data and used clustering to analyze “Book reading action patterns.” Goda et al. (2015) collected data from an e-learning system and used statistical methods to analyze the “Learning pace patterns” of the learners. In the present study, an e-book system was used to collect data, and the clustering method was used to analyze “Backtrack reading patterns.”

To summarize, in previous studies, the researchers did not analyze learning behavior patterns using hidden measures (i.e., backtracking reading patterns). Furthermore, a few of the researchers used a marginal amount of data. The present study, on the other hand, focused on using e-books to collect student learning log data without leakage and visualizing the HLB. We then analyzed the learning behavior patterns using a large amount of data, which were objectively accumulated. The objectives of this study are listed below:

1. Identifying the types of HLB from the log data.
2. Analyzing students’ possible behavior patterns by employing cluster analysis.
3. Determining the correlations between students’ learning behavior patterns and learning achievements.

3. Method

3.1. Participants and context

The data used in this study were collected during an information science course in October 2014 at Kyushu University in Japan. The learning goal of the course was to teach students the basic principles of information and communication technology, including information, calculation, communication, intelligence, and the algorithms of computer science.

The students were providing the teaching materials for the subsequent class and were asked to prepare the lesson before the subsequent class. A hundred and eight freshmen (aged 18–19) attended the course. Ten students were removed from the study sample after data processing, including drop-out students among others. The data from the remaining 98 students were used. Among these, 22 (22.4%) were female and 76 were male (77.6%). Prior to entering the class, they had no previous experience of using BookLooper, an e-book system that records students’ learning behaviors when they read e-books. Of these 98 students, 3 were from the School of Education, 10 were from the School of Letters, 13 were from the Faculty of Science, 3 were from the School of Medicine, and 69 were from the Faculty of Engineering.

3.2. Data collection and measures

The data were collected from BookLooper. As illustrated in Figure 1, the instructors and students could access these two systems using their smartphone or laptop from all locations within or outside the campus. Through BookLooper, they could perform actions such as “open learning content,” “turning to the next page,” “returning to a previous page,” “adding a bookmark,” “adding a marker,” and “writing a memo.” All actions using BookLooper were recorded in a database (Figure 1).

Table 3 presents a sample of reading action logs. Each data log contains date, time, user ID, learning content ID, page number, user action, and other data. BookLooper was used in the information science course. A total of 400,000 records were gathered from October 1, 2014 to January 31, 2015.

Data processing: The following data processing rules were determined by five teaching-material creators who have over 10 years of teaching experience:

(a) Invalid reading time: If a student spends less than 5 s on one page, then, the student did not read the page.
(b) Invalid record: If the time variation between two actions is longer than 20 min, then, the record is invalid. It implies that the student did not read the contents as he/
she did not conduct any action within 20 min.

(c) Invalidity preview: If a student did not preview the lesson (read the learning content before class) up to 3 min before the class, then, he/she is considered to have not done a preview of the lesson.

The e-book measures: The measures from the BookLooper data were Reading Pages (RP), Preview Times (PT), Read Time (RT), Number of “Next” (NN), and Number of “Prev” (NP):

- NN: The number of times a student turns to subsequent pages.
- NP: The number of times a student returns to previous pages.
- PT: The number of times a student previews the lesson before class.

All the teaching materials were uploaded to BookLooper. Students could preview the learning content before class. The reading action logs for “Action Time” listed the students who engaged in this behavior, and we calculated the number of times this occurred. RT was used to determine whether students previewed their lessons.

- RP: The total number of pages that a student read. The reading action logs for “Page No.” and “Action Time” listed the number of pages the students read. Many of them repeatedly read specific pages.
- RT: The total time spent reading the learning contents. The reading action logs “Action Time” listed the lengths of time students spent reading the learning content. RT was calculated on an hourly basis.
- Backtrack reading rate (BRR): BRR is a hidden measure, which is also calculated using NN and NP. This will be described in the following section.

