A page jump recommendation model based on digital textbook contents and student log data

Wenhao WANG [a], Natsumi YAMAMOTO [a], Fuzheng ZHAO [a], Etsuko KUMAMOTO [b], Zicheng KANG [c], Chengjiu YIN [d*]

[a] *Graduate School of System Informatics, Kobe University, Japan*
[b] *Information Science and Technology Center, Kobe University, Japan*
[c] *Faculty of Systems Science and Technology, Akita Prefectural University, Japan*
[d] *Research Institute for Information Technology, Kyushu University, Japan*
*yin.chengjiu.247@m.kyushu-u.ac.jp

Abstract: In the analysis of student learning data, researchers found that students often switch pages but cannot accurately find the target page. Therefore, in this study, a page recommendation model to provide students with the target page for jumping to improve the efficiency of student page jumping behavior was developed. The recommendation model linearly combines the content-based recommendation model using term frequency–inverse document frequency (TF-IDF) and the recommendation model based on student log data, and determines the optimal weight of linear combination through gradient descent. Finally, the accuracy of the model was evaluated by comparing it with expert-generated recommendation outcomes.

Keywords: Page jump recommendation, TF-IDF, Log data

1. Introduction

Online learning has been on the increase in the last two decades (Martin, 2020). As an important part of online learning, e-books, and other digital teaching materials are gradually replacing traditional paper textbooks due to their low cost and better portability (Wang J, Shimada A, 2023). However, research has found that some students can remember less content when reading textbooks on screens than when reading paper textbooks, and their understanding of the content is shallower than when reading paper textbooks (Lauterman T, Ackerman R, 2014). At the same time, the sudden outbreak of COVID-19 has forced schools around the world to adopt online teaching methods on a large scale. Research by Emma Dorn and Bryan Hancock (2020) shows that COVID-19 has caused significant disruption to the education system and students' academic performance has declined to varying degrees.

Researching and improving online learning has become an important factor in improving students' online learning outcomes (García-Morales, 2021). In the analysis of online learning behavior, Yin et al. (2019) found that students with better academic performance have a significantly higher Backing Track Rate (BTR) through the analysis of learning behavior patterns based on student behavior data. Haverkamp (2020) also showed through research that backtracking behavior has a uniquely positive effect on improving content understanding when reading digital materials. However, in the analysis of learning data from 122 students, we noted that more than 40% of backtracking behavior requires more than two jumps to navigate to the target page, indicating that students' backtracking behavior is inefficient.

To improve the efficiency of backtracking behavior and provide students with navigation for page jumping, this study proposes a page recommendation model that combines content-based and user data-based methods to help students find the target page for jumping and improve learning efficiency when reading e-books.

2. Literature review

2.1 TF-IDF algorithm

TF-IDF is a commonly used weighting technique for information retrieval and text mining, which can reflect the importance of a word to a document or a corpus.

The TF-IDF algorithm was first proposed by Sparck Jones (1972) to evaluate the performance of information retrieval systems. Later, it is widely used in various natural language processing tasks. For example, Wu et al. (2008) applied TF-IDF to relevance decision, Shahzad Qaiser et al. (2018) used the TF-IDF algorithm to examine the relevance of keywords and documents in the corpus, Zhou et al. (2020) applied TF-IDF to cluster news.

Due to its simple mathematical calculation formula, low computational complexity, and relatively good accuracy in classification (Wu, 2018) this study also uses this algorithm in the recommendation model.

2.2 student log data analysis in Education

With the development of web-based education, educational data mining has become a promising field (Romero, 2007), and many studies have been conducted on mining student log data. In mining student log data, Yin et al. (2017) mined learning behavior patterns of different types of students from data of students reading digital books; Zhao et al. (2021) predicted students' academic performance using 6 prediction algorithms based on student data from an e-book system; Riestra-González et al.(2021) predicted student performance in the early stages of a course by analyzing log data from an LMS system; Gobert (2013) evaluated students' abilities in science inquiry through their log data.

In the research direction of recommending something to students based on student log data, Aher (2012) recommended online courses through Moodle system's student course browsing data; Reddy (2016) recommended the learning sequence of courses based on the courses and students' grades.

These studies all focus on the personalized course or learning path recommendations for students, while this paper focuses on recommending pages of digital books at a more specific level.

