Construction of a Literature Review Support System Using Latent Dirichlet Allocation

Yusuke Kometani(✉) and Keizo Nagaoka

School of Human Sciences, Waseda University,
2-579-15 Mikajima, Tokorozawa, Saitama 359-1192, Japan
kometani@aoni.waseda.jp, k.nagaoka@waseda.jp

Abstract. The role of universities in imparting knowledge to students is declining as e-learning and massive open online courses become widespread, and it seems likely that eventually only seminar activities will remain on university campuses. Prof. Nagaoka, Waseda University in Japan, previously proposed the importance of making seminar activities the core of university education, considering them as a "university within a university," and furthermore proposed the concept of a seminar management system (SMS). Following this proposal, we report on the development of a literature review support system using latent Dirichlet allocation as one aspect of an SMS.

Keywords: Seminar activity · University within a university · SMS (seminar management system) · Literature review support system · Latent Dirichlet allocation

1 Introduction

Since the early 2000s, different methods for providing universal access through distance education, such as the OpenCourseWare program and MOOC (massively open online courses), have rapidly gained prominence, and universities worldwide have been pressed to change with the times. Although it seems that in the near future most lecture-type classes are likely to be offered through distance education to off-campus locations, discussion- and participatory-type lessons are still mainly performed at university campuses, and they continue to require in-person attendance, even in Japanese University.

One type of educational model is centered on seminar activities. In particular, each seminar activity offered by a university instructor should fill a role in a larger framework that defines the curriculum. Prof. Keizo Nagaoka have proposed that such seminar activities should support the particular educational philosophy established for a "university within a university," and furthermore proposed the concept of an integrated Seminar Management System (SMS) (Kometani and Nagaoka 2015, Nagaoka and Kometani 2016).

In this report, we propose a system for supporting literature review in undergraduate research as one aspect of an SMS for supporting seminar activities. The proposed system is based on an electronic portfolio (e-portfolio), and uses the latent Dirichlet allocation (LDA) (Blei et al. 2003) and linguistic topic models to create feedback for students.

S. Yamamoto (Ed.): HIMI 2016, Part II, LNCS 9735, pp. 159–167, 2016.
DOI: 10.1007/978-3-319-40397-7_16

A key concept is that students can effectively discover their research interests and research questions through reflection on the materials obtained through their literature review and the LDA-based analysis results. Accordingly, it is expected that using LDA for analyzing the literature review should result in improved student attitude toward seminar activities and research. The main aims of this research are to develop a literature review support system that is appropriate for studying materials gained through the literature review process and to examine whether topic analysis using LDA effectively supports the literature review. We examine the utility of using a literature review support system by analyzing the study materials as well as questionnaire survey results.

2 Design of Literature Review Support System

2.1 Process of Literature Review

In designing a literature review support system, we first define the process of literature review (Fig. 1). To create a review, students should first read many articles critically. For each article, students make their own summary as a study material. Second, when writing the literature review, students should organize and relate the contents of articles to create a narrative and narrow down their topics. Finally, students should decide their research questions. If the students cannot determine their research question after this three-step process, they should repeat the entire process.

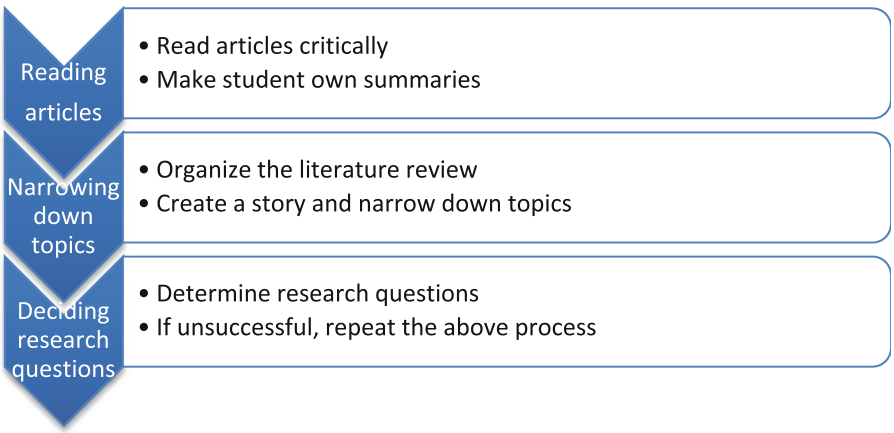

Fig. 1. Literature review process

2.2 Target of Our System and Difficulties Faced in Literature Review

In previous research, many methodologies have been suggested to support students in reading an article critically.

