A Journal Article Selection based on a Combination of Scanning and Skimming Techniques

Nantapong Keandoungchun and Nithinant Thammakoranonta*

School of Applied Statistics, National Institute Development of Administration, Bangkok, Thailand

* Corresponding author. E-mail address: nithinan@as.nida.ac.th

Received: 5 January 2021; Revised: 15 March 2021; Accepted: 22 March 2021; Available online: 28 May 2021

Abstract

Typically, an academic writing style of journals is complicated because those journals use complex vocabularies and complex sentences. It might be hard for students to read documents in detail in short period of time. Therefore, this research proposes a Journal Article Selection (JAS) based on a combination of scanning and skimming techniques to reduce number of documents sufficient to support unskilled readers. JAS is developed by python on Google Colaboratory (Bisong, 2019). The designed JAS is assessed in three aspects which are the optimal number of topics and subtopics by considering the coherence value and the relevant terms within topic, the correctness of model and the percentage of the document reduction. The experimental results revealed that there are three layers of topic model: Medical, Business Law and Computer are in layer 1, Social Network and Information System are in layer 2, and Process Model, General Problem and Artificial Intelligence are in layer 3. On other aspects, the proposed model is achieved in 77.36 average percentage of F-measure and 94.15 average percentage of unnecessary document reduction. In conclusion, it can be concluded that JAS can reduce documents sufficient for readers to read in short period of time.

Keywords: Journal Article Selection, Topic Modeling, Scanning, Skimming

Introduction

Most of documents are usually written in English language such as novels, magazines, and journals. Readers who want to have knowledge from these documents need to use different reading styles because of their different nature. For example, novel readers need to read in detail for gathering the whole story, while magazine readers may skip through the pages to see an interested article. In case of academic journal, students need to read a lot of journals to gain knowledge and use it for their research. However, academic writing style of journals is complicated. They used complex vocabularies and complex sentences (Cahyono, 1997). It might be hard for students to read documents in detail in short period of time. In fact, all students may improve their background knowledge and their reading skills in order to speed up their reading. Reading experts normally use reading techniques to identify only specific important area before reading it in detail.

There are two well-known reading techniques. First, scanning which is a technique that uses keywords to detect information in specific area. Second, skimming which is a quick reading to find out what topic of the document is in order (Ismail, Syahruzah, & Basuki, 2017). Those reading techniques can support many readers to reduce their reading time. Unfortunately, the efficiency and effectiveness of these reading techniques are depended on reading skill of readers. Especially, in Thai culture, there are some students who weak on the English reading skill (Sittirak & Pornjamroen, 2009). It is also hard for them to improve their reading skill in short period of time.

Scanning and skimming reading techniques can be applied with machine learning for helping readers to drop out unnecessary information and select only specific interested area. Unskilled readers will take less time for reading only specific areas than reading the whole paper. Scanning or searching method, in term of

machine learning, is to identify a location of keywords from unknown sources. Nowadays, there are many searching algorithms such as linear search, binary search or brute force search (Usman, Bajwa, & Afzal, 2014). Only linear search is employed in this paper.

This research is also interested in the skimming technique because it tries to understand what documents refer to. The skimming technique can be applied with machine learning, which have two approaches called supervised learning and unsupervised learning. Some of related work employed supervised learning (e.g., Decision Tree and Naïve Bayes) (Ide & Véronis, 1998) and apply labelled corpus as a training set. Others use unsupervised learning (e.g., Latent Dirichlet Allocation (LDA)), to identify patterns in large data without using labelled corpus (Navigli, 2009).

Although supervised learning can provide more accuracy than unsupervised one, but it must spend more time to label on the training corpus. It is too cumbersome to label every word with a topic. Therefore, topic modeling can handle this issue by automatically labelling the topics of a set of documents. Topic modeling is an unsupervised machine learning technique that can scan a set of documents, detect words within them, and automatically cluster word groups and similar expressions that best characterize a set of documents. Latent Dirichlet Allocation (LDA) is one of well-known topic modeling, that can use to identify topics with single layer. Unfortunately, LDA is unsuitable for identifying topics and their subtopics in several layers. Hence, this research proposed a modified version of LDA to support multi layers called Multi-Layer Latent Dirichlet Allocation (Multi-Layer LDA), which enhance an ability to identify topics and subtopics in any layers.

