JASPER: Journal Article Selection Program for Non-native English Readers

Nantapong Keandoungchun [1], Jitimon Angskun [2,*], and Thara Angskun [2]

[1] School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok 10600, Thailand
[2] Institute of Digital Arts and Science, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
Email: nantapong.kean@kmutt.ac.th (N.K.); jitimon@g.sut.ac.th (J.A.); angskun@sut.ac.th (T.A.)
*Corresponding author

***Abstract*—Typically, reading a journal article can be time-consuming, mainly for non-native English readers, because academic writing usually uses complicated vocabulary and sentences. Therefore, this paper proposes a Journal Article Selection Program for Non-native English Readers (JASPER) for selecting journal articles from abstracts using scanning and skimming techniques. JASPER employs linear searching as a scanning technique and a novel multi-layer Latent Dirichlet Allocation (LDA) as a skimming technique. It automatically classifies journal articles into multi-layer topics and selects only articles with related topics to reduce the number of articles readers must read. JASPER is evaluated in terms of accuracy and efficiency using journal articles on Computer Science topics. It achieved an average of 82.62% of the F-measure. It can also reduce the number of journal articles by an average of 98.68%. Therefore, JASPER can practically reduce the number of journal articles for non-native English readers.**

***Keywords*—journal article selection, Journal Article Selection Program for Non-native English Readers (JASPER), article reduction, multi-layer Latent Dirichlet Allocation (LDA), multi-layer topic modeling**

I. Introduction

Knowledge extraction from various sources (such as books, journals, or academic articles) is essential to help people identify the knowledge needed to carry out such tasks. For example, in terms of education, students need to read many journal articles to draw on their knowledge for tasks such as homework and research. However, these articles are usually written in English. Unfortunately, most Thai students still lack English reading skills [1]. In addition, the difficulty and complexity of the academic paper (i.e., the difficulty of the content or the complexity of words or sentences) [2, 3] is another factor that makes those students challenging to read and understand both English and the content of various articles at the same time. Moreover, there is also a limitation in terms of duration per semester, making it difficult for most students to improve within that short period. Hence, this research focuses on how to extract knowledge using human reading processes and how to design an application that will help extract that necessary knowledge to reduce spending time reading unnecessary articles.

There are several ways to process human reading: SQ3R (survey, question, read, recite, review) [4, 5], SQ4R [6], and SQ5R [7] (more on SQ3R in "relate") and SQ5R (in addition to SQ3R's "record" and "respond") sections). In the first three parts of the SQ3R process, students must thoroughly survey, question and read to extract knowledge. Reciting and reviewing is just a process for storing and organizing that knowledge to be ready for use. The other complementary approaches are similar and rely mainly on in-depth human reading skills to drive. However, this research focuses on finding and using that knowledge to reduce the number of unnecessary journal articles. Then surveying, questioning, and reading are the most critical parts. Since these are processes that allow students to search for and ask questions of interest, this allows those students to reduce journal articles that are inconsistent with their interests. This research examines techniques for exploring the contents more quickly, namely scan reading and skimming [8]. The scanning reading technique focuses mainly on locating keywords, while the skimming approach focuses on finding the main heading. Skimming requires more skill from the reader than scanning because the reader must understand both English and the content sufficiently to extract exciting topics from the content. Especially if there are many articles, it takes more effort and time.

Therefore, this research examines how machine learning can be applied to these two techniques, especially skimming, by allowing the system to assist in the survey question and read process of SQ3R [4]. This research mainly focuses on the part of skimming to extract knowledge. Therefore, linear search, the well-known method for stationary scanning techniques, was chosen. However, there are several machine learning methods, both supervised and unsupervised [9, 10], that can be applied to skimming techniques. Supervised learning relies on labeled data and uses other data as input to build the model. The resulting model is highly accurate but comes at the cost of extensive data preparation. On the other hand, unsupervised learning uses data association methods or clustering of the data without tagging data but requires expert analysis and interpretation.

Manuscript received March 9, 2023; revised May 8, 2023; accepted July 24, 2023; published January 18, 2024.

doi: 10.12720/jait.15.1.79-86

Regarding skimming, supervised learning requires topics to be defined first and then applied to the tag data. In contrast, unsupervised can be used to topic modeling, a method for finding topics considering words in each article. Therefore, unsupervised learning is more consistent with this research.

