A Classification of a Scene in a Field note Using Topic Model

Yamada, T.
Historiographical Institute, the University of Tokyo, 3-1, Hongo 7, Bunkyo-ku, Tokyo, Japan
E-mail: t_yamada@hi.u-tokyo.ac.jp

Abstract

In this study, we propose a method that can represent and analyze the spatial features of scenes in historical materials (field notes) related to area studies. To promote area studies research, we introduce a method to construct text database of area research resources using semantic web technologies. To improve accessibility and deepen the understanding of an area using a field note, we also introduce Latent Dirichlet Allocation (LDA) method. We constructed a text database using a field note written by Yoshikazu Takaya, a prominent researcher in Southeast Asian area studies. We show an experimental result on detected 30 topics from the constructed database. In this paper, we inspect the detection results and describe the advantages of the proposed method.

1. Introduction

Recently, area studies have seen remarkable progress because researchers can search and analyze large volumes of data easily and quickly using information technology, such as web technology, data analysis, and data engineering. To promote such analyses, researchers have published various databases related to area studies, such as catalogs, images, statistical data, and spatial and temporal data. For example, the Center for Integrated Area Studies, Kyoto University (CIAS)[1] published 42 databases related to area studies, and an overview of these databases has been published (Tanigawa and Yamamoto, 2013). These databases primarily comprise catalogs of books and historical materials, photographs, movies and sounds related to a landscape and an event in a given area, and statistical data of an area's feature. However, databases of the text of books and historical materials related to an area are not available. We believe that text data are an essential area studies resource. For example, field note text can include descriptions of sights, scenes, and customs, as well as latent topics or subjects that can be key elements to characterize an area.

In this study, we propose a method to construct a database of area studies text resources. Field notes are an important text resource because they can include valuable information; however, field notes are rarely shared. To improve accessibility and improve the understanding of an area based on field notes, we propose text analysis and topic detection methods. We prepared a field note database in which the data unit is a description of a sight or a scene. We used latent Dirichlet allocation (LDA) to detect latent topics. In LDA, each text can be considered a mixture of various (latent) topics, and each topic can be considered a mixture of various words. The remainder of this paper is organized as follows. The features and structure of the field notes used in our analysis are described in Section 2. Section 3 describes the construction of the field note text database and the workflow of our text analysis method. The results of the field note text analysis are presented in Section 4 and discussed in Section 5. The conclusions are given in Section 6.

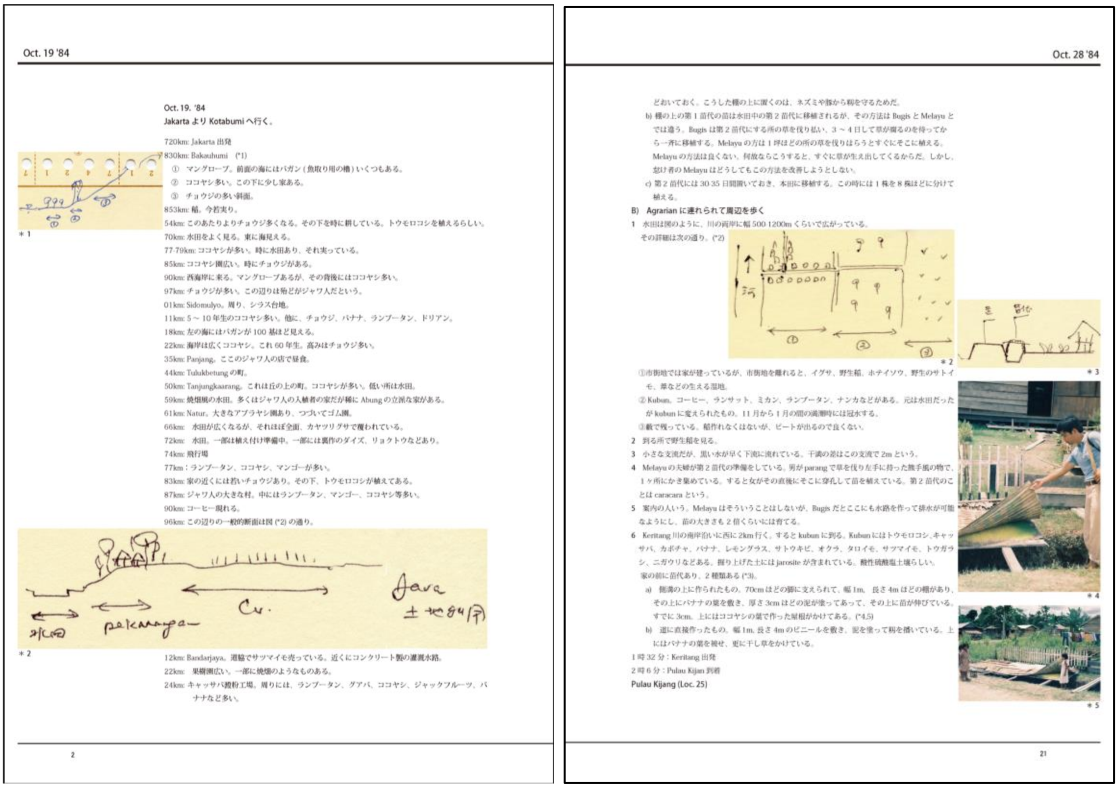
2. Field Notes

We used "the field note collection 2 Sumatra," which is part of the "Area Studies Archives: Assembled Field notes" (Takaya, 2012) authored by Yoshikazu Takaya, a prominent Southeast Asian area studies researcher. The field notes are from a field survey conducted from October 19, 1984 to January 18, 1985 on Sumatra. Figure 1 shows sample pages from Takaya's field notes. The field notes consist of text, sketches, and photographs of each visited area. Note that the original field notes have been edited. The edited field notes include text transcribed from the original field notes. The edited field notes comprise 165,757 characters (197 pages).

3. Text structure and Text Analysis

In this section, we describe the construction of the text database and the text analysis method. An overview of the workflow is shown in Figure 3.

