International Journal of Institutional Research and Management
International Institute of Applied Informatics
2017, Vol.1, No.1, P.103 – 116

Visualization for University Brand Image Clustering: Comparison between Male and Female Students

Yukari Shirota [*], Setsuko Katayama[*], Takako Hasimoto[†]
Basabi Chakraborty [‡]

Abstract

In this paper, visualization of the results of clustering the brand images of different universities has been presented. Generally high school students care a lot about the brand images of the universities. The brand image is an important criterion for their selection of a particular university to continue their studies. The brand images of universities have been analyzed here by using Bayesian inference. Bayesian inference is widely used in various application fields such as topic extraction. An analysis by topic model has been conducted so far in topic extraction. However, as far as we know, no research that applies the topic model for the clustering of university brand images is available. This is the first application of the topic model in clustering brand images of universities. In this paper, we present the simple topic model visualization tool that we have developed and, as an example, we have applied our developed tool to visualize the clustering of university brand image. When high school students select a university, their university choices greatly depend on the universities' brand images. So the university public relations section needs the survey of the brand images. The clustering results are helpful for the public relations section of the universities to fix up their publicity strategy. The clustering results show that there is a class of most popular and most difficult universities for high school students and there is another class of public universities which can be characterized by high level education and low cost tuition. The outcome of the clustering is quite consistent with our expectation. In this paper comparison of university choices between male and female students is also presented.

Keywords: Bayesian inference, MCMC, visualization, simple topic model, Gibbs sampler, university brand image, clustering.

1 Introduction

Bayesian inference is widely used in decision processes during data analysis in various application fields [1-3]. The goal of Bayesian inference is to maintain a full posterior probability distribution over a set of random variables. However use of this distribution involves computation

[*] Gakushuin University, Tokyo, Japan
[†] Chiba University of Commerce, Chiba, Japan
[‡] Iwate Prefectural University, Iwate, Japan

of complicated integrals. Sampling algorithms based on Monte Carlo Markov Chain (MCMC) techniques are popularly used to avoid complicated calculations. Gibbs sampling is one MCMC technique suitable for this task. Previously we developed a tool based on Gibbs sampler to visualize the Markov Chain Monte Carlo (MCMC) process for simple topic model and used the tool as the teaching material [4].

We have done researches on topic extraction by using Latent Dirichlet Allocation model [5-7] and applied the model in various text mining applications [8-12]. The initial purpose of developing the visualization tool was to provide a teaching material for our students. The mathematical process in Bayesian inference is difficult to understand for our students. They are unfamiliar to Markov chain and conditional probability and so it is difficult for them to derive posterior probability. Therefore we developed the visualization tool when we teach the MCMC algorithm in our classes. However we have found that the Gibbs sampler visualization tool had been useful not only for teaching purpose, but also it is valuable in practical use. By using the tool, we can see whether the Markov chain becomes an invariant status or not. In other words, we can visually see whether the burn-in period we have set is sufficiently long or not. Another advantage of the tool is that we can see interactively and visually to which class the target document should belong after LDA analysis.

In this paper, we would like to present the results of our experiments on university brand image clustering using the MCMC visualization tool. When high school students select a university for their admission, the brand image of the university have a profound effect on their selection. We can assume that the high school students are brand-loyal. We are not sure whether almost all students have similar university brand images. However, we assume that their images do not differ largely from others. We have also conducted the comparative analysis of university brand images perceived by male and female students separately. Here we used the approach of presenting our results of experiments to the students using visualization tools as before because we found that it is much easier for students to grasp the algorithm and the mathematical process behind it with visualization of the process.

As the previous work in DSIR 2015, we used the same approach as we used to teach Latent Semantic Analysis and its mathematical process[13]. We did the clustering for getting the groups of similar university brand image. We used the reliable data by Recruit Co. [14] as the input data. This data is widely used among university office staff in Japan. Although the survey results may be quite different from the actual/real values of the universities, the survey of the current situation of the high school students could be extracted, which is considered to be important. Such analysis is useful for the university public relations section.

In the next section, we shall explain the simple topic model and Gibbs sampler that we used in the visualization tool. In Section 3, we present the visualization of the clustering and explain the results. Then in Section 4, we shall compare results obtained for male and female students and in Section, 5 we describe some of the related works. In the last section, we conclude the paper.

2 Simple Topic Model and Gibbs Sampler

In this section, we shall explain the simple topic model and Gibbs sampler in brief. The topic model has been used as popular Latent Dirichlet Allocation (LDA) topic model [5-7]. First we would like to visualize the topic model. However, the visualization is too complicated for students to understand. Instead of that, we used the simple topic model. For the Markov chain Monte Carlo (MCMC) implementation of the simple topic model, we used the Gibbs sampler.

2.1 Simple Topic Model

In the visualization material of Gibbs sampler, we adopted, instead of the topic model, a simple topic model (or mixture of unigrams). A simple topic model is a simple modified version of the topic model with the limitation that a document has only one topic [15, 16]. There, one topic has its word distribution.

The most effective point of the visualization for students is the changes of the word distribution of a topic. When the document is given a new topic ID, all the words of the document will move to the new topic word group. On the screen, we can see the word moving between the previous topic group and the new topic group. The animation of word moves can make students understand the clustering process of the Gibbs sampling.

