Behaviormetrika (2017) 44:425–444
DOI 10.1007/s41237-017-0027-7

ORIGINAL PAPER

Diverse reports recommendation system based on latent Dirichlet allocation

Masaki Uto[1] · Sébastien Louvigné[1] · Yoshihiro Kato[2] · Takatoshi Ishii[3] · Yoshimitsu Miyazawa[4]

Received: 24 April 2017 / Accepted: 3 July 2017 / Published online: 13 July 2017

Abstract This paper presents a proposal for system supporting learners in improving their report-writing skills by recommending reports from previous learners. The proposed system recommends reports that share similar subjects but which have different structures, expressions, and originality based on the distributions of words and subjects within the reports, as estimated using latent Dirichlet allocation (LDA). An important assumption made for this study is that reports with different word distributions tend to include different structures, expressions, and originality when they share similar subjects. Based on that assumption, the system selects and recommends reports that have dissimilar word distributions but which share similar subject distributions with a learner's report. The proposed

Communicated by Neil Rubens.

The original version of this article was revised due to a retrospective Open Access order.

✉ Masaki Uto
uto@ai.lab.uec.ac.jp

Sébastien Louvigné
louvigne@ai.lab.uec.ac.jp

Yoshihiro Kato
y-kato@mail.benesse.co.jp

Takatoshi Ishii
t.ishii@rs.tus.ac.jp

Yoshimitsu Miyazawa
miyazawa@u-gakugei.ac.jp

[1] The University of Electro-Communications Chofu, Tokyo, Japan

[2] Benesse Education Research and Development Institute, Tokyo, Japan

[3] Tokyo University of Science, Tokyo, Japan

[4] Tokyo Gakugei University, Tokyo, Japan

system is expected to enhance learning of various writing skills from other learners. Finally, this paper demonstrates the effectiveness of the proposed system through actual data experiments.

Keywords Recommender system · Topic model · Latent Dirichlet allocation · Writing skills · E-learning

1 Introduction

Studies of higher education have recently emphasized the importance of writing skills (Britt et al. 2004; Villalón et al. 2008; Azilawati et al. 2009; Calvo et al. 2011; Uto and Ueno 2015). However, beginning learners often find difficulty in learning how to write academic reports by themselves. Therefore, to support novice writers, this study develops an adaptive recommendation method that provides reports from earlier learners based on an apprenticeship approach.

Recent learning theories have been transiting progressively and increasingly towards Vygotsky's social constructivist approaches (Vygotsky 1978). Vygotsky modeled human knowledge construction, as presented in Fig. 1. The model emphasizes the importance of supporting processes to understand learning objects, instead of emphasizing a basic transfer of knowledge. Vygotsky claimed that beginner learners gradually acquire meta-psychological skills such as attention, self-reflection, attitude, motivation, and passion by receiving adaptive support from experts based on the learner abilities. According to this model, the degrees of interest and passion for teaching, perspectives, and ethics from experts become a part of the culture of learning experience. Learners can acquire this cultural background and this psychological view of the learning contents. In addition, rather than receiving unilateral support from experts, beginners consciously learn from others through observation, imitation, and comparison. Autonomous development will occur through interaction with others without direct support received from experts. Gradual and autonomous self-development via observation, imitation, and comparison is a fundamentally important aspect of learning (Ueno 2015).

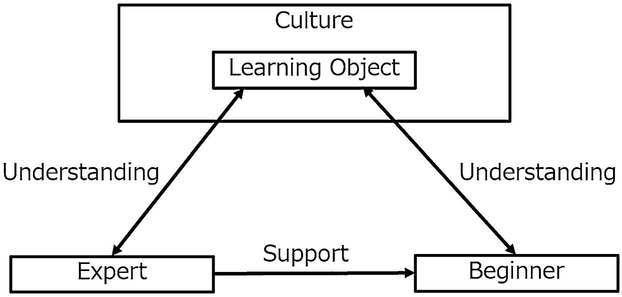


Fig. 1 Vygotsky learning model

Based on the idea of *learning from others*, this study proposes a system providing support for improving report-writing skills. Concretely, this study develops a recommender system that provides past learners' reports.

Several systems providing support for report writing have been developed to date. For example, systems providing automatic feedback related to spelling, style, plagiarized sentences, and citations have been developed (e.g., Macdonald et al. 1982; Britt et al. 2004; Villalón et al. 2008). Other studies (e.g., O'Rourke and Calv 2009; Aluisio et al. 2001; Shibata and Hori 2002; Feltrim and Teufel 2004; Uto and Ueno 2011) have examined systems visualizing the logical structure of reports and have suggested strategies to improve them. Other systems (e.g., Verheij 2005; Azilawati et al. 2009; Uto and Ueno 2015) visualize the argument structure of reports and provide feedback for revision based on the Toulmin model (Toulmin 1958), which is known as a normative model of argumentation.

However, these systems have limited provision of support in terms of the writing style and formal structure. The proposed approach incorporates recommendations of reports from other learners. This approach supports learning of other writing methods from others through comparison of one's own reports to others' reports. This objective raised the question of which report to recommend.

Many recommender systems have been developed in the domain of research in education (e.g., Lu 2004; Tang and McCalla 2005; Yang et al. 2009; Abel et al. 2010; Huang et al. 2009; Ueno and Uto 2011). They recommend widely various personalized contents such as learning objects, resources, learning process, and academic papers using statistical methods. For example, Lu (2004) developed a framework that determines student needs using a multi-attribute evaluation method, and which recommends learning materials based on the needs. Tang and McCalla (2005) proposed collaborative filtering techniques for recommending learning materials based on learners' interests and background knowledge. Yang et al. (2009) proposed an adaptable recommendation framework that summarizes and recommends multimedia contents. Abel et al. (2010) studied different recommendation strategies on a discussion board for a personalized framework. Furthermore, Huang et al. (2009) developed a Markov-based recommendation system modeling the learning process and recommending learning paths. Ueno and Uto (2011) proposed a recommender system for recommending e-portfolios considering similarity among learners' e-portfolios. Some other recommender systems recommend academic and scientific papers based on cosine similarity between TF and IDF vectors (e.g., McNee et al. 2002; Bollacker et al. 1999).

The current recommender systems recommend contents that share high degrees of similarity with those of a target learner. However, effective learning can be expected only to a slight degree by providing only similar contents (Ueno 2014). To improve learning efficiency, recommender systems should also provide unexpected contents that differ from what learners already know (Manouselis et al. 2012).