[1] In January 2017, CIAS changed to Center for Southeast Asian Studies (CSEAS)



Oct. 19 '84

Oct. 19. '84
Jakarta より Kotabumi へ行く。

720km: Jakarta 出発
830km: Bakauhumi (*1)
① マングローブ。前面の海にはバガン（魚取り用の櫓）いくつもある。
② ココヤシ多い。この下に少し家ある。
③ チョウジの多い斜面。
853km: 稲。今若実り。
54km: このあたりよりチョウジ多くなる。その下を時に耕している。トウモロコシを植えるらしい。
70km: 水田をよく見る。東に海見える。
77-79km: ココヤシが多い。時に水田あり、それ実っている。
85km: ココヤシ園広い。時にチョウジがある。
90km: 西海岸に来る。マングローブあるが、その背後にはココヤシ多い。
97km: チョウジが多い。この辺りは殆どがジャワ人だという。
01km: Sidomulyo。周り、シラス台地。
11km: 5～10 年生のココヤシ多い。他に、チョウジ、バナナ、ランブータン、ドリアン。
18km: 左の海にはバガンが 100 基ほど見える。
22km: 海岸は広くココヤシ。これ 60 年生。高みはチョウジ多い。
35km: Panjang。ここのジャワ人の店で昼食。
44km: Tulukbetung の町。
50km: Tanjungkaarang。これは丘の上の町。ココヤシが多い。低い所は水田。
59km: 焼畑風の水田。多くはジャワ人の入植者の家だが稀に Abung の立派な家がある。
61km: Natur。大きなアブラヤシ園あり、つづいてゴム園。
66km: 水田が広くなるが、それほぼ全面、カヤツリグサで覆われている。
72km: 水田。一部は植え付け準備中。一部には裏作のダイズ、リョクトウなどあり。
74km: 飛行場
77km：ランブータン、ココヤシ、マンゴーが多い。
83km: 家の近くには若いチョウジあり。その下、トウモロコシが植えてある。
87km: ジャワ人の大きな村。中にはランブータン、マンゴー、ココヤシ等多い。
90km: コーヒー現れる。
96km: この辺りの一般的断面は図 (*2) の通り。

＊1

＊2

12km: Bandarjaya。道脇でサツマイモ売っている。近くにコンクリート製の灌漑水路。
22km: 果樹園広い。一部に焼畑のようなものある。
24km: キャッサバ澱粉工場。周りには、ランブータン、グアバ、ココヤシ、ジャックフルーツ、バナナなど多い。

2

Oct. 28 '84

どおいておく。こうした櫓の上に置くのは、ネズミや豚から籾を守るためだ。
b) 櫓の上の第 1 苗代の苗は水田中の第 2 苗代に移植されるが、その方法は Bugis と Melayu とでは違う。Bugis は第 2 苗代にする所の草を伐り払い、3～4 日して草が腐るのを待ってから一斉に移植する。Melayu の方は 1 坪ほどの所の草を伐りはらうとすぐにそこに植える。Melayu の方法は良くない。何故ならこうすると、すぐに草が生え出してくるからだ。しかし、怠け者の Melayu はどうしてもこの方法を改善しようとしない。
c) 第 2 苗代には 30-35 日間置いておき、本田に移植する。この時には 1 株を 8 株ほどに分けて植える。

B) Agrarian に連れられて周辺を歩く
1 水田は図のように、川の両岸に幅 500-1200m くらいで広がっている。
その詳細は次の通り。(*2)

＊2

＊3

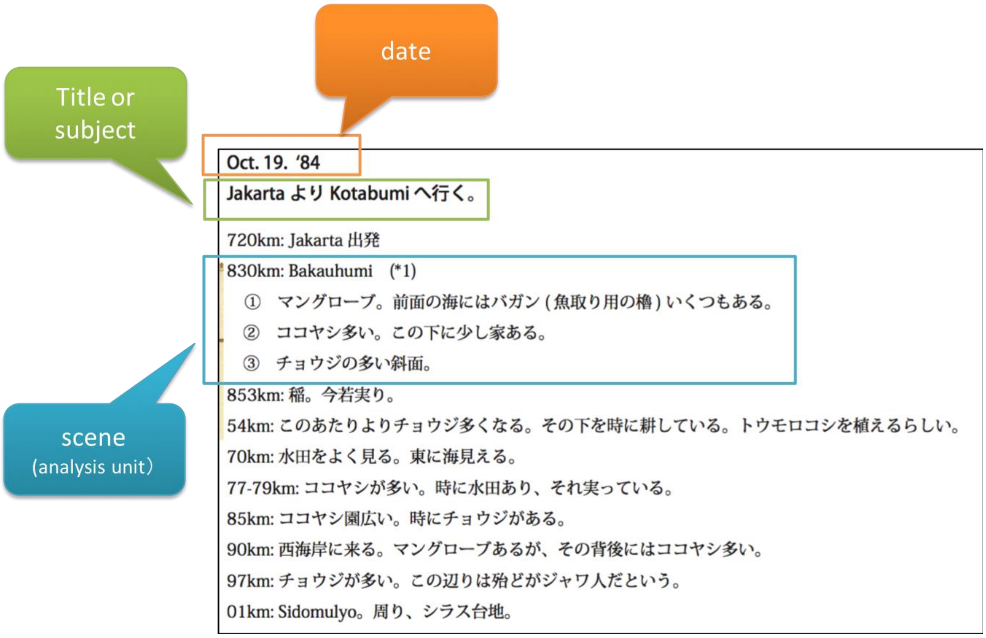
①市街地では家が建っているが、市街地を離れると、イグサ、野生稲、ホテイソウ、野生のサトイモ、葦などの生える湿地。
② Kubun。コーヒー、ランサット、ミカン、ランブータン、ナンカなどがある。元は水田だったが kubun に変えられたもの。11 月から 1 月の間の満潮時には冠水する。
③藪で残っている。稲作れなくはないが、ピートが出るので良くない。
2 到る所で野生稲を見る。
3 小さな支流だが、黒い水が早く下流に流れている。干満の差はこの支流で 2m という。
4 Melayu の夫婦が第 2 苗代の準備をしている。男が parang で草を伐り左手に持った熊手風の物で、1ヶ所にかき集めている。すると女がその直後にそこに穿孔して苗を植えている。第 2 苗代のことは caracara という。
5 案内の人いう。Melayu はそういうことはしないが、Bugis だとここにも水路を作って排水が可能なようにし、苗の大きさも 2 倍くらいには育てる。
6 Keritang 川の南岸沿いに西に 2km 行く。すると kubun に到る。Kubun にはトウモロコシ、キャッサバ、カボチャ、バナナ、レモングラス、サトウキビ、オクラ、タロイモ、サツマイモ、トウガラシ、ニガウリなどある。掘り上げた土には jarosite が含まれている。酸性硫酸塩土壌らしい。
家の前に苗代あり、2 種類ある (*3)。
a) 側溝の上に作られたもの。70cm ほどの脚に支えられて、幅 1m、長さ 4m ほどの棚があり、その上にバナナの葉を敷き、厚さ 3cm ほどの泥が塗ってあって、その上に苗が伸びている。すでに 3cm。上にはココヤシの葉で作った屋根がかけてある。(*4,5)
b) 道に直接作ったもの。幅 1m、長さ 4m のビニールを敷き、泥を塗って籾を播いている。上にはバナナの葉を被せ、更に干し草をかけている。
1 時 32 分：Keritang 出発
2 時 6 分：Pulau Kijan 到着
Pulau Kijang (Loc. 25)