3. Design and development of page jump recommendation model

In this section, the specific calculation method of TF-IDF in content-based recommendation models, the method of mining page relevance from student log data, and the construction method of hybrid models are mainly introduced.

3.1 Finding page associations based on textbook content

This study uses TF-IDF and cosine similarity to evaluate the relevance of pages. TF-IDF is an analytical method for evaluating the importance of words in a document. TF represents the frequency of a word in a document, and IDF represents the uniqueness of a word in a document (Ramos J, 2003). By combining the two, TF-IDF takes into account both the number of occurrences and the unique value of a word, providing a more comprehensive evaluation of the association between a word and a text. The TF-IDF value is calculated using the following formula:

$$TF(E|t,p) = \frac{Number\ of\ word\ t\ in\ page\ p\ in\ document\ E}{Total\ words\ of\ page\ p\ in\ E}$$

$$IDF(E|t) = \log_2\left(\frac{pc}{f(E|t)}\right)$$

$$TFIDF(E|t,p) = TF(E|t,p) * IDF(E|t)$$

Table 1. *Formula of TF-IDF*

E	E-book number
pc	Total number of pages of E
p	A page in E
t	A word
f(E|t)	Number of pages that t appear in E

Through the TF-IDF value of words, each page of the textbook can be simply vectorized, and then the cosine similarity between texts can be calculated. The closer the cosine similarity between two vectors is to 1, the more similar the texts are. Repeating the calculation process we can get a similarity matrix R_α between texts

3.2 Mining page relevance based on students' learning data

According to the research results of Chiba (2017), the reading speed evaluation index words per minute (wpm) for students who use English as a second language when reading English materials is 144.42 wpm. Based on this wpm, the time T_r required for a student to read a page can be inferred from wpm and the total number of words on the page W_s by this formula:

$$T_r = \frac{W_s}{wpm/_{60s}}$$

When a student performs a jump page behavior, a page sequence will be generated. Considering that the student will not read all the content on the target page completely, we use 0.8 times the reading time T_r of that page as the standard duration for that page.

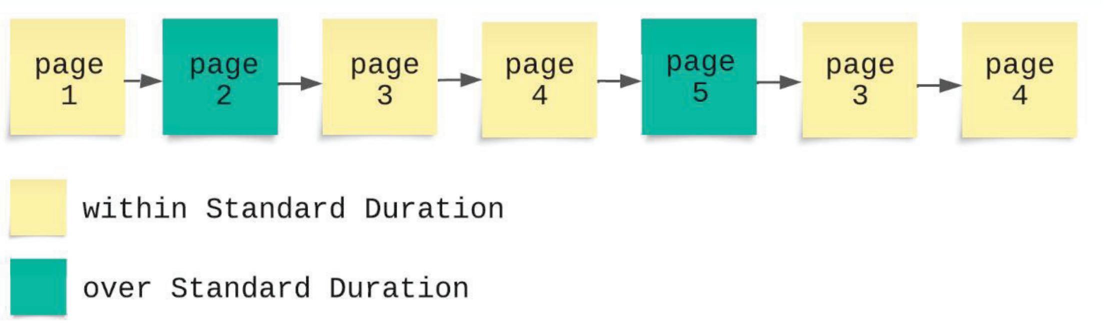


Figure 1. The method of judging page relevance.

$$\begin{array}{c} \\ P.1 \\ P.2 \\ P.3 \\ P.4 \\ P.5 \end{array}
\begin{array}{c} \begin{array}{ccccc} P.1 & P.2 & P.3 & P.4 & P.5 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 3 & 2 \\ 0 & 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 4 & 0 & 0 & 0 \\ 3 & 0 & 2 & 0 & 0 \end{pmatrix} \end{array}
\longrightarrow
\begin{array}{c} \begin{array}{ccccc} P.1 & P.2 & P.3 & P.4 & P.5 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.6 & 0 & 0.3 \\ 0.5 & 0 & 0 & 0 & 0.5 \\ 0.2 & 0.8 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 & 0 \end{pmatrix} \end{array}$$

Figure 2. The text similarity matrix R_β.