This study presents a proposal of a method to recommend reports that share similar subjects but which are produced from diverse structures, expressions, and originality. This proposed method uses latent Dirichlet allocation (LDA: Blei et al. 2003), which can estimate the distribution of subjects (designated as *topics*) and words within a corpus of reports. For this study, we assume that reports with dissimilar word distributions tend to include different structures, expressions, and originality when they have similar subjects. Here, *structure* means text organization and argumentative structure of a report, *expression* represents vocabularies and phrases used to convey ideas in a report, and *originality* denotes how unusual or novel a claim and contents of a report are. Based on that assumption, the proposed system recommends reports that have dissimilar word distributions but which share similar subject distributions with those of a learner's report. Learning from diverse others is known to be more effective than learning from similar others (Ueno 2014). Therefore, the proposed recommender system is expected to improve report-writing skills effectively in terms of structure, expression, and originality.

This paper demonstrates the effectiveness of the proposed method based on results of actual data experiments.

2 LMS Samurai

This study uses learner data stored in the learning management system (LMS) called *Samurai*, which has been developed for many years by Ueno (2004a). The LMS has been used with numerous e-learning courses.

LMS Samurai presents content sessions tailored for 90-min classes. Fifteen of these content sessions constitute a two-unit course. Each session provides instructional text screens, instructional images, instructional videos, and practice tests. Learners choose from the array of contents and watch the lesson. How learners respond to the sessions and how long it takes them to complete the lesson are stored automatically in the learning history database. Those data are analyzed using various data mining techniques to facilitate learning processes (e.g., Ueno 2004a, b; Ueno and Uto 2011).

Report assignments are usually given during a course. LMS Samurai has a discussion board system that enables learners to submit reports and which enables them to assess and discuss with one another. The learner who submitted the report can take the ratings and comments into consideration and rework them. The rating data are used for calculating the average rating score of the report, estimating latent ability of learners (Uto and Ueno 2016), and recommending e-portfolios (Ueno and Uto 2011).

For our report recommendation, we use the report data stored in the system.

3 Latent Dirichlet allocation

This section presents a description of the topic model used for our recommendation method. The topic model estimates topics in a document from the occurrence frequency of words based on the assumption that certain words will appear depending on the potential topics of the text. Latent Semantic Analysis (LSA: Deerwester et al. 1990), Probabilistic Latent Semantic Indexing (PLASI: Hofmann 1999), and latent Dirichlet allocation (LDA: Blei et al. 2003) are regarded as examples of topic models. The LDA was chosen in this study for its higher accuracy in estimating topics (Blei et al. 2003).

The graphical model of LDA is presented in Fig. 2. Here, K stands for the number of topics, D is the number of documents, and V signifies the vocabulary size in the documents. In addition, W represents observed words in each document, and Z represents the topic allocation for each word. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$ denotes a set of D multinomial distribution over the K topics (designated as *topic distribution*). $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$ represents a set of K multinomial distributions over V vocabulary words (designated as the *word distribution*). The topic distribution $\boldsymbol{\theta}$ shows what topic tends to be generated in each document. The word distribution $\boldsymbol{\phi}$ shows how vocabulary words are used in each topic. Both α and β are parameters of the Dirichlet prior distribution. They are designated as hyperparameters.

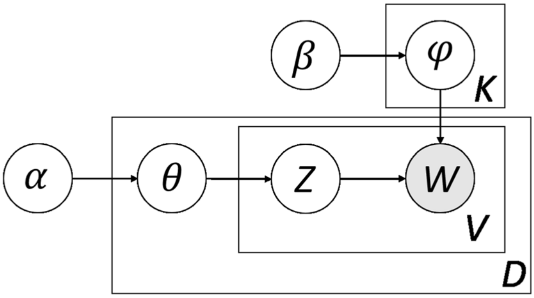
The document generation process in LDA consists of the following steps.

1. For each document $d \in \{1, \ldots, D\}$, the topic distribution $\boldsymbol{\theta}_d = (\theta_{1d}, \ldots, \theta_{Kd})$ is generated from the K-dimensional Dirichlet prior distribution with hyperparameter α. Here, $\theta_{kd} \in \boldsymbol{\theta}_d$ represents the occurrence probability of kth topic in the dth document.
2. A topic $z_{di} \in \{1, \ldots, K\}$ that corresponds to the ith word in document d (the word is designated as w_{di}) is selected with the probabilities given by the topic distribution $\boldsymbol{\theta}_d$.
3. Each word w_{di} is drawn from the word distribution $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kV})$ given the corresponding topic z_{di}. Here, ϕ_{kv} represents the probability of the vth vocabulary word in the kth topic. In addition, $\boldsymbol{\phi}_k$ is generated from the V-dimensional Dirichlet prior distribution with hyperparameter β.

The parameters θ_{kd} and ϕ_{kw} can be estimated using a collapsed Gibbs sampler (Griffiths and Steyvers 2004) given the data that consist of occurrence counts of words in each document.

The number of topics K is generally estimated from data based on model selection techniques. The Akaike Information Criterion (AIC: Akaike 1974) and Bayesian Information Criterion (BIC: Schwarz 1978) are well-known model selection criteria. However, unfortunately, AIC and BIC, which assume regularity for an approximation, are inapplicable to learning LDA problems that have no regularity (Watanabe 2010, 2013). Therefore, the log marginal likelihood and

Fig. 2 Graphical model representation of LDA



perplexity have traditionally been used in LDA (Griffiths and Steyvers 2004; Wallach et al. 2009; Taddy 2012; Blei et al. 2003). The optimal number of topics can be estimated by maximizing the log marginal likelihood or minimizing the perplexity. The log marginal likelihood based estimation is expected to provide higher performance than perplexity, because it has the asymptotic consistency, which is a property by which the estimates converge to the true value as the sample size goes to infinity (Griffiths and Steyvers 2004; Taddy 2012; Buntine 2009; Wallach et al. 2009). Although the log marginal likelihood cannot be solved analytically, because it requires calculation of the posterior with all possible instances of topics over words, we can estimate them approximately (Griffiths and Steyvers 2004; Taddy 2012; Buntine 2009; Wallach et al. 2009). The well-known approximation is the harmonic mean method using MCMC samples of the topic allocation (Griffiths and Steyvers 2004).

Another approach for ascertaining the number of topics is using the hierarchical Dirichlet process (HDP: Teh et al. 2004), which is a nonparametric extension of LDA. The HDP assumes an infinite number of topics and concretizes a finite number of them. Although HDP can estimate the number of topics as part of posterior inference (Blei et al. 2010), the estimation is known to be sensitive to hyperparameters of HDP (Buntine and Mishra 2014). In addition, HDP is more complicated to implement than LDA is. For those reasons, this study estimates the number of topics using the maximum log marginal likelihood approach.