This research aims to support unskilled readers to read a lot of papers within short period of time. A journal article selection has been proposed based on combination of scanning and skimming techniques to reduce number of documents or paragraphs. To evaluate the model, this research collected journals from three distinct areas, which are Medical, Business Law and Computer. Furthermore, finding more subtopics is also evaluated with computer documents.

Methods and Materials

The proposed Journal Article Selection (JAS) is a framework that combine scanning and skimming techniques, which are human reading techniques, to identify specific areas which students should be concerned. JAS is developed by python on Google Colaboratory (Bisong, 2019). Technically, the JAS applied linear search and Multi-Layer LDA as scanning and skimming technique, respectively. The proposed framework consists of three main components which are Knowledge base, Preprocessing, Scanning and skimming

Knowledge base

Knowledge base is designed based on multi-layer topic model to contain essential information and knowledge for the journal article selection. Knowledge base consists of three components which are corpus, documents, topic model. Corpus in this research is list of words obtaining from WordNet. Documents component contains list of documents and their corpus. Documents component is acquired from Computer journals in ScienceDirect online databases between 2018 and 2020. Finally, topic model component consists of topics and their subtopics, probability of each topic within each document, and probability of each word within each topic.

Pre-processing

Pre-processing is a process to transform collected documents into word sequences which are a group of words with their position. Pre-processing is composed of two subprocesses as shown in Figure 1.

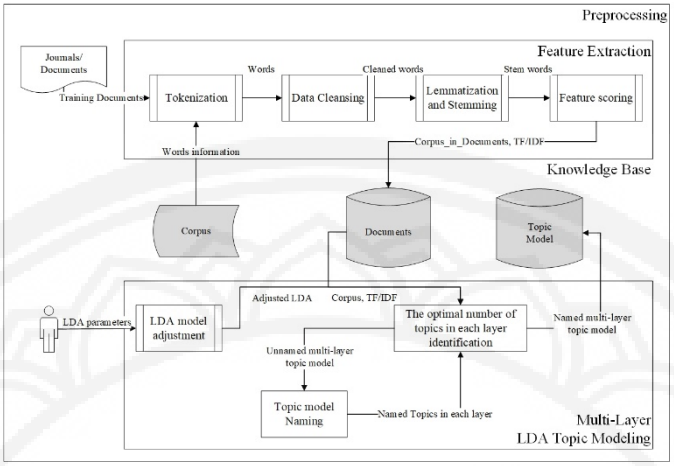


Figure 1 Preprocessing of Journal Article Selection (JAS)

Feature Extraction

This subprocess is used for extracting features from all documents or journals as the following steps.

The first step is tokenization. This step is to split documents into paragraphs, sentences, and words. Second, data cleansing step is to clean some unnecessary data such as all stop words and some too short words. Third, lemmatization and stemming step is to lemmatize words to its simple form such as changing verbs in past tenses into present tenses. In addition, it also applied Porter's stemming algorithm to transform words to stem, which is their root form. Finally, feature scoring step is performed by applying a bag of words for counting word frequency and TF-IDF approaches to score weight of each corpus. Then, storing their scores into knowledge base.

Multi-Layer Latent Dirichlet Allocation (Multi-Layer LDA)

Topic modeling is a task of discovering the hidden "topic" that occurring in documents. Latent Dirichlet Allocation (LDA), a generative statistical model, is one of the most popular topic modeling. However, traditional LDA is not suitable for this research because LDA normally uses to identify topics in one dimension. Hence, this research purposed Multi-layer LDA, which is modified version of traditional LDA to identify the optimal number of subtopics of each topic in each layer. Multi-layer LDA represents topics and subtopics in hierarchy structure as shown in Figure 2.

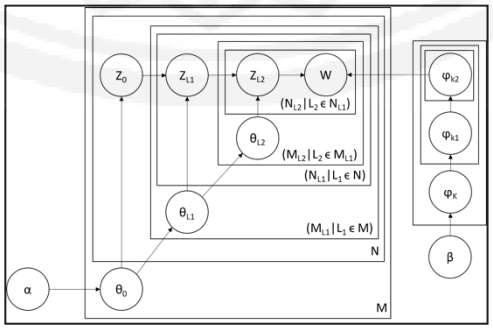


Figure 2 Multi-layer LDA Topic Modeling Notation with 3 layers

The parameters illustrated in Figure 2 are described where

L denotes the layer L.

M_L denotes the number of documents in level L.

N_L denotes number of words in each document in level L.