However, few studies have focused on supporting students as novice researchers aiming to narrow down topics. Therefore, our system targets supporting the "narrowing down of topics" stage in Fig. 1.

To narrow down topics, students should relate the contents of each article and create structures. However, it is difficult for students to relate the contents of articles for two reasons:

1. Students who are novices in the research area do not have enough knowledge, making it difficult to understand which articles are similar and to find relations between articles easily.
2. Students do not know how to organize a literature review, as it is difficult to define a well-organized literature review.

2.3 Functions of System

To address the above two difficulties, we propose two functions:

1. Feedback on similarity between students' own summaries of articles
2. Feedback on the literature review structure

The key idea is to use LDA, which is one type of linguistic topic model (Blei et al. 2003). It can be used for estimating the implicit topics of documents. We believe that showing the implicit topics estimated from articles can help support students who do not have sufficient knowledge.

As support function 1, to display the similarities and differences among articles clearly, we propose the student summaries' similarity feedback function using LDA and multi-dimensional scaling (MDS). Figure 2 shows the concept of this function. The topic distribution of each article can be calculated using LDA. We use the divergence between the distributions as an indicator of the similarity of the articles. By calculating the similarity (distance) of each pair of articles, we can obtain a distance matrix. Finally, the similarity of articles can be visualized on an x-y coordinate system using MDS. It is expected that this function will make students aware of which articles are related, thus helping them in writing their literature review.

As support function 2, to display the structure of the literature review, we propose a literature review structure feedback function using the LDA results. Figure 3 shows the concept of this function. The review sentences can be divided into bags of words. Here, each bag of words belongs to one topic, and it is calculated using LDA. As a result, we can obtain a topic occurrence pattern. We assume that there are different patterns between well-trained researchers and novice researchers. Therefore, collecting good patterns will be useful for students in organizing their literature review. A graphing function is created using the data series shown in Fig. 3, and a clear structure is displayed to the students.

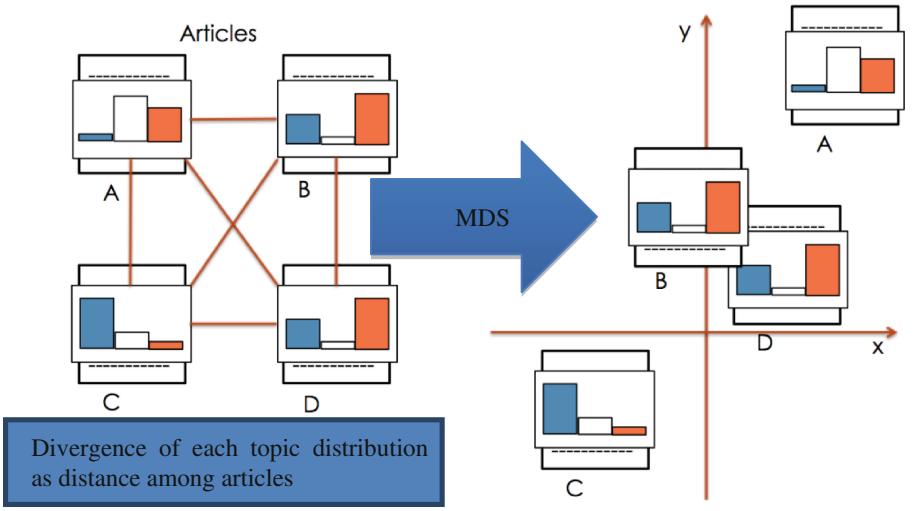

Fig. 2. Concept of student summaries' similarity feedback function using LDA and MDS