4 Similarity calculation using Jensen–Shannon divergence

A unique feature of LDA is that it can accommodate the topic distributions and the word distribution separately. This section defines the similarity between the topics of documents and defines the similarity of words in documents using LDA.

In this paper, we define the dissimilarity of the topic and that of words between documents using Jensen–Shannon divergence. This index takes a minimum value of 0 when two probability distributions are consistent. It returns a large positive value that reflects the degree to which the two distributions differ.

Letting the Kullback–Leibler divergence be KLD(), the Jensen–Shannon divergence of topic distributions between document d and d' is expressed as

$$T_{\mathrm{JSD}}(d, d') = \frac{1}{2}\mathrm{KLD}(\boldsymbol{\theta}_d \| \boldsymbol{m}_{dd'}) + \frac{1}{2}\mathrm{KLD}(\boldsymbol{\theta}_{d'} \| \boldsymbol{m}_{dd'}), \tag{1}$$

where

$$\mathrm{KLD}(\boldsymbol{\theta}_d \| \boldsymbol{m}_{dd'}) = \sum_{k=1}^{K} \theta_{kd} \ln\left(\frac{\theta_{kd}}{m_{kdd'}}\right). \tag{2}$$

In those equations, $m_{kdd'} = (\theta_{kd} + \theta_{kd'})/2$, and $\boldsymbol{m}_{dd'} = [m_{1dd'}, \ldots, m_{Kdd'}]$. Distance $T_{\mathrm{JSD}}(d, d')$ becomes 0 when two documents share the same topic distribution. Using this method, one can find reports that have a similar topic to that of the target report.

The Jensen–Shannon divergence of word distributions between documents d and d' is expressed as

$$W_{\mathrm{JSD}}(d, d') = \frac{1}{2}\mathrm{KLD}(\boldsymbol{\eta}_d \| \boldsymbol{l}_{dd'}) + \frac{1}{2}\mathrm{KLD}(\boldsymbol{\eta}_{d'} \| \boldsymbol{l}_{dd'}), \tag{3}$$

where

$$\mathrm{KLD}(\boldsymbol{\eta}_d \| \boldsymbol{l}_{dd'}) = \sum_{v=1}^{V} \eta_{dv} \ln\left(\frac{\eta_{dv}}{l_{vdd'}}\right). \tag{4}$$

In those equations, $\boldsymbol{\eta}_d = [N_{d,v=1}/N_d, \ldots, N_{d,v=V}/N_d]$, $N_{d,v}$ stands for the occurrence frequency of vocabulary word v in document d, with $N_d = \sum_{v=1}^{V} N_{d,v}$. In addition, $l_{vdd'} = (\eta_{vd} + \eta_{vd'})/2$, and $\boldsymbol{l}_{dd'} = [l_{1dd'}, \ldots, l_{Vdd'}]$. This equation evaluates the distance of the word distributions between two documents. This index is useful to identify reports made from contents that are as different as possible from contents of the target report.

5 Report recommendation system

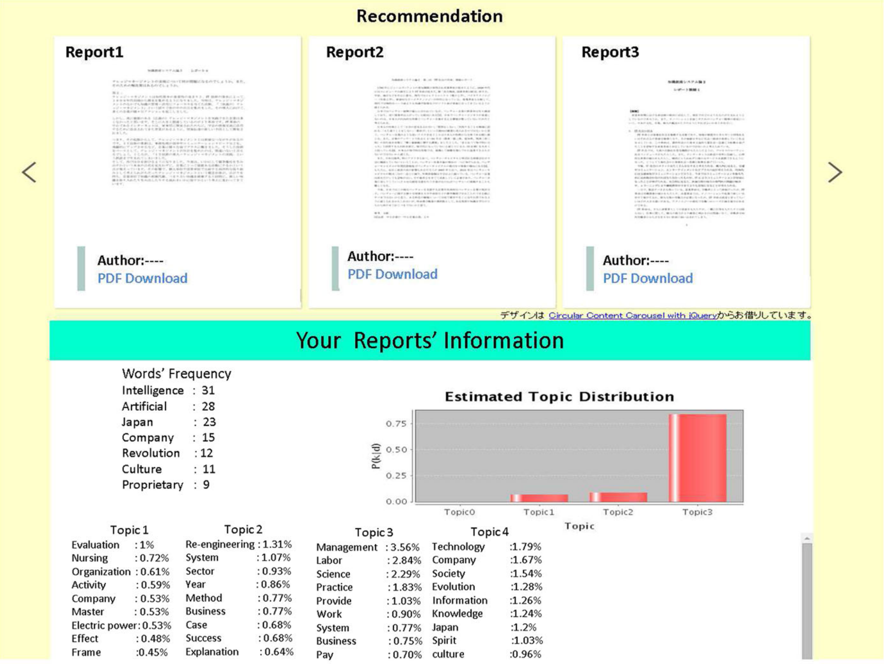
In this section, we propose a system of recommending past reports to learners using the LDA model. To enhance deeper learning of report writing, the proposed system recommends reports on a similar subject, but reports derived from structures, expressions, and originality that are as different as possible, rather than simply recommending reports including similar words. This study assumes that reports produced from dissimilar word distributions tend to have different structures, expressions, and originality when they share similar topic distributions. Based on that assumption, the proposed system recommends reports sharing high similarity of topic distributions and high dissimilarity of word distributions with a learner's report.

The algorithm of the recommendation mechanism is presented in Algorithm 1. Through the algorithm, the proposed system recommends S reports selected from a corpus of reports $\boldsymbol{X} = \{d_1, \ldots, d_D\}$ to a learner given a report of a learner d_x. In the algorithm, the topic distribution $\boldsymbol{\theta}$ and the word distribution $\boldsymbol{\phi}$ of the reports corpus $\boldsymbol{X}$ are first estimated. Then, the topic distribution $\boldsymbol{\theta}_{d_x}$ of the learner's report d_x is estimated by the collapsed Gibbs sampler, given the estimated word distribution $\hat{\boldsymbol{\phi}}$. The third step extracts N reports from the corpus of reports $\boldsymbol{X}$ with the smallest Jensen–Shannon divergence $T_{\mathrm{JSD}}(d_x, d_i)$ between d_x and $d_i \in \boldsymbol{X}$. Then, the fourth step selects M reports from the extracted N reports with the largest divergence of word distribution $W_{\mathrm{JSD}}(d_x, d_i)$. Finally, from the M reports, S reports that have high peer assessment scores are selected and recommended to the learner. N, M and S are restricted to $N > M > S$. In this paper, we fixed the parameters as $N = 15$, $M = 10$, and $S = 4$.

```
input  : Learner's report: $d_x$
         Corpus of reports : $\boldsymbol{X} = \{d_1, \cdots, d_i, \cdots, d_D\}$
output: Report recommended to learners: $\boldsymbol{Y}$
Estimate the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using the corpus $X$.
Estimate the topic distribution $\boldsymbol{\theta}_{d_x}$ of learner's report $d_x$ given $\hat{\boldsymbol{\phi}}$.
Create $\boldsymbol{Y}$ by extracting $N$ reports from $\boldsymbol{X}$ with smallest dissimilarity $T_{JSD}(d_x, d_i)$
  for each $d_i \in \boldsymbol{X}$.
Update $\boldsymbol{Y}$ by extracting $M$ reports from $\boldsymbol{Y}$ with largest dissimilarity $W_{JSD}(d_x, d_i)$
  for each $d_i \in \boldsymbol{Y}$.
Update $\boldsymbol{Y}$ by extracting $S$ reports from $\boldsymbol{Y}$ with the highest peer assessment scores.
return $\boldsymbol{Y}$.
```