α denotes the parameter of the Dirichlet prior on the per-document topic distributions.

β denotes the parameter of the Dirichlet prior on the per-topic word distributions.

θ_{Ld} denotes the topic distribution for document d in level L.

φ_{kL} denotes the word distribution for topic k in level L.

Z_{Ldn} denotes the topic for the n-th word in document d in level L.

W_{Ldn} denotes the specific word.

Scanning and Skimming Models

Scanning and skimming models are processes to reduce enormous lists of documents or journals. The steps of the scanning and skimming models are shown as Figure 3.

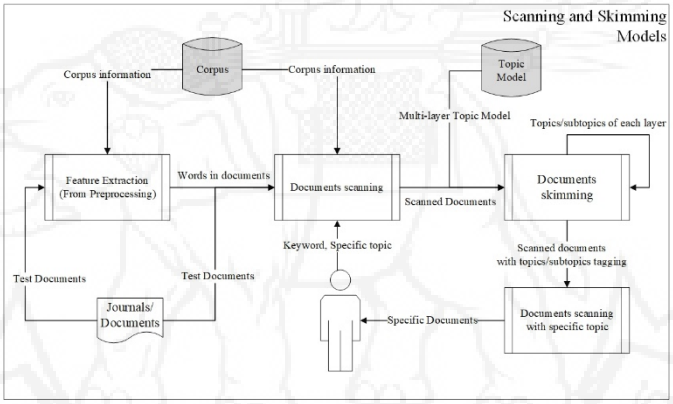


Figure 3 Scanning and Skimming Models

As showing to Figure 3, unseen documents is operated by feature extraction to acquire words in each document. After that, when readers input keyword and topic, the model will apply scanning function with keyword to reduce a ton of documents before applying skimming function to find out topics and subtopics on the scanned documents. Finally, the model will apply the scanning function again with specific topic and return the specific documents to readers.

The specific topic of each document is selected by a scoring measure. The scoring measure is calculated from equation (1).

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_{dn}|\theta_d)\, p(w_{dn}|Z_{dn},\beta) \right) d\theta_d \qquad (1)$$

where

$p(D|\alpha,\beta)$ denotes the probability of topics in each document with the parameter α and β.

$p(\theta_d|\alpha)$ denotes the probability of the topic distribution for document d with the parameter α.

$p(Z_{dn}|\theta_d)$ denotes the probability of the topic for specific word with the topic distribution for document d.

$p(w_{dn}|Z_{dn},\beta)$ denotes the probability of specific word with the topic of that word and the parameter β.

Results and Discussion

This research collected data from 978 journals which consist of 578 Computer journals, 200 Business Law journals and 200 Medical journals. As mentioned above, this research further explores on Computer journals. After that, there are three interested journals, which are Information System (IS), Artificial Intelligence (AI) and Social Network (SN) with 191, 200 and 187 documents, respectively have been explored. According to Dobbin and Simon (2011), they stated that the optimal proportion of case for the training set is to be in range of 40 to 80 percentage. Then, the designed model (JAS) uses 75 and 25 percentage of overall journals in the dataset as a training set and a test set, respectively. They are randomly selected.

The research was assessed in three aspects which are the optimal number of topics, the correctness of model and the percentage of the document reduction. The results of three aspects are presented and discussed in the next section.

Evaluation of the Optimal Number of Topics

Topic modeling refers to the task of discovering the hidden "topic" that appear in a collection of documents. The challenge is how to extract good quality of topics that are clear and meaningful as well as finding the optimal number of topics. In the experiment of this research, JAS need to identify the optimal number of topics in each layer by using multi-layer LDA. JAS consists of three layers as shown in Figure 4. In addition, the experiment in this research determines amount of iteration of Multi-Layer LDA to be 50 iterations per layer. The results of each layer are described in next section.