3.3. Backtrack reading rate

As mentioned earlier, certain learning behaviors are hidden and unconscious, and it is therefore challenging to determine the frequency with which individuals engage in these unobserved learning behaviors (Arroyo & Woolf, 2005); therefore, we classified learning behaviors into two categories: OLB and HLB. Arroyo and Woolf (2005) pointed out that OLB—the outward learning behavior of students, such as “problem-solving time,” mistakes, and help requests—is convenient to record. On the other hand, HLB—the inward learning behavior of students, such as learning attitude, habits of preparing for the lesson, and learning methods—is unobserved and challenging to identify.

In the present study, we refer to e-book log entries such as “spending time reading pages,” “turning to the next page,” and “reading pages” as OLB as they are conveniently recorded, while we refer to learning behaviors such as “preparing for the lesson” and “backtrack reading” as HLB. HLB needs to be calculated by using OLB measures.

Data visualization technologies can be used to determine reading strategies by reading the action logs of e-books. By using visualization technologies, we visualized students’ actions using the “Action time,” “Page No.,” “Prev,” and “Next” logs. We observed that several students recorded a number of “Prev” actions, indicating their frequent review of previous pages; they often backtracked while reading. We thus refer to them as backtracking learners. Meanwhile, other students displayed higher number of “Next” actions, indicating that they read the pages of the learning content in sequence (Yin et al., 2015).

We define Equation (1) to compare the number of “Prev” and “Next” actions in order to calculate the BRR of backtrack learners, where BRR is a decimal number from 0 to 1. If the number of “Prev” actions is 0, then, the PR is 0; if the number of “Next” actions is equal to that of “Prev,” then, the PR is 1.

\[
BRR = \frac{num(Prev)}{num(Next)}
\]

num(Next) represents number of “Next”; num(Prev) represents number of “Prev”

3.4. Pretest and post-test achievement

Two examination scores were used in this study. The students were asked to take a few quizzes at the beginning of the course, and the scores in these quizzes were treated as their pretest achievement. The students were also asked to take a few quizzes during the 10 lessons of the course; the sum of these quiz scores was treated as their PTA.

Pretest achievement: At the beginning of the course, the students took quizzes that assessed their fundamental knowledge of the course topic. The quizzes were administered using Moodle (a free and open-source software learning-management system), and students could answer the questions multiple times. A record of the first attempt was used to calculate the quiz scores. These scores (ranging from 0 to 10) were used as a pretest achievement baseline.

Post-test achievement: We then administered quizzes at the end of the 10 course lessons. The sum of the 10 scores was used as the PTA rating (ranging from 0 to 100).

3.5. Correlation analysis

The PTA was used to measure the learning outcomes. We analyzed a number of variables that could affect PTA, including the behaviors and their related variables (time spent reading pages, reading pages). SPSS (IBM SPSS Statistics, New York, USA) was used to determine the partial correlation of PTA with these variables. Table 4 lists the results, which indicate that the variable PTA exhibits a significant positive correlation with BRR, PT, RP, and RT. In addition, based on the results of partial correlation, a $k$-means clustering analysis was conducted to cluster students into groups to analyze the features of the learning behaviors in each group.

3.6. K-means clustering

To further understand students’ possible behavior patterns, cluster analysis was employed. Students were clustered into groups according to the similarity in their learning behaviors. We then analyzed the learning behavior features in each group.
The main problem of k-means clustering is the determination of k and selection of cluster centers. Firas-Matinez et al. (2007) analyzed users’ similar behaviors by k-means clustering. This method was used in the present study also. Equations 2 and 3 were used to determine k:

\[ y_i = \min \left( b_{i,m}, \ m = 1, \ldots, k \right) - d_i \]

\[ \max(d_i, \min \left( b_{i,m}, \ m = 1, \ldots, k \right)) \]

\[ q_k = \frac{\sum_{i=1}^{N} y_i}{N} \]

The initial cluster centers were randomly selected. Given the randomness of the original centers, k-means was run 100 times for each value of k. A minimized distance was selected from each data to its cluster center. Values of k = 2, …, 9 were assigned, and an algorithm was run using the Euclidean distance.