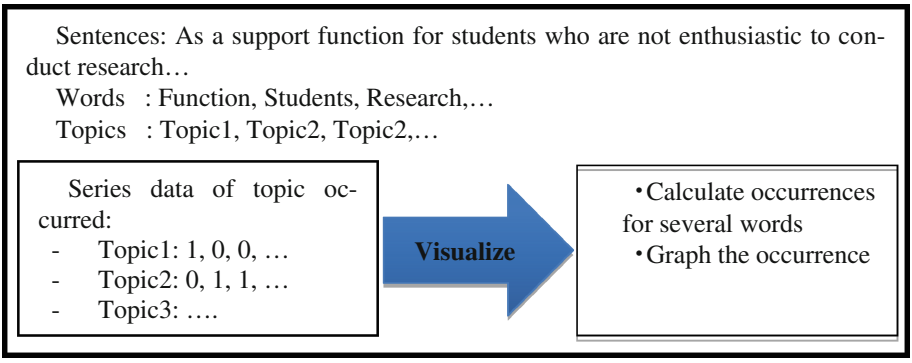

Fig. 3. Concept of literature review structure feedback function using LDA result

3 System Development

We developed a system prototype (Figs. 4, 5, 6, 7, 8 and 9). Figure 4 shows the system configuration. A learner summarizes his or her own article (learner summary) and a review in the client-side user interface, and those are saved into a database (DB). The system reads these document data from the DB, uses morphological analysis to divide them into morphemes for each document, and eliminates stop words. Bag-of-words expressions are derived for each document. Further, only a noun and an adjective are selected in this research. Next, topics for each word are estimated from the bag-of-words data. Topic distributions of learner summaries and topic occurrence patterns of the literature review are calculated using word–topic correspondence. These results are returned to the client-side user interface.