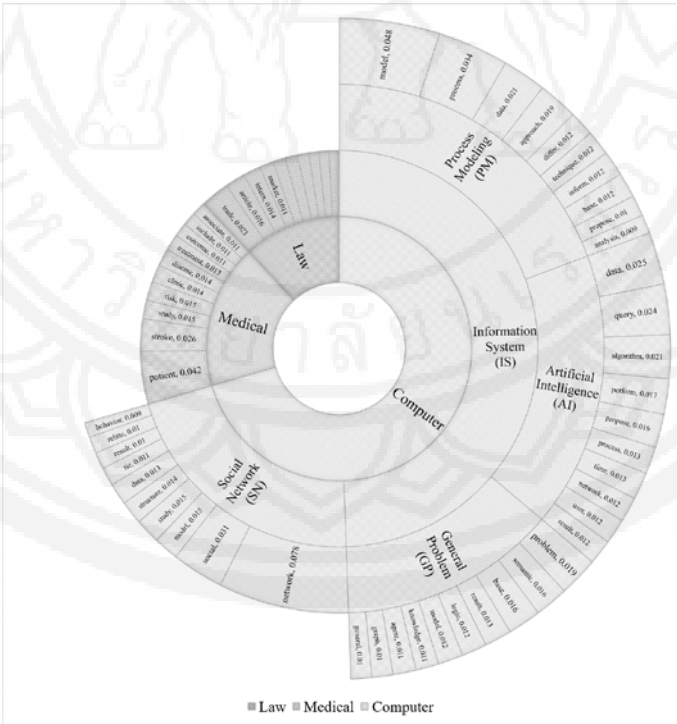


Figure 4 3-Layers Topic Model

The Optimal Number of Topics in Layer 1

To identify the number of topics, this research uses an LDA topic modeling to create topic models with two to ten topics. The number of topics (k) is considered from investigating the coherence value and the dependency of topics as illustrated in Figure 5 and Figure 6, respectively.

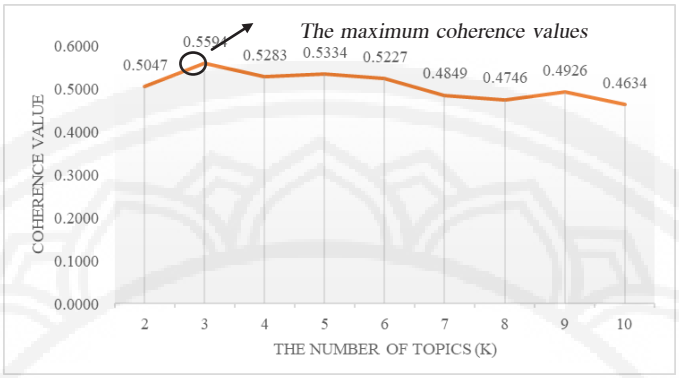


Figure 5 Coherence Values of each Topic Model in Layer 1

According to Figure 5, the coherence value provides an assessment tool for topic modeling. The results reveal that the 3-topic model has the maximum coherence value of 0.5594 and it slightly declined as number of topics increased.

Others, JAS evaluation is to examine the produced topics and the relevant terms (keywords). Each bubble on the plot in Figure 6 represents a topic. A good topic model will have huge and non-overlapping bubbles scattered. As shown in Figure 6, the 3-topic model produces independent topics, while the rest are dependence. Therefore, 3-topic model is the optimal number of topics for applying on the JAS. The detail of 3-topic model is shown as Figure 7.

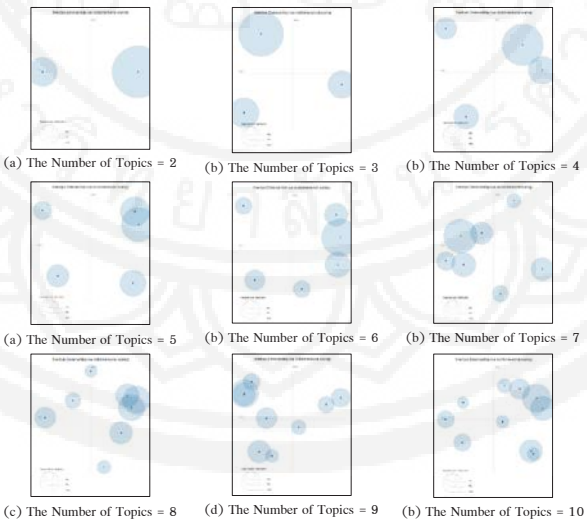


Figure 6 Topics Represented by Scatter Graph in Layer 1

Figure 7 visualizes the topics and the strongly relevant terms of the 3-topic model. This figure lists the top-30 most relevant terms for Topic 1.