Figure 2 presents the evolution of the quality of the partitions obtained for the values of k tested. The optimum partition was obtained with k = 4 because the q-value (q_k) here was the largest. Therefore, the students were clustered into four groups. Four variables were used to cluster the students: RP, BRR, PT, and RT. All the variables had varied ranges: RP ranged from 10 to 540 pages, BRR ranged from 0 to 0.68, PT ranged from 1 to 7 times, and RT ranged from 0.18 to 21.62 hr. To eliminate the influence of each factor over another (i.e., to provide equal chances to the variables), these variables were transformed to within a specific range [0, 1] by min–max normalization, which enhanced the accuracy and efficiency of mining algorithms involving the distance measurements (Han et al., 2011).

4. Results

4.1. Clusters

Table 5 presents the center, mean values, and standard deviations of each cluster as well as comparisons of the post-hoc tests (Scheffe). Clusters 1–4 (C1, C2, C3, C4) included 25, 29, 14, and 30 students, respectively.

k-means clustering can divide students into various groups; nonetheless, it is still difficult to say that there are statistically significant between their variables. In order to examine the inter-cluster variations, one-way analysis of variance (ANOVA) was conducted for each measure with C4 as a between-subject factor (the data of the four clusters satisfied the ANOVA requirements).

Statistical differences were observed in the BRR among the four clusters (F (3,94) = 36.53, p < 0.001) as well as with regard to RT (F (3,94) = 63.53, p < 0.001), PT (F (3,94) = 120.29, p < 0.001), and RP (F (3,94) = 77.37, p < 0.001).

4.2. Relationships between clusters and learning outcomes

As mentioned above, a pretest was carried out to evaluate the students’ basic knowledge of information science, the topic of the course. A one-way ANOVA was performed on the pretest results, which revealed no significant differences (F (3, 94) = 1.486, p = 0.224 > 0.05) among the pretest results of the students in the four groups.

The PTA was used as the PTA rating. An analysis of covariance (ANCOVA) was performed on the post-test results in which the pretest was the covariance. These results together constituted the dependent variable. The various learning behavior clusters (four groups) together constituted the fixed factor and were used to test the relationships among the post-test results of the three groups.

<table>
<thead>
<tr>
<th>PTA (Post-test achievement)</th>
<th>RP</th>
<th>PT</th>
<th>RT</th>
<th>BRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation coefficient</td>
<td>0.728</td>
<td>0.417</td>
<td>0.681</td>
<td>0.652</td>
</tr>
<tr>
<td>Significance probability</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

RP: reading pages; RT: reading time; PT: preview times; BRR: backtrack reading rate.

Table 4. Partial correlation result.

Figure 2. k-Value determined by q-value.
The outcomes indicate statistical differences in learning efficacy among the clusters ($F(3, 94) = 21.045, p < 0.05$). The ANCOVA results are presented in Table 6; these demonstrate that the learning achievements of the students of C1 group were significantly better than those of C3 and C4, whereas no significant differences were revealed among the students in the C2 group.

4.3. Learning behavior patterns

Figure 3 illustrates four graphs of the learning behaviors with regard to reading the e-book. The graphs visualize the students’ actions using the “Action time,” “Page No.,” “Prev,” and “Next” logs. Each node of the graphs represents a page of the e-book. “P” is the abbreviation of “Page,” and the number after “P” is the page number; for e.g., P1 implies Page 1, and P2 implies Page 2. The node color has no significance as it is generated randomly.

The nodes are sorted by the order in which they are accessed. If a page is accessed first, then, it is displayed at the front; therefore, the nodes are not sorted by page number (it does not appear to exhibit sequence); this problem is corrected in the next version.

There are two types of lines between the nodes: green and blue. Each green line implies turning to the subsequent page, and each blue line implies returning to the previous page. Each line represents one access time; therefore, a larger number of lines indicates a higher frequency of reading the page.

