A Topic Model for Clustering Learners Based on Contents in Educational Counseling

Takatoshi Ishii(✉), Satoshi Mizoguchi, Koji Kimita, and Yoshiki Shimomura

Department of System Design, Tokyo Metropolitan University, Asahigaoka 6-6, Hino-Shi, Tokyo 191-0065, Japan
t_ishii@tmu.ac.jp

Abstract. For improving the quality of education, we need to analyze the interaction among the learners and teachers. For example, we empirically know that an agreement among the learner and teacher on the point of learning motivation makes good lecture. For this purpose, this paper aim to characterize the interactions based on the contents in the interactions. This paper employs a topic model for characterizing the interactions. Topic model is a method for estimating topic (theme or subject) in documents and clustering the documents based on estimated topics. By using topic model, this paper analyzes contents in actual educational counseling.

Keywords: Educational engineering · Service engineering · Natural language processing · LDA

1 Introduction

Recently, various senses of values are appearing in society. Accompanying this situation, in higher education, the diversity of learners is increasing. Hereafter, the institution of the higher education must adjust their education for the various learners in order to improving their international competitiveness. For example, the universities in japan have implemented new entrance examination for measuring the ability that cannot be measured in paper test, web based learning/examination system, and new course for adult students who participate part-time. In this manner, the institutions of the higher education are needed to improve their faculty.

In order to improve the faculty for new learners, we employ the framework from service engineering, because we regard education as a kind of service. In service engineering, "service" is defined as the "application of competences (knowledge and skills) for the benefit of another entity or the entity itself" (Lusch and Vargo, 2014). In education, the learners gain knowledge and skills from activities with the faculty. In addition, the faculty gains know-hows (e.g., how to talk for learners' understanding). In this situation, faculty competence and learners competence is applied to the activity in the institution. In addition, the activity in the institution contributes for learners and faculty. Therefore, we regard the education in universities as a kind of service.

Recently, in service engineering, the important philosophy named Service Dominant Logic (SDL) is proposed (Lusch and Vargo, 2014). SDL has a concept for value

S. Yamamoto (Ed.): HIMI 2015, Part II, LNCS 9173, pp. 323–331, 2015.
DOI: 10.1007/978-3-319-20618-9_32

(bene-fit) called as “value in use” or “value in context”. In SDL, the value of the service (and products) is created in the uses of the service (and products). In addition, this value depends on the context of uses, where context is defined as a “set of unique actors with unique reciprocal links among them”. The value of service has two following features: (1) the value is co-created by the actors, and (2) the value is mutually constitutive. For instance, in a lecture, the value for learners is getting knowledge, and the value for teachers is getting experience for how to make good lecture. In this situation, the learners and teachers get the value mutually and these values are created among the interaction among the learners and teachers. Thus, for improving the value of target service, it is needed to consider the context of service and the interaction among the actors.

Accordingly, in order to improve the value of faculty, it is needed to consider the new learners’ context and interaction among learners and teachers. For this final goal, the purpose of this paper is to propose a method that senses the learners’ context from the interaction among the learners and teachers. In many case on the education, this interaction is described in natural language. For example, discussion, descriptions on the blackboard, mini-tests, homework, and the response for the homework are described by natural language. However, qualitative analysis (e.g., conversation analysis) by hand requires huge cost. In addition, digital data of interactions on Learning Management System: LMS are increased by the progress of information technologies.

From this background, this paper attempts analysis of those interactions without human hand. To achieve this, this paper applies a natural language processing method for characterizing the interactions. To be more precise, this paper applies Latent Dirichlet Allocation (LDA) (Blei et al. 2003) that is a method for estimating the topic in the documents, and for clustering those documents based on estimated topic. By using LDA, this paper can characterize the interactions based on the estimated contents. In this paper, to demonstrate the usability of the method, we show an application to analyze a counseling data.

2 Latent Dirichlet Allocation (LDA) (Blei et al. 2003)

This paper employs Latent Dirichlet Allocation (LDA) (Blei et al. 2003) for analyzing contents in interaction. This chapter introduces Latent Dirichlet Allocation (LDA) that is one of topic model. There are some topic models: Latent Semantic Analysis (LSA) (Deerwester et al. 1990), Probabilistic LSA (pLSA) (Hofmann 1999), Latent Dirichlet Allocation (LDA) (Blei et al. 2003), and more. Topic model is a method for clustering documents based on the topics (subjects or contents) in the documents. From these topic models, this study employs LDA for analyzing the interaction. LDA assumes and models that the document includes some abstract “topics” that are subject or contents in the documents. LDA estimates the topic of each word from bias of word frequency in the documents. By aggregating the topics of words in the document, LDA estimates the topic allocation of each document.

