Shih, J. L. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

# Identifying Reading Styles from E-book Log Data

**Ivica BOTICKI [a*], Hiroaki OGATA[b], Karla TOMIEK[a], Gokhan AKCAPINAR[b], Brendan FLANAGAN[b], Rwitajit MAJUMDAR[b] & Nehal HASNINE[c]**
[a]*Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia*
[b]*Kyoto University, Japan*
[c]*Tokyo University of Agriculture and Technology, Japan*
*ivica.boticki@fer.hr

**Abstract:** In this paper, a model for identifying e-book reading style is proposed and applied onto a learning log dataset. Learning log data available as non-structured data source is processed to identify patterns of reading exhibited by users using three main structures: reading sessions, reads and passages. These structures are used to extract information on users' reading style to be used as part of user modeling process. The proposed model is applied on a set of log data generated by university students during one semester of digital resource use. The findings show students adopt predominantly receptive reading style, while responsive style occurs rarely. Further analysis revealed no significant relationships between reading style variables and student academic success for the Architecture course indicating the variables of responsive and receptive reading bring new information as part of user modeling.

**Keywords:** Reading styles, log data, user modelling, e-books

## 1. Introduction

E-books have become a popular medium for content delivery and are widely being accepted as an alternative to traditional paper-based sources. There is a number of advantages to e-books such as the convenience of use, greater interactivity etc. that make them valuable educational tools and resources today (Jagušt, Botički, & So, 2018). Nevertheless, more research is needed to ascertain how they affect users, and how can their interactive features be used in supporting learners.

There are two main log data subsets to be used as part of this study: the data on user reading and navigation within e-books, and the data on user interaction with e-books, such as making bookmarks, notes or memos. In order to prepare non-structured learning log data for the analysis, a process model with structures called sessions, reads and passages is utilized. Such data structures are used to model specific properties of e-book reading usage, facilitating further high-level e-book reading style analysis. It is to be noted that reading styles, as adopted in this paper, will be tied to contextual nature of e-book reading process. Such an approach is necessary to distinguish between potentially different conditions that arise from the course within an e-book content was being read, particular students using the e-book, or specific e-book resource being used. The contextual nature of the model is to allow for deeper understanding of the specifics of e-book use and its relationship with the identified user reading style.

## 2. Theoretical Background

Recognizing reading styles has been of great interest even before the appearance of cutting-edge technologies such as e-books and learning log mechanisms. Researchers had been trying to understand how humans read and thereby often relied on the use of methods such as the observation of reader physiognomic behavior, surveys and questionnaires, interview and diary studies (Freeman & Saunders, 2016; Marshall, 2009). Nevertheless, the issue of interfering with so called silent reading had presented a big methodological obstacle (Pugh, 1979), as reading is a personal and a reflective process.

Learning data logs are quantitative educational data, and they are used to meet the following objectives in the learning and teaching domains (Ogata et al., 2017): (1) Learning: Analyzing the details of behavior of "active learners" to make the students more active; and based on the relationships between log patterns and academic achievements, detecting the students who may drop out and those who will perform excellently; and (2) Teaching: Based on the logs made during a class session, improving course designs, which include collaborative learning and flipped classroom approaches and are based on the students' patterns of viewing e-books (e.g., understanding which page was frequently viewed), improving teaching materials and the structure of the e-books.

We adopt the approach developed in (Freeman & Saunders, 2016), where reading style is generated from learning log data. The authors build on work by Thayler at al. (Thayer et al., 2011) to further interpret reading styles originally defined in (Pugh, 1979) and devise the following classification:

- *Receptive reading* - reading sequentially from beginning to end with little variation in pace, to find out what an author has to say
- *Responsive reading* - active engagement with arguments in the text, with frequent changes of pace, pauses, rereading

In (Thayer et al., 2011) specific kind of e-books used as part of the study "were well suited to receptive reading, searching, and scanning, but did not support responsive reading and skimming well at all". We further explore the notion of receptive and responsive reading by using an e-book platform developed at Kyoto university (Flanagan & Ogata, 2018) with added interactive features, such as adding bookmarks, memos and markers.

## 3. Learning Log Processing Model

The process is based on the work presented in (Freeman & Saunders, 2016) and has several steps used to extract user sessions which represent contiguous block of e-book usage. Sessions are then used to create higher-level structures such as reads, passages and passage pairs. These are input into the algorithms for reading style identification, to be subsequently used as part of user modeling.