＊4

＊5

21

Figure 1: Example field notes (left: p.2; right: p.21)



Oct. 19. '84
Jakarta より Kotabumi へ行く。

720km: Jakarta 出発
830km: Bakauhumi (*1)
① マングローブ。前面の海にはバガン（魚取り用の櫓）いくつもある。
② ココヤシ多い。この下に少し家ある。
③ チョウジの多い斜面。
853km: 稲。今若実り。
54km: このあたりよりチョウジ多くなる。その下を時に耕している。トウモロコシを植えるらしい。
70km: 水田をよく見る。東に海見える。
77-79km: ココヤシが多い。時に水田あり、それ実っている。
85km: ココヤシ園広い。時にチョウジがある。
90km: 西海岸に来る。マングローブあるが、その背後にはココヤシ多い。
97km: チョウジが多い。この辺りは殆どがジャワ人だという。
01km: Sidomulyo。周り、シラス台地。

Figure 2: Extraction of metadata from a field note

3.1 *Scene*

The structure of the field notes is shown in Figure 2. In each field note, the investigation of a scene or sight is described in chronological order. The field note includes a title or subject (green rectangle) and a date (orange rectangle). The description of each investigation is separated into a scene (blue rectangle). Note that the description of the scene includes distance data (e.g., 720 km), which is the distance value of the trip meter given by a vehicle's odometer. The remainder of the field note describes the scene investigated on the given date.

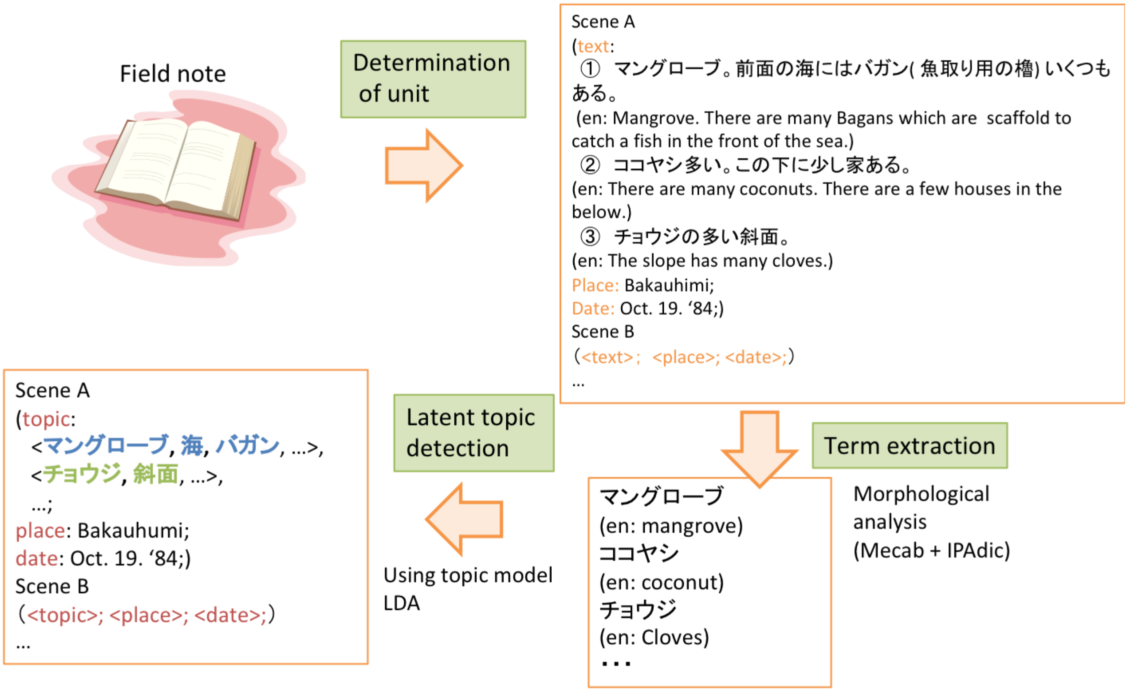
Thus, we used “scene” as the data unit of the text database. If a place name is included in the description of a scene, we can know the location. Determining the exact position is difficult because the latitude and longitude values are not given; however, the location can be approximated. If a place name is not included, the given scene is considered to be located between the previous and next scenes. Thus, the scene data comprise text, place, and date information.

3.2 Topic Model

To characterize a scene, we done word segmentation and part-of-speech tagging (POS tagging), and, based on the results, we created a bag-of-words. Note that we used MeCab[2], a well-known Japanese word segmentation library, and IPADic[3], one of dictionaries available for MeCab. We targeted nouns and adjectives for extraction. Note that pronouns, suffixes, adverbs, adjective stems, conjunctional, and nonautonomous were excluded. In addition, we chunked consecutive nouns and suffixes that occurred immediately after the extracted nouns and “[a-zA-Z]+” sequences. Four nouns (“ランブータン (Rambutan),” “ジャックフルーツ (Jack fruit),” “キャッサバ (cassava),” and “サゴヤシ (sago palm)”) failed in the word segmentation and the POS tagging even though they occur frequently in the field notes. Therefore, they were included in the MeCab dictionary. To express extracted terms and their occurrence frequency, we output the result as bag-of-words[4].