Latent Dirichlet Allocation (LDA) is one of the most well-known topic modeling methods, which is a generative statistical model [11, 12]. However, traditional LDAs are limited in terms of the number of layers that cannot be used to model multi-layer topics. Therefore, this research has modified the multi-layer LDA [13, 14] by improving the function of specifying the optimal number of sub-topics of all topics in each layer and the criteria for determining them. Therefore, this research proposed a journal article selection program, Journal Article Selection Program for Non-native English Readers (JASPER), based on the proposed multi-layer LDA topic model as a skimming technique. Thus, JASPER enables non-skilled readers to read many journal articles by reducing the number of journal articles.

The main contributions of this research are to identify the knowledge as the multi-layer topic model by using the proposed multi-layer LDA topic modeling and to use the knowledge for reducing the number of unnecessary documents. This research consists of four sections: Section I is the introduction. Section II is the JASPER Architecture, which describes the whole process of this research. After that, the multi-layer topic model will be shown as the result of JASPER in Section III. Finally, the summary of JASPER will be described in the Section IV.

II. JASPER ARCHITECTURE

A. Data Collection

Computer journals in ScienceDirect online databases between 2018 and 2020 are collected as preliminary documents. This process is to acquire essential knowledge for journal article selection. The accumulated knowledge comprises a corpus, documents, and a multi-layer topic model. The corpus is a list of words obtained from WordNet [15], while the documents contain journal articles and their corpus. Finally, the multi-layer topic model comprises topics, subtopics, the probability of each word within each topic, and the probability of each topic within each document.

B. Pre-processing

In pre-processing, collected journal articles are transformed into words. Pre-processing comprises five subprocesses: tokenization, data sanitization, stemming lemmatizing, and feature scoring.

The first step starts with tokenization, dividing a journal article into paragraphs, sentences, and words, and then cleaning up some unnecessary information, such as some words that are too short and all the stop words. Stemming, or lemmatizing, was then performed using Porter's algorithm [16] as the next step to transform the word into its root form or meaningful base forms, such as changing verbs from the past tense to the present tense. The final step is feature scoring, which uses word bags to count word frequencies, and the TF-IDF [17] method scores the weights of each corpus. These TF-IDF scores are then stored in the knowledge base and used in the following process.

C. Multi-layer LDA Topic Modeling

Topic modeling is a statistical model for identifying hidden "topics" in each document. Generally, Latent Dirichlet Allocation (LDA) is the most popular topic modeling. However, the traditional LDA is used to determine the single-layer topic model. This research proposes a modification of the conventional LDA by adding the ability to identify topics in the form of a hierarchical structure known as multi-layer LDA topic modeling. The multi-layer LDA topic modeling is illustrated as a hierarchical structure of topics and subtopics, as depicted in Fig. 1.

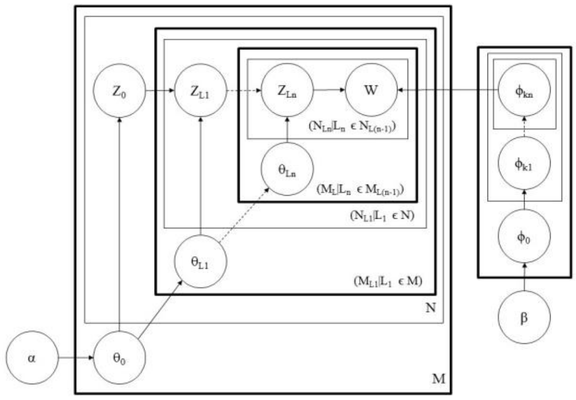


Fig. 1. Notation of Multi-layer LDA topic modeling.

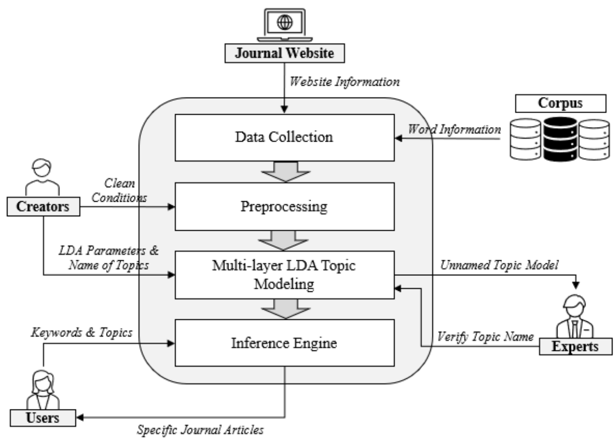


Fig. 2. JASPER architecture.

