Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

Investigating Students' e-Book Reading Patterns with Markov Chains

Gökhan AKÇAPINAR[a, b*], Rwitajit MAJUMDAR[a], Brendan FLANAGAN[a] & Hiroaki OGATA[a]

[a] *Academic Center for Computing and Media Studies, Kyoto University, Japan*

[b] *Department of Computer Education & Instructional Technology, Hacettepe University, Turkey*

*akcapinar.gokhan.2m@kyoto-u.ac.jp

Abstract: In this paper, we analyze students' e-book reading patterns by using Markov Chains (MCs). We used click-stream data of 236 students while they read 7 different contents shared by the instructor across different weeks of the course. To analyze reading patterns, we first clustered students independently based on their interaction with each content. We grouped students in *None, Low, Medium*, and *High* clusters. Then by using MCs, we calculated cluster transition probabilities between different contents. We also visualized these patterns and applied a prediction algorithm to predict students' reading patterns. Results revealed that students are likely to follow the same reading patterns across the semester. In other words, if a student reads less in the first content, s/he is likely to read less during the rest of the semester. We also found that transition data could be used to predict students' further reading behaviors. The developed model can be used to plan an intervention system for at-risk students. Visualization of these transitions may help a teacher to understand how well students use contents.

Keywords: e-book, sequential behavior analysis, markov chains, clustering, learning analytics, educational data mining

1. Introduction

Digital learning materials are widely used in online learning to deliver lecture contents to the students. Along with their advantages for the students and instructors, digital materials have made it possible to collect a vast amount of data regarding students' reading behaviors. Such log data was previously used by the researchers to visualize patterns in class preparation and review of the students (Ogata et al., 2018).

In this paper, students' e-book reading logs are analyzed with Markov Chains (MCs) to understand their e-book reading behaviors. To analyze students' reading patterns, we first clustered them independently based on their activities (e.g. session count, total time, event count, etc.) for each content then analyzed their transitions with the help of MCs.

1.1 BookRoll System

BookRoll is an e-book reader, currently used in Kyoto University to support in-class teaching. It is integrated into the university's Learning Management System (LMS), hence students can access it through LMS. Instructors are using BookRoll to support their classes for uploading their slides or giving students extra reading materials. In BookRoll, there are marker features to highlight different sections of a reading material in yellow and/or red color. Additionally, memos can be created to take notes. Students can also bookmark pages or use the full-text search function to find the pages that contain information they are looking for. Learning materials can be uploaded to BookRoll in PDF format, and it supports various devices as it can be accessed through a standard web browser.

Students' reading logs are recording in the learning analytics system developed by the Flanagan and Ogata (2017). This system records all the anonymized learning logs into the integral Learning Record Store (LRS) as xAPI statements. Any action taken by students while using the

BookRoll system (e.g. next, previous, page jump, highlight, adding a memo, search, etc.) is recorded.

1.2 Student Behavior Analysis

Previously, researchers used to partial correlation analysis (Yin et al., 2015) and lag-sequential analysis (Yin et al., 2017) to understand students' behavioral patterns while reading e-books. Markov Chains (MCs), however, are mainly used for profiling users' web navigation behaviors, especially to predict their next clicks (Sarukkai, 2000). In educational settings, MCs previously used for the learning sequence recommendation to the students in an adaptive learning environment (Yueh-Min, Tien-Chi, Kun-Te, & Wu-Yuin, 2009). Marques and Belo (2011), used MCs to analyze students' Moodle logs and established students' usage profiles for online learning. Hlosta et al. (2014) used MCs for analyzing student activities in the online virtual learning environment to identify the behavior of at-risk students. Caprotti (2017) used MCs to visualize logpaths of students in an online calculus course and analyzed the study strategies of the students to construct a learning resources recommendation engine able to suggest the best learning resources.

In this study, we formulized students' reading patterns as MCs to understand their content transition behaviors. We also visualized these patterns and applied a prediction algorithm to predict their further reading behaviors. To achieve this, we first clustered students based on their interaction data in each content, and then analyzed transitions between these clusters by using MCs. This method helped us to understand how students' reading patterns changes across the different contents and how this data can be used to predict at-risk students.