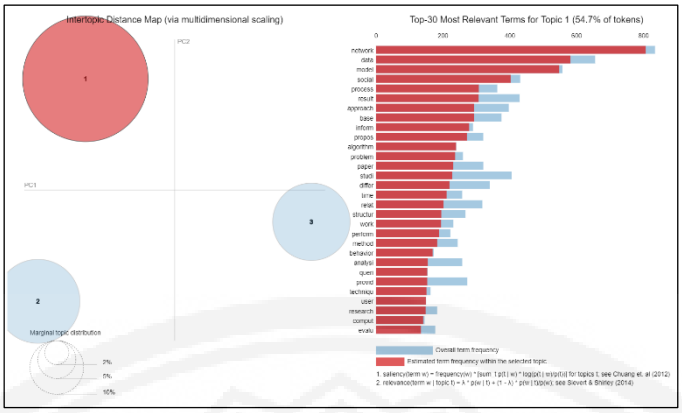


Figure 7 The Detail of 3-Topic Model in Layer 1

Unfortunately, the result only shows unnamed topics with their related terms. The relevant terms for each topic and their probability are represented as shown in equation (2), (3) and (4).

Topic 1 (Business Law) = 0.023*"trade" + 0.016*"article" + 0.014*"intern" + 0.011*"market" + 0.010*"economy" + 0.010*"agreement" + 0.009*"country" + 0.008*"state" + 0.008*"develop" + 0.008*"approach" (2)

Topic 2 (Medical) = 0.042*"patient" + 0.026*"stroke" + 0.015*"risk" + 0.015*"study" + 0.014*"disease" + 0.014*"clinic" + 0.013*"treatment" + 0.011*"associate" + 0.011*"include" + 0.011*"outcome" (3)

Topic 3 (Computer) = 0.030*"network" + 0.022*"data" + 0.021*"model" + 0.015*"social" + 0.012*"process" + 0.012*"result" + 0.011*"base" + 0.011*"approach" + 0.010*"inform" + 0.010*"propose" (4)

According to equation above, the relevant terms of each topic are considered by experts in each area to reveal that topic 1, topic 2 and topic 3 refer to Business Law, Medical and Computer topics, respectively.

The Optimal Number of Topics in Layer 2

In layer 1, documents are in field of Medical, Business Law and Computer. Layer 2 is drill-down into Computer topics. In this layer, this research will take an action to identify only subtopics of Computer topics. The result indicated that the optimal number of Computer topic is two subtopics with max coherence value of 0.3799. Hence, those two subtopics are Social Network (SN) and Information System (IS) as shown in Equation (5) and (6).

Topic 1 (Social Network (SN)) = 0.078*"network" + 0.031*"social" + 0.015*"study" + 0.015*"model" + 0.014*"structure" + 0.013*"data" + 0.011*"tier" + 0.010*"relate" + 0.010*"result" + 0.009*"behavior" (5)

Topic 2 (Information System (IS)) = 0.021*"model" + 0.017*"data" + 0.015*"process" + 0.013*"approach" + 0.013*"base" + 0.013*"propose" + 0.012*"algorithm" + 0.011*"problem" + 0.011*"result" + 0.009*"query" (6)

The Optimal Number of Topics in Layer 3

The same as layer 2, this step will deep down to identify more subtopics of those subtopics that acquired in layer 2.

The Optimal Number of Topics in Layer 3.1

Firstly, this part will try to identify subtopics of Social Network. The result shows that SN has two subtopics with max coherence value of 0.3971. Although, the result that SN has two subtopics, however, when considering on the relevant term of those two subtopics, it shows that both key relevant terms are similarly as shown in Equation (7) and (8). In this case, this research assumes that there are no more subtopics.

Topic 1 = 0.057*"network" + 0.022*"social" + 0.020*"data" + 0.019*"model" + 0.017*"behavior" + 0.013*"result" + 0.012*"differ" + 0.011*"level" + 0.010*"study" + 0.010*"active" (7)

Topic 2 =0.073*"network" + 0.030*"social" + 0.015*"study" + 0.015*"structure" + 0.013*"relate" + 0.012*"community" + 0.012*"model" + 0.012*"group" + 0.011*"individual" + 0.009*"tie" (8)

The Optimal Number of Topics in Layer 3.2

In this layer, the result reveal that six subtopics is the best with max coherence value of 0.3360. However, when considering the dependency of topics, there are some overlapping. In this case, the next priority, the 3-topic model will be considered instead of 6-topic model. Therefore, the 3-topic model with coherence value 0.3248 is more suitable to be used in this research. They are Process Modeling (PM), General Problem (GP) and Artificial Intelligence (AI) as shown in Equation (9), (10) and (11).