We have selected a simple topic model because the simple topic model shares basic model concept as the topic model and then we think that understanding of the simple topic model can lead smoothly to understanding of the topic model in particular. Another reason of the selection was that the topic model was too complicated to be visualized on the screen. In the simple topic model, one document can have several topics. The topic distribution of the document was too complicated owing to the limitation of the space.

2.2 Gibbs Sampler

In Markov chain Monte Carlo (MCMC) process, the sequence is constructed so that, although the first sample may be generated from the prior sample, successive samples are generated from distributions that probably get closer and closer to the desired posterior [17]. In Gibbs sampler, we iterate over each of the unsolved variables, sampling a new value for each variable using all other (n – 1) variables [17]. In case of the topic extraction of documents, a topic identification of each documents is in turn decided from other (n – 1) documents' current status.

We visualize the meaning of this as shown in Figure 1. The spheres represent documents. In Figure 1, there are fifty documents and the number of topics is five. The height of the sphere represents the topic ID. On the radius, a topic probability distribution function of the document is shown. The highest value topic is selected as the new topic ID of the target document.

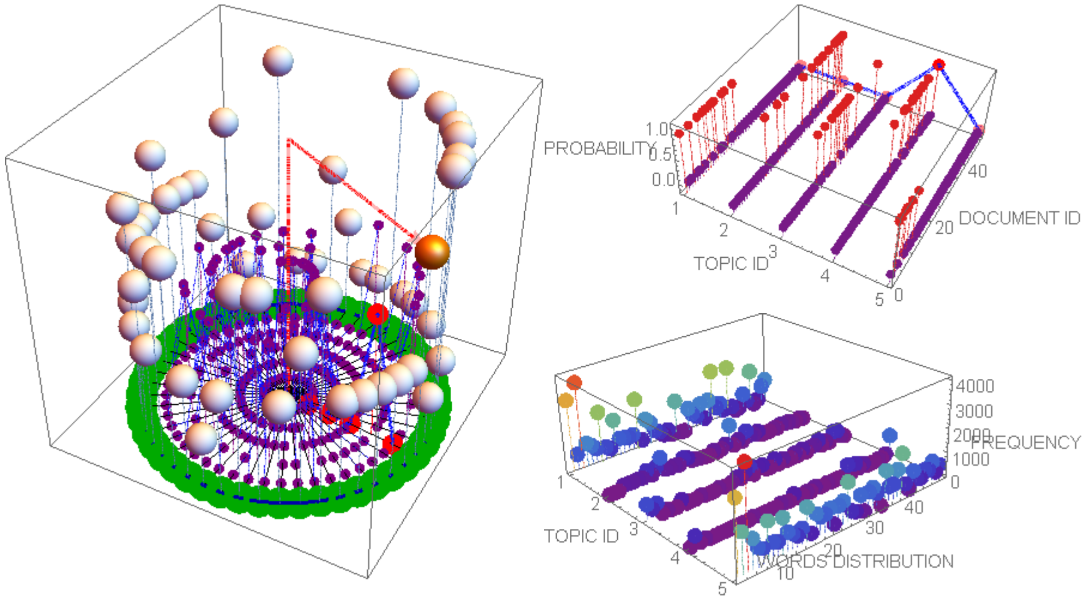

Figure 1-1: The visualization of the university brand image clustering. The number of repetitions are 10 rounds. The right sided upper figure illustrates the topic probability density of the individual document. In the figure, the target document has the highest probability at the topic 4. The right sided lower figure illustrates the word distribution probability of individual topics. In the figure, the topic 1 and 5 have the most frequencies, compared with the other topics.