### *3.1 Sessions*

In order to allow for high-level analysis of learning log data, sessions, reads, passages and paired passages are used (Freeman & Saunders, 2016). Sessions occur naturally as users progress in reading an e-book. Every session presents one time-limited usage of an e-book in which its reader traverses e-book pages or interacts with the e-book by adding bookmarks, notes, memos and similar. Figure 2 gives examples sessions generated from e-book usage log data.

As illustrated in Figure 2, sessions are composed of continuous events that last for a limited period of time and include any possible sequence of page changes: they can represent reading page by page, include jumps forward or backwards etc. Session is a useful structure since it organizes learning log data according to user interaction with a specific e-book resource in a limited amount of time and usually within the same reading context (i.e. covering 10 pages in a mathematics book on Thursday evening).

### *3.2 Reads and Paired Passages*

Reads are used to merge sessions data coming from one user reading one resource. They are used to model the overall reading of a single resource done by one user and indicate reading jumps rather than concrete resource pages. Figure 3 shows reads structure created upon sessions laid out in Figure 2.

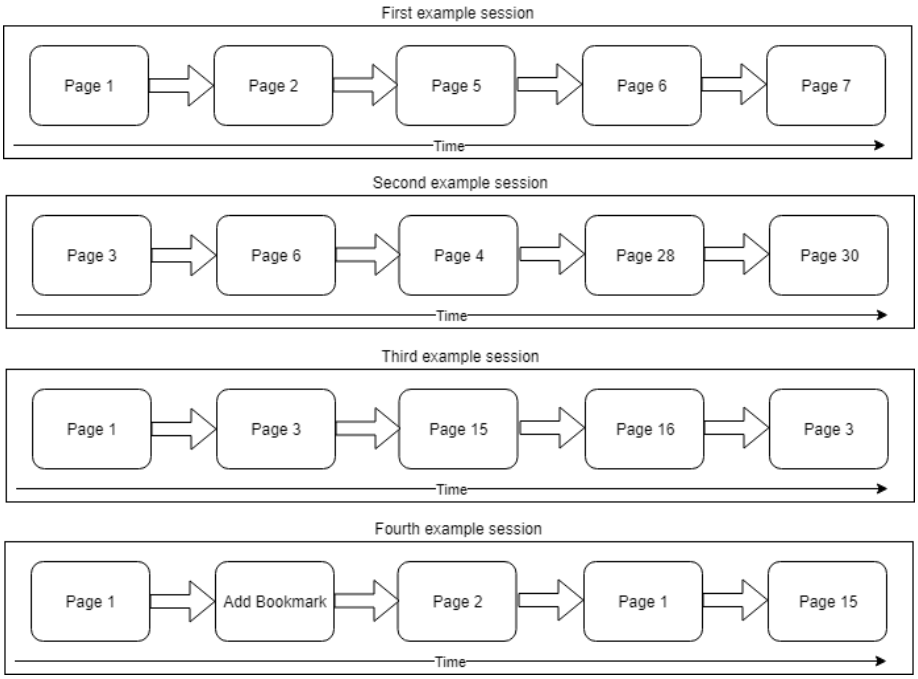


*Figure 2.* Example sessions extracted from learning logs.

In order to interpret the reads data in terms of reading dynamics (consecutive reading or jumps), interpretation notations names passages and paired passages are used (Freeman & Saunders, 2016). This notation focuses on the length and orientation of jumps differentiating between short, long, forward and backward jumps. Table 1 lists passages and paired passages created from reads in Figure 3.