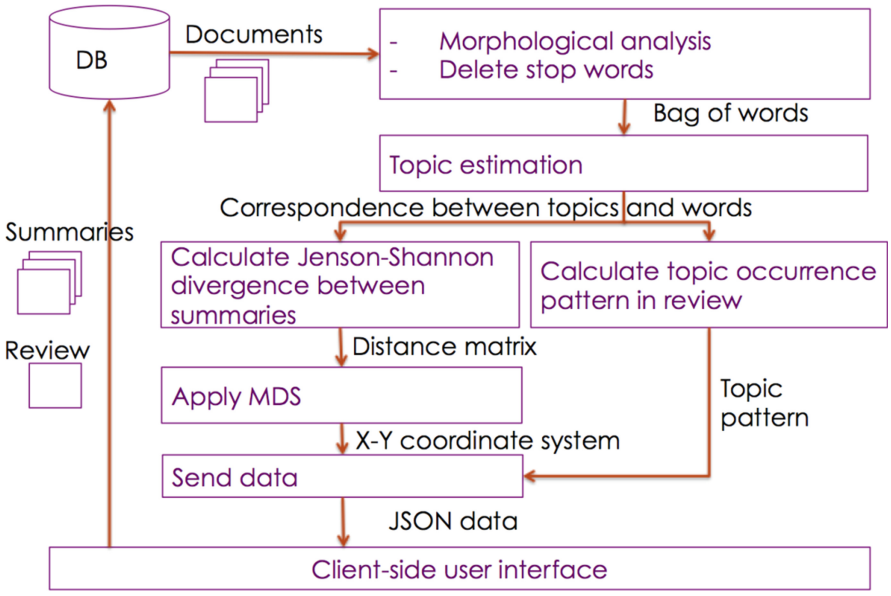

Fig. 4. System configuration

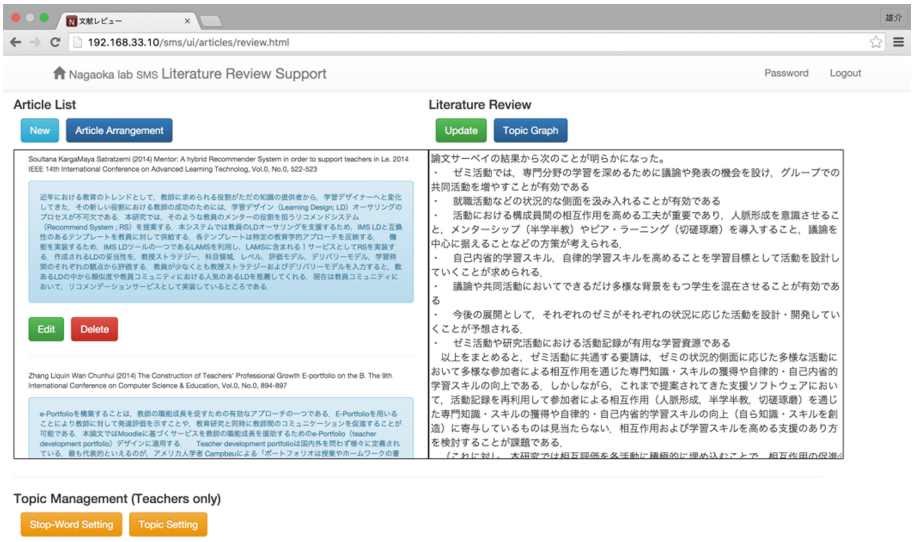

Fig. 5. User interface of the developed system

Figure 5 shows the client-side user interface. The article list is shown on the left side and the literature review editor is shown on the right side. Teachers can use a button at the bottom to call up the topic management dialog. It is possible to create a new thesis using the learner summary by clicking the “New” button of the article list.

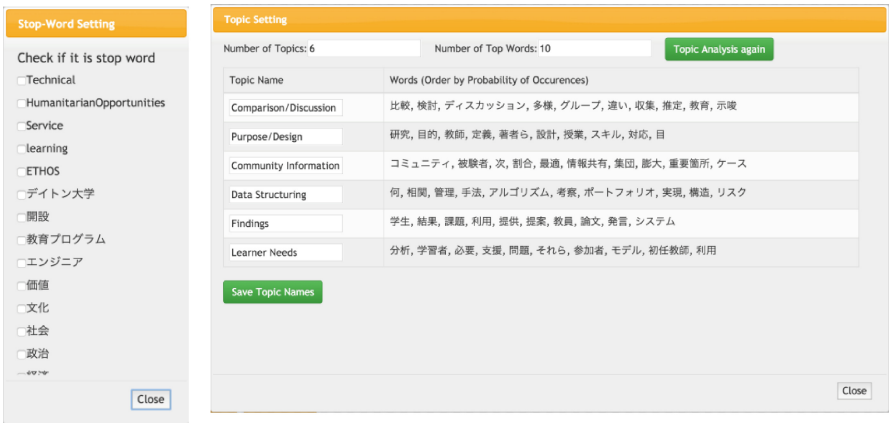

Fig. 6. User interface of stop-word setting dialog and topic setting dialog (teacher use only)