In Fig. 2, multi-layer LDAs are parameterized using author-defined combinations of α and β. Afterward, JASPER identifies each topic's optimal number of subtopics in each layer based on the harmonic mean of Coherence Value (CV) and Log-Likelihood Value (LLV). However, the multi-layer topic model remains anonymous, requiring experts in computer science to identify them. Finally, the resulting multi-layer topic model is stored in the designed database.

where

α is the parameter of the Dirichlet prior to the per-article topic distributions.

β is the parameter of the Dirichlet prior to the per-topic word distributions.

L is layer L.

M_L is the number of articles in layer L.

N_L is the number of words in layer L.

θ_{La} is the topic distribution of article a in layer L.

φ_{kL} is the word distribution of topic k in layer L.

Z_{Lan} is the topic of the n-th word in the article a in layer L.

W_{Lan} is the n-th word in the article a in layer L.

D. Inference Engine

When users input their interesting keywords, topics, or subtopics to JASPER, JASPER performs keyword searching to detect the areas. It applies multi-layer LDA topic modeling to identify the topic of each document. In this process, topics and subtopics in each layer are classified by skimming pre-processed journal articles using a multi-layer topic model. After that, to assign a specific topic or subtopics to each journal article, JASPER calculates the scoring criteria for each layer and then selects the topic for each journal article. The scoring criteria are calculated from Eq. (1).

$$p(A|\alpha,\beta)=\prod_{a=1}^{M}\int p(\theta_a|\alpha)\left(\prod_{n=1}^{N_a}\sum_{Z_{an}} p(Z_{an}|\theta_a)\,p(w_{an}|Z_{an},\varphi)P(\varphi|\beta)\right)d\theta_a d\varphi \tag{1}$$

where

α is the parameter of the Dirichlet prior to the per-article topic distributions.

β is the parameter of the Dirichlet prior to the per-topic word distribution.

A is the specific article.

M is the number of articles.

N is the number of words in each article.

θ_a is the topic distribution of article a.

φ_k is the word distribution of topic k.

Z_{an} is the topic of the n-th word in article a.

W_{an} is the n-th word in article a.

TABLE I. EXAMPLE OF SELECTING SUBTOPICS IN A LAYER WITH THREE SUBTOPICS

Articles	Scores			Selected Subtopic
	Subtopic 1	Subtopic 2	Subtopic 3	
Article 1	0.58	0.35	0.17	*Subtopic 1*
Article 2	0.08	0.00	0.92	*Subtopic 3*
Article 3	0.11	0.82	0.07	*Subtopic 2*

From the calculation of Eq. (1), the probability of each topic in each layer is calculated, and the maximum value is used as a criterion for selecting the appropriate topic for that article in each layer. An example of selecting subtopics in a layer with three subtopics is shown in Table I.

III. RESULT AND DISCUSSION

The experimental results in this research consist of three aspects as follows.

A. Results of the Multi-layer LDA Modeling

For creating the multi-layer topic model, this research uses the harmonic mean of coherence value and log-likelihood value to determine each topic's appropriate number of subtopics in each layer. The experimental result indicates three layers of the multi-layer topic model, as illustrated in Fig. 3.

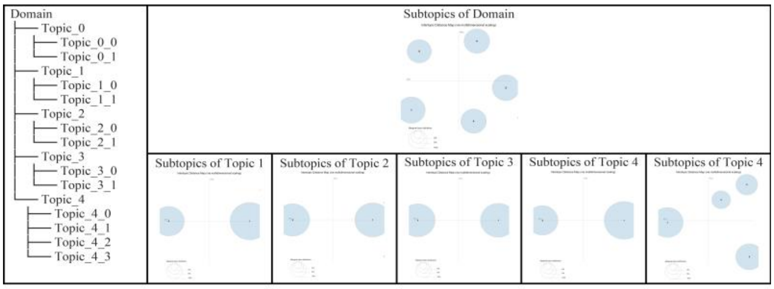
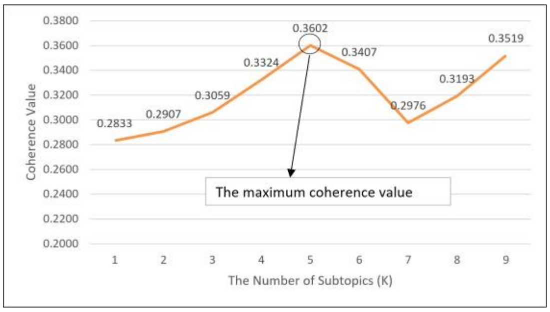
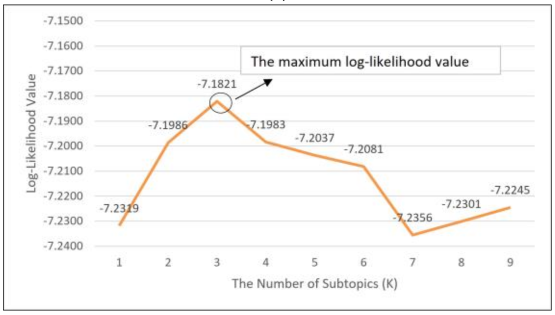


Fig. 3. Multi-layer topic model of JASPER.