3.3 Topic Model

We used LDA (Blei et al., 2003) to detect topics in the field notes. LDA treats a set of terms subject to statistical co-occurrence as latent topics. LDA assumes there are multiple topics in a given scene and models the distributions of these topics. Figure 4 shows a graphical representation of LDA, where the blue disk indicates the observation variable and the white disks indicate unknown variables.



Scene A
(text:
① マングローブ。前面の海にはバガン(魚取り用の櫓) いくつもある。
(en: Mangrove. There are many Bagans which are scaffold to catch a fish in the front of the sea.)
② ココヤシ多い。この下に少し家ある。
(en: There are many coconuts. There are a few houses in the below.)
③ チョウジの多い斜面。
(en: The slope has many cloves.)
Place: Bakauhimi;
Date: Oct. 19. ‘84;)
Scene B
（<text>；<place>; <date>;）
…

マングローブ
(en: mangrove)
ココヤシ
(en: coconut)
チョウジ
(en: Cloves)
…

Scene A
(topic:
<マングローブ, 海, バガン, …>,
<チョウジ, 斜面, …>,
…;
place: Bakauhumi;
date: Oct. 19. ‘84;)
Scene B
（<topic>; <place>; <date>;）
…

Figure 3: Text database creation and text analysis workflow

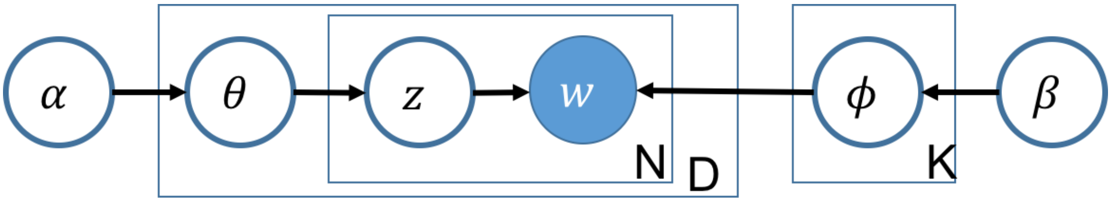


Figure 4: Graphical model of LDA

[2] http://taku910.github.io/mecab/
[3] https://github.com/neologd/mecab-ipadic-neologd
[4] https://en.wikipedia.org/wiki/Bag-of-words_model

The rectangles indicate iterative processes, and the numbers in the lower right of the rectangles (N, D and K) indicate the number of times the process was repeated. Here w, i.e., the only observed variable, denotes the term extraction result (Section 3.1), z denotes topics, θ is the topic distribution for the scenes, and ϕ is the term distribution in all topics. In addition, α and β are LDA hyperparameters. When the number of scenes is D and the number of topics is K, θ_d (i.e., the topic distribution of scene $d \in D$) and ϕ_k (i.e., the term distribution of topic $k \in K$) are generated as follows:

$$\theta_d \sim Dir(\alpha)\ (d = 1, \dots, D),$$
$$\phi_k \sim Dir(\beta)\ (d = 1, \dots, K)$$

Equation 1

Here, $Dir\ (\cdot)$ represents the Dirichlet distribution. The topic $z_{d,i}$ is generated as follows:

$$z_{d,i} \sim Multi\left(\theta_{z_{d,i}}\right)\ (i = 1, \dots, N_d)$$

Equation 2

Here, $Multi(\cdot)$ is the multinomial distribution and N_d is the number of terms in scene d. In addition, $w_{d,i}$ can be generated as follows:

$$w_{d,i} \sim Multi\left(\phi_{z_{d,i}}\right)$$

Equation 3

Note that the predictive distribution of LDA cannot be calculated analytically; thus, an approximation algorithm that can efficiently calculate the posterior distribution of LDA should be introduced. The variational Bayesian (VB) (Blei et al., 2003), collapsed Gibbs sampling (CGS) (Griffiths, 2004), and collapsed VB (CVB) (The et al., 2006b) methods are well-known inference methods for LDA. In this study, we used CGS because it realizes direct sampling of $z_{d,i}$ in formula (2) by marginalizing θ and ϕ. CGS also has the following advantages:

- Because sampling of θ_d and ϕ_k in formula (1) is not required, the implementation of CGS is fairly simple than that of the VB method.
- Learning in CGS requires a significant amount of iterative processing; however, the calculation cost per calculation can be reduced considerably compared with the VB and CVB methods.
- The prediction performance of CGS is comparable to that of the CVB method and is better than the VB method (Asuncion et al., 2009).

Figure 5 shows the CGS procedure used in this study.

```
Initialize $\alpha$ and $\beta$
Initialize $z$
Set $S$ : the number of sampling
for s = 1, …, $S$ do
    for $d$ = 1, … , $D$ do
        for $i$ = 1, … , $N_d$ do
            Sample $z_{d,i}$
            Update $N_{d,z_{d,i}}$
        end for
    end for
    Update $\alpha$ and $\beta$
end for
```

Figure 5: Collapsed Gibbs sampling procedure

In this process, $N_{d,j}$ is the number of terms assigned to topic j in scene d. $z_{d,i}$ can be sampled as follows:

$$z_{d,i} \sim Multi\left(p\left(z_{d,i} \middle| W, Z_{\backslash d,i}\right)\right)$$
$$\propto\ \left(N_{d,i} + \alpha\right)\frac{N_{k,w_{d,i}} + \beta}{N_k + \beta V}.$$

Equation 4

Here, W represents the terms in all scenes, V represents the number term types in all scenes, and α and β are parameters in the Dirichlet distribution. Wallach (2009) reported that the performance of LDA can be improved when the value of α is not uniform ($\alpha_k \neq \alpha_l, k \neq l$) and the value of β is uniform ($\beta_1 = \beta_2 = \cdots = \beta_K$). Note that we set these hyperparameters according to the literature (Wallach et al., 2009) as follows:

$$\alpha_k^{new} = \alpha_k \frac{\sum_D \Psi\left(N_{d,k} + \alpha_k\right) - D\Psi(\alpha_k)}{\sum_d \Psi(N_d + \sum_{k'} \alpha_{k'}) - D\Psi(\sum_{k'} \alpha_{k'})}$$

Equation 5

$$\beta^{new} = \beta \frac{\sum_k \sum_V \Psi\left(N_{k,v} + \alpha_k\right) - D\Psi(\alpha_k)}{V \sum_k \Psi(N_k + \beta V) - KV\Psi(\beta V)}$$

Equation 6

Here, $\Psi(\cdot)$ is a digamma function.