Algorithm 1: Report recommendation mechanism.

We implemented the report recommendation mechanism in LMS *Samurai*. The system interface is shown in Fig. 3. In the system, a learner first submits the learner's own report. Then, the system selects reports from a corpus of reports based on Algorithm 1, and displays them at the top of the system. Learners can download and browse the recommended reports. By clicking on the recommended report, the system shows the writer's information and the topic distribution of the report. The system also displays statistics for the input report. Specifically, the topic distribution, the word distribution for each topic, and ranking of word appearance are displayed. By comparing the topic distribution of the input report and the



Recommendation

Report1

Author:----
PDF Download

Report2

Author:----
PDF Download

Report3

Author:----
PDF Download

デザインは Circular Content Carousel with jQueryからお借りしています。

Your Reports' Information

Words' Frequency

Intelligence	:	31
Artificial	:	28
Japan	:	23
Company	:	15
Revolution	:	12
Culture	:	11
Proprietary	:	9

Topic 1		Topic 2		Topic 3		Topic 4	
Evaluation	: 1%	Re-engineering	: 1.31%	Management	: 3.56%	Technology	:1.79%
Nursing	: 0.72%	System	: 1.07%	Labor	: 2.84%	Company	:1.67%
Organization	: 0.61%	Sector	: 0.93%	Science	: 2.29%	Society	:1.54%
Activity	: 0.59%	Year	: 0.86%	Practice	: 1.83%	Evolution	:1.28%
Company	: 0.53%	Method	: 0.77%	Provide	: 1.03%	Information	:1.26%
Master	: 0.53%	Business	: 0.77%	Work	: 0.90%	Knowledge	:1.24%
Electric power	: 0.53%	Case	: 0.68%	System	: 0.77%	Japan	:1.2%
Effect	: 0.48%	Success	: 0.68%	Business	: 0.75%	Spirit	:1.03%
Frame	:0.45%	Explanation	: 0.64%	Pay	: 0.70%	culture	:0.96%

Fig. 3 Report recommendation system

recommended reports, learners can ascertain which topics are similar, and can see the word groups that are included in the topic.

6 Evaluation

This section presents evaluation of the effectiveness of the proposed recommender system. As described herein, we assume that recommending a report with similar topics but different word distributions is more effective for improving report-writing skills. To evaluate that hypothesis, we consider the following alternative settings:

1. Recommending reports that have word distributions as similar as possible to those of the learner's report for similar topics.
2. Recommending reports that have word distributions as similar as possible to those of the learner's report for dissimilar topics.
3. Recommending reports that have word distributions as different as possible from those of the learner's report for dissimilar topics.

With similar topic distributions and word distributions, only similar papers are expected to be recommended, which produces a lack of diversity. However, when the recommended reports share low similarity in topic distributions with the learner's report, irrespective of the similarity of the word distributions, it becomes irrelevant to the learner's topic, which also reduces efficiency.

As suggested by the points presented above, this experiment assumes the following recommendation methods corresponding to the alternative hypothesis and that of this paper.

- Method A: recommending reports with high similarity in topic distributions and in word distributions.
- Method B (proposed method): recommending reports with high similarity in topic distributions but with a low similarity in word distributions.
- Method C: recommending reports with low topic similarity but still sharing high similarity in word distributions.
- Method D: recommending reports with low topic similarity and low similarity in word distributions.
- Method E: recommending reports with high cosine similarity between TF and IDF vectors.
- Method R: recommending reports randomly.

Here, Method E was introduced as a conventional recommendation method that recommends reports with high cosine similarity of the TF–IDF value (Bollacker et al. 1999; McNee et al. 2002). In addition, we introduced Method R, recommending reports randomly, to confirm the effectiveness of similarity-based recommendations.

6.1 Corpus of reports

As the corpus of reports used for the experiment, we used reports stored in the LMS *Samurai*. Specifically, this study used 90 reports submitted for an assignment "Describe prior knowledge production technique and its problems in the company" in the Master's course lecture of "Knowledge production system theory". The course comprises 15 lectures, which provide instructions in various management theories such as scientific management theory, leadership theory, knowledge management, education, and learning theory. The report assignment was related to all those themes. Therefore, reports with various subjects can be submitted in the assignment.

This experiment used the reports only for the single assignment. It will be difficult to compare the writing skills if we used reports for different assignments because differences among reports depend both on the writing skills and the given assignments.

For our research purposes, LDA analysis using all the vocabulary words would not be appropriate, because the existence of many common or stop words often introduces meaningless or uninterpretable topics (Schofield et al. 2017). Therefore, in this experiment, to remove common or stop words, the instructor and the tutor selected 382 representative words of the course. Specifically, they were asked to select keywords from the list of vocabularies, which were created from the corpus of reports and which were sorted by occurrence frequency. Here, we instructed them not to choose common and stop words. We used the selected words as vocabulary words for LDA analysis.

The reports in the corpus are written in Japanese. Words in a Japanese text are not separated by spaces. Therefore, we leave a space between words for reports using the Japanese morphological analyzer MeCab (Kudo et al. 2004). Then, we counted the vocabulary words for the reports.

For our LDA analysis, we first estimated the number of topics from the actual data using the maximum log marginal likelihood method described in Sect. 3. For calculations, we set the hyperparameters to $\alpha = 1/K$ and $\beta = 1/KV$, as used in common LDA software (Taddy 2012).