Topic 1 (Process Modeling (PM)) = 0.048*"model" + 0.034*"process" + 0.021*"data" + 0.019*"approach" + 0.012*"base" + 0.012*"inform" + 0.012*"technique" + 0.012*"differ" + 0.010*"propose" + 0.009*"analysis" (9)

Topic 2 (General Problem (GP)) = 0.019*"problem" + 0.016*"base" + 0.016*"semantic" + 0.013*"result" + 0.012*"model" + 0.012*"logic" + 0.011*"agent" + 0.011*"knowledge" + 0.010*"general" + 0.010*"graph" (10)

Topic 3 (Artificial Intelligence (AI)) = 0.025*"data" + 0.024*"query" + 0.021*"algorithm" + 0.017*"perform" + 0.016*"propose" + 0.013*"time" + 0.013*"process" + 0.012*"result" + 0.012*"user" + 0.012*"network" (11)

Evaluation of Model Correctness

The evaluation of model correctness is to measure performance of the JAS using precision, recall and F-measure as shown in equation (12), (13) and (14).

$$Precision = \frac{TP}{(TP+FP)} \times 100\% \quad (12)$$

$$Recall = \frac{TP}{(TP+FN)} \times 100\% \quad (13)$$

$$F\text{-}measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (14)$$

where

TP (True Positive) is the number of positive documents that are correctly selected.

FP (False Positive) is the number of negative documents that are incorrectly selected.

FN (False Negative) is the number of positive documents that are incorrectly selected.

TN (True Negative) is the number of negative documents that are correctly selected.

This research assign topics into test set which obtained from expert in each area (Business Law, Medical and Computer). Topic model of JAS consists of three layers, which obtained from previous step. Hence, this research evaluates its correctness of each layer using precision, recall and F-measurement.

The summary of model evaluation

The experimental results reveal that the model achieves 73.88, 82.60 and 77.36 average percentage of overall Precision, Recall and F-measure, respectively. Overall precision, recall and F-measure are calculated by computing with precision, recall and F-measure of topics without subtopics in every layer as shown in Table 1.

Table 1 Evaluation of Model Correctness

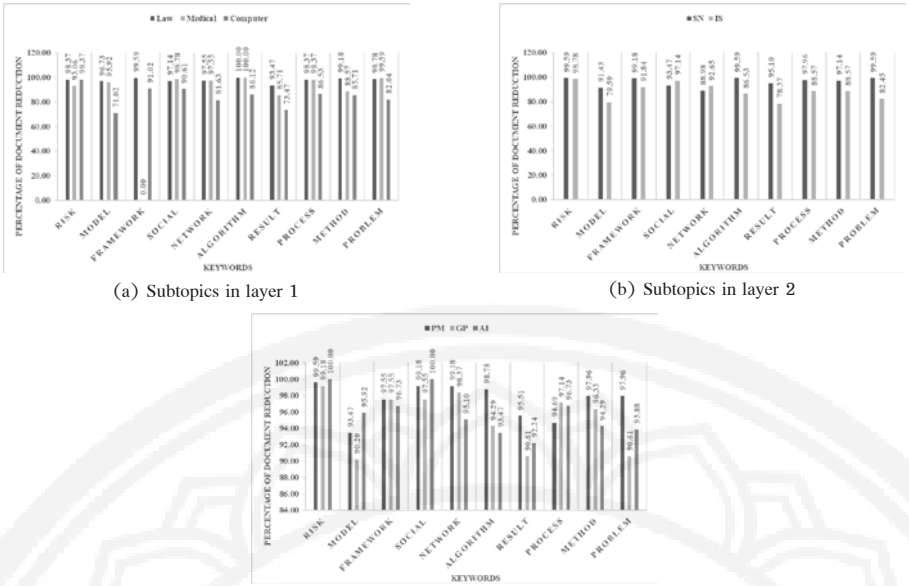
Scoring Measure	Precision	Recall	F-measure
Layer I			
Business Law	66.67	52.00	58.43
Medical	82.46	94.00	87.85
Layer II (Computer)			
Social Network	80.56	80.56	80.56
Layer III (Information System (IS))			
Process Modeling (PM)	56.67	85.00	68.00
General Problem (GP)	65.96	86.11	74.70
Artificial Intelligence (AI)	86.49	94.12	90.14
Weighted Average	73.88	82.60	77.36

According to Table 1, it reveals that Business Law and Process Modeling (PM) have less precision, recall and F-measure than others. That means there are some documents of Business Law in field of medical or computer and PM in field of GP or AI. Hence, this causes the model correctness drop only on Business Law and PM.