When the student's stay time on a certain page in the page sequence is greater than or equal to the standard duration of that page, it is considered that the student has found the target page of this jump, that is, the page is related to the starting page of the jump behavior, and it is reflected in the text similarity matrix R_β as an increase of 1 in the corresponding element value. Then each element is divided by the sum of the elements in its row and normalized to update the matrix R_β.

3.3 The construction of the recommendation model

This study combines the content-based text similarity matrix R_α and the user data-based text similarity matrix R_β linearly according to the following formula to obtain the text similarity matrix R(α,β) of the hybrid model, where α and β are obtained by gradient descent. Then sorting each row of the R(α,β) matrix in descending order we can obtain the recommendation list for each page.

$$R(\alpha,\beta) = \alpha R_\alpha + \beta R_\beta$$

This paper uses the following formula (Bartell, 1994) to find the optimal values of the parameters α and β.

$$J_i(\alpha,\beta) = \frac{\sum_{j \succ q_i k} [R_{ij}(\alpha,\beta) - R_{ik}(\alpha,\beta)]}{\sum_{j \succ q_i k} |R_{ij}(\alpha,\beta) - R_{ik}(\alpha,\beta)|}$$

$$\arg\min J = -\frac{1}{|Q|}\sum J_i(\alpha,\beta)$$

$j \succ q_i k$ means that for page i, the user is more likely to jump to page j than to page k. R_{ij} is the similarity value between page i and page j, while R_{ik} is the similarity value between page i and page k. Q={q_i} is the recommendation list for the textbook made in advance by experts, where q_i is the recommendation list for page i. When the order of the recommendation list generated by the model is completely consistent with the order of the expert recommendation list, J=-1.

By comparing the order of the recommendation list generated by the model with the expert-generated recommendation list, the α and β that minimize J are the optimal parameter values.

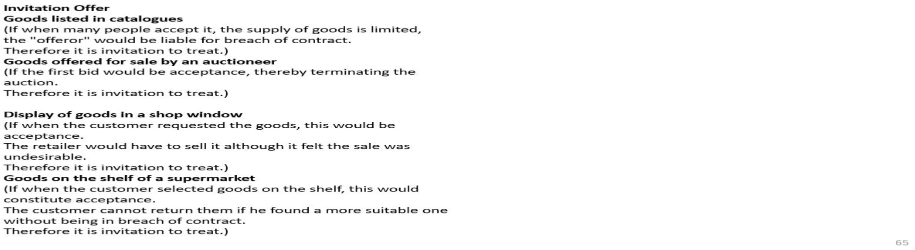
4. Model evaluation and result

4.1 Model evaluation method

In 2.3, part of the expert-generated recommendation list is used as the training set and the recommendation list R_m generated by the hybrid model is determined. Another part of the expert-generated recommendation list is used as the validation set R_e. When one of the top three recommended pages given by R_m is included in the top three recommended pages of R_e, the model recommendation is considered successful. The evaluation index accuracy is obtained by calculating the quotient of the number of successful pages and the total number of pages.

4.2 Model evaluation result

In this study, we collected data from 122 university students who use English as a second language studied the *"Commercial Law*" digital textbook in English and mined the associations between texts from this data.



Invitation Offer
Goods listed in catalogues
(If when many people accept it, the supply of goods is limited,
the "offeror" would be liable for breach of contract.
Therefore it is invitation to treat.)
Goods offered for sale by an auctioneer
(If the first bid would be acceptance, thereby terminating the
auction.
Therefore it is invitation to treat.)

Display of goods in a shop window
(If when the customer requested the goods, this would be
acceptance.
The retailer would have to sell it although it felt the sale was
undesirable.
Therefore it is invitation to treat.)
Goods on the shelf of a supermarket
(If when the customer selected goods on the shelf, this would
constitute acceptance.
The customer cannot return them if he found a more suitable one
without being in breach of contract.
Therefore it is invitation to treat.)

65

Figure 3. The text image of "*Commercial Law*".

Table 2. *Samples of log data*

User ID	Operation	Date	Page No
admin	NEXT id:12 page:39	2016/10/12 14:00:31	39
demo	NEXT id:12 page:45	2016/10/12 14:01:05	45
admin	PREV id:12 page:41 bookpage:41	2016/10/12 14:00:34	41

Through gradient descent, the final determined values of α and β are 135.545422448513 and 116.968136487428 respectively.