2. Method

2.1 Data Collection and Feature Extraction

In this paper, more than 250 thousand lines click-stream data of 236 students who registered Information Basics course were analyzed. The course was offered for the 1st year undergraduate university students. Students used the BookRoll system to access course materials uploaded by the instructor in weekly or biweekly intervals. From these data, features were extracted to reflect students' reading patterns. These features include the number of sessions, number of the unique days of access, total time they spent on the system, total number of events, total number of short events (event duration < 3s), total number of next, previous, jump events, and total number of memos, markers and bookmarks. Since there was not enough data related to students' jump, marker, memo, bookmark actions, those features were excluded from the analysis. Descriptive statistics for the remaining features for each content are given in Table 1.

Table 1 Descriptive Statistics of Features in Dataset

		Mean (SD)						
e-Book	**N**	**Session**	**UniqueDay**	**TotalTime**	**TotalEvent**	**ShortEvent**	**Next**	**Prev**
Content 1	215	2.75 (1.64)	2.28 (1.28)	25.1 (24.0)	119 (108)	86.8 (88.6)	88.8 (75.1)	23.7 (32.3)
Content 2	223	3.33 (1.76)	2.62 (1.29)	66.5 (49.5)	222 (163)	158 (133)	142 (100)	68.3 (60.9)
Content 3	217	3.12 (1.77)	2.55 (1.24)	43.9 (34.1)	162 (116)	115 (94.2)	114 (75)	41.1 (41.2)
Content 4	213	2.82 (1.65)	2.34 (1.17)	51.8 (43.1)	174 (129)	122 (103)	117 (81)	50.2 (47.3)
Content 5	202	3 (1.69)	2.47 (1.18)	34.7 (37.8)	188 (155)	143 (135)	147 (107)	33.6 (52.6)
Content 6	203	2.44 (1.55)	2.02 (1.05)	26.3 (32.5)	109 (98)	79.6 (81)	83.5 (67.4)	19.7 (33.1)

Content 7	206	3.16 (1.83)	2.59 (1.33)	45.5 (41.3)	206 (175)	151 (149)	134 (105)	65.4 (70.6)

2.2 Clustering

Clustering analysis was used to cluster students based on their reading activities by following the process given in Figure 1. According to this process, first, related data filtered, then attributes are using in the analysis were selected. Since the attributes are different in scale, all of them were normalized before the clustering algorithm was applied. X-Means algorithm (Pelleg & Moore, 2000) was used for clustering and this process was applied seven times independently for each content. Sankey diagram was used to visualize the transitions across the clusters across seven contents (Figure 2).

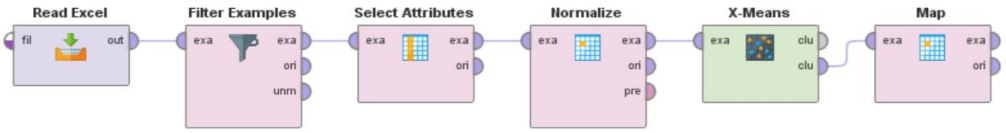

Figure 1. Clustering process

2.3 Markov Chains Analysis

Markov Chains (MCs) Analysis was used to calculate content transition probabilities of the students. The MCs is a popular method for modeling sequential data and previous studies have already shown its effectiveness in educational settings to understand how students used the resources that teachers share with them (Marques & Belo, 2011). MCs analyses were conducted in R (R Core Team, 2017) with the help of click-stream (Scholz, 2016) and markovchain (Spedicato, 2017) packages. MCs results were presented in the forms of state transition diagrams and sequence prediction analysis.

3. Results

3.1 Cluster Analysis

Students were clustered based on their reading patterns. Cluster numbers are automatically determined by the X-means algorithm. X-means determined 3 or 4 clusters in each data. Since cluster analysis applied independently for each content, obtained clusters are mapped to pre-defined clusters (LOW, MEDIUM, HIGH) manually. Otherwise, the comparison between clusters in different content could not be possible. For the mapping, cluster centers analyzed and appropriate mapping is decided by the authors for each content. After mapping was done, we had three clusters for each content. Along with these three clusters, cluster NONE created manually and students who do not have any activity in a specific content assigned to this cluster. Numbers of students in each cluster for each content can be seen in Table 2.