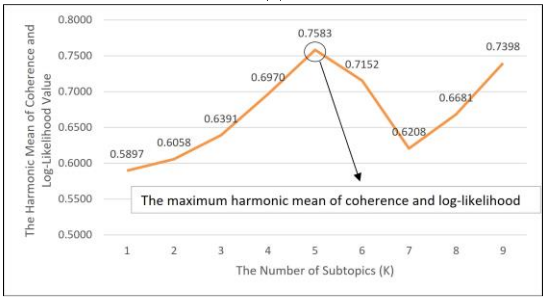
For an example of the first layer, there is only one domain topic: Computer Science. Moreover, JASPER uses the criteria to determine the suitable subtopics as follows. Initially, the system considers the Coherence Value (CV), the Log-Likelihood Value (LLV), and the harmonic mean of the two values, as shown in Fig. 4. A CV generally represents a suitable topic format, while a larger LLV is less likely to have topic overlaps. However, this research focuses on finding a multi-layer topic model. Hence the topics must be transparent and independent. Therefore, the harmonic mean of the two values is used to calculate that equilibrium point and determine the optimal number of subtopics. For example, Fig. 4(a) shows the highest CV of 0.3602, which yields five subtopics, while Fig. 4(b) shows the highest LLV of −7.1821, which generates three subtopics. The two criteria lead to different results. Then when considering the two values together, the maximum harmonic mean is 0.7583, of which five subtopics are the optimal number, as shown in Fig. 4(c). This research combines the CV and LLV using the harmonic mean of CV and LLV in Eq. (2).



(a)



(b)



(c)

Fig. 4. The criteria values of the Computer Science topic: (a) Coherence Values (CV), (b) Log-Likelihood Values (LLV), (c) Harmonic means of CV and LLV.

$$Harmonic\ Mean\ of\ CV\ and\ LLV = \frac{2 \times CV \times LLV}{CV + LLV} \tag{2}$$

Secondly, the JASPER examines the produced topics and their relevant terms (keywords). Fig. 5 visualizes the five subtopics of the Computer Science topic and shows that each subtopic has independent words. It also offers an example of the top 30 most relevant terms of the first subtopic.

However, even though the resulting subtopics remain unnamed, the related words and their probabilities for each subtopic are available. For example, each subtopic's top 10 relevant terms are shown as Eqs. (3)–(7).

Subtopic 1 (CA) = Time (0.013) + Propose (0.011) + Network (0.010) + System (0.010) + User (0.009) + Perform (0.009) + Data (0.008) + Paper (0.008) + Application (0.008) + Base (0.007) (3)

Subtopic 2 (DS) = Feature (0.015) + Data (0.015) + Learn (0.013) + Base (0.010) + Propose (0.010) + Approach (0.008) + Result (0.007) + Query (0.007) + Network (0.007) + Paper (0.007) (4)

Subtopic 3 (SDD) = Model (0.030) + Process (0.017) + Software (0.015) + Develop (0.014) + Analysis (0.011) + Approach (0.011) + Test (0.010) + Language (0.009) + Design (0.008) + Method (0.008) (5)

Subtopic 4 (IS) = Problem (0.018) + Program (0.015) + Algorithm (0.014) + Model (0.012) + Base (0.010) + Optimize (0.009) + Propose (0.009) + Compute (0.009) + Constraint (0.008) + Time (0.008) (6)

Subtopic 5 (AI) = Propose (0.024) + Method (0.024) + Algorithm (0.016) + Estimate (0.014) + Model (0.012) + Base (0.012) + Data (0.010) + Perform (0.010) + Image (0.009) + Result (0.009) (7)

As the equation above, the multi-layer topic model in the first layer is constructed with five subtopics: First is Computer Architecture (CA), followed by Data Science (DS), Software Design and Development (SDD), and Information Systems (IS) respectively, and finally, Artificial Intelligence (AI). These names are considered by experts based on the high to low probability of relevant terms for each subtopic. Eventually, when the system finds appropriate subtopics, it repeats these steps on every subsequent subtopic in every layer until the layer limit is reached or until no further subtopics are available (in technical, it means there is only one subtopic).