Figure 3 (Figure 3: C1): C1 students exhibited significantly higher tendency to spend time reading numerous pages of the learning content compared with the C3 and C4 students (RP, RT: 1 > 3, 1 > 4). With regard to the frequency of previewing the lesson for class, C1 exhibited a higher frequency of previewing the lesson for class compared with C3 and C4 (PT: 1 > 2, 1 > 3, 1 > 4).
frequency than those of C2, C3, and C4 (PT: 1 > 2, 1 > 3, 1 > 4) and a higher learning achievement than those of C3 and C4 (PTA: 1 > 3, 1 > 4). C1 demonstrated a significantly higher backtrack rate than that of C4 (BRR: 1 > 4).

Of all the clusters, C1 exhibited excellent scores for all the variables. This implies that C1 students repeatedly previewed lessons and spent time reading content. These behaviors contributed to their high learning achievement (mean = 93.70, SD = 5.50). Thus, for its behavioral patterns, C1 is characterized as the "preview and diligent group."

C2 (n = 29, Figure 3: C2): C2 students also exhibited a significantly higher tendency to spend time reading and learning numerous pages of content compared with C3 and C4 students (RP, RT: 2 > 3, 2 > 4). C2 students reported significantly higher learning achievement than C4 students (PTA: 2 > 4). Although no significant difference was observed in the BRR between C2 and C4, C2 students spent significantly more time reading the learning content and thus obtained higher learning achievement (mean = 89.84, SD = 6.79). Accordingly, for its learning behavior patterns, C2 is deemed the “diligent group.”

There is no significant difference in the BRR of C1 and C2, with C2 students spending an equivalent amount of time as C1 for reading the learning content. This substantiates the similar learning achievement in both the clusters.

C3 (n = 14, Figure 3: C3): The reading time of C3 is not comparable to that of C1 and C2 (RP, RT: 3 < 1, 3 < 1). They did not preview lessons before class unlike C1 (PT: 3 < 1). Their learning achievement was significantly lower compared with C1 (PTA: 3 < 1). However, C3 exhibits a significantly higher difference for the backtrack rate compared with C1, C2, and C4 (BRR: 3 > 1, 3 > 2, 3 > 4). The C3 learning achievement is statistically higher than that of C4 (PTA: 3 > 4) and comparable to that of C2. Therefore, C3 can be identified as the "efficient group."

C3 demonstrated the highest BRR. C3 students reviewed previous pages numerous times and spent relatively shorter time learning; however, these students previewed lessons for class at a marginal frequency. Consequently, their learning achievement (mean = 80.22, SD = 8.79) was not the highest.

C4 (n = 30, Figure 3: C4): The BRR of C4 students exhibits a significant difference compared with the C1 and C3 students (BRR: 4 < 1, 4 < 3). Among all the students, those of C4 exhibited the least learning achievement (PTA: 4 < 1, 4 < 2, 4 < 3) as well as the lowest total count of reading pages and reading time (RP, RT: 4 < 1, 4 < 2, 4 < 3). Compared with C1 students, C4 students previewed the lessons at a marginal frequency (PT: 4 < 1). Therefore, owing to their behavior of only reading the pages of the learning content in sequence, C4 is identified as the "poor performance group."

C4 scored the smallest PR. C4 students read pages in sequence, and they spent minimal amount of time reading the learning content. Thus, they obtained the lowest performance (mean = 65.68, SD = 22.23).

### 4.4. Cluster comparison

Table 7 presents the cluster comparison results wherein C1 and C2 students are compared. Only the “time of previewing the lesson for class” exhibited a significant difference (PT: 1 > 2) between the two groups; no significant difference was observed in RP, RT, BRR, or SC. Thus, these students’ learning achievements were not affected by PT in the cases wherein reading times, reading pages, and BRRs were not significantly different.

Comparing the students of C3 and C2, significant differences were observed in the BRR (BRR: 3 > 2), Reading Pages (RP: 3 < 2), and reading time (RT: 3 < 2), although not in their learning achievement. C3 students tended to frequently review previous pages, clocked a shorter time for reading, and obtained satisfactory learning achievement (similar to C2 students). This observation demonstrates that BRR exerts a significant positive influence on learning effectiveness, and it aids students to manage their time for efficient learning. BRR exhibits a relevant correlation with learning efficiency and is thus a “good” reading strategy.