Table 2 Numbers of Students in each Cluster and their Percentage

Book	N	NONE	LOW	MEDIUM	HIGH
Content 1	236	21 (8.9%)	96 (40.7%)	74 (31.4%)	45 (19.1%)
Content 2	236	13 (5.5%)	67 (28.4%)	90 (38.1%)	66 (28%)
Content 3	236	19 (8.1%)	71 (30.1%)	84 (35.6%)	62 (26.3%)
Content 4	236	23 (9.7%)	73 (30.9%)	88 (37.3%)	52 (22%)
Content 5	236	34 (14.4%)	83 (35.2%)	76 (32.2%)	43 (18.2%)
Content 6	236	33 (14%)	146 (61.9%)	47 (19.9%)	10 (4.2%)
Content 7	236	30 (12.7%)	71 (30.1%)	61 (25.8%)	74 (31.4%)

At the end of the analysis, data file was generated that contained all students' clusters for each content (e.g. Student 1: HIGH, HIGH, MEDIUM, HIGH, MEDIUM, HIGH, HIGH). This transition data visualized as Sankey diagram to show all transitions (Figure 2). When the Figure 2 analyzed, we observed that students tend to follow the same reading patterns regardless of time and context of the content. Since contents uploaded by the instructor in different weeks of the course, our results also represent students' reading patterns over time.

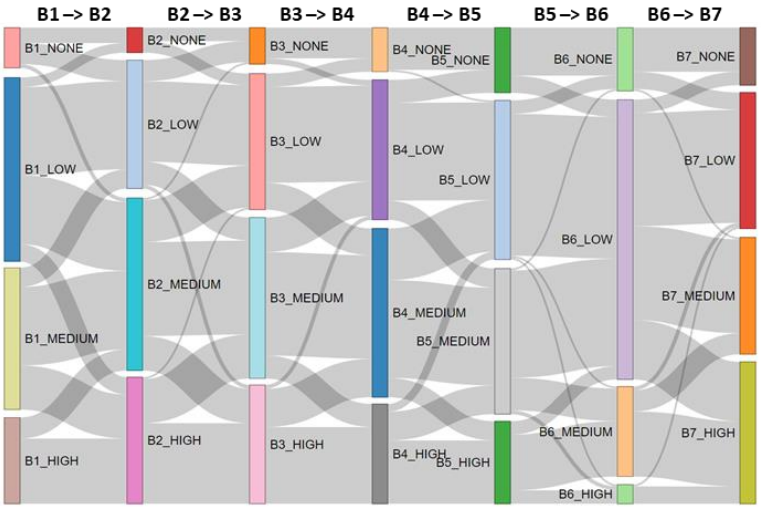

Figure 2. Visualization of students' reading patterns across 7 contents

For the content 6, we observed that majority of the students (61.9%) are in the LOW cluster (see Table 2). This pattern is different than other contents. At present, we do not have enough data to analyze reasons of this pattern, however, instructor of the course might have insight about the course when s/he looks at this graph. Although it makes easy to understand general content usage patterns of the students, this graph alone cannot be used to identify at-risk students; therefore, we used MCs to see probability of the transitions.

3.2 Markov Chains Analysis Results

State transition diagrams are generally used to represent transition matrices produces by MCs. For the seven different content, it is possible to create twenty-one different transition matrices, we reported two of the in this paper (see Figure 3) as an example. The Figure 3-a shows the students' transitions between Content 1 and Content 2, and Figure 3-b shows the students transition between Content 1 and Content 7. These two graphs were selected since both are important to observe the changes in students' reading behavior after the first content.