For example, there is an example document: “I want to learn English. Because I’m interest in European football, so I want to go Spain for watching football games.” In this example, the result of LDA indicates that the word: “English” and “learn” have a

topic of “English learner”, and the other word: “football” and “game” have a topic of “football”. Thus, this document has two main topics: “English learner” and “football”.

LDA assumes that each document has multiple topics. On the other hand, LSA and pLSA assume that the document consist of one abstract topic. Thus, the assumption of LDA is fitter actual document, and has better accuracy than LSA and pLSA.

In our study, we assume that free descriptions for interaction include some topics of contents. For example, a counselling log about learning motivation will include two types of contents “as an English learner” and “as a football fun”. LDA is able to model such case that various topics are included in a document. Therefore, in this study, LDA is employed for estimating topics in interaction.

3 A Topic Model for Clustering Learners Based on Contents in the Educational Counseling

In this chapter, we introduce an application of topic model for analyzing interaction among learners and teachers. An overview of this application is shown in Fig. 1. The first, we obtained learning counseling data. The next, we analyzed those data with LDA.

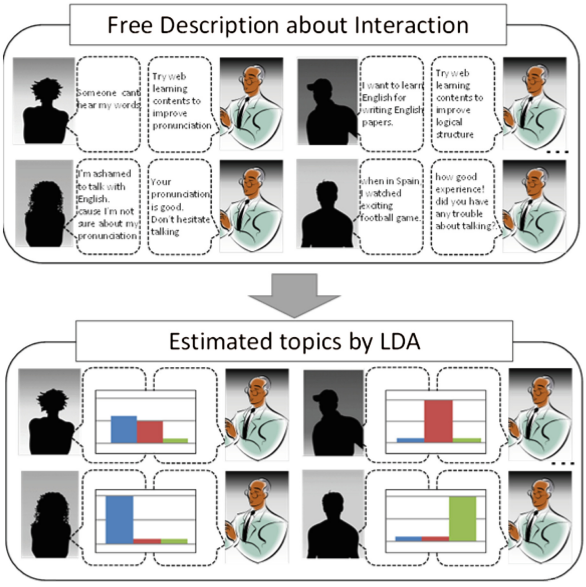

Fig. 1. Overview of the proposed method

3.1 Obtained Data

The first, we obtained the data of interaction. As this data, we employed educational counseling data for an English lecture for Japanese students. This lecture was given in 2014 for engineering students in Japan. The number of students is 23. The counseling

is conducted for 17 learners in Japanese. We got the 17 voice data from 17 counseling, and then converted the voice data to 17 text data. Thus, we got 17 text data from 17 learners. This text data included speeches from learners and teachers. We used verbs and nouns for analysis. The statistics of the 17 text data is shown in Table 1.

Table 1. The statistics of the 17 text data

Number of words	139834
Number of vocabulary	7044
Average of # words	8225.5
Standard division of # words	2558.8

3.2 The Result of LDA for the Obtained Data

The next, we analyzed the obtained text data based on LDA. The results are shown in Tables 2 and 3.

Table 2 shows the estimated topic allocation (rate) for each data. In addition, the cells with Top 30 % values are painted red. For example, the column of ID03 in the table shows that ID03 has 12.2 % of topic 1, 82.1 % of topic 2 and 4.7 % of topic 7. From this result, the main subject in the counseling is topic 2 because each data has over 70 % of topic 2.

Table 2. Estimated topic allocation for each counseling data

ID	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
ID01	13.4%	82.0%	1.0%	0.0%	3.6%	0.0%	0.0%
ID02	0.4%	74.4%	0.0%	0.0%	0.7%	24.5%	0.0%
ID03	12.2%	82.1%	0.5%	0.4%	0.0%	0.0%	4.7%
ID04	0.0%	72.8%	0.5%	0.0%	25.2%	0.0%	1.5%
ID05	1.1%	77.7%	2.1%	0.0%	11.2%	0.1%	7.7%
ID06	0.0%	78.0%	0.0%	0.0%	1.7%	20.2%	0.0%
ID07	23.1%	76.1%	0.0%	0.0%	0.2%	0.6%	0.0%
ID08	0.0%	62.6%	0.2%	37.1%	0.0%	0.0%	0.1%
ID09	0.0%	79.0%	1.4%	0.0%	19.6%	0.0%	0.0%
ID10	0.0%	75.4%	1.8%	0.0%	0.4%	0.0%	22.5%
ID11	0.7%	79.6%	0.0%	0.9%	17.8%	0.0%	0.9%
ID12	0.0%	80.5%	0.7%	0.0%	11.6%	7.1%	0.0%
ID13	0.1%	76.5%	0.0%	0.0%	0.0%	0.0%	23.3%
ID14	16.6%	77.3%	0.0%	0.0%	5.8%	0.0%	0.3%
ID15	1.9%	75.3%	0.0%	0.1%	1.0%	21.7%	0.0%
ID16	3.6%	72.9%	0.0%	0.0%	22.0%	1.5%	0.0%
ID17	0.0%	64.9%	35.1%	0.0%	0.0%	0.0%	0.0%