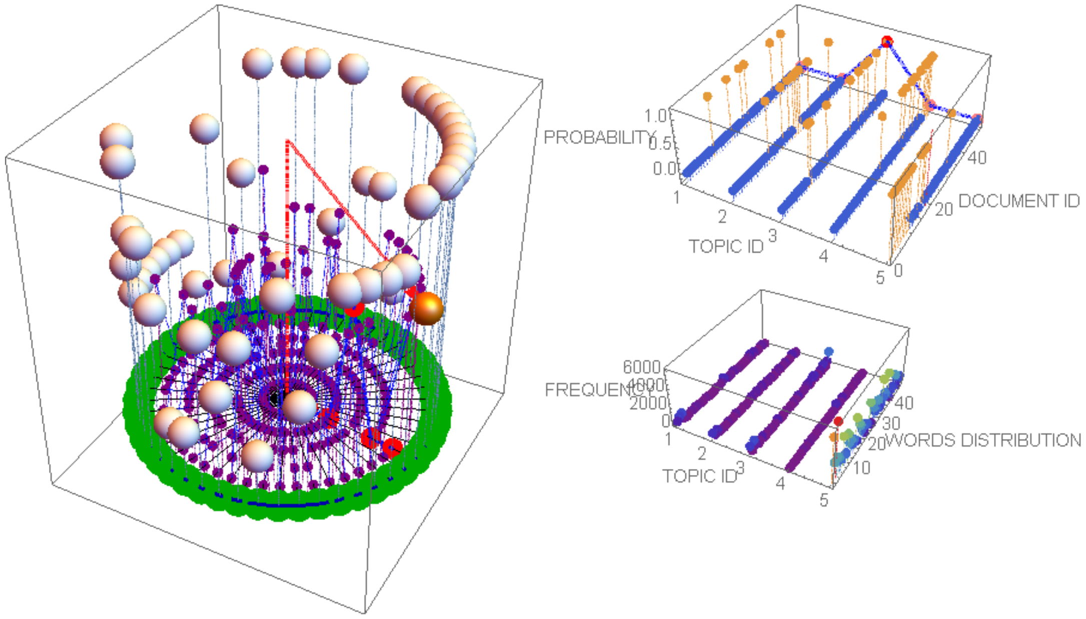

Figure 1-2: The visualization of the university brand image clustering. The number of repetitions are 20 rounds. The right sided lower figure illustrates the word distribution probability of individual topics. In the figure, the topic 5 has the most frequencies, compared with the other topics. The highest word distribution of topic 5 can be seen later through every status. The topic 5 included the most popular universities such as University of Tokyo.

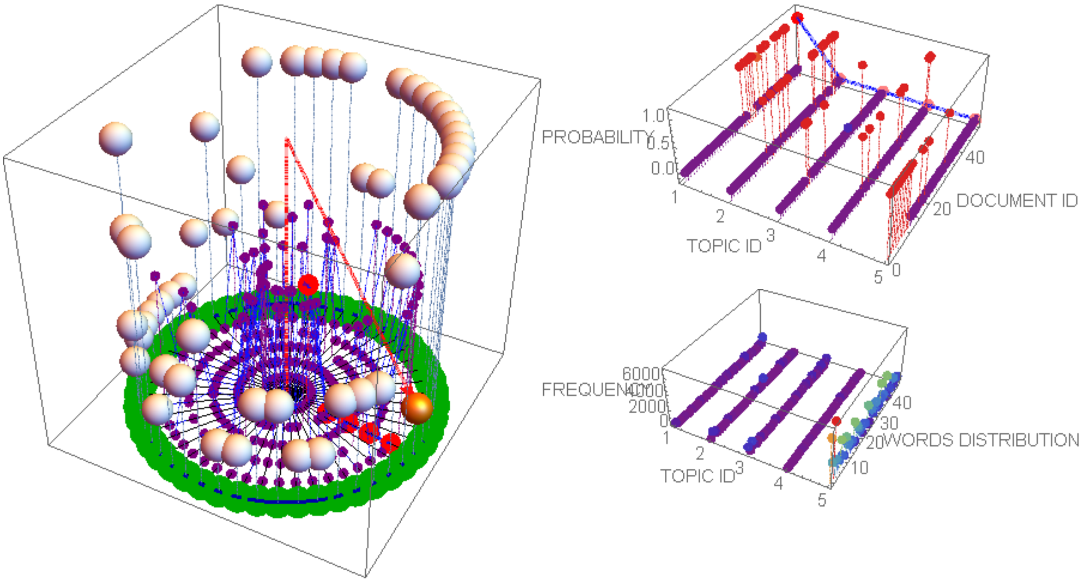

Figure 1-3: The visualization of the university brand image clustering. The number of repetitions are 30 rounds. Seeing the left image, we can see that the clustering has been almost fixed; there is no great change between the figure after 20 rounds. The right sided lower figure illustrates the word distribution probability of individual topics. In the figure, the topic 5 has kept the most frequencies, compared with the other topics. The highest word distribution of topic 5 can be seen later through every status.

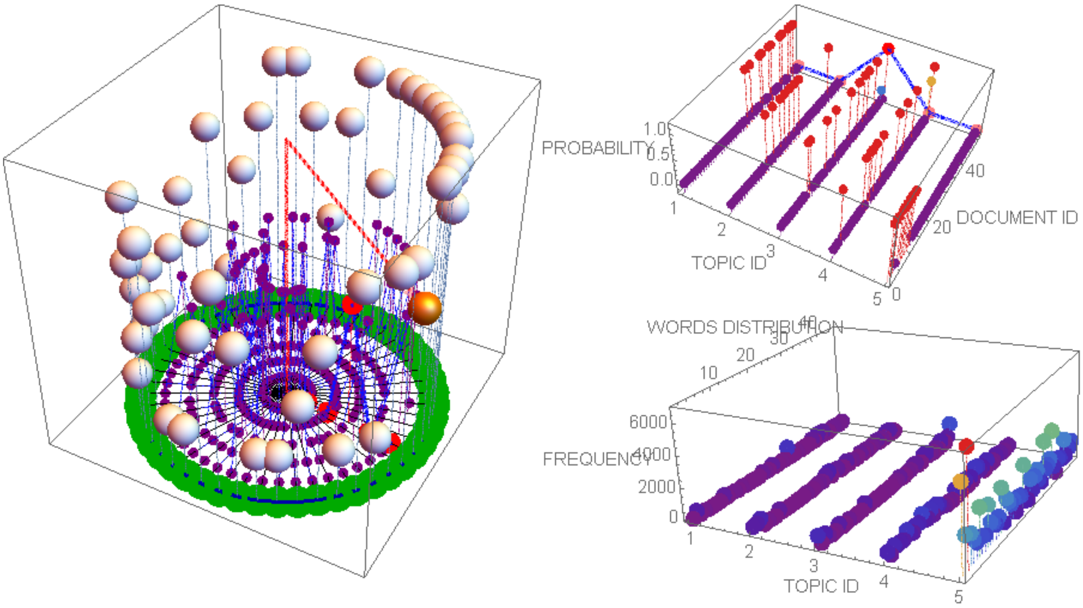

Figure 1-4: The visualization of the university brand image clustering. The number of repetitions are 40 rounds. Seeing the left image, we can see that the clustering has been slightly modified from the figure after 30 rounds. They are fluctuations in the equilibrium state.