Evaluation of the Percentage of the Document Reduction

This evaluation is to assess the effectiveness of the model in the term of document reduction. This evaluation uses 10 test case consisting of keywords and topics which are collected from users in order to measure the number of document reduction.

The experimental results reveal that the journal article selection (JAS) can reduce unnecessary documents and show only documents related to a specific topic. Figure 8 depicts unnecessary document reduction of each topic in layer 1 with different keywords.



(a) Subtopics in layer 1

(b) Subtopics in layer 2

(c) Subtopics in layer 3

Figure 8 Evaluation of the Percentage of the Document Reduction in each layer

Figure 8 reveals the percentage of document reduction in each layer. For instance, a keywork "algorithm" can reduce 100.00, 100.00 and 86.12 percentage of Business Law, Medical and Computer, respectively, in layer 1 as illustrated in Figure 8(a). In this case, it guarantees that the keyword "algorithm" does not related to Business Law and Medical. While, with topic the keyword "algorithm" can reduce the number of Computer documents into 86.12 percent. Moreover, with specific topic "information System" in layer 2, the percentage of document reduction is increased to 86.53 percentage as shown in Figure 8(b). In layer 3, it is also increased to 98.78, 94.29, and 93.47 with specific topic "Process Modeling", "General Problem" and "Artificial Intelligence", respectively as shown in Figure 8(c).

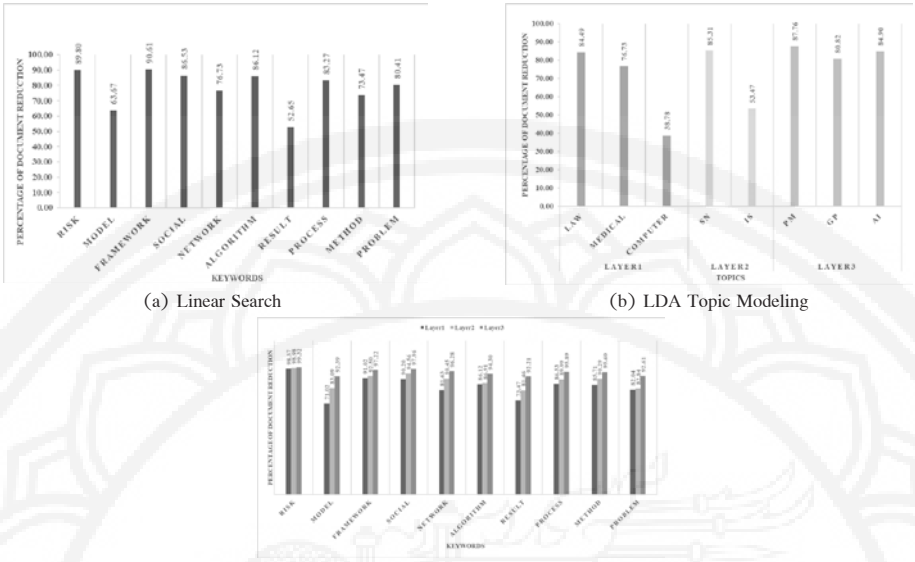
The average document reduction of overall 10 test cases in layer 1 is 97.18, 89.64 and 81.00 percentage of Business Law, Medical and Computer documents, respectively. In layer 2, it reduces 95.28 and 85.72 percentage of Social Network and Information System documents, respectively. In layer 3, there are 96.74, 93.95 and 95.03 percentage of Process Modeling, General Problem and Artificial Intelligence, respectively.

Finally, the proposed model (JAS) is achieved in 94.15 average percentage of unnecessary journal reduction which can approximately reduce 230 documents out of 245 documents. It means that readers need to read only 15 documents due to limited time. In addition, Tenopir, Mays, and Wu (2011) reports that most people spent 29.3 minutes per article reading, and also spent 17.25 average hours for reading weekly. Therefore, most people must read approximately at least 34 articles per week. Therefore, this research can summarize that JAS can reduce documents sufficient for readers to read in their course.