3.4 Scene Representation

We represent a scene using the Resource Description Framework (RDF) data model (W3C, 1999), and the data are stored in our database. The RDF is a general method for conceptual descriptions in web resources. Currently, the RDF is an important semantic web (W3C, 2015) technology represented by Linked Open Data (LOD) (Berners-Lee, 2006). In the RDF data model, a resource relationship is represented by a statement about the resource in an expression of the form subject-predicate-object (a triple). Here, the "subject" denotes the resource, the "object" denotes a value which is related to the "subject", and the "predicate" denotes the traits or aspects of the resource and expresses a relationship between the "subject" and "object."

Figure 6 shows an example RDF graph of a scene, where red arrows indicate a "predicate." For the data representation, we introduced vocabulary sets from the Dublin Core Element Set (DC) (DCMI, 2012), where the DC prefixes are "dc" and "dcterms." In Figure 6, the subject of the scene is represented as "dc:title," the date is represented as "dcterms:tempral," the place name is represented as "dcterms:spatial," and the description is represented as "dc:description." The "subject" of the scene data is indicated by the blue ellipse in the upper left and can be represented using a Uniform Resource Identifier. For the scene representation, we also introduced an original vocabulary set whose prefix is "fn." With the "fn" vocabulary set, we can represent the identifier of the description (as "fn:descId") and detected latent topics (as "fn:topicClass" and "fn:term"). Here, "fn:topicClass" indicates the topic number classified by LDA, and "fn:term" indicates the term by term extraction method (Section 3.2).

4. Experiment

4.1 Experimental Setup

We obtained the terms from the field note data using the term extraction method described in Section 3.1. Here, the number of terms was 19,287 among 5,666 term types. We assumed that the number of topics in the field note was 30. We were able to detect topics using CGS (Section 3.3). Note that determining the sampling frequency in CGS is difficult.

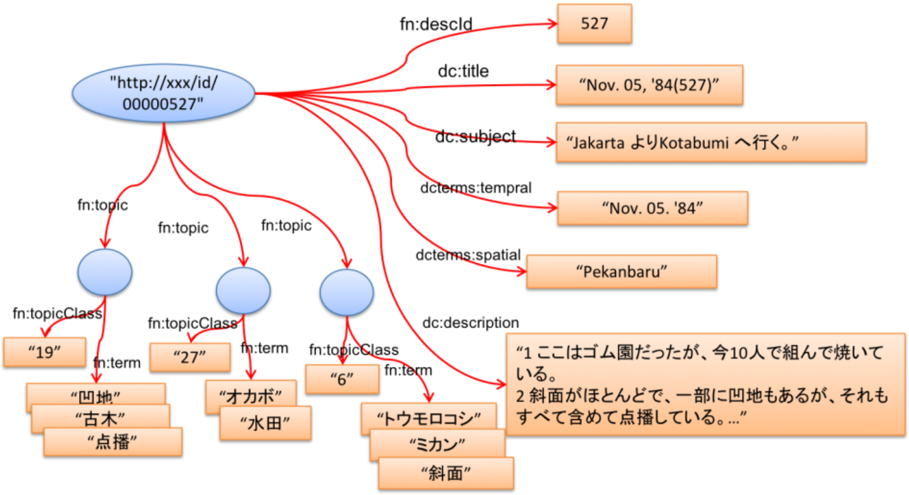


Figure 6: Scene data by RDF graph

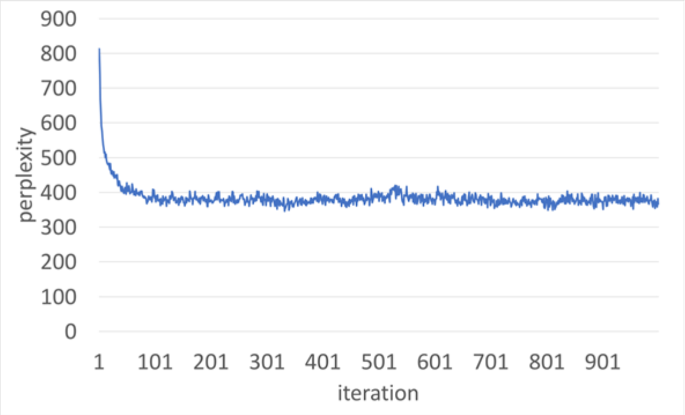


Figure 7: Procedure of collapsed Gibbs sampling

To evaluate the performance of LDA, we conducted a preliminary experiment to calculate the perplexity. Test data is needed to calculate the perplexity, however, there is no text of the same place and the same time written by the same author. Therefore, we treated one scene data removed from learning data as test data and could calculate perplexity of the situation. The calculation was applied to all scenes, and the average was taken as the perplexity in the iteration. The results are shown in Figure 7. As can be seen, perplexity becomes stable at approximately 100 sampling iterations; however, tremor behavior continued until 800 sampling times. Thus, we decided to use an average of 900 to 1,000 sampling times as the topic detection parameters.

4. 2 Scene Topics

Figure 8 shows various terms and their corresponding frequency in ascending order (top 10) for each topic detected from all scenes in the field note. We can understand the feature for each topic by terms belonged the topic. For example, Topic 7 has a list of terms that includes "水田 (paddy field)," "池 (pond)," "魚池 (fish pong)", "集落 (village)," etc. Topic 5 includes "オカボ (upland rice)," "多い (many)," "トウモロコシ (corn)," "コーヒー (coffee)," etc. We can easily understand the differences between these topics by comparing these term lists. However, we could not easily grasp topic differences in some cases. Here, there are two main reasons for this: (1) the meaning of topics could not be ascertained and (2) certain terms appear across multiple topics. The former case (e.g., Topics 2, 22, and 29) occurs when it is very difficult to understand the topics. LDA is a very simple algorithm for detecting topics by term co-occurrences; thus, it is not always possible to ascertain the meaning of term co-occurrences.