In MCMC, after enough time repetition of the substitutions, the target distribution P(x) becomes the invariant distribution of the Markov chains. The invariant distribution means that when we generate a sample x from the distribution P, substituting x by x', by the Gibbs update operation $P(x_i'|x_1,\cdots,x_{i-1},\quad x_{i+1},\cdots,x_N)$ then, the distribution of x' is again inclined to become P as follows [18] :

$$\sum_{x_i}\{P(x_i'|x_1,\cdots,x_{i-1},\quad x_{i+1},\cdots,x_N)\times P(x_1,\cdots,x_{i-1},\ x_i,\quad x_{i+1},\cdots,x_N)\}$$

$$= P(x_i'|\ x_1,\cdots,x_{i-1},\ \ x_{i+1},\cdots,x_N)\times P(x_1,\cdots,x_{i-1},\quad x_{i+1},\cdots,x_N)$$

$$= P(x_1,\cdots,x_{i-1},\qquad x_i',\ \ x_{i+1},\cdots,x_N)$$

In visualization of the Gibbs sampling algorithm, our challenge was to adopt the cylinder shape. To emphasis that the other (n – 1) documents' status, their topic IDs and word frequencies make the decision of the left document and that the changes will be conducted in turn, the cylinder shape was effective. The students of our classes could easily understand the corresponding mathematic expression after watching the graphics.

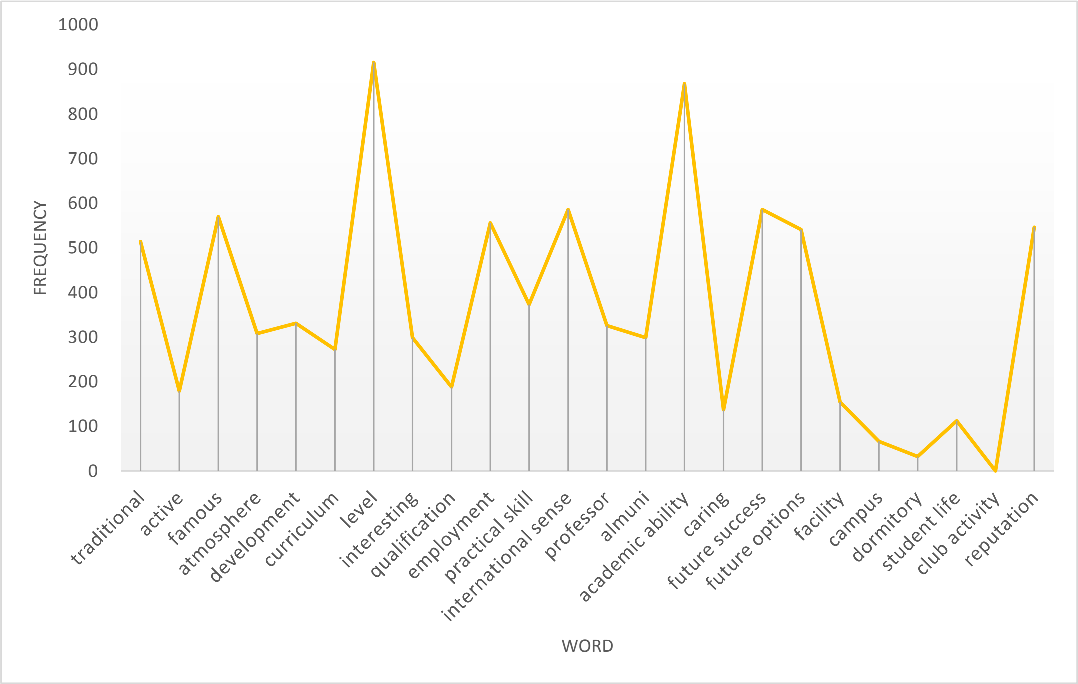

Figure 2: Word distribution of class 4 for the total data. The representative words are "educational level", "academic skill", "low tuition", and "intelligent". The class 4 include many public universities so we can understand the word appearance in the word distribution as high frequency ones.