Figure 4 shows the log marginal likelihoods for the respective numbers of topics. The vertical axis shows values of the log marginal likelihood; the horizontal axis shows the number of topics. Figure 4 shows that the log marginal likelihood reached its maximum when the number of topics was 4. Therefore, we set the optimal number of topics as $K = 4$.

Then, we estimated the word and topic distributions given $K = 4$. Table 1 shows the top 10 most frequently used vocabulary words in the respective topics. According to the estimated word distribution, one can interpret the topics, as shown below. Topic 1 is Scientific Management Theory. Topic 2 is the Industrial Revolution. Topic 3 is knowledge management. Topic 4 is engineering. They are the important subjects in the course.

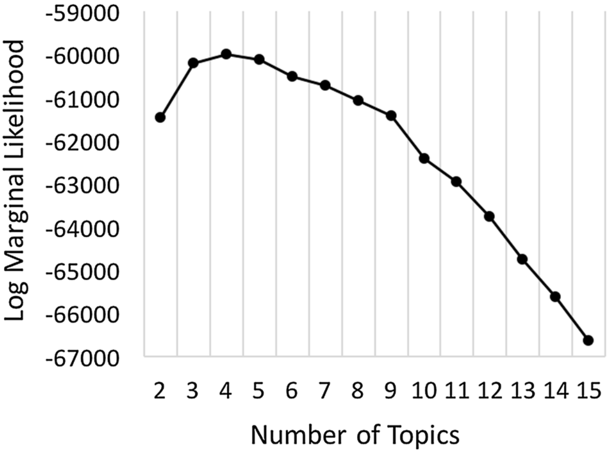


Fig. 4 Log marginal likelihood for respective number of topics

Table 1 Top 10 words in topics

	Words (probability)
Topic 1	Labor (0.105), science (0.09), Taylor (0.075), work (0.071), practice (0.043), standard (0.026), advantage (0.025), disadvantage (0.025), management (0.023), resource (0.023)
Topic 2	Technology (0.069), company (0.058), industrial (0.057), society (0.053), revolution (0.045), venture (0.043), innovation (0.036), Japan (0.034), culture (0.028), cultivate (0.023)
Topic 3	Years (0.036), management (0.031), success (0.029), company (0.026), method (0.026), business (0.024), case (0.024), process (0.023), cost (0.022), product (0.022)
Topic 4	System (0.061), information (0.048), requirement (0.03), work (0.026), problem (0.026), reports (0.022), creation (0.019), efficiency (0.019), time (0.018), possibility (0.016)

6.2 Experimental procedure

Using the corpus of reports, we conducted the following experiment to evaluate the effectiveness of the proposed system.

1. 60 university students were recruited as participants.
2. They were divided randomly into six groups corresponding to recommendation methods A, B, C, D, E, and R.
3. Participants were asked to create a report for the assignment "Describe prior knowledge production technique and its problems in a company". We provided reference materials that were used during the lecture. The participants completed reports using the materials and the internet. We designate the reports that the participants first created as *pre-reports*.
4. After loading pre-reports, the system asked participants to read the recommended reports; then, they were asked to add modifications to their pre-reports before the next submission. The revised reports are designated as *post-reports*.

In this experiment, the numbers of valid responses were seven people in groups for recommendation Methods A and C, six people in the groups for Methods D and E, and five people in groups for Methods B and R.

6.3 Recommended reports

To confirm whether the recommendations worked as expected, or not, we first calculated the similarity of topic distributions between pre-reports and recommended reports for each recommendation method. The similarity is defined as the inverse of $T_{\mathrm{JSD}}(d, d')$ between two reports d and d'. The average values are shown in Fig. 5. Figure 5 shows that Methods A, B, and E recommended reports with high topic similarity, whereas Methods C, D, and R recommended reports with low topic similarity.

We next calculated the similarity of the word distributions between pre-reports and recommended reports. The similarity is defined as the inverse of $W_{\mathrm{JSD}}\,(d, d')$ between two reports d and d'. The average values are presented in Fig. 6. Figure 6 shows that the proposed method B recommended reports with lower word similarity than Method A or E, as expected. The proposed method revealed higher word similarity than either Method C or D, which recommend reports with different topic distributions. Although the word similarity can be decreased by selecting reports with different topic distributions, the purpose of the proposed method is not merely minimizing the word similarity. The proposed method selects reports with as different word distributions as possible while sharing similar topic distributions. Therefore, the word similarity of the proposed method is necessarily not lower than that of Method C or D.

Finally, to demonstrate that the proposed method selected reports with similar topic distributions and dissimilar word distributions, we calculated the word

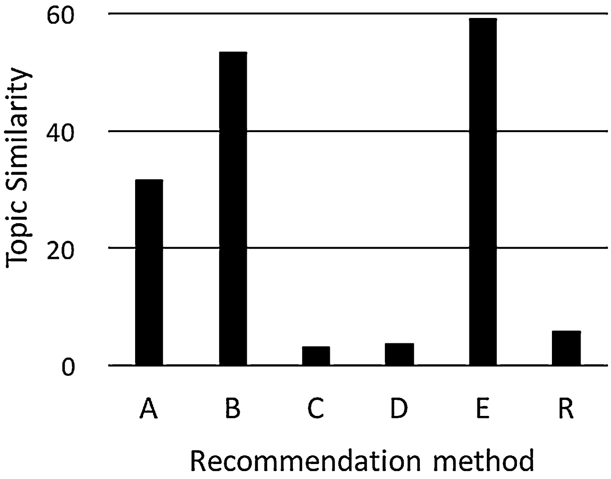


Fig. 5 Topic similarity between pre-reports and recommended reports

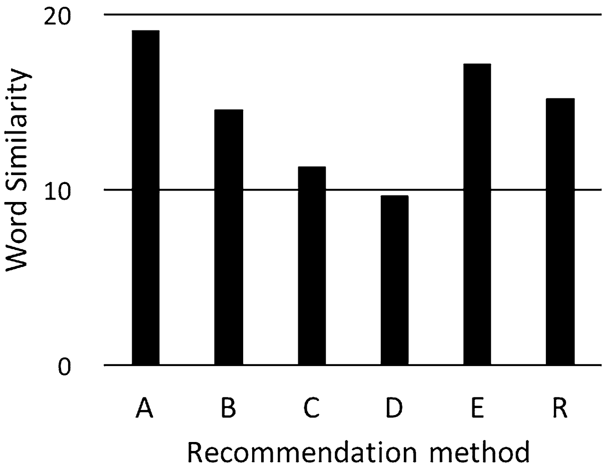


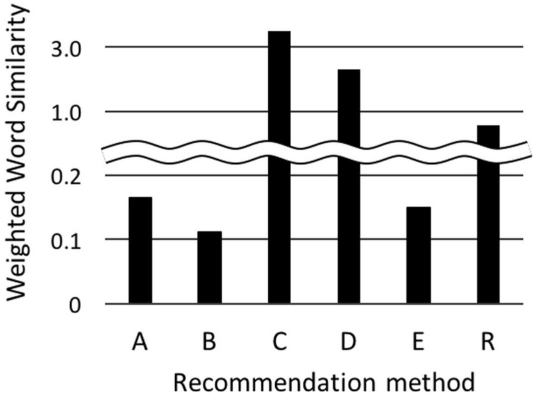
Fig. 6 Word similarity between pre-reports and recommended reports