To demonstrate the second reason, we focus on some topics related to "水田." For example, Topics 1, 9, 25, and 30 include "水田" as a term; however, the features of these topics are quite different because the co-occurring terms of each topic differ. In other words, the difference in features indicates that the meaning differs depending on the situation. However, understanding the differences of topics is difficult when the co-occurrences of a term increase in the topics. For example, Topics 24 and 30 both contain the terms "ゴム (rubber)[5] and "ゴム園 (rubber field)," etc. (some terms are not shown in Figure 8). As a result, it is difficult to clarify the differences between these two topics. Therefore, we performed hierarchical clustering for the topics and visualized the result as a dendrogram (Amorim, 2015). We used the weight $weight(t_{i,k})$ of term k in topic i as the element of the feature vector of the topic, and calculated using Ward's criterion as the linkage criterion between topic clusters. Here, $weight(t_{i,k})$ represents the weight of term k in topic i, and we used the frequency of the term as the value. The results are shown in Figure 9. According to the results, we can estimate the following:

- Topics 24 and 30 may relate to rubber (tree), a rubber field, and a field around a rubber field.
- Topics 5, 27, 8, and 12 may relate to the situation and use of farmland around a village, where Topics 8 and 12 may relate to sight of waterside relative to Topics 5 and 27.
- Topics 14 and 28 may relate to the state of a town.
- Topics 11, 18, and 17 relate to land use (including coco and sago palm) on the waterside.
- Topics 1 and 25 may relate to the appearance of agricultural work.
- Topics 7, 26, and 9 may relate to a wetland and paddy field.

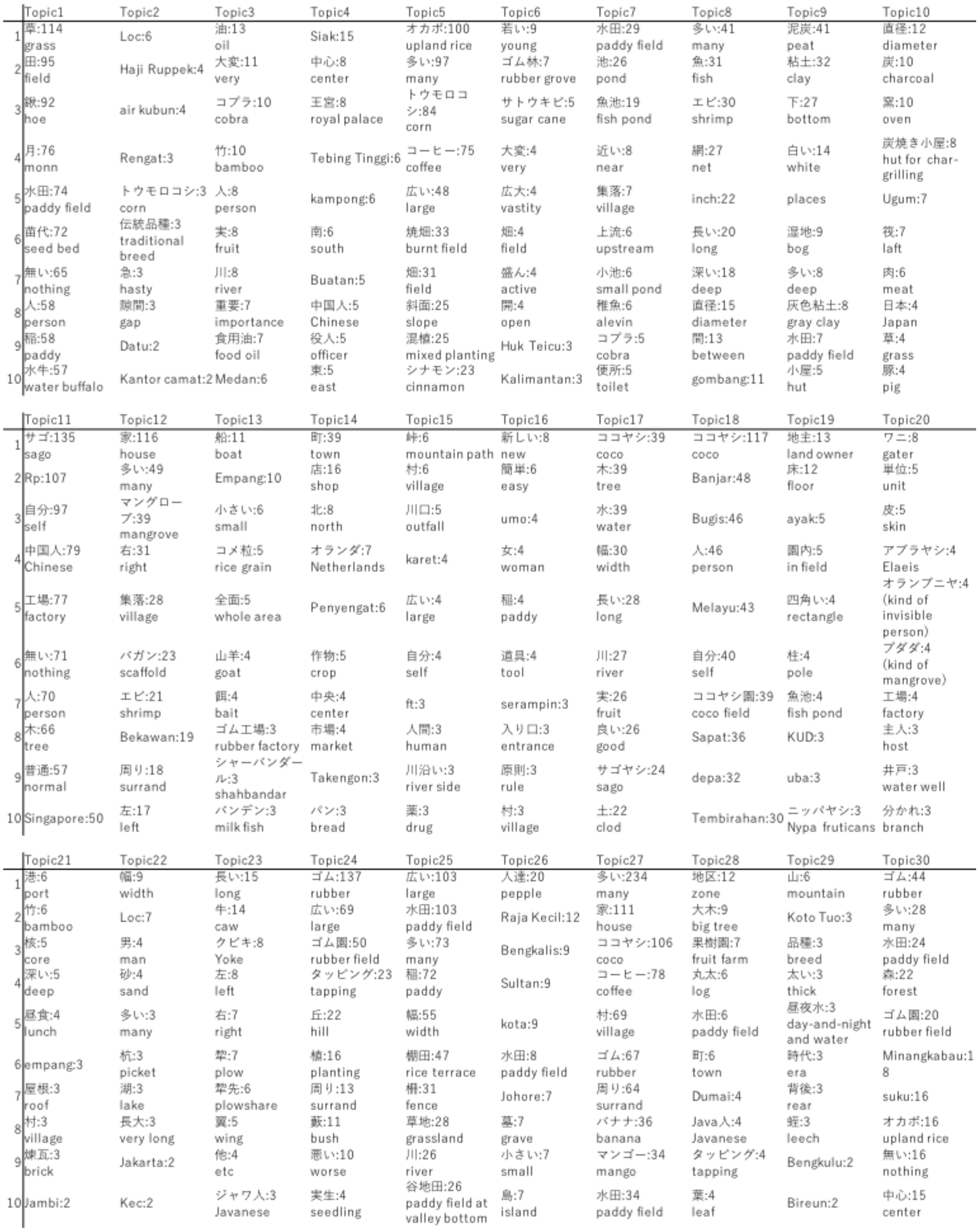
A serious analysis of topic detection will appear as a research result of area studies.

4.3 Topics of Terms

Figure 10 shows the variance of topics in the top 15 frequent terms and their total number. From the results shown in Figure 8, we can grasp the features of each topic using the terms detected as the topic. Figure 10 shows an assigned topic of a term for all scenes. According to the results, we can ascertain the following:

- Almost of term "サゴ (sago palm)" were detected as Topic 11, which may be related to waterside land use.
- The term "家 (house)" is assigned to Topics 12 and 27. There are some terms assigned to Topic 12, such as "多い (many)" and "周り (around)." Some terms, such as "多い (many)," "ココヤシ (coco palm)," "水田 (paddy field)," "ゴム (rubber)," "コーヒー (coffee)," and "周り (around)" are assigned to Topic 27. The detection results indicate that the meaning of "家 (house)" changes depending on the given scene.
- "水田 (paddy field)" is assigned to various topics, such as Topics 1, 7, 25, 27, and 30. We infer that a paddy field is an important item that can characterize a landscape or scene on Sumatra.

[5] The “rubber” means rubber as an agricultural crop.