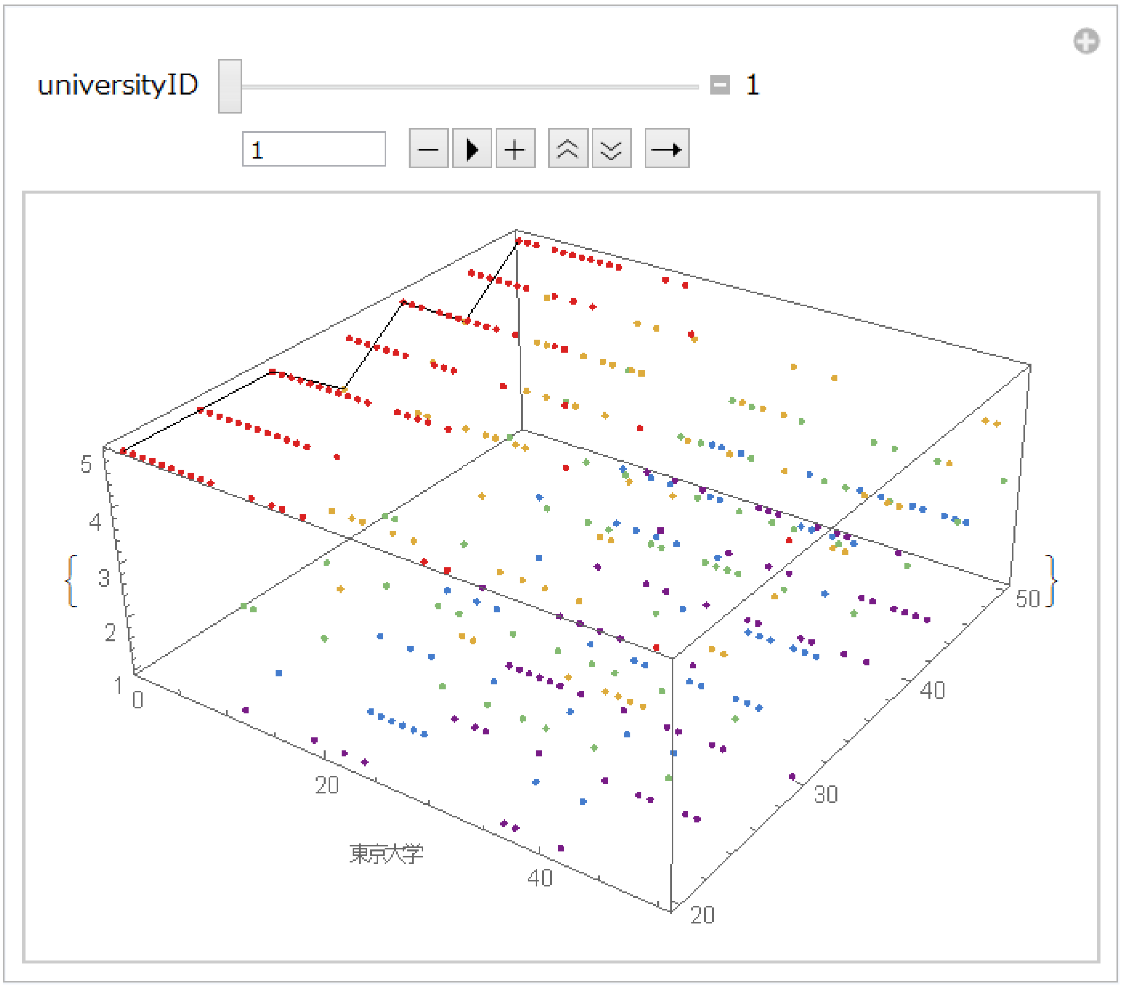

Figure 3: In the equilibrium state University of Tokyo is fluctuating between class 5 and class 4.

In MCMC simulations, we discard the burn-in period data because the data has not yet been invariant. The evaluation of burn-in period is difficult. Although it is up to the applications, in general, we think that 25 to 100 repetitions may be sufficient enough for considering burn-in period. That is an empirical knowledge. The MCMC visualization tool might produce quite well evaluation on this decision when the burn-in period is.

3 Visualization of the brand image clustering

In this section, we present the method of clustering of university brand images. The survey data [14] that we used offers various kinds of questionnaire data. Among them, we have used the questionnaire data of Kanto region as our data. The questionnaire includes 47 brand image attributes that characterize the positive image of a university. Among those 47 attributes, 32 attributes correspond to “functionality” of the university and 15 attributes consider the “sensibility” of the university as universities consider those attributes as their own characteristics. The examples of former attributes are “location”, “career” and “facilities” and the latter attributes are “traditional”, “famous”, and “posh”. In text mining process, we prepare a bag of words (BOW) for a document and count the frequency of each word appeared in the document. In this case, a university is supposed to be a document and an attribute of the university is supposed to be a word.

The data used[14] gives the top 15 ranking data of universities on each attribute. In the ranking list, each university has a score from 0 to 100 %. For example, concerning the attribute "traditional", University of Tokyo is the first one and the value is 71.4 which means that the 71.4% of the sample population of high school students think that the university is traditional and has a long history. Therefore the sum of the top 15 universities' score values may be over 100. In the analysis, we use the Kanto region questionnaire data because we are interested in the Tokyo area. Then, we selected the top 50 universities by the total score of the 47 attributes. The BOW matrix size is 50 (universities) times 47 (attributes).

Figure 1 presents the visualization process. In the following explanation, we use terms "university" and "attribute" instead of "document" and "word". There are four images in Figure 1 which correspond to ones after every 10 rounds from 10 to 40 rounds. One round has 50 repetition. There are 50 universities and each university is represented as a sphere. The height of the sphere shows the class ID. In this sample, the total number of the classes is set to be five. In the visualization tool, the user can move the status back and forth, step by step, by slider mover so that they can see the changes.