The comparison between C3 and C1 students exhibits significant differences in BRRs (BRR: 3 > 1), Reading Pages (RP: 3 < 1), reading times (RT: 3 < 1), and learning achievement (PTA: 3 < 1). The observations establish that although C3 students demonstrated an adequate reading style, it is necessary for them to spend more time reading the learning content in order to ensure higher learning achievement. This implies that an adequate reading style and sufficient learning time are simultaneously required.

### 5. Discussion

Analyzing student learning behaviors has been recognized as a critical issue in education research. This study aimed to determine meaningful measures from e-book materials and to employ these measures in the analysis of students’ learning-behavioral patterns. There are several innovations in this study, as discussed in the following subsections.

#### 5.1. Collecting data objectively

A majority of previous studies asked learners to provide feedback consciously, for e.g., by using questionnaires to collect data; consequently, the research results are likely to be affected by the learners’ personal subjective factors (e.g., Goda et al., 2015). In this study, an e-book system, BookLooper, was adopted. With the help of this system, students’ reading behaviors can be automatically recorded in their daily academic life. This implies that the data of the present work were objectively
collected so as to prevent subjective factors that are likely to affect the authenticity of the data.

5.2. Using visualization technologies to find HLB

The approach presented in this article is novel because similar studies have tended to focus on analyzing learning behavior patterns using VLB (Cheng & Tsai, 2014; Goda et al., 2015; Ho et al., 2013). In contrast to those studies, our research focused on using HLB, which is unobserved and challenging to identify. In order to discover HLB, we proposed using data visualization technologies to visualize the students’ actions and to determine the HLB.

5.3. Demonstrating the merit of HLB

This research was conducted to investigate significant learning behavior patterns and their effects on course achievement. A partial correlation analysis was conducted to identify the correlation of learning achievement with other variables such as the number of pages read, number of times a lesson was previewed before class, total time spent reading the learning content, and BRRs. The results demonstrate that these variables exhibit significant correlations with students’ learning achievement. Based on these results, BRR, PT, reading pages (RP), reading time (RT), and learning achievement were selected as the variables for k-means clustering.

A critical observation emerged from the analyses: “Backtrack-reading learning behavior was observed to exhibit merit; it can aid students to save time while studying.” This result supports the observations of previous research (Asarta & Schmidt, 2013; Jo & Kim, 2013; You, 2016) that emphasized the quality of learning behaviors rather than the quantity of learning.

It is noteworthy that backtrack reading learning behavior can be linked to a review learning strategy of allotting time to committing information to long-term memory (Lindsey, Shroyer, Pashlery, & Mozer, 2014). Furthermore, backtrack reading learning behavior can be linked to a reflection learning strategy of associating current knowledge with previous knowledge (Costa & Kallick, 2008). This observation can be used for enhancing the design of e-books. Between the related knowledge, teachers can add a link between the e-book pages to aid students in the performance of backtrack reading.

5.4. Proposing an algorithm for supporting group learning

The presented research proposed an algorithm to determine HLB and suitable measures and cluster students into varied groups; effectively, the proposed algorithm can be used for supporting group learning.

The formation of a learning group is a challenging problem, termed the Group Formation Problem. It is defined as the challenge of optimizing learning-group formation from a specified set of peers to form a balanced quality of the constructed groups (Konert, Burlak & Steinmetz, 2014). It is convenient to manually build learning-groups in the classroom; however, in the case of unsupervised learning environments (e.g., e-learning environments), manual group formation by lecturers is impractical, and algorithmic solutions provide significant support and are urgently required (Konert et al., 2014). Our algorithm can be considered as a type of algorithmic solution to the Group Formation Problem.

5.5. Recommendations for developing personalized e-books

Based on the analysis results, the following recommendations are provided to researchers who intend to develop personalized e-books to accommodate the requirements of students in type of groups:

(1) For the “preview and diligent group,” i.e., those students who tend to repeatedly preview lessons and spend time reading content, it would be appropriate to provide supplementary or extensive reading materials by incorporating a few “to know more” links on the relevant e-book pages. As these students study intensely and usually exhibit higher learning outcomes, the provision of additional learning materials or hints could further promote their learning performance.