similarity weighted by the inverse of the topic similarity. Specifically, the index is defined as $W_{\mathrm{JSD}}(d,d')^{-1}/T_{\mathrm{JSD}}(d,d')^{-1} = T_{\mathrm{JSD}}(d,d')/W_{\mathrm{JSD}}(d,d')$ between two reports d and d'. We refer to the index as *weighted word similarity*. Weighted word similarity takes a low value as the word similarity decreases and the topic similarity increases. If the proposed method worked appropriately, then it is expected to reveal the lowest values.

Figure 7 presents average values of the index for the respective recommendation methods. According to Fig. 7, the proposed method B presented the lowest value. The result shows that the proposed method recommended reports sharing similar topic distributions and dissimilar word distributions.

In summary, the experimentally obtained results presented in this subsection verified that the proposed method and the alternative methods worked as expected.

Fig. 7 Word similarity weighted by the inverse of topic similarity



6.4 Characteristics of reports with similar topic distributions

This study assumes that the topic distributions estimated by LDA reflect the subjects within reports. To evaluate the assumption, the following questionnaire was presented to the participants at the end of the experiment. "Recommended report was associated with the subject of your report." The answers for the item were given on the following five-point scale: (1) do not agree at all; (2) do not really agree; (3) not sure; (4) agree a little; and (5) strongly agree.

Table 2 presents the results. Furthermore, we conducted Kruskal–Wallis test to evaluate the differences in the average values. The result showed that a significant difference ($\chi^2(5) = 15.59$, $p < 0.01$) was found among the methods. Therefore, we conducted multiple comparisons using the Steel–Dwass method. Although the multiple comparison showed no significant difference, the scores of Methods A, B, and E took higher values than Method C, D, or R.

Results showed that topic distributions reflected the subjects within reports.

6.5 Characteristics of reports with different word distributions

This study examines the hypothesis that reports with dissimilar word distributions have different structures, expressions, and originality when they share similar topics. To evaluate the hypothesis, we conducted the following experiment.

We first randomly selected 10 reports written by the participants. Then, for each participant report, 5 reports in the corpus which sharing the smallest similarity of the topic distributions with the given report were extracted. Then, each report pair, consisting of an extracted report and the corresponding participant's report, was evaluated by two experts based on the following evaluation items.

1. How different are the *structures* of the reports?
2. How different are the *expressions* of the reports?
3. How different are the *originality* of the reports?

Each item was evaluated using the following five-point rating scale: (1) extremely similar; (2) similar; (3) neither; (4) dissimilar; and (5) extremely dissimilar. Before the evaluation, we explained the means of the structure, expression, and originality to the experts.

For each evaluation item, we calculated Spearman's rank correlation coefficient R between the word dissimilarities $W_{\mathrm{JSD}}()$ and the average scores. If the correlation reveals a high value, then the hypothesis of this study can be supported.

Table 2 Average and standard deviation (in parentheses) values for the questionnaire related to subject similarity between own report and recommended reports

Method A	Method B	Method C	Method D	Method E	Method R
4.29 (1.03)	4.20 (0.40)	2.57 (1.29)	2.17 (1.07)	4.00 (1.00)	3.00 (0.71)

The Spearman correlation analysis revealed significant correlation $R = 0.586$ for the evaluation of the *structure* ($p < 0.01$), $R = 0.652$ for the *expression* ($p < 0.01$), and 0.559 for the *originality* ($p < 0.01$). Scatter plots for respective evaluation items are presented in Fig. 8. In the figures, the vertical axis shows average scores for the respective evaluation items; the horizontal axis shows the word dissimilarity. Regression lines are shown as dashed lines. As the figures and the correlation analysis results confirm, reports sharing similar topics but made from different word distributions tend to have different structures, expressions, and originality.

6.6 Quality of pre-reports and post-reports

This subsection presents a description of how the proposed recommendation is effective for improvement of report-writing skills. For that purpose, pre-reports and post-reports written by participants in each group were evaluated by two experts based on the following evaluation items.

- *Q1*: How was the structure of the report?
- *Q2*: How was the expression of the report?
- *Q3*: How was the originality of the report?

Each item was evaluated using the following five-point rating scale: (1) do not agree at all; (2) do not really agree; (3) not sure; (4) agree a little; and (5) strongly agree. Before the evaluation, we explained the means of the structure, expression, and originality to the experts. Furthermore, the evaluations were conducted by not informing the experts which report is a pre-report or post-report.

The results are presented in Table 3. To evaluate the differences in the average values among the recommendation methods, we conducted Kruskal–Wallis test. The χ^2 value and the p value for each evaluation item are presented in Table 3.

Results for pre-reports in the table show that no significant difference was found. However, the results of post-reports show that significant differences were found between the recommendation methods for all evaluation items. Multiple comparisons using the Steel–Dwass method for each item of post-reports revealed the following results.

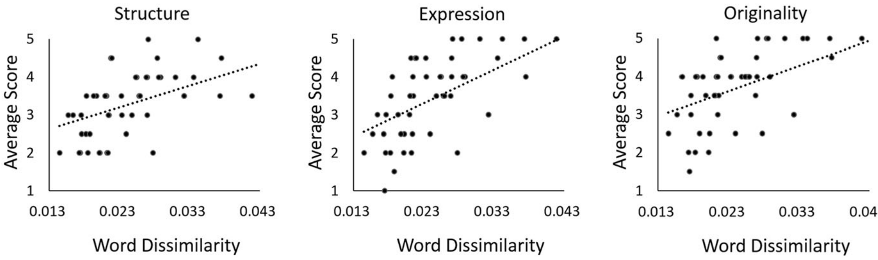


Fig. 8 Relations between the word dissimilarity and average scores for evaluation of structure, expression, and originality differences

Table 3 Average and standard deviation (in parentheses) values of expert evaluation for pre-report and post-reports