Table 1. The class and its attributes.

classID	#of universities	features
1	10	famous/club-activities
2	11	specializatione/qualification
3	8	posh/traditional/facsionable
4	8	high-level-education/academic-achievement/reasonable-tuition
5	13	most-popular-difficult/famous/traditional/intelligent

The topic distribution of a document is also shown in Figure 1 which represents the distribution of a university's class ID. A university has a class ID probability distribution and, among the five probability values, the maximum value class is selected and the university becomes the member of the class, because in a simple topic model, a document has only one topic. In other words, seeing the word distribution of the target document, the most similar word distribution class is selected for the target document. The target university becomes a member of the most similar attribute distribution class.

In the result, after 30 rounds, the Markov chain has obtained the invariant distribution status. In Figure 1, we can see that there is almost no change between the class members, compared with one after 40 rounds. The classes obtained just after 50 rounds are shown in Table 1. The class 5 includes the universities having the top scores such as University of Tokyo, Waseda university and Keio university. We can infer that the class 5 is the first group for aspiring high school students because of the highest scores.

The class 4 attributes also includes universities that offer high level education to students (See Figure 2). Other features of class 4 are academic skills and low tuitions. Almost all class 4 members are national universities or governmental public ones. The class 3 characterizes posh universities with traditional atmosphere. Among the eight universities in class 3, five are women's universities. The class 2 universities characterize some specialization. Among the 11 members of class 2 universities, there are three medical universities, one pharmaceutical, three engineering/agricultural, and one teacher's training university. The Buddhist priest, bonze qualification education is offered by one of them. The class 1 universities represent skill in sports and club activities. The class 1 universities include various kinds of universities.

In the invariant distribution status, the universities with top scores such as Waseda, and Keio do not change the class IDs. However, University of Tokyo fluctuates in the equilibrium state between class 5 and class 4 (See Figure 3). This is because University of Tokyo has two aspects which are the most popular university (class 5) and the national university (class 4). As we use the simple topic model, the fluctuation occurs. If we use the topic model instead of the simple topic model, there would be no fluctuations like this. We can see that some other universities without high scores also fluctuate between two classes. For example, universities A and B fluctuate between class 5 and 3. This fluctuation could be interpreted as the universities have two aspects which are the most popular university (class 5) and the posh-traditional-fashionable university (class 3). Other fluctuations could be also explained similarly.

4 Comparison Between Male and Female Students

In this section, we will present the comparative study of the results of clustering the university ratings (brand image) data by male and female students separately. The previous section represented the results obtained for all the students together. In the section, we shall illustrate the results for male and female students separately. For all the students, we set the number of round in clustering as 50, here we set it as 60.This is because the convergence in this case is slower than the previous analysis with all the data. The number of classes is five. The classification results are shown in Table 2. The division/integration relationships among the three cases are shown in Figure 4. The thick arrow depicts the dominant class correspondences. The small change is expressed by thin arrows in the figure.

Table 2: Comparison among three cases

	total		male		female	
	classID	number	classID	number	classID	number
famous/club-activities	#1	10	#5	9	#2	8
specialization	#2	11	#3	7	#4	9
					#5	9
posh-traditional	#3	8	#4	14	#3	8
high-level-educaion	#4	8	#2	9	#1	16
most-popular-difficult	#5	13	#1	11		
		50		50		50

The most popular class (total class 5) is almost same in both the male and the female cases.

The male most popular class is class 1 and the female most popular class is class 1. In case of results for the female students' data, the following two division/integration are shown (See Table 2):

- Integration of the most-popular-difficult class and high-level-education class.
- Division of the specialization class to
  (1) sports science, music, and art universities, and
  (2) medical and pharmacology and English language universities.

The division of the specialization class in the female case indicates that female high school students might be more interested in sports, art and music related universities than male high school students. As shown in Figure 4, the two specialization classes by female students are not the division of the class 2 for the mixed data, but other universities from other classes join in these two classes; the female class 5 partly consists of the class 2 for the mixed data and the three new universities which did not appear in the case of mixed data.

The integration of the most-popular-difficult class and high-level-education class in case of female data can be interpreted as the appearance of the national university class as the combination of them. It can be interpreted as that the female students' primary criterion is the low tuition; the high-level education class are mainly national universities where the tuition is lower than the private universities.

Let us check the posh-traditional class in details. The number of the universities in the class is 8 (for mixed), 14 (for male) and 8 (for female). However, in the result of the male case, the ratio of women's universities decreases; they are 4/14=28.5% (male), 5/8=62.5% (total), and 7/8=87.5% (female). We can conclude that the male students are not interested in women's universities and that universities which they think as posh universities include many coeducational ones. In addition, the large number 14 for the male class 2 suggest that the male students are not so much interested in posh and fashionable universities and the class included other type of universities. On the other hand, the female students are likely to evaluate highly the posh universities and their posh class (female class 3) includes new fashionable universities that were not appeared in the mixed and the male results. The female high school students might put more importance on the fashionable attributes.