Table 3. The representative 5 words of 5 topics

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7
Research	English	United kingdom	United kingdom	Review	United kingdom	Aerospace
Graduate school	Talk	Friends	Motivation	Unit	Football	Technical term
Society	Understand	Dormitory	Speech	Quality	World cup	Presentation
Infrastructure	Pronunciation	Company	Sing	Studio session	Challenge	Paper
Museum	Question	English conversation	Golf	Click	Game	Oral

Table 3 shows the representative 5 words for each topic from LDA analysis. From this table, we can understand what word appears with high probability in each topic. For example, in the topic 7, the words "Aerospace" and "technical term" appeared. Therefore, we can understand the topic 7 is about the research of aerospace engineering.

4 Consideration

4.1 Meaning of Each Topic

The first, we need to consider the meaning of each topic. The most frequent topic was topic 2, and this topic was regarded as "talking about English lecture". The words "talk" and "pronunciation" are subjects of talking in English, and the words "understand" and "question" have relations to instructional activity. Therefore, topic 2 was regarded to "talking about English lecture". Of course, all of the counseling has topic 2 with high rate, because this is the main theme in the counseling. In the same manner, topic 1 was as "talking about academic", topic 3 was as "talking about studying in abroad", topic 4 was as "talking about learning motivation for English", topic 5 was as "talking about learning contents", topic 6 was as "talking about European football", and topic 7 was as "talking about the research of aerospace engineering"

4.2 Clustering the Counseling Data

Moreover, we can cluster the 17 counseling data based on contents similarity. This clustering did not require human hand. (Only the meaning for clusters requires human power.) We regarded topic rate as feature vector, and calculate similarity (i.e., cosine similarity, and J-S divergence as distance).

In this result, 17 counseling data were clearly divided to 6 clusters. The counseling data were divided by including - excepting topic 2. For example, the biggest cluster is the set of ID04, ID05, ID09, ID11, ID12, and ID16. Each data in the cluster includes topic 5. In short, these learners talked about web-learning contents in the counseling. Therefore, summarizing these results, LDA characterized each counseling data based on contents of counseling.

4.3 Application for Improving Faculty

From these result and consideration, we could understand what contents were talked about in each data. For example, 6 learners (ID04, ID05, ID09, ID11, ID12, and ID16) talked about web learning contents, and they probably interest in learning on the web contents. Other 4 learners (ID01, ID02, ID07, and ID14) and 2 learners (ID10 and ID13) talked about their research on graduate school. In addition, the former group (ID01, ID02, ID07, and ID14) talked about infrastructure, and the latter group (ID10 and ID13) talked about aerospace. These were probably their majors. This analysis found out a part of learners context. This result is expected to be utilized for improving faculty.

The last, we checked the relationship among this clustering result and other statistics those are expected to characterize learning activity. To be more precise, we checked the interrelation among the clustering result, performance measured by an examination, personality, interest for web contents, and the ability for self-regulated learning. Table 4 shows all learners' performance, personality, interest, the ability for self-regulated learning, and results of the clustering result by LDA application.

Table 4. The clustering based on counseling contents, performance, personality, interest, and the ability for self-regulated learning for each learners.

ID	Performance	Personality	Self-regulated	Interest	LDA
ID01	2	A	A	B	B
ID02	-	D	C	B	C
ID03	1	A	C	A	B
ID04	2	B	A	A	A
ID05	0	B	A	C	A
ID06	1	A	D	B	C
ID07	0	A	D	C	B
ID08	1	C	A	B	F
ID09	0	D	B	A	A
ID10	0	C	B	D	D
ID11	1	-	-	-	A
ID12	4	A	B	-	A
ID13	0	A	B	B	D
ID14	0	B	D	A	B
ID15	1	D	C	D	C
ID16	0	A	A	C	A
ID17	0	C	D	A	E

The "performance" is indicated the deference of before and after test grade. The learners took the OPIc examination (The American Council on the Teaching of Foreign Languages) twice: before and after the counseling. The "personality" indicates clustering result of personality. The learners were characterized using big 5 personality test (Murakami and Murakami, 2001). We divided learners to 4 cluster based on features of

big 5 test. The "self-regulated" indicates the ability of self-regulated learning. We employed the item set and the scoring in (Gouda et al. 2012), and we clustered learners to 4 cluster based on the scores. The "interest" indicates the clustering of interest for web learning contents. We employ ARCS model (Keller 2010) for characterizing the interest for web contents. For detecting interrelation, we removed three data: ID02, ID11 and ID12 that included missing values.