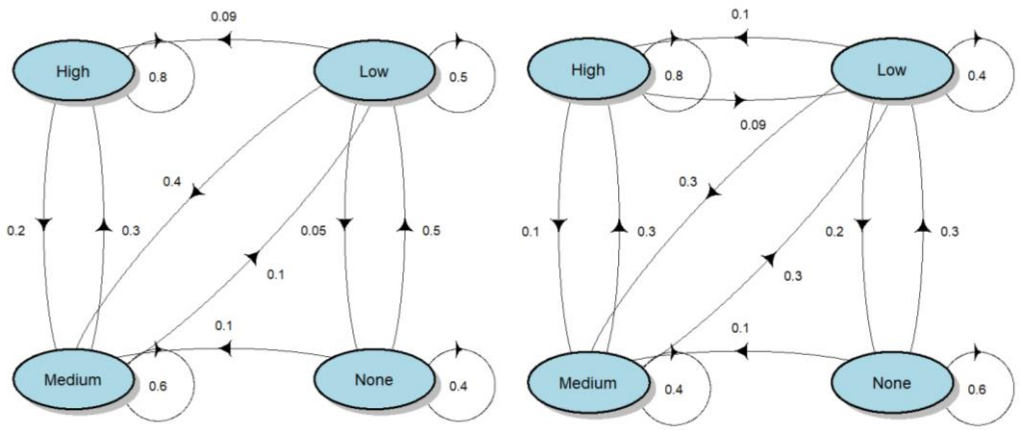

Figure 3. State Transition Diagrams: (a) Content 1 to Content 2 (left), (b) Content 1 to Content 7 (right)

When we analyzed the state transition diagrams, we found that, from Content 1 to Content 2 there was no transition between HIGH to LOW, HIGH to NONE, and MEDIUM to NONE groups and only 10% of the students moved from NONE to MEDIUM and 9% of the students moved from LOW to HIGH group. Similar patterns were observed from Content 1 to Content 7. Here again, there was no transition between HIGH to NONE, and MEDIUM to NONE groups and only 10% of the students moved from LOW to HIGH and 10% of the students moved from NONE to MEDIUM group. On the contrary, in this case we found that 9% of students moved from HIGH to LOW group. In both cases, we observed that students mostly tended to follow similar patterns since within group transition had the highest probability for each group. On the other hand, 60% of the students who did not read the Content 1, did not read the Content 7 either and 30% of them moved to the LOW group.

Prediction algorithm was also applied based on MCs result to predict the 2nd and 3rd order cluster state of a student. Example patterns and their probabilities are shown in Table 3. Since prediction probabilities decrease after the 3rd order, we limited our results at the 3rd order. Patterns with the highest probability are reported for each case. According to these patterns, if a student in the NONE or LOW group for the first content, the student will most likely be in the LOW group for the next (with a probability of 0.5) and following content (with a probability of 0.3). On the other hand, if a student in the MEDIUM or HIGH group in the first content, the student will most likely be in the same group for the next and following content with a probability of 0.6 and 0.8.

Table 3 Predicted Patterns and their Probabilities

Pattern*	**Probability**
B1NONE → B2LOW	0.5
B1NONE → B2LOW → B3LOW	0.3
B1LOW → B2LOW	0.5
B1LOW → B2LOW → B3LOW	0.3
B1MEDIUM → B2MEDIUM	0.6
B1MEDIUM → B2MEDIUM → B3MEDIUM	0.3
B1HIGH → B2HIGH	0.8
B1HIGH → B2HIGH → B3HIGH	0.5

* Initial State → 2nd Order Prediction → 3rd Order Prediction

4. Conclusion

In this paper, we analyzed students' e-book reading patterns with the help of MCs. We first grouped students based on their activity level in different contents of the course, and then we analyzed their transition probabilities between the groups in different contents. We also visualized these transitions as state transition diagrams and investigated the predictive usage of these patterns. Our results showed that, without any intervention, students' transition probability between NONE, LOW and MEDIUM, HIGH groups are low. Moreover, most of the students follow the same pattern that they followed in the first content. This finding can be used to identify at-risk students and planning interventions for them.

We analyzed students' reading patterns regardless of their academic performance. For this reason, we can only speculate students in the NONE and LOW groups are probably at-risk, since engagement is one of the important factors for success in online learning (Hrastinski, 2009), and low engagement is related to low academic success (Akçapınar, Altun, & Aşkar, 2016). However, it is important to relate these patterns with academic performance to give students more accurate feedback about their learning.