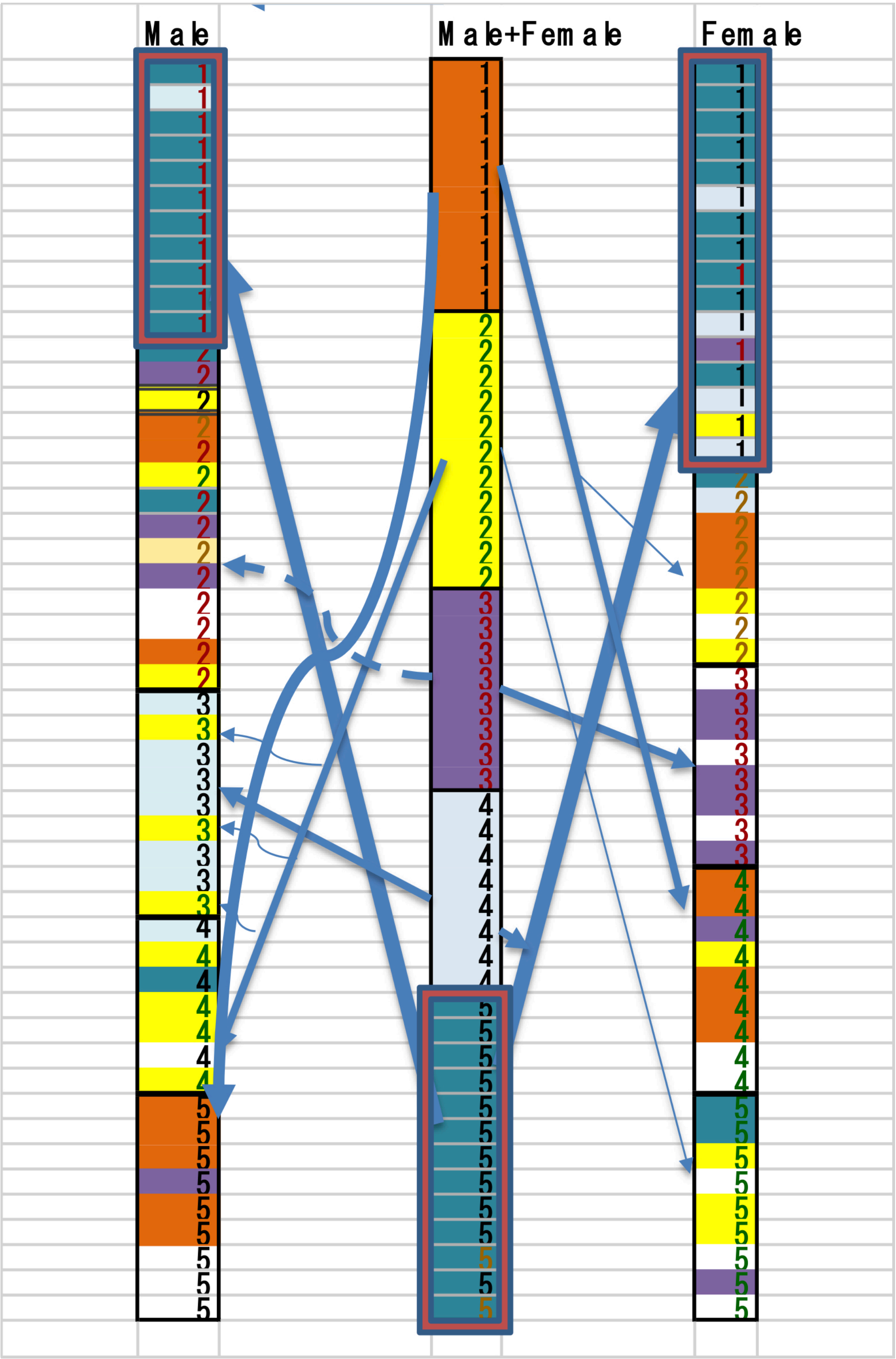

Figure 4: The division/integration relationships among the male, total, and female cases.

5 Related Work

In the section, we present a brief survey of the existing researches on visual reasoning of the topic model. Sato’s textbook on the topic model explains the mathematical processes carefully[19]. Regarding visualization of Gibbs sampling, the most famous book would be the illustration by David JC MacKay[20]. The illustration shows alternate updates of just two variable density function P(x1, x2). In the widely used text book by Bishop also, the same kind of illustration on P(x1, x2) which is a correlated Gaussian, is available[15, 21]. Computer graphics of these two variable density functions have been published such as one in the Mathematica online manual page titled “howto/PerformAMonteCarloSimulation.” However, we need to analyze multivariable cases consisting of more than three variables, because the two variable density function case does not express the intrinsic meaning of Gibbs sampling enough.

In the field of the topic model visualization, there are many visualizations[22]. However, they focus on relationships between a topic and word distribution. There is no visualization to explain Gibbs sampler on the topic model, focusing on the explanation of the mathematical process. On the other hand, our developed tool focuses on the visualization of the inherent mathematical process and it is useful to evaluate the invariant status.

7 Conclusions

This paper presents an application of Bayesian inference on the clustering of university brand images. Many university ranking data are available depending on different attributes of the university. The feature of the paper is that we used Bayesian inference. In the analysis, we used our developed visualizing tool. The tool conducts the MCMC simulation with Gibbs sampler and a simple topic model. The advantage of the visualization tool is interactive operation. The user can increase and decrease in turn the repetition times and see the status step by step with the visualization tool. The reverse operation can also be done.

Then, we can see the convergence behavior to the invariant Markov chain process and the member changing status among the classes. By using the students’ familiar university brand images, showing the invariant distribution Markov process enables their mathematical skills to grow up. In our classes, many students understood the MCMC behavior after they used the visual material. The visual material is available as a Wolfram CDF from our website (http://wwwcc.gakushuin.ac.jp/~20010570/mathABC/SELECTED/).