	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
1	草:114 grass	Loc:6	油:13 oil	Siak:15	オカボ:100 upland rice	若い:9 young	水田:29 paddy field	多い:41 many	泥炭:41 peat	直径:12 diameter
2	田:95 field	Haji Ruppek:4	大変:11 very	中心:8 center	多い:97 many	ゴム林:7 rubber grove	池:26 pond	魚:31 fish	粘土:32 clay	炭:10 charcoal
3	鍬:92 hoe	air kubun:4	コブラ:10 cobra	王宮:8 royal palace	トウモロコシ:84 corn	サトウキビ:5 sugar cane	魚池:19 fish pond	エビ:30 shrimp	下:27 bottom	窯:10 oven
4	月:76 monn	Rengat:3	竹:10 bamboo	Tebing Tinggi:6	コーヒー:75 coffee	大変:4 very	近い:8 near	網:27 net	白い:14 white	炭焼き小屋:8 hut for char-grilling
5	水田:74 paddy field	トウモロコシ:3 corn	人:8 person	kampong:6	広い:48 large	広大:4 vastity	集落:7 village	inch:22	places	Ugum:7
6	苗代:72 seed bed	伝統品種:3 traditional breed	実:8 fruit	南:6 south	焼畑:33 burnt field	畑:4 field	上流:6 upstream	長い:20 long	湿地:9 bog	筏:7 laft
7	無い:65 nothing	急:3 hasty	川:8 river	Buatan:5	畑:31 field	盛ん:4 active	小池:6 small pond	深い:18 deep	多い:8 deep	肉:6 meat
8	人:58 person	隙間:3 gap	重要:7 importance	中国人:5 Chinese	斜面:25 slope	開:4 open	稚魚:6 alevin	直径:15 diameter	灰色粘土:8 gray clay	日本:4 Japan
9	稲:58 paddy	Datu:2	食用油:7 food oil	役人:5 officer	混植:25 mixed planting	Huk Teicu:3	コブラ:5 cobra	間:13 between	水田:7 paddy field	草:4 grass
10	水牛:57 water buffalo	Kantor camat:2	Medan:6	東:5 east	シナモン:23 cinnamon	Kalimantan:3	便所:5 toilet	gombang:11	小屋:5 hut	豚:4 pig

	Topic11	Topic12	Topic13	Topic14	Topic15	Topic16	Topic17	Topic18	Topic19	Topic20
1	サゴ:135 sago	家:116 house	船:11 boat	町:39 town	峠:6 mountain path	新しい:8 new	ココヤシ:39 coco	ココヤシ:117 coco	地主:13 land owner	ワニ:8 gater
2	Rp:107	多い:49 many	Empang:10	店:16 shop	村:6 village	簡単:6 easy	木:39 tree	Banjar:48	床:12 floor	単位:5 unit
3	自分:97 self	マングローブ:39 mangrove	小さい:6 small	北:8 north	川口:5 outfall	umo:4	水:39 water	Bugis:46	ayak:5	皮:5 skin
4	中国人:79 Chinese	右:31 right	コメ粒:5 rice grain	オランダ:7 Netherlands	karet:4	女:4 woman	幅:30 width	人:46 person	園内:5 in field	アブラヤシ:4 Elaeis
5	工場:77 factory	集落:28 village	全面:5 whole area	Penyengat:6	広い:4 large	稲:4 paddy	長い:28 long	Melayu:43	四角い:4 rectangle	オランプニヤ:4 (kind of invisible person)
6	無い:71 nothing	バガン:23 scaffold	山羊:4 goat	作物:5 crop	自分:4 self	道具:4 tool	川:27 river	自分:40 self	柱:4 pole	ブダダ:4 (kind of mangrove)
7	人:70 person	エビ:21 shrimp	餌:4 bait	中央:4 center	ft:3	serampin:3	実:26 fruit	ココヤシ園:39 coco field	魚池:4 fish pond	工場:4 factory
8	木:66 tree	Bekawan:19	ゴム工場:3 rubber factory	市場:4 market	人間:3 human	入り口:3 entrance	良い:26 good	Sapat:36	KUD:3	主人:3 host
9	普通:57 normal	周り:18 surrand	シャーバンダール:3 shahbandar	Takengon:3	川沿い:3 river side	原則:3 rule	サゴヤシ:24 sago	depa:32	uba:3	井戸:3 water well
10	Singapore:50	左:17 left	バンデン:3 milk fish	パン:3 bread	薬:3 drug	村:3 village	土:22 clod	Tembirahan:30	ニッパヤシ:3 Nypa fruticans	分かれ:3 branch

	Topic21	Topic22	Topic23	Topic24	Topic25	Topic26	Topic27	Topic28	Topic29	Topic30
1	港:6 port	幅:9 width	長い:15 long	ゴム:137 rubber	広い:103 large	人達:20 pepple	多い:234 many	地区:12 zone	山:6 mountain	ゴム:44 rubber
2	竹:6 bamboo	Loc:7	牛:14 caw	広い:69 large	水田:103 paddy field	Raja Kecil:12	家:111 house	大木:9 big tree	Koto Tuo:3	多い:28 many
3	核:5 core	男:4 man	クビキ:8 Yoke	ゴム園:50 rubber field	多い:73 many	Bengkalis:9	ココヤシ:106 coco	果樹園:7 fruit farm	品種:3 breed	水田:24 paddy field
4	深い:5 deep	砂:4 sand	左:8 left	タッピング:23 tapping	稲:72 paddy	Sultan:9	コーヒー:78 coffee	丸太:6 log	太い:3 thick	森:22 forest
5	昼食:4 lunch	多い:3 many	右:7 right	丘:22 hill	幅:55 width	kota:9	村:69 village	水田:6 paddy field	昼夜水:3 day-and-night and water	ゴム園:20 rubber field
6	empang:3	杭:3 picket	犂:7 plow	植:16 planting	棚田:47 rice terrace	水田:8 paddy field	ゴム:67 rubber	町:6 town	時代:3 era	Minangkabau:18
7	屋根:3 roof	湖:3 lake	犂先:6 plowshare	周り:13 surrand	柵:31 fence	Johore:7	周り:64 surrand	Dumai:4	背後:3 rear	suku:16
8	村:3 village	長大:3 very long	翼:5 wing	藪:11 bush	草地:28 grassland	墓:7 grave	バナナ:36 banana	Java人:4 Javanese	蛭:3 leech	オカボ:16 upland rice
9	煉瓦:3 brick	Jakarta:2	他:4 etc	悪い:10 worse	川:26 river	小さい:7 small	マンゴー:34 mango	タッピング:4 tapping	Bengkulu:2	無い:16 nothing
10	Jambi:2	Kec:2	ジャワ人:3 Javanese	実生:4 seedling	谷地田:26 paddy field at valley bottom	島:7 island	水田:34 paddy field	葉:4 leaf	Bireun:2	中心:15 center