	Pre-reports			Post-reports		
	Structure	Expression	Originality	Structure	Expression	Originality
Method A	2.29 (0.70)	2.43 (0.50)	2.00 (0.00)	2.43 (0.73)	2.71 (0.70)	2.00 (0.00)
Method B	2.60 (0.49)	2.80 (0.40)	2.20 (0.40)	4.20 (0.40)	4.00 (0.00)	3.20 (0.40)
Method C	2.86 (0.83)	2.86 (0.64)	2.29 (0.45)	2.14 (0.99)	2.86 (0.64)	1.86 (0.64)
Method D	2.50 (0.50)	2.67 (0.47)	2.17 (0.37)	2.50 (0.50)	2.67 (0.47)	2.17 (0.37)
Method E	2.67 (0.47)	2.50 (0.50)	2.00 (0.00)	2.67 (0.47)	3.00 (0.00)	2.17 (0.37)
Method R	2.20 (0.40)	2.20 (0.40)	2.20 (0.40)	2.20 (0.40)	2.20 (0.40)	2.20 (0.40)
χ^2-value	3.57	5.52	3.61	15.23	17.45	16.98
p value	0.61	0.36	0.61	<0.01	<0.01	<0.01

Degrees of freedom of the Kruskal–Wallis tests are 5

For the evaluation of the *structure*, the proposed method B showed significant differences with a significance level of 5% with respect to Methods A ($p = 0.037$), D ($p = 0.048$), and E ($p = 0.046$), and 10% with respect to Methods C ($p = 0.080$) and R ($p = 0.059$). No significant difference was found between the other methods. Results show that reading the reports sharing similar topics but with diverse structures is more effective for participants to improve the structures of their reports.

Regarding the evaluation of *expression*, the proposed method B showed significant differences with a significance level of 10% with respect to Methods A ($p = 0.087$), C ($p = 0.083$), and 5% with Methods D ($p = 0.037$), E ($p = 0.019$), and R ($p = 0.045$). No significant difference was found between other methods, which demonstrates that participants improved their expressions more by reading reports with greater diversity of expressions under similar topics.

Finally, for the evaluations of *originality*, the proposed method B showed significant differences with a significance level of 5% with respect to Methods A ($p = 0.015$) and R ($p = 0.045$), and 10% with respect to Methods C ($p = 0.082$), D ($p = 0.090$), and E ($p = 0.091$). No significance was found for other techniques. This result showed that participants were able to improve the originality of their reports by reading reports that shared the similar topics but with different origins.

6.7 Amount of correction

We calculated the number of revised sentences produced by the participants. Table 4 shows the results. We also presented the χ^2 values and the p value of the Kruskal–Wallis test in the table. The results show that a significant difference was found among the recommendation methods. Therefore, we conducted multiple comparisons using the Steel–Dwass method. Consequently, the proposed method B showed significant differences with a significance level of 5% with respect to Methods A ($p = 0.048$) and C ($p = 0.048$), and 10% with respect to Methods D ($p = 0.062$), E ($p = 0.064$), and R ($p = 0.074$). No significant difference between other methods was found.

Table 4 Average and standard deviation (in parentheses) values of the number of revised sentences made by participants, and answers for questionnaire related to report elaboration

	Number of revised sentences	Questionnaire
Method A	3.00 (2.19)	3.86 (1.24)
Method B	14.80 (2.19)	4.40 (0.49)
Method C	2.71 (2.28)	3.00 (0.93)
Method D	2.83 (2.84)	3.33 (0.94)
Method E	4.67 (3.92)	3.67 (0.94)
Method R	1.00 (1.82)	3.80 (1.09)
χ^2-value	16.18	7.29
p value	<0.01	0.20

Degrees of freedom of the Kruskal-Wallis tests are 5

The results yielded the following interpretations for the respective recommendation methods:

- Proposed Method B showed the greatest amount of corrections from pre-reports to post-reports. Recommendation of reports with similar topics but with different word distributions was able to enhance the corrections.
- Methods A and E showed fewer corrections, mostly because of the recommended reports including too few differences from the participants' reports.
- Methods C and D also showed few corrections, which indicates that recommending reports sharing different topics cannot enhance revisions.
- Method R showed fewer corrections, because it often recommends reports with different topics or with similar word distributions.

Furthermore, the following questionnaire related to the elaboration of reports was presented to participants at the end of the experiment. "By reading the recommended report, I was able to improve my report." The responses to the item were made using the following five-point scale: (1) do not agree at all; (2) do not really agree; (3) not sure; (4) agree a little; and (5) strongly agree.

The last column of Table 4 shows the results. Although the Kruskal–Wallis test showed no significant difference, the proposed method B received the highest evaluation scores.

As results presented in this subsection show, we confirmed that the proposed method enhanced an increase of the amount of corrections and that the participants were aware of them. We can assume that the number of corrections reflects the magnitude of learning from others. Therefore, we conclude that the proposed method can improve the learning effectiveness of report-writing skills.

7 Conclusion

We proposed a recommender system based on LDA that provides useful reports from previous learners to support learning from others in writing reports. This system first identified previous learners' reports that share similar topics with a learner's report by estimating the similarity of the topic distributions. Then, from reports with similar topics, the system selected reports derived from different structures, expressions, and originality from those of the learner's report. Those reports were chosen based on the dissimilarity of the word distributions, assuming that reports with dissimilar word distributions tend to have different structures, expressions, and originality when they have similar topics.

From results of experiments conducted using actual data, we demonstrated the effectiveness of the proposed method. Specifically, the results of the experiments showed the following features of the proposed system. (1) Reports with similar topic distributions and dissimilar word distributions tend to include different structures, expressions, and originality. (2) The proposed system can enhance reflection and learning of report-writing skills in terms of structure, expression, and originality.

This study does not claim that learners can acquire report-writing skills through the proposed system. A long-term experiment to evaluate whether the system can enhance skill acquisition is a remaining task of this study.

Another future task is to extend the proposed system using extension models of LDA that incorporate auxiliary information corresponding to structures, expressions, and originality of reports. Many LDA extensions that incorporate auxiliary information were proposed earlier (e.g., Flaherty et al. 2005; Blei et al. 2003; Rosen-Zvi et al. 2004; Iwata et al. 2013). The use of auxiliary information generally improves the accuracy of topic estimation better than using only the occurrence frequency of words. Although this study demonstrated that the LDA provides sufficient performance for our research purpose, those extension models might further improve the performance.