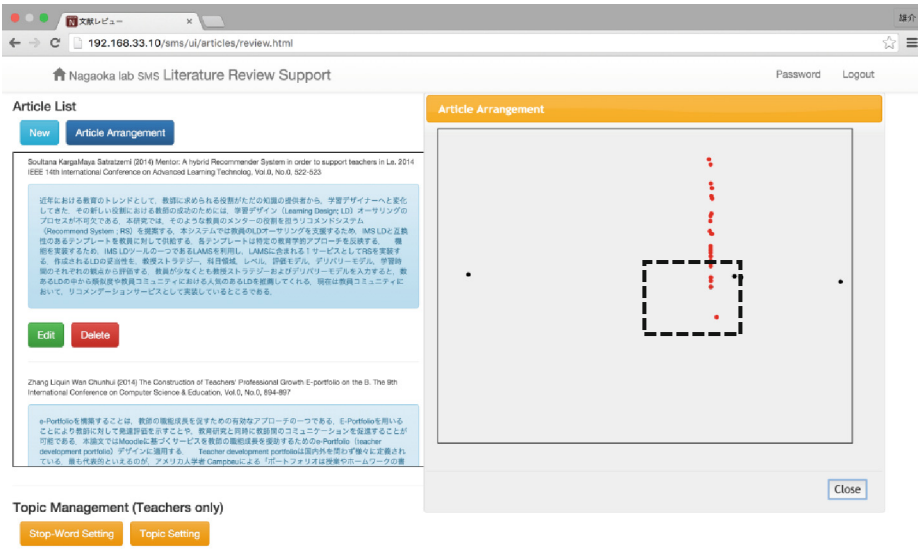

Fig. 7. User interface of the article arrangement dialog

Clicking the “Edit” button allows the registered article information to be edited, and clicking the “Delete” button deletes the article information. Clicking the “Article Arrangement” button allows use of the article arrangement function (Fig. 2). Clicking the “Topic Graph” button shows a visualization of the topic pattern (Fig. 3).

Figure 6 shows the stop-word setting dialog and the topic setting dialog that only teachers are allowed to use. This is important for effectively guiding students toward clarifying whether words are important in the research area. The system thus allows seminar instructors to select stop-words to reflect the intention. Furthermore, topics that can be estimated from summaries are changed if the number of learner summaries

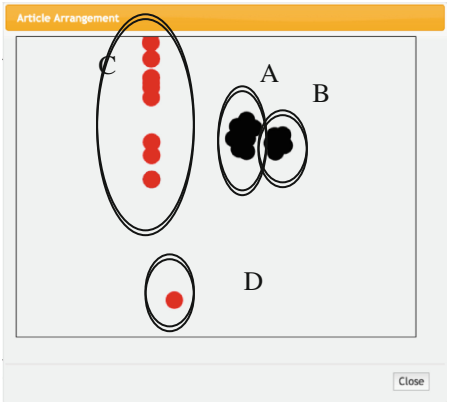

Fig. 8. Focus function of the article arrangement dialog

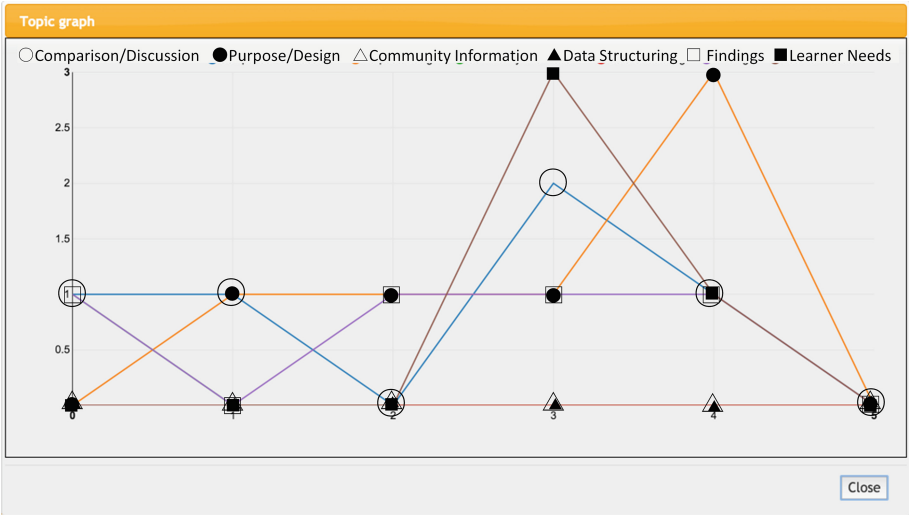

Fig. 9. Topic graph dialog

increases. The topic management screen is prepared for reanalysis and updating the number of topics and topic name.

Figure 7 shows the article arrangement screen. Each dot in the right-side scatter graph is equivalent to one article. When a dot is clicked, information regarding the corresponding article is shown in the article list on the left side. This allows comparison between articles with similar topics. Panning and zooming is possible, allowing more detailed observation of the dot distribution. As an example, Fig. 8 shows a magnification of the area enclosed by a dotted line in Fig. 7. Section 4 presents concrete results.

Figure 9 is a topic-occurrence pattern chart (topic graph) of the literature review. The review text is divided into five parts, and the occurrence frequency of the topic is

calculated by respective parts. We can observe a variation of topic occurrence according to the position in review. This shows that it is possible to visualize the structure of the literature review as a pattern of topic occurrences. Section 4 presents a more detailed consideration of Fig. 9.

