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

Developing E-Book Page Ranking Model for Pre-Class Reading Recommendation

Christopher C.Y. YANG[a*], Gökhan AKÇAPINAR[b,c], Brendan FLANAGAN[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[c]*Department of Computer Education and Instructional Technology, Hacettepe University, Turkey*
*yang.yuan.57e@st.kyoto-u.ac.jp

Abstract: In this paper, we propose an E-Book Page Ranking (EBPR) method to rank e-book pages from the original learning material automatically. The proposed method ranks all the e-book pages by the class probabilities retrieved from machine learning models. The top-ranked e-book pages are then selected to form the pre-class reading (preview) recommendation. The proposed method extracts image features and text features from e-book page contents as well as the e-book usage features from students' previous reading logs. In this paper, we test the performance of the proposed model with two different cases, with and without past e-book usage data. The experimental results showed the improvability of the model after taking into account learners' past e-book usages.

Keywords: E-book page ranking, educational recommender systems, e-book usage logs, preview reading

1. Introduction

Preparing for lectures is crucial for learners and has numerous advantages over their learning. A study conducted by Dreher and Sammons (1994) indicated that students who previewed the learning material before the class were able to answer the questions in exams better than students who did not preview. However, learners are not always willing to prepare for the class, especially when learning materials are too long for them. Another study conducted by Shimada et al. (2017) revealed that giving learners a subset of learning material (important pages) instead of the whole content, increased their preview behaviors and overall learning performances. In the context of e-learning, the characteristics of important document page/e-book page and the associated content features have been mentioned and applied in several articles (Neto et al., 2002; Shimada et al., 2017). Additionally, in this paper, we assumed that the learners' past e-book usage that related to the target e-book material should be taken into account when considering e-book preview recommendation. The proposed method can be potentially anticipated to well reduce the overhead for course preview material creation on e-book system. In this study, we attempt to answer the following two research questions:

1. What is the best-performed machine learning algorithm for the proposed E-Book Page Ranking (EBPR) method?
2. Can we improve the performance of EBPR by adding features to the model related to learners' past e-book usage?

2. Machine Learning-Based E-Book Page Ranking Method

2.1 Data Collection and Feature Extraction

The entire performing process is described as follows: we use BookRoll system (Ogata et al., 2015) which is a digital textbook reading system that contains plenty of functions such as page turning, internal learning content searching, page jumping, annotation creation, annotation transfer across

e-book revisions (Yang et al., 2018), etc. Learners' reading behaviors while using BookRoll store in the database of BookRoll. In this study, two types of data were collected from BookRoll as the input of the proposed method. The first type of data is text contents and image contents from the original e-book material that contains 91 e-book pages in BookRoll named *Semantic Web Services* with 38 learners enrolled under the period of 3 weeks in the previous semester. The second type of data is the recorded 10147 e-book reading events that related to the same e-book material. The reading events tracked by BookRoll has been described in the previous article (Ogata et al., 2015). In the feature extraction, we applied text processing techniques, image processing algorithms including background subtraction method and inter-frame difference method, as well as the educational data mining methods to extract text, image, e-book usage features from the collected e-book page contents and students' e-book reading events as described in Table 1.

Table 1

Description of the Extracted Features

Category_Index	**Feature Name**	**Feature Description**
Text_01	*TotalChar*	Total characters in a page
Text_02	*AvgTFIDF*	Average of TFIDF (Term Frequency-Inverse Document Frequency) value
Text_03	*Similarity to title*	Cosine similarity to the title of the content
Text_04	*Similarity to keywords*	Cosine similarity to the keywords of the content
Text_05	*Page-Page cohesion*	Sum of cosine similarities to the rest pages
Text_06	*Punctuation*	Total occurrence of punctuations in a page
Image_01	*Background subtraction*	Foreground pixels in a page
Image_02	*Background subtraction + Inter-frame difference*	Absolute foreground pixel differences with previous page and next page (choose the higher value)
Usage_01	*Marker*	Total number of marker added in a page
Usage_02	*Memo*	Total number of memo added in a page
Usage_03	*Bookmark*	Total number of bookmark added in a page
Usage_04	*UniqueVisit*	Total number of learners visited a page
Usage_05	*TotalTime*	Total reading time on a page
Usage_06	*AvgTime*	Average reading time on a page per learner
Usage_07	*TotalEvent*	Total clicking events in a page
Usage_08	*AvgEvent*	Average clicking events in a page per learner