References

Abel F, Bittencourt II, Costa E, Henze N, Krause D, Vassileva J (2010) Recommendations in online discussion forums for e-learning systems. IEEE Trans Learn Technol 3(2):165–176

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Aluisio SM, Barcelos I, Sampaio J, Oliveira ON (2001) How to learn the many unwritten "rules of the game" of the academic discourse: a hybrid approach based on critiques and cases to support scientific writing. In: Proc. IEEE International Conference on Advanced Learning Technologies, pp 257–260

Azilawati J, Chee YS, Ho CML (2009) Fostering argumentative knowledge construction through enactive role play in Second Life. Comput Educ 53(2):317–329

Blei D, Carin L, Dunson D (2010) Probabilistic topic models: a focus on graphical model design and applications to document and image analysis. IEEE Signal Process Mag 27(6):55–65

Blei DM, Jordan MI (2003) Modeling Annotated Data. In: Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, pp 127–134

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bollacker KD, Lawrence S, Giles CL (1999) A system for automatic personalized tracking of scientific literature on the web. In: Proc. Fourth ACM Conference on Digital Libraries, pp 105–113

Britt MA, Wiemer-Hastings P, Larson AA, Perfetti CA (2004) Using intelligent feedback to improve sourcing and integration in students' essays. Int J Artif Intell Educ 14:359–374

Buntine W (2009) Estimating likelihoods for topic models . In: Proc. Asian Conference on Machine Learning: Advances in Machine Learning, pp 51–64

Buntine WL, Mishra S (2014) Experiments with non-parametric topic models. In: Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 881–890

Calvo RA, O'Rourke ST, Jones J, Yacef K, Reimann P (2011) Collaborative writing support tools on the cloud. IEEE Trans Learn Technol 4(1):88–97

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

Feltrim VD, Teufel S (2004) Automatic Critiquing of Novices' Scientific writing using argumentative zoning. In: Proc. AAAI spring symposium exploring affect and attitude in text

Flaherty P, Giaever G, Kumm J, Jordan MI, Arkin AP (2005) A latent variable model for chemogenomic profiling. Bioinformatics 21(15):3286–3293

Griffiths TL, Steyvers M (2004) Finding scientific topics. In: Proc. National Academy of Sciences of the United States of America, pp 5228–5235

Hofmann T (1999) Probabilistic latent semantic indexing. In: Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 50–57

Huang Y-M, Huang T-C, Wang K-T, Hwang W-Y (2009) A Markov-based recommendation model for exploring the transfer of learning on the Web. J Educ Technol Soc 12(2):144–162

Iwata T, Yamada T, Ueda N (2013) Modeling noisy annotated data with application to social annotation. IEEE Trans Knowl Data Eng 25(7):1601–1613

Kudo T, Yamamoto K, Matsumoto Y (2004) applying conditional random fields to japanese morphological analysis. In: Proc. Conference on Empirical Methods in Natural Language Processing, vol 4, pp 89–96

Lu J (2004) Personalized e-learning material recommender system. In: Proc. International Conference on Information Technology for Application, pp 374–379

Macdonald N, Frase L, Gingrich P, Keenan S (1982) The Writer's Workbench: computer aids for text analysis. IEEE Trans Commun 30(1):105–110

Manouselis N, Drachsler H, Verbert K, Duval E (2012) Recommender systems for learning. Springer

McNee SM, Albert I, Cosley D, Gopalkrishnan P, Lam SK, Rashid AM, Riedl J (2002) On the recommending of citations for research papers. In: Proc. ACM Conference on Computer Supported Cooperative Work, pp 116–125

O'Rourke ST, Calvo RA (2009) Analysing semantic flow in academic writing. In: Proc. Artificial Intelligence in Education, pp 173–180

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: Proc. the 20th Conference on Uncertainty in Artificial Intelligence, pp 487–494

Schofield A, Magnusson M, Mimno D (2017) Pulling out the stops: rethinking stopword removal for topic models. In: Proc, the 15th Conference of the European Chapter of the Association for Computational Linguistics: vol 2, Short Papers, pp 432–436

Schwarz G (1978) Estimating the dimensions of a model. Ann Stat 6(2):461–464

Shibata H, Hori K (2002) A framework to support writing as design using multiple representations. In: Proc. Asia Pacific Conference on Computer–Human Interaction

Taddy M (2012) On estimation and selection for topic models. In: Lawrence ND, Girolami MA (eds) Proc. International Conference on Artificial Intelligence and Statistics, vol 22, pp 1184–1193

Tang TY, McCalla G (2005) Smart recommendation for an evolving e-learning system: architecture and experiment. Int J ELearn 4(1):105–129

Teh YW, Jordan MI, Beal MJ, Blei DM (2004) Hierarchical Dirichlet processes. J Am Stat Assoc 101

Toulmin SE (1958) The use of argument. Cambridge University Press

Ueno M (2004a) Data mining and text mining technologies for collaborative learning in an ILMS "Samurai". In: Proc. IEEE International Conference on Advanced Learning Technologies, pp 1052–1053

Ueno M (2004b) On-line contents analysis system for e-learning. In: Proc. IEEE International Conference on Advanced Learning Technologies, pp 762–764

Ueno M (2014) ePortfolio system using past learners' history data. J Jpn Soc Inf Knowl 24(4):414–423. doi:10.2964/jsik

Ueno M (2015) Support of learning from the others. J Jpn Soc Artif Intell 30(4):469–472

Ueno M, Uto M (2011) Learning community using social network service. In: Proc. Web Based Communities and Social Media. Proc. web based communities and social media, pp 109–119

Uto M, Ueno M, (2011) Article structure construction support system by Bayes code. IEICE Trans Inf Syst J94-D(12):2069–2081

Uto M, Ueno M (2015) Academic writing support system using Bayesian Networks. In: Proc. IEEE International Conference on Advanced Learning Technologies, pp 385–387

Uto M, Ueno M (2016) Item response theory for peer assessment. IEEE Trans Learn Technol 9(2):157–170

Verheij B (2005) Evaluating arguments based on Toulmin's scheme. Argumentation 19(3):347–371

Villalón J, Kearney P, Calvo RA, Reimann P (2008) Glosser: enhanced feedback for student writing tasks. In: Proc. IEEE International Conference on Advanced Learning Technologies, pp 454–458

Vygotsky LS (1978) Mind in Society: the development of higher psychological processes. MAHarvard University Press, Cambridge

Wallach HM, Murray I, Salakhutdinov R, Mimno D (2009) Evaluation methods for topic models. In: Proc. International Conference on Machine Learning, pp 1105–1112

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J Mach Learn Res 3571–3594

Watanabe S (2013) A widely applicable Bayesian information criterion. J Mach Learn Res 14(1):867–897

Yang JC, Huang YT, Tsai CC, Chung CI, Wu YC (2009) An automatic multimedia content summarization system for video recommendation. Educ Technol Soc 12(1):49–61