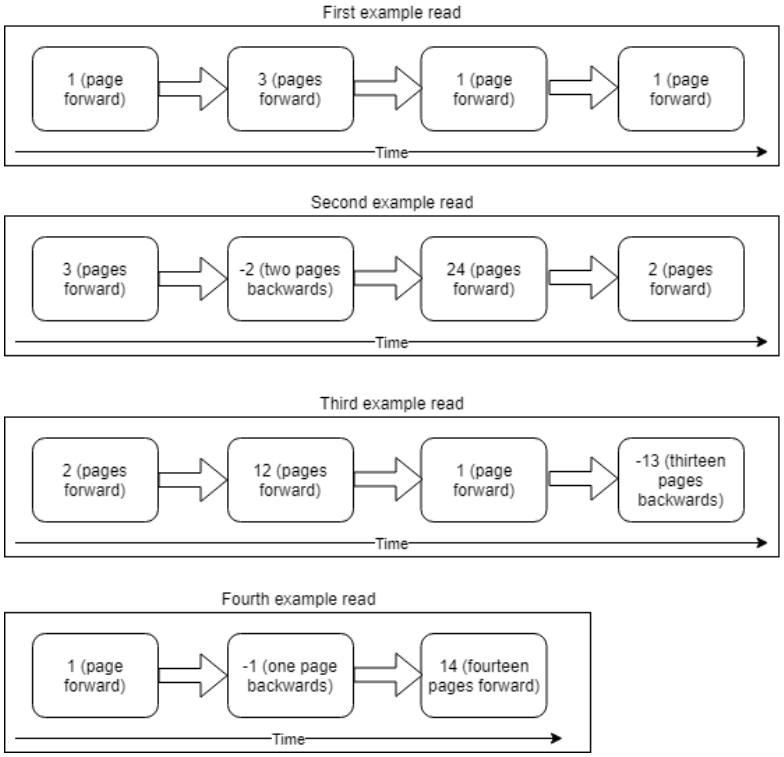


*Figure 3:* Reads created from sessions from Figure 2.

Passages are shortened representation of reads and they model five main jump categories: FOR (one or two pages skipped), SJF (small jump forward), BJF (big jump forward), SJB (small jump backwards) and BJB (big jump backwards). The first number of each passage is the number of pages

jumped. The second is the number of continuous pages read after the jump. Paired passages combine passages into pairs whereby immediate change in reading direction can be identified.

Table 1

*Passages and paired passages created from reads shown in Figure 3*

| Example | Passages | Paired passages |
|---|---|---|
| First example | FOR(1,1), SJF (3,3) | (FOR-SJF) |
| Second example | SJF(3,1), SJB(-2,1), BJF(24,2) | (SJF-SJB), (SJB-BJF) |
| Third example | FOR(2,1), BJF(12,2), BJB(-13,1) | (FOR-BJF), (BJF-BJB) |
| Fourth example | FOR(1,2), BJF(14,1) | (FOR-BJF) |

### *3.3 Extracting Reading Styles*

By following the proposed classification of reading styles, two variables are being modeled: the receptive reading variable and the responsive reading variable. Paired passages are primarily used to identify the level of receptive reading of students by calculating the proportion of forward-oriented paired passages in reading an e-book resource, which are generated from all recorded sessions in which user interact with the resource.

*Receptive reading (x)= number of forward oriented paired passages (x) / total number of paired passages (x)*
*where x stands for a specific time period.*

Responsive reading is extracted from e-book session data by using information on interactive events which appear during e-book session (i.e. creating memos, bookmarks or makers). It is modeled using the following formula:

*Responsive reading (x) = (number of memo operations (x) + number bookmark operations (x) + number of marker operations (x)) / total number of pages read (x)*
*where x stands for a specific time period.*

## 4. Case Study

In this study, data sample coming from the BookRoll e-book learning system developed at Kyoto University in Japan [1] is analyzed. The presented process model is applied onto one semester of usage log data generated by 407 students from 2017-09-09 to 2018-02-05. There are in total 6546 sessions generated from NoSQL log entries recorded while the students engaged in educational activities with 48 resources (e-books). The identified sessions were used to generate 1486 reads structures, whereby the number of associated users was reduced to 331 and the number of resources to 33 (the reduction took place due to elimination of trivial sessions by some users who did not engage significantly in e-book usage).

Table 2 shows the detailed analysis data generated during the process model steps execution. The results indicate that the average student read period (time from the first to last access to a resource) was about 4 days, with the average resource read period being around 3 days. It is to be noted that reads are continuous structures that map to a specific resource and that one resource is usually taught during one week of lectures. All read periods lengths come with high standard deviations indicating high between-student differences in total period spent on a resource.

In terms of the receptive reading style variable, values are uniform across general, user and resource categories (means range from 0.83 to 0.86) with observably smaller standard deviation values. This indicates reading is done in fairly linear fashion with relatively small proportion of large jumps across the reading resource. On the other hand, the responsive reading variable values indicate there is

about one interactive event per 10 pages of student reading, with large observed between-student differences. Correlation analysis was done to test for possible relationships between the variables of read period time, receptive, responsive reading, and the final course grade. Results are shown in Table 3 and Table 4.

Table 2

*The detailed analysis results while applying the process model, including user reading style data*

| | AVG | STDEV |
|---|---|---|
| Reads per student | 4.49 | 2.75 |
| Reads per resource | 45.00 | 50.48 |
| Reads per student per resource | 1 | 0 |
| Read period (s) | 340,367 | 970,262 |
| Read period per student (s) | 400,334 | 820,189 |
| Read period per resource (s) | 236,437 | 391,596 |
| Receptive reading | 0.85 | 0.16 |
| Receptive reading per student | 0.83 | 0.13 |
| Receptive reading per resource | 0.86 | 0.10 |
| Responsive reading | 0.01 | 0.09 |
| Responsive reading per student | 0.01 | 0.1 |
| Responsive reading per resource | 0.02 | 0.04 |

Table 3

*Results of correlation analysis for resource read period time, and receptive and responsive reading style variables*

| | Receptive reading | Responsive reading |
|---|---|---|
| Resource read period | -0.054* | 0.023 |
| Receptive reading | 1 | -0.033 |

*p<0.05

The correlation analysis results indicate all correlation values are low and that there is only one significant correlation coefficient (between receptive reading and resource read time), however of weak power r=-0.054.