In the clustering, we selected the number of topics as five for the mixed data. We interpreted the class #5 as the most popular universities, the class #4 as the national university group, the class #3 as the posh and traditional university class, the class #2 as the specialization featured universities, and the class #1 as the famous universities. The almost word frequencies are collected on the most popular class #5. It is interesting that University of Tokyo shares characteristics of the two classes which are the most-popular class and the national university class; University of Tokyo fluctuates between the two classes. The alternate class change in the equilibrium status of the MCMC sequence shows the toggling of the University of Tokyo in the two classes.

Comparison of university choice between male and female high school students was conducted using the simple topic model. As the mixed (both included) data did, the most popular class was found in each clustering. The differences between male and female high school students are found in posh-traditional classes. In the case of male students analysis result, there is a class which seems to be a posh-traditional class. However, the male students are not interested in women's universities and that universities which they think as posh universities include many coeducational ones. In addition, the male students are not so much interested in posh and fashionable universities and the class included other type of universities. On the other hand, the female students are likely to evaluate highly the posh universities and their posh class (female class 3) includes new fashionable universities that were not appeared in the mixed and male data results. The female high school students seem to put more importance on the fashionable attributes.

In the paper, we presented the university brand image clustering using Bayesian inferences. We think that Bayesian inference approach is useful in the field of brand image evaluation among customers and more analysis should be conducted. As we conduct topic extraction using the topic model, we conducted analysis brand image clustering from the questionnaire data of customers. We would like to continue conducting university brand image analysis.

Akcknowledgement

We thank Prof Tetsuji Kubotyama (Computer Centre, Gakushuin University) for his wide range of knowledge about machine learning that helps our research. This research was partly supported by funds from Gakushuin University Computer Centre projects from 2015-2016. We sincerely express our gratitude to the Centre for its support.

References

[1] P. D. Hoff, *A First Course in Bayesian Statistical Methods*: Springer, 2010.

[2] M. D. Lee, and E.-J. Wagenmakers, *Bayesian Cognitive Modeling: A Practical Course*: Cambridge University Press, 2014.

[3] J. Kruschke, *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*: Academic Press, 2014.

[4] Y. Shirota, T. Hashimoto, and B. Chakraborty, "Visual Materials to Teach Gibbs Sampler," *International Journal of Knowledge Engineering,* vol. 2, pp. 92-95, 2016.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[6] D. Blei, and J. Lafferty, "Dynamic topic models," *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[7] D. M. Blei, "Probabilistic topic models," *Commun. ACM,* vol. 55, no. 4, pp. 77--84, 2013/07/01, 2012.

[8] T. Hashimoto, T. Kuboyama, and Y. Shirota, "Graph-based Consumer Behavior Analysis from Buzz Marketing Sites," *Proc. of 21st European Japanese Conference on Information Modelling and Knowledge Bases, Estonia, June 6-10*, 2011.

[9] Y. Shirota, T. Kuboyama, T. Hashimoto, S. Aramvith, and T. Chauksuvanit, *Study of Thailand People Reaction on SNS for the East Japan Great Earthquake - Comparion with Japanese People Reaction* -, p.^pp. 62: Research Institute for Oriental Cultures Gakushuin University, 2015.

[10] Y. Shirota, T. Hashimoto, and S. Tamaki, "MONETARY POLICY TOPIC EXTRACTION BY USING LDA －JAPANESE MONETARY POLICY OF THE SECOND ABE CABINET TERM －," *Proc. of IIAI International Congress on Advanced Applied Informatics 2015, 12-16 July, 2015, Okayama, Japan*, pp. 8-13, 2015.

[11] Y. Shirota, T. Hashimoto, and T. Sakura, "Topic Extraction Analysis for Monetary Policy Minutes of Japan in 2014," *Advances in Data Mining: Applications and Theoretical Aspects*, Lecture Notes in Computer Science P. Perner, ed., pp. 141-152: Springer International Publishing, 2015.

[12] T. Hashimoto, and Y. Shirota, "Framework of an Advisory Message Board for Women Victims of the East Japan Earthquake Disaster," *Prof. of JADH2013 (Japanese Association for Digital Humanities), Sept 19-21, Kyoto*, pp. 31-32, 2013.

[13] Y. Shirota, and B. Chakraborty, "Visual Explanation of Mathematics in Latent Semantic Analysis," *Proc. of IIAI International Congress on Advanced Applied Informatics 2015, 12-16 July, 2015, Okayama, Japan*, pp. 423-428, 2015.

[14] Recruit Co., "Survey of University Brand Power 2015," *Recruit Co. College Management,* vol. 194 / Sep. - Oct. 2015, pp. 6-44, 2015 (written in Japanese).

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*: Springer, 2006.

[16] D. Koller, and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*: The MIT Press, 2009.

[17] Y. Iba, *Bayesian Statistics and Statistical Physics*: Iwanami-shoten, 2003, (written in Japanese).

[18] I. Sato, *Statistical Latent Semantic Analysis Based on Topic Model*: Corona Publishing Co., 2015, (written in Japanese).

[19] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*: Cambridge university press, 2003.

[20] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, third edition ed.: Chapman & Hall/CRC, 2004.

[21] A. J.-B. Chaney, and D. M. Blei, "Visualizing Topic Models," *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012.*