4 Results of Trial Use of the System

We experimentally apply the system to a set of seminar activities. At present, eight students (two women, six men) are participating in these activities. They are fourth-year university students who need to complete their graduation thesis this winter.

However, it is necessary to consider initial use of this system in actual situations. If only students make article summaries, there might be anxiety regarding obtaining enough sample data and a sufficient number of articles for a literature review. We therefore suggest that seminar instructors initially add self-created article summaries to the system. This allows students creating a literature review to quote not only their own summaries, but also those by the instructor and other students. Of course, students who have read many articles can still use only their own data.

We analyzed 123 articles in this experimental trial, 101 of which were summaries by one of the authors. The eight students read the other 22 articles.

The technique of using information criteria is a general method for selecting a suitable number of topics. In this research, however, instructors can limit the number of topics to allow seminar teachers to reflect their intention. We therefore selected a number of topics that give results that make sense according to the authors' interpretation, namely 6. Topic analysis resulted in estimated topics being "Comparison/Discussion," "Purpose/Design," "Community Information," "Data Structuring," "Findings," and "Learner Needs." as a result of the topic analysis. "Comparison/Discussion," "Purpose/Design," "Findings," and "Learner Needs" are topics related to article composition, and "Community Information" and "Data Structuring" seem related to article content.

Figures 7 and 8 show the results of article arrangement using these topics. Several clusters can be seen in Fig. 8. Clusters A and B were articles registered by students, and C and D were registered by an author. Both A and B are about communication, but their detailed contents are different. Cluster A includes articles focusing on improving presentation skills, while those in B focus on more practical needs such as career education and work after graduation. The C group is near A and B, but the article contents regard interaction supports between learners and teachers, such as discussion or mentoring. Interaction support can be considered a communication support, but with a different support target than in A or B. The D group consists of only one article, and it is located far from A, B, and C. The article content describes the relation between understanding and lecture evaluations of students in a lecture course, so it differs from A, B, and C. These results show that the article arrangement function grouped articles with similar content. It is useful for students to identify articles in similar categories but with somewhat different targets. This function can thus support creation of literature reviews.

Figure 9 shows the results of the topic graph function. One of the authors collected articles related to seminar activity and wrote reviews for them. We used these reviews for analysis. The first part of each review summarizes knowledge obtained from reading the article, the middle part describes social needs, and the last part describes purposes appropriate for enhancing seminar activity research. In the chart, "Findings" appears first, then "Learner Needs" and "Purpose/Design." While we have presented only one case, we can see that the topic occurrence pattern reflects the actual contents. When learners see this pattern, it will trigger reflection on the structure of the literature review. Furthermore, there is a possibility that learning from others can be induced, such as learning how to structure the literature review by seeing topic graphs of master learners such as graduate students and instructors.

5 Conclusion

Following the SMS concept proposed by Nagaoka, here we proposed a literature review support system as one aspect of a support function. Features include (1) topic models from LDA, (2) displaying article similarity using topic distributions of learner summaries and MDS, and (3) visualization of the literature review structure as an occurrence pattern of topics in the literature review. Trial use of the system shows that these functions can be useful supports for creating a literature review.

In future work we will continue using this system to accumulate learner summaries, evaluate the support and learning effects, and enhance the model. We presume that methods for choosing stop words and topic estimation parameters will influence the analysis results, so we will investigate what support can be realized using this system.

References

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Kometani, Y., Nagaoka, K.: Development of a seminar management system. In: Yamamoto, S., de Oliveira, N.P. (eds.) HIMI 2015. LNCS, vol. 9173, pp. 350–361. Springer, Heidelberg (2015). doi:10.1007/978-3-319-20618-9_35

Nagaoka, K., Kometani, Y.: Seminar Activity as Center of University Education - SMS: Seminar Management System, Proposal and State of Development. Research Report, Japan Society for Educational Technology, JSET15-1 (2016) (in Japanese, printing)