For the purposes of in-depth analysis of reading style at a specific course level (Botički, Budišćak, & Hoić-Božić, 2008), a course taught in the logged time period was chosen – the Architecture course. Although the final course grade in the Architecture course correlated with the overall read period time, which is in line with prior research on usage time of e-books (Oi, Okubo, Shimada, Yin, & Ogata, 2015), no significant and strong correlations were identified between the final grade and the receptive and responsive reading variables.

Table 4

*Results of the correlation analysis: final grades in the Architecture course, and receptive and responsive variables*

| | Final grade |
|---|---|
| Resource read time | 0.237**/0.352** |
| Receptive reading | 0.077/-0.077 |
| Responsive reading | -0.034/-0.234 |

**p<0.001

There was observable non-significant negative correlation between the final course grades and the responsive reading variable possibly indicating students with such reading style perform poorly on

the final exam, but more research is needed to confirm this (units of are analysis one resource per one student in the Architecture course/all resources per one student in the Architecture course).

## 5. Conclusions

In this paper learning log data was analyzed via the adopted process model in order to extract usage sessions, reads, read passages and read passage pairs. Both sessions and read passage pairs were then used as means of defining reading style variables, with the final aim of including these variables in user modeling.

The results of the analysis indicate that students predominantly utilize linear, forward-oriented reading of e-book materials and have relatively low interaction with the material, although interactive features of e-books are readily available for their use. Further correlation analysis indicates that the receptive and responsive reading variables do not correlate with the student final course grades.

Limitations of the study stem from the characteristics of the data sample being used, which covers only undergraduate students in academic environments. Future research will focus on alternative approaches to reading style variables extraction including the use of sub-symbolic classifiers (Pessa, n.d.) and on identifying possible educational activities with e-books which rely on the usage of reading style variables to enhance some elements of the educational process.

## Acknowledgements

This work is supported by the JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

## References

Botički, I., Budišćak, I., & Hoić-Božić, N. (2008). Module for online assessment in AHyCo learning management system. *Novi Sad J. Math*, *38*(2), 115–131.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management and E-Learning*, *10*(4), 469–484.

Freeman, R. S., & Saunders, E. S. (2016). E-Book Reading Practices in Different Subject Areas: An Exploratory Log Analysis. In *In Academic E-Books: Publishers, Librarians, and Users* (pp. 223–248). Purdue University Press.

Jagušt, T., Botički, I., & So, H. J. (2018). Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computers and Education*, *125*, 444–457. https://doi.org/10.1016/j.compedu.2018.06.022

Marshall, C. C. (2009). *Reading and Writing the Electronic Book*. *Synthesis Lectures on Information Concepts, Retrieval, and Services* (Vol. 1). https://doi.org/10.2200/S00215ED1V01Y200907ICR009

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., … Hirokawa, S. (2017). Learning analytics for E-book-based educational big data in higher education. In *Smart Sensors at the IoT Frontier* (pp. 327–350). https://doi.org/10.1007/978-3-319-55345-0_13

Oi, M., Okubo, F., Shimada, A., Yin, C., & Ogata, H. (2015). Analysis of preview and review patterns in undergraduates' e-book logs. *Proceedings of the 23rd International Conference on Computers in Education, ICCE 2015*, 166–171. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84979681388&partnerID=40&md5=a0cb568ab33b35c83c91ad19e635aa28

Pessa, E. (n.d.). Symbolic and subsymbolic models, and their use in systems research. *Systems Research*, *11*(3), 23–41. https://doi.org/10.1002/sres.3850110303

Pugh, A. K. (1979). Styles and Strategies in Silent Reading. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing of Visible Language* (pp. 431–443). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-0994-9_27

Thayer, A., Lee, C. P., Hwang, L. H., Sales, H., Sen, P., & Dalal, N. (2011). The Imposition and Superimposition of Digital Reading Technology: The Academic Potential of E-readers. *CHI 2011, Session: Reading and Writing*, 2917–2926. https://doi.org/10.1145/1978942.1979375