2.2 Modeling, E-Book Page Ranking, and Generation of E-Book Preview Recommendation

To train the machine learning models we asked one lecturer to label important e-book pages in his learning content (gold-standard). The lecturer labeled 45 out of 91 pages as important pages (recommended for preview). Rest of the pages considered as less important for preview. We formed our problem as a binary classification problem, however, we obtained class probabilities as the e-book page ranking output instead of class labels, which gave us the flexibility of ranking pages based on their importance. The top-ranked e-book pages are then selected to form the pre-class reading recommendation.

3. Test Results

To evaluate the ranking performance of the proposed EBPR method, we first selected the top-ranked 45 e-book pages from each model (the same number of pages are labeled as important page by the lecturer). Model performances were evaluated by using metrics such as precision, recall, and Area Under the Curve (AUC) along with the 3-folds cross validation. To evaluate the improvability of models after taking into account learners' past e-book usages, we conducted two different experiments which are past e-book usage feature exclusion and inclusion, respectively. As shown in Table 2, in the first

experiment, e-book usage features were excluded from the training process, resulting of the best-performed model *MultilayerPerception* with precision 0.667, recall 0.667, and AUC 0.67. In the second experiment, e-book usage features were included from the training process of models, the observed result in this experiment indicates that the best-performed model was still *MultilayerPerception* with precision 0.756, recall 0.756, and AUC 0.758.

Table 2

Performance of each model (excluding / including students' past e-book usage logs)

Model	**Precision**	**Recall**	**AUC**
DTNB	0.422 / 0.556	0.422 / 0.556	0.429 / 0.560
JRip	0.489 / 0.556	0.489 / 0.556	0.494 / 0.560
RandomForest	0.556 / 0.600	0.556 / 0.600	0.560 / 0.604
J48	0.533 / 0.511	0.533 / 0.511	0.538 / 0.516
BayesNet	0.422 / 0.422	0.422 / 0.422	0.429 / 0.429
GaussianNaïveBayes	0.467 / 0.467	0.467 / 0.467	0.472 / 0.472
LogisticRegression	0.622 / 0.733	0.622 / 0.733	0.626 / 0.736
MultilayerPerception	**0.667** / **0.756**	**0.667** / **0.756**	**0.670** / **0.758**

4. Conclusion and Future Work

In this paper, we proposed a machine learning-based E-Book Page Ranking (EBPR) method for the recommendation of e-book preview that can be integrated into any e-book systems. We ranked the input e-book pages by the predicted class probabilities accordingly. After the ranking process, the top-ranked e-book pages are selected to form the recommendation of preview material for learners before a class. We compared several classification models and reported the best-performed model *MultilayerPerception*, which answered our first research question. We evaluated the performances of e-book page ranking in different conditions through two experiments. The statistical results shown in Table 2 reported that when including learners' past e-book usage features, the overall performance of e-book page ranking will be improved, which answered our second research question. In the future, we will look for more e-book page samples from lecturers in different domains as the training samples to investigate and evaluate the ranking performance of the model and the most important features.

Acknowledgments

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0.

References

Dreher, M. J., & Sammons, R. B. (1994). Fifth graders' search for information in a textbook. *Journal of Reading Behavior*, *26*(3), 301-314.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002, November). Automatic text summarization using a machine learning approach. In *Brazilian Symposium on Artificial Intelligence* (pp. 205-215). Springer, Berlin, Heidelberg.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In International Conference on Computer in Education (ICCE 2015) (pp. 401-406).

Shimada, A., Okubo, F., Yin, C., & Ogata, H. (2017). Automatic Summarization of Lecture Slides for Enhanced Student Preview–Technical Report and User Study–. *IEEE Transactions on Learning Technologies*, *11*(2), 165-178.

Yang, C.C.Y., Flanagan, B., Akcapinar, G., & Ogata, H. (2018). Maintaining reading experience continuity across e-book revisions. *Research and practice in technology enhanced learning*, *13*(1), 24.