(2) For the “diligent group,” i.e., those students who tend to spend time reading and exhibit satisfactory learning outcomes, it is inferred that the current e-book content satisfies their learning requirements and capability; therefore, no further modification of the e-book is required.

(3) For the “efficient group,” i.e., those students who tend to spend less time reading and exhibit satisfactory learning performance, it would be appropriate to incorporate additional interactive functions to guide and encourage them to spend additional time reading so as to achieve higher learning outcomes.

(4) For the “poor performance group,” that is, those students who spend marginal time reading and exhibit low learning outcomes, additional supports are required. In addition to incorporating additional interactive functions to encourage them to read, it is necessary to assist them in organizing the learning content by integrating graphical-knowledge-construction tools such as concept maps or knowledge maps into their e-books (Hwang, Sung, & Chang, 2016).

6. Conclusions

The goal of this exploratory study was to investigate whether unobserved learning behaviors can be identified from e-book log data and to determine the types of learning behavior patterns that are positive and significantly related to learning outcomes. To achieve these objectives, we used visualization technologies to identify unobserved learning behaviors. We then analyzed their learning behavior patterns using k-means clustering technologies.

The students were grouped into four clusters of varying learning behavior patterns: the “preview and diligent group,” “efficient
group,” “diligent group,” and “poor performance group.” The following observations emerged from the analysis results:

1. Backtrack reading learning behavior has merit because it can aid students in saving time while studying.
2. A reasonable learning behavior complemented by sufficient learning time can yield excellent learning results.
3. Students’ learning achievements are not affected by content-PT in the event that they spend sufficient time reading the learning content.

7. Limitations and future research

In this study, the number of participants was relatively small. We are thus currently collecting data from other courses and intend to increase the number of participants in a future study. We also intend to incorporate other measures from BookLooper and the Moodle log data. Furthermore, we analyzed students’ general learning behaviors in this study. A future effort is likely to delve into specific cases of learning behaviors among students. Such a study is also likely to differentiate learning behaviors adopted by students for various types of learning content.

Notes


Funding

A part of this research work was supported by the Grant-in-Aid for Scientific Research No. 16H03078 and No. 16H06304 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

References


international conference on intelligent tutoring systems (pp. 20–28). Berlin/Heidelberg, Germany: Springer-Verlag.


About the Authors

Chengjiu Yin is an associate professor of the Information Science and Technology Center, Kobe University, Japan. He received his PhD degree in Computer Science and Information Engineering from Tokushima University in Japan. His current research focuses interests on Educational Data Mining. He is a member of IEEE, JSET, and IPSJ.

Masanori Yamada is an associate professor in Faculty of arts and science, Kyushu University. He received PhD in Human System Science from Tokyo Institute of Technology in 2008. He is engaged in research and development of CSCL for project-based learning, based on social and educational psychology theories.

Misato Oi is a research fellow at Faculty of Arts and Science, Kyushu University, Japan. She received her Doctor of Philosophy from Nagoya University in 2010. Her research includes Computer Supported Ubiquitous and Mobile Learning, bilingualism, and non-verbal communication.

Atsushi Shimada received the DE degrees from Kyushu University in 2007. He is an associate professor of Faculty of Information Science and Electrical Engineering, Kyushu University. He also works as a JST-PRESTO researcher since 2015. His current research interests focused on image processing, pattern recognition, and learning analytics.

Fumiya Okubo is an assistant professor at Faculty of Arts and Science, Kyushu University, Japan. He received his PhD in Science from Waseda University in 2014. His research includes formal language theory, automata theory, and computation theory.

Kentarō Kojima is an associate professor in Faculty of arts and science, Kyushu University. He received MS and PhD from Kyushu University in 2005 and 2008, respectively. His recent research interests are Physics education research and development of CSCL.

Hiroaki Ogata is a full professor at Faculty of Academic Center for Computing, Kyoto University, Japan. He was a visiting researcher of Center of Lifelong Learning and Design, the University of Colorado at Boulder, USA from 2001 through 2003. His research includes CSUML, CSCL, CSCW, CALL, and Learning Analytics.