The hybrid model, the recommendation model based on user data and the content-based recommendation model were evaluated according to the method in 3.1, and the results are shown in the table.

Table 3. *The accuracy of models*

Category	Accuracy
hybrid model in this study	0.64
user data-based model only	0.61
content-based model only	0.37

The accuracy of the recommendation list generated using only user data is 0.61, the accuracy of the recommendation list generated using only textbook content is 0.37, and the accuracy of the model linearly combined through gradient descent reaches 0.64. The results prove that the recommendation list generated by the hybrid model is more accurate and effective than using only teaching material content or user data analysis alone.

5. Conclusions

This study linearly combined the recommendation model based on textbook content and the recommendation model based on user data, and determined the optimal weights of the two through gradient descent, resulting in a hybrid recommendation model for the page. After comparing with expert data, the accuracy of the hybrid model was evaluated, proving the accuracy of the page recommendation model is higher than that of a single model.

On the other hand, the poor performance of the recommendation model based solely on content indicates that there is still much room for improvement in the processing methods of the textbook content. And due to the lack of expert-generated recommendation lists and student learning logs, the overall performance of the model is not satisfactory.

Therefore, more research is needed to improve the recommendation model, and the effectiveness of the recommendation system still needs to be verified in actual learning environments.

Acknowledgements

This research was partially supported by Grants-in-Aid for Scientific Research No. 21H00905 from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) in Japan.

References

Martin, F., Sun, T., & Westine, C. D. (2020). A systematic review of research on online teaching and learning from 2009 to 2018. *Computers & education*, *159*, 104009.

Wang, J., Shimada, A., Oi, M., Ogata, H., & Tabata, Y. (2023). Development and evaluation of a visualization system to support meaningful e-book learning. *Interactive Learning Environments*, *31*(2), 836-853.

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, *35*, 455-463.

Dorn, E., Hancock, B., Sarakatsannis, J., & Viruleg, E. (2020). COVID-19 and student learning in the United States: The hurt could last a lifetime. *McKinsey & Company*, *1*, 1-9.

García-Morales, V. J., Garrido-Moreno, A., & Martín-Rojas, R. (2021). The transformation of higher education after the COVID disruption: Emerging challenges in an online learning scenario. *Frontiers in psychology*, *12*, 616059.

Yin, C., Ren, Z., Polyzou, A., & Wang, Y. (2019). Learning behavioral pattern analysis based on digital textbook reading logs. In *Distributed, Ambient and Pervasive Interactions: 7th International Conference, DAPI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21* (pp. 471-480). Springer International Publishing.

Haverkamp, Y. E., & Bråten, I. (2022). The Role of Strategic Backtracking When Reading Digital Informational Text for Understanding. *Literacy Research and Instruction*, 1-16.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 1-37.

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25-29.

Zhou, Z., Qin, J., Xiang, X., Tan, Y., Liu, Q., & Xiong, N. N. (2020). News text topic clustering optimized method based on TF-IDF algorithm on spark. *Computers, Materials & Continua*, *62*(1), 217-231.

Wu, H., & Yuan, N. (2018, May). An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information. In *Proceedings of the 3rd International Conference on Intelligent Information Processing* (pp. 211-215).

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

Zhao, F., Kumamoto, E., & Yin, C. (2021, July). The effect and contribution of e-book logs to model creation for predicting students' academic performance. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 187-189). IEEE.

Riestra-González, M., del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, *163*, 104108.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521-563.

Yin, C., Uosaki, N., Chu, H. C., Hwang, G. J., Hwang, J. J., Hatono, I., & Tabata, Y. (2017, December). Learning behavioral pattern analysis based on students' logs in reading digital books. In *Proceedings of the 25th international conference on computers in education* (pp. 549-557).

Aher, S. B., & Lobo, L. M. R. J. (2012). Course recommender system in E-learning. *International Journal of Computer Science and Communication*, *3*(1), 159-164.

Reddy, S., Labutov, I., & Joachims, T. (2016, April). Learning student and content embeddings for personalized lesson sequence recommendation. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 93-96).

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29-48).

Chiba, Katsuhiro. (2017). The effect of extensive reading on English reading speed. *Bulletin of the Faculty of International Studies at Bunkyo University*, *28*(1), 57-65.

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University* (pp. 173-181). Springer London.