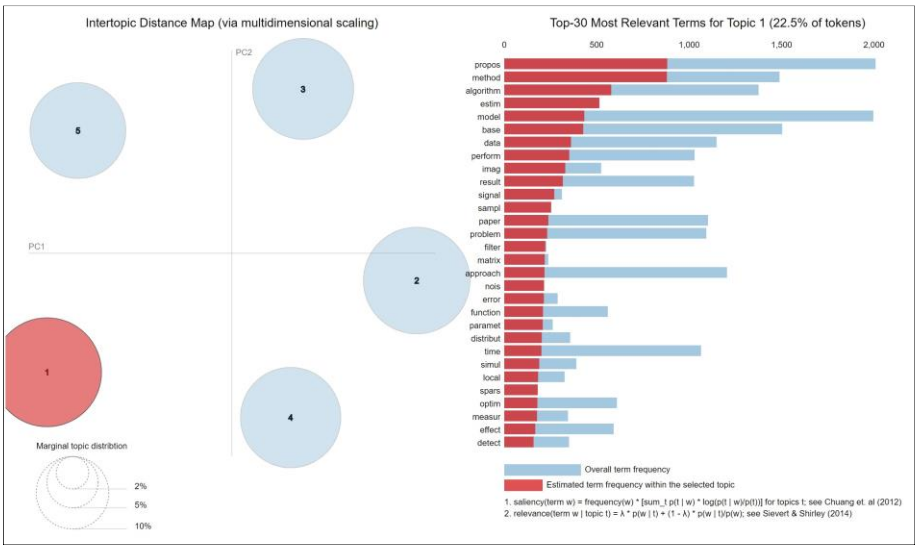


Fig. 5. Relevant terms of the first subtopic of the Computer Science topic.

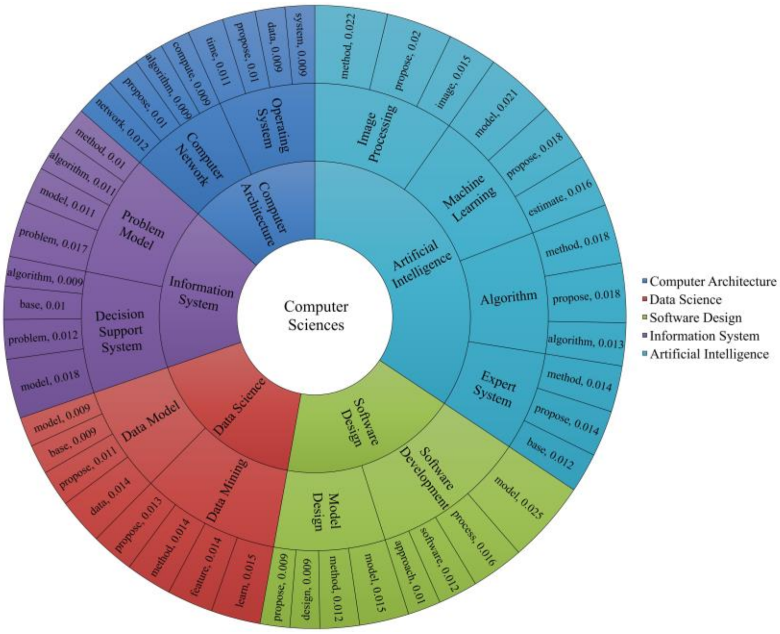


Fig. 6. Three-layer topic model of the Computer Science topic.

The experimental results in all layers showed that Computer Science topics consisted of 3 layers, as shown in Fig. 6.

B. *Model Accuracy*

JASPER's multi-layer topic model consists of three layers derived from the previous step. Therefore, to assess the accuracy at each layer, this research used accuracy, recall, and F-measure [18], well-known performance measures. In this research, the test set was used by computer science experts to evaluate the model's validity. The results of accuracy evaluation, recall, and F-measurement are shown in Table II.