Visualization is also important to help instructors to interpret learning data easily (Coffrin, Corrin, Barba, & Kennedy, 2014; Cristóbal & Sebastián, 2017; Majumdar & Iyer, 2016). State transition diagrams generated by MCs may help a teacher to understand how well students use contents. Moreover, graphical output provides support to instructors in planning interventions (Hlosta et al., 2014).

In future studies, we are planning to use MCs model obtained here to predict students' content engagement levels at the beginning of the course. We are also planning to design interventions for the students who are in the LOW and NONE group to help them to change their reading behaviors. To help instructors to interpret results obtained here, we are planning to implement a Stratified Attribute Tracking (SAT) Diagram developed by Majumdar and Iyer (2016).

Acknowledgements

This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

References

Akçapınar, G., Altun, A., & Aşkar, P. (2016). Çevrimiçi Öğrenme Ortamındaki Benzer Öğrenci Gruplarının Kümeleme Yöntemi ile Belirlenmesi [Identifying Different Student Profiles in an Online Learning Environment With Cluster Analysis]. Educational Technology Theory and Practice, 6(2).

Caprotti, O. (2017). Shapes of Educational Data in an Online Calculus Course. Journal of Learning Analytics, 4(2), 76–90.

Coffrin, C., Corrin, L., Barba, P. d., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. Paper presented at the Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, Indianapolis, Indiana, USA.

Cristóbal, R., & Sebastián, V. (2017). Educational data science in massive open online courses. Wiley Inter disciplinary Reviews: Data Mining and Knowledge Discovery, 7(1), e1187. doi:doi:10.1002/widm.1187

Flanagan, B., & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy. Paper presented at the 25th International Conference on Computers in Education, ICCE 2017, New Zealand.

Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z., & Wolff, A. (2014). Modelling student online behaviour in a virtual learning environment. Paper presented at the Machine Learning and Learning Analytics workshop at The 4th International Conference on Learning Analytics and Knowledge (LAK14), 24-28 March 2014, Indianapolis, Indiana, USA, Indianapolis, Indiana, USA. http://oro.open.ac.uk/40670/

Hrastinski, S. (2009). A theory of online learning as online participation. Computers & Education, 52(1), 78–82. doi:https://doi.org/10.1016/j.compedu.2008.06.009

Majumdar, R., & Iyer, S. (2016). iSAT: a visual learning analytics tool for instructors. Research and Practice in Technology Enhanced Learning, 11(1), 16. doi:10.1186/s41039-016-0043-3

Marques, A., & Belo, O. (2011). Discovering Student Web Usage Profiles Using Markov Chains. Electronic Journal of e-Learning, 9(1), 63–74.

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., . . . Hirokawa, S. (2018). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In H. Yasuura, C.-M. Kyung, Y. Liu, & Y.-L. Lin (Eds.), Smart Sensors at the IoT Frontier (pp. 327–350). Cham: Springer International Publishing.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. Paper presented at the Icml.

R Core Team. (2017). R: A language and environment for statistical computing: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains1This work was done by the author prior to his employment at Yahoo Inc.1. Computer Networks, 33(1), 377–386. doi:https://doi.org/10.1016/S1389-1286 (00)00044-X

Scholz, M. (2016). R Package clickstream: Analyzing Clickstream Data with Markov Chains. Journal of Statistical Software, 74(4), 1–17. doi:10.18637/jss.v074.i04

Spedicato, G. A. (2017). Discrete Time Markov Chains with R. The R Journal, 9(2), 84–104.

Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., & Ogata, H. (2015). Identifying and Analyzing the Learning Behaviors of Students Using e-Books. Paper presented at the 23rd International Conference on Computers in Education, ICCE 2015.

Yin, C., Uosaki, N., Chu, H.-C., Hwang, G.-J., Liu, G.-Z., Hwang, J.-J., & Tabata, Y. (2017). Learning Behavioral Pattern Analysis based on Students' Logs in Reading Digital Books. Paper presented at the 25th International Conference on Computers in Education, ICCE 2017, New Zealand.

Yueh-Min, H., Tien-Chi, H., Kun-Te, W., & Wu-Yuin, H. (2009). A Markov-based Recommendation Model for Exploring the Transfer of Learning on the Web. Journal of Educational Technology & Society, 12(2), 144–162.