Figure 2 shows scatter plots and Spearman's rank correlation coefficient for each pair among the values. This result was output by R (The R Project for Statistical Computing). Unfortunately, we could not find out the pair having strong correlation. This result possibly bring out this analysis do not have utilize for characterize educational activity. (The number of data may be insufficient to draw conclusions. From this data, we could not find out strong interrelation among other 4 statistics either.) Therefore, the future work includes following two tasks: (1) to find the educational activity (or statistics) related with this clustering result, and (2) to validate those relation by obtains sufficient data.

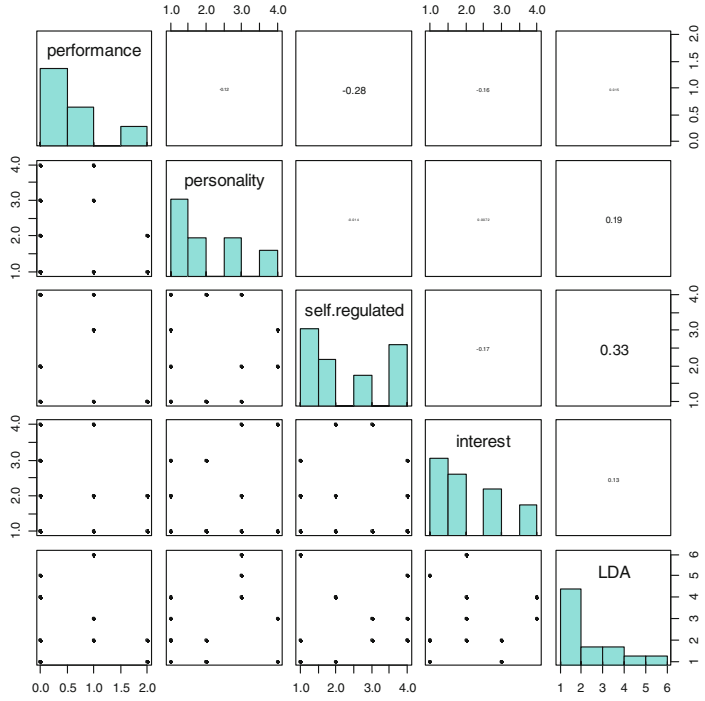

Fig. 2. The result of scatter plots and Spearman's rank correlation coefficient for each pair of values in Table 4.

Moreover, the future work includes finding the interaction contents that have strong interrelation for the educational activity. There are variations of LDA processing the documents with tags. For example, there is a Labeled-LDA model (Ramage 2009). In These models, the model estimate topics from not only words frequency but also tags.

In addition, we can indicate topic meaning by tagging documents. (These models are called as "semi-supervised model"). Those models is expected to characterize contents in interaction based on target statistics. For example, if the documents are tagged on the point of performance ("up", "stay", and "down"), these models will provide what contents (or words) are included in each performance group. Therefore, the future work includes (3) to implement semi-supervised LDA model for improving utility.

5 Conclusion

In this paper, for analyzing actual educational counseling, we applied LDA that is a natural language processing method. This application provided the result of clustering based on contents in the counseling. For this clustering, this application did not require human's hand. Thus, this application is a way to reduce cost of qualitative analysis for the counseling. LDA can apply other various data (e.g., discussion on lecture or LMS, and/or a submitted homework). These data of interactions increase from now on. Thus, qualitative analysis supported by machine learning becomes more important.

Future work will include following 3 tasks: (1) to find the educational activity (or statistics) related with this clustering result, and (2) to validate the relation by obtains sufficient data, and (3) to implement semi-supervised LDA model for improving utility.

Acknowledgement. This research is supported by Service Science, Solutions and Foundation Integrated Research Program (S3FIRE), Research Institute of Science and Technology for Society (RISTEX), Japan Science and Technology Agency (JST).

References

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

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)

Gouda, Y., Yamada, M., Kato, H., Matsuda, T., Saito, Y., Miyagawa, H.: The scale for self-adjusting learning in asynchronous distributed e-learning. Kumamoto University education annual report, vol. 15, pp. 9–20 (2012) (in Japanese)

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

Keller, J.M.: The ARCS Model Approach, p. 345. Springer, New York (2010)

Lusch, R.F., Vargo, S.L.: Service-Dominant Logic: Premises, Perspectives, Possibilities. Cambridge University Press, Cambridge (2014)

Murakami, Y., Murakami, C.: Handbook of Big Five Personality Test. Gakugeitosho, Tokyo (2001). (in Japanese)

Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP 2009 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1 (2009)

The American Council on the Teaching of Foreign Languages (n.d.). Oral Proficiency Assessments (including OPI & OPIc). http://www.actfl.org/professional-development/assessments-the-actfl-testing-office/oral-proficiency-assessments-including-opi-opic. Accessed March 2015

The R Project for Statistical Computing (n.d.). http://www.r-project.org/. Accessed March 2015