TABLE II. THE EVALUATION OF TOPIC MODEL ACCURACY

Scoring Measure	Weight	Precision	Recall	F-measure
Layer 1	*An evaluation of this layer is unnecessary because there is only one topic.*			
Layer 2 (Computer Science)				
− Computer Architecture (CA)	0.2083	0.7877	0.9200	0.8487
− Data Science (DS)	0.1717	0.8854	0.8252	0.8543
− Software Design and Development (SDD)	0.2383	0.8936	0.8811	0.8873
− Information System (IS)	0.1400	0.8875	0.8452	0.8659
− Artificial Intelligence (AI)	0.2417	0.9197	0.8690	0.8936
Weighted Average of Layer 2		**0.8756**	**0.8717**	**0.8721**
Layer 3 (Computer Architecture)				
− Operating System (OS)	0.0950	0.8654	0.7895	0.8257
− Computer Network (CN)	0.1133	0.6702	0.9265	0.7778
Layer 3 (Data Science)				
− Data Model (DMD)	0.0867	0.8478	0.7647	0.8041
− Data Mining (DMN)	0.0850	0.8600	0.8269	0.8431
Layer 3 (Software Design and Development)				
− Model Design (MD)	0.0667	0.7600	0.9500	0.8444
− Software Development (SDV)	0.1717	0.9011	0.7961	0.8454
Layer 3 (Information System)				
− Problem Model (PM)	0.1000	0.8654	0.7500	0.8036
− Decision Support System (DSS)	0.0400	0.7500	0.8750	0.8077
Layer 3 (Artificial Intelligence)				
− Expert System (ES)	0.0150	0.7000	0.7778	0.7368
− Algorithm (A)	0.0800	0.9302	0.8333	0.8791
− Image Processing (IP)	0.0283	0.8000	0.9412	0.8649
− Machine Learning (ML)	0.1183	0.8750	0.7887	0.8296
Weighted Average of Layer 3 (Overall)		**0.8378**	**0.8250**	**0.8262**

According to Table II, the experimental results show that the model accuracy rate decreases accordingly for each layer, which can be investigated from the weighted average of the overall F-measurements in Layer 2 and Layer 3, 87.21%, and 82.62%, respectively. Since the previous layer is invalid, the next layer will also be invalid. Hence the accuracy decreases by layer. However, the overall weighted average F value of 82.62% is sufficient to adopt a multi-layer topic model to reduce the number of unnecessary journal articles, although it still needs some fixing, as described in the next section.

C. Percentage of Journal Article Reduction

JASPER usage starts when users enter their interesting keywords, topics, or subtopics into JASPER. This assessment aims to evaluate the model's effectiveness in reducing the number of journal articles. This assessment is based on 15 keywords from the collection related to each topic or subtopic. These keywords were collected from five volunteers who wanted to read the journal articles.

After users enter keywords, the JASPER uses a scanning function to retrieve the journal articles by applying the linear search. Then it uses a skimming technique by applying the multi-layer LDA topic modeling to find the topic of journal articles.

This research evaluates the retrieved journal article reduction by using a percentage value. A percentage is a simple measure that indicates how much a given value compares to the actual value. The percentage of journal articles reduction (PER_{Reduce}) is shown in Eq. (8).

$$PER_{Reduce} = \frac{Article_{Actual} - Article_{Retrived}}{Article_{Actual}} \times 100 \quad (8)$$

where

PER_{Reduce} is a percentage of journal articles reduction.

$Article_{Actual}$ is the number of actual journal articles.

$Article_{Given}$ is the number of given journal articles.

The percentage of journal article reduction in each keyword can be summarized in Fig. 7. This figure shows the PER_{Reduce} of each keyword in each layer. The journal article's reduction percentage slightly increases in each layer. The deeper the layer, the more journal articles will get pruned. The average rate of journal article reduction of Layers 1, 2, and 3 are 92.34%, 97.87%, and 98.68%, respectively.

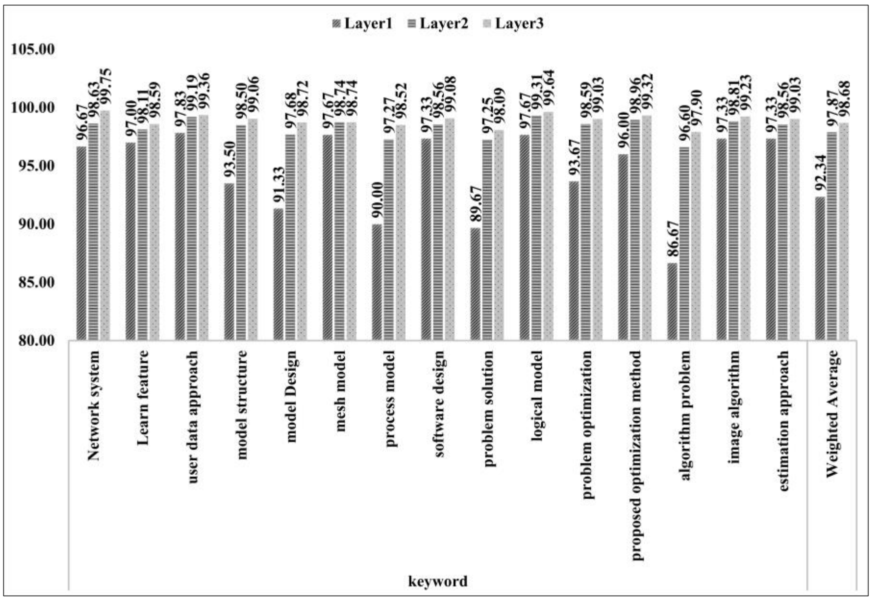


Fig. 7. The evaluation of the percentage of journal article reduction.