The Comparison of Linear Search, Multi-layer LDA and JAS

This research tries to compare JAS with other methods which are linear search and multi-layer LDA. On comparison, this research uses the same environment as when evaluating the JAS model. This comparison employs the percentage of document reduction as measurement. Linear search can reduce documents in average

of 78.33 percentage as shown in Figure 9(a). Multi-layer LDA topic modeling can achieve 82.56 percentage as shown in Figure 9(b). Finally, JAS reveals that the percentage of document reduction gets a high value of 94.15 percentage as shown in Figure 9(c).



(a) Linear Search (b) LDA Topic Modeling

(c) JAS

Figure 9 The Comparison on Percentage of the Document Reduction

Conclusion and Suggestions

This research aims to design and develop a journal article selection for supporting unskilled readers. It begins with analyzing the problem of reading journals or documents, followed by surveying the related work about scanning and skimming techniques. The results of studying present the problem statements that are necessary for designing and developing the proposed model. Problem statements is to complexed academic writing style of journals which might be hard for unskilled students to read documents in detail in short period of time.

Therefore, this research proposed a journal article selection based on the combination of scanning and skimming techniques. The proposed model is called a journal article selection (JAS) to reduce the number of documents. The JAS combines scanning and skimming techniques, which are human reading techniques, to identify specific areas which students should be concerned. The scanning technique applies linear search to detect areas which keywords are belong. Skimming technique in this research applied multi-layer LDA topic modeling to identify topics and subtopics of each document.

After the model is developed, the designed JAS is assessed in three aspects which are the identification of topic model, the correctness of model, and the percentage of the document reduction. The research findings are summarized as follows.

1) Evaluation of the identification of topic model for applying on the JAS, this research found that the topic model has three layers. Firstly, the topic in layer 1 consists of Business Law, Medical, Computer. Next, this research only focuses on computer area. There are two subtopics of computer topic in layer 2 which are Information System (IS) and Social Network (SN). The Social Network is assumed that there are no subtopics

because of its similarity of the relevant terms of their subtopics. Hence, only IS topic will further investigate on more subtopics. The IS subtopics are Problem Modeling (PM), General Problem (GP) and Artificial Intelligence (AI).

2) Evaluation of the correctness of model, the experimental results reveal that the model achieves 73.88, 82.60 and 77.36 average percentage of overall Precision, Recall and F-measure, respectively.

3) Evaluation of the percentage of the document reduction, the proposed model (JAS) is achieved in 94.15 average percentage of unnecessary journal reduction.

4) Comparison of Linear Searching, LDA Topic Modeling and JAS, this research shows that linear searching, LDA Modeling and JAS are achieved in 78.33, 82.56 and 94.15 percentage of document reduction, respectively. Therefore, the designed JAS achieve the best result in term of number of unnecessary document reduction.

In conclusion, the proposed model (JAS) can be achieved in 94.15 average percentage of unnecessary journal reduction, which can approximately reduce 230 documents out of 245 documents. It means readers need to read only 15 documents instead of 245 documents. In addition, Tenopir et al. (2011) reports that most people spent 29.3 minutes per article reading, and also spent 17.25 average hours for reading weekly. Therefore, most people must read approximately at least 34 articles per week. It can be concluded that JAS can reduce documents sufficient for readers to read in short period of time.

References

Bisong, E. (2019). *Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley. CA: Apress. https://doi.org/10.1007/978-1-4842-4470-8_7

Cahyono, B. Y. (1997). Effectiveness of Journal Writing in Supporting Skills in Writing English Essay. *The Journal of Education, 4*, 310-318.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics, 4*(1), 31. https://doi.org/10.1186/1755-8794-4-31

Ide, N., & Véronis, J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics, 24*(1), 1-40.

Ismail, H., Syahruzah, J. K., & Basuki. (2017). Improving the Students' Reading Skill through Translation Method. *Journal of English Education, 2*, 124-131.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys, 41*(2), 10:11-10:69. http://doi.acm.org/10.1145/1459352.1459355

Sittirak, N., & Pornjamroen, S. (2009). The Survey of British English and American English Vocabulary Usage of Thai Students. *Songklanakarin Journal of Social Sciences and Humanities, 15*(4), 559-575.

Tenopir, C., Mays, R., & Wu, L. (2011). Journal Article Growth and Reading Patterns. *New Review of Information Networking, 16*(1), 4-22. https://doi.org/10.1080/13614576.2011.566796

Usman, M., Bajwa, Z., & Afzal, M. (2014). Performance Analysis of Searching Algorithms in C#. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2*(XII), 511-513.