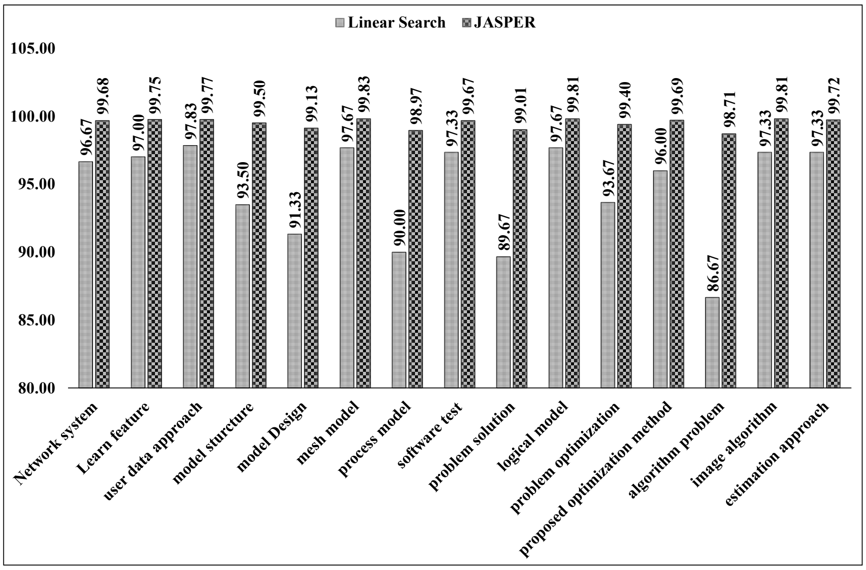


Fig. 8. The comparison of JASPER and linear search.

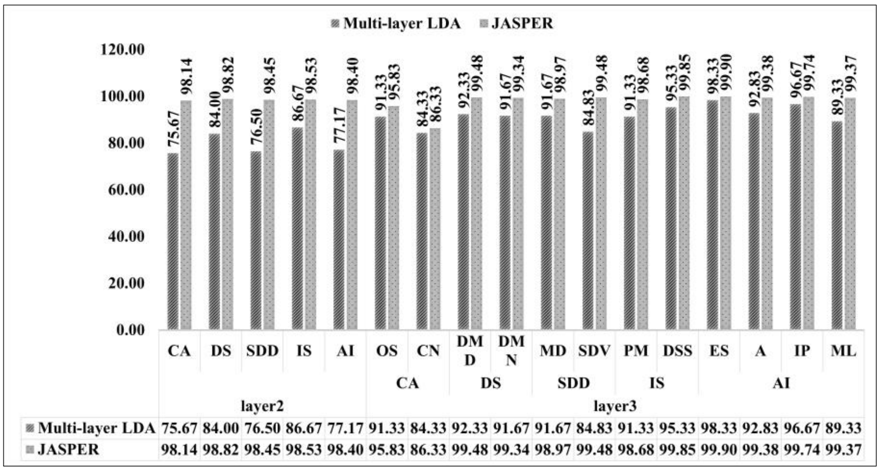


Fig. 9. The comparison of JASPER and multi-layer LDA modeling.

Finally, JASPER summarized the weighted average of the overall journal article reduction at 98.68%, calculated from all topics without subtopics in each layer. Readers must read only eight journal articles related to that keyword and topic.

Additionally, Tenopir *et al*. [19] reported that most people spent 29.3 min per article reading and an average of 17.25 h per week reading. Therefore, most people need to read at least 34 articles per week. Accordingly, readers only spend about once a week reading those articles. Thus, this research summarized that JASPER could sufficiently reduce the number of journal articles for readers.

D. Comparison of JASPER and Other Techniques

The JASPER combines both scanning and skimming techniques. Two techniques are used for comparing the JASPER: a linear search (scanning) and a multi-layer LDA topic modeling (skimming). This research uses the percentage of journal articles reduction as a comparison value. The comparison of JASPER and the solely linear search technique is shown in Fig. 8. The comparison of JASPER and the sole multi-layer LDA modeling technique is demonstrated in Fig. 9. According to Fig. 8 and Fig. 9, the results reveal that JASPER performs better than those two traditional techniques.

IV. CONCLUSION

This research focuses on the problems of reading journal articles of students or researchers who need more English reading skills. They can only read a few journal articles within a limited time. Therefore, this research proposed JASPER, a journal article selection program for non-native English readers. The JASPER aims to reduce the number of unrelated journal articles by combining linear search and multi-layer LDA topic modeling.

Applying the topic modeling revealed that the Computer Science topic could be classified into three layers. There will be a few topics at each layer: for layer 1, there is one topic: Computer Science. It can be divided into five sub-topics at Layer 2: CA, DS, SDD, IS, and AI. These sub-topics can be further subdivided into Layer 3. Layer 3 has 2, 2, 2, 2, and 4 sub-topics of each topic in Layer 2.

The JASPER is evaluated in model accuracy and percentage of journal article reduction. The evaluation results revealed that the proposed journal selection model achieved 82.62% of the F-measure, and the average rate of journal article reduction was 98.68%.

Nevertheless, this research has some limitations. Journal articles applicable to this research are still limited to computer science articles, and their quantity is still limited. In addition, the topic names derived from this research also require experts to assist in reviewing them before deployment.

Some improvements could be made shortly. Future work can increase the number of journal articles or other disciplines of journal articles. In addition, developing an engine that can help interpret knowledge instead of using experts may create a thesaurus or ontology to increase the program's efficiency even further.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

N.K. formulated research questions. T.A. involved intellectual contribution, model analysis, and model evaluation. J.A. and N.K. generated figures and tables. Writing and reviewing were done by all authors. All authors had approved the final version.

REFERENCES

[1] N. Sittirak and S. Pornjamroen, "The survey of British English and American English vocabulary usage of Thai students," *Songklanakarin Journal of Social Sciences and Humanities*, vol. 15, no. 4, pp. 559–575, 2009.

[2] B. Y. Cahyono, "Effectiveness of journal writing in supporting skills in writing English essays," *The Journal of Education*, vol. 4, pp. 310–318, 1997.

[3] S. Yulianti, S. Nuraeni, and A. Parmawati. "Improving students' writing skill using brains writing strategy," *Project (Professional Journal of English Education)*, vol. 2, no. 5, 714, 2019.

[4] I. N. Aziz, "Implementation of SQ3R method in improving the students' basic reading skill," *Journal of Education*, vol. 5, no. 1, pp. 97–106, 2019.

[5] G. R. Anjuni and R. Cahyadi, "Improving students' reading comprehension through SQ3R (Survey, Question, Read, Recite and

Review) technique," *Project (Professional Journal of English Education)*, vol. 2, no. 1, pp. 1–6, 2019.
[6] M. Başar and M. Gürbüz, "Effect of the SQ4R technique on the reading comprehension of elementary school 4th grade elementary school students," *International Journal of Instruction*, vol. 10, no. 2, pp. 131–144, 2017.
[7] T. Sangcharoon and W. Sripetpu, "Reading and writing skills development: Use of SQ5R technique," in *Proc. the 2nd International Conference on Humanities and Social Sciences*, 2010.
[8] H. Ismail, J. K. Syahruzah, and Basuki, "Improving the students' reading skill through translation method," *Journal of English Education*, vol. 2, pp. 124–131, 2017.
[9] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.
[10] M. Srinivas, G. Sucharitha, and A. Matta, *Machine Learning Algorithms and Applications*, Hoboken: Wiley-Scrivener, 2021.
[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[12] H. Jelodar, Y. Wang, C. Yuan, X. Feng, and X. Jiang, "Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, pp. 15169–15211, 2019.
[13] N. Keandoungchun and N. Thammakoranonta, "A design of a journal article selection model based on abstract using combination of scanning and skimming techniques," Doctoral dissertation, NIDA, 2021.
[14] N. Keandoungchun and N. Thammakoranonta, "A journal article selection based on combination of scanning and skimming techniques," *Naresuan University Journal: Science and Technology*, vol. 30, no. 1, pp. 30–41, 2022.
[15] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.
[16] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
[17] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.
[18] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall, and F-score, with implication for evaluation," in *Proc. the 27th European Conference on Advances in Information Retrieval Research*, Santiago de Compostela, Spain, 2005, pp. 345–359.
[19] C. Tenopir, R. Mays, and L. Wu, "Journal article growth and reading patterns," *New Review of Information Networking*, vol. 16, no. 1, pp. 4–22, 2011.