Figure 8: Topic detection results

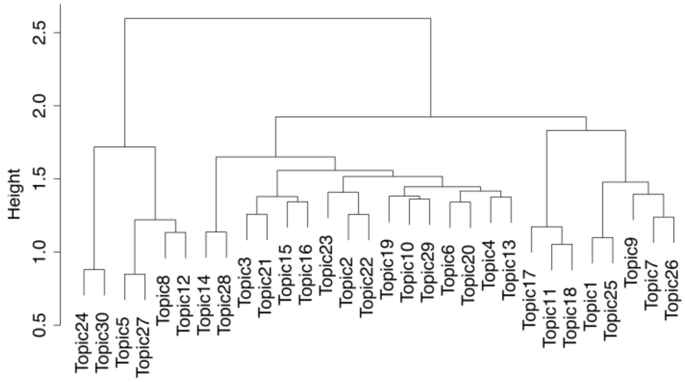
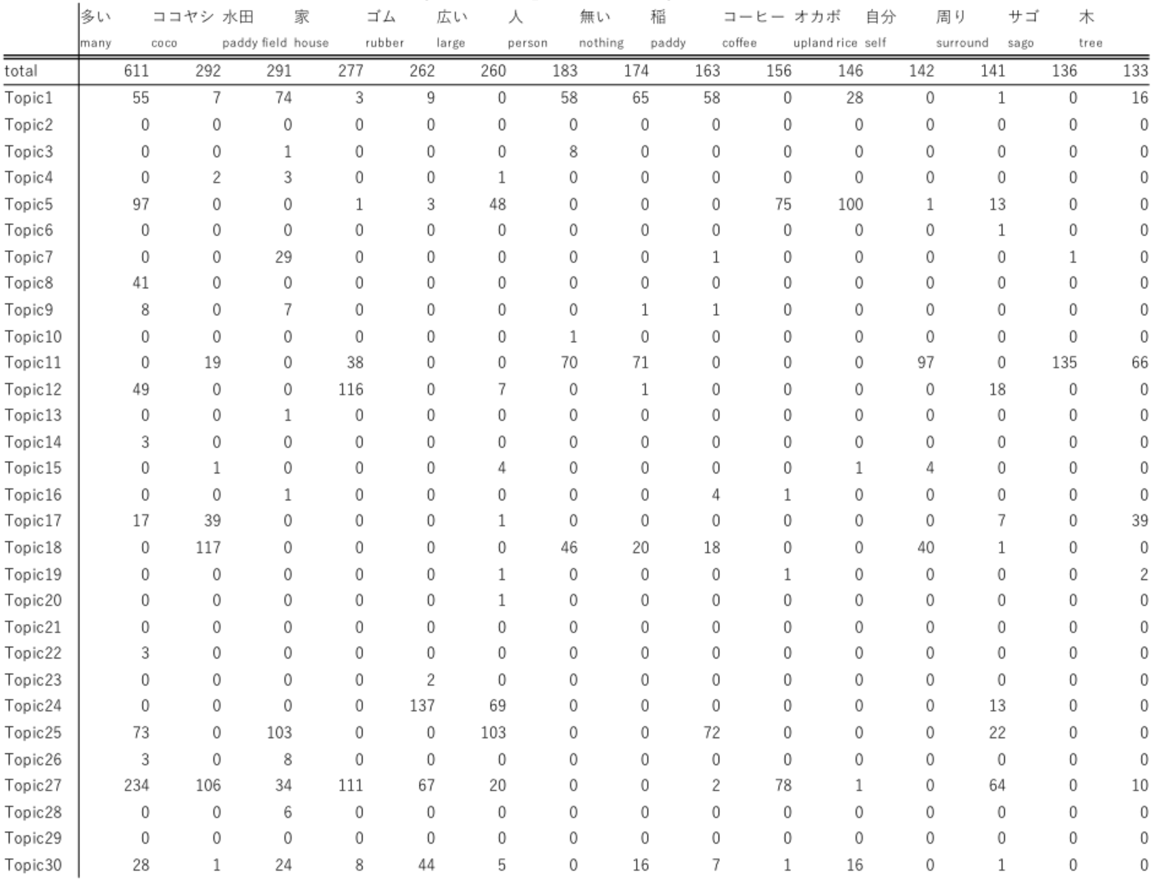
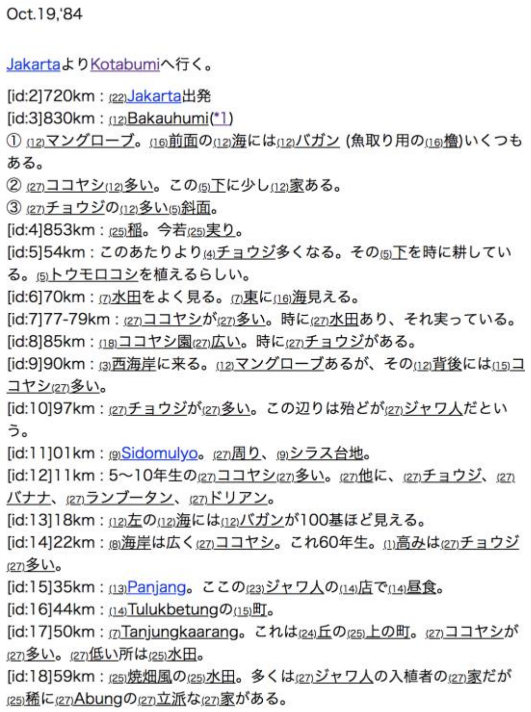


Figure 9: Topic clustering results



	多い many	ココヤシ coco	水田 paddy field	家 house	ゴム rubber	広い large	人 person	無い nothing	稲 paddy	コーヒー coffee	オカボ upland rice	自分 self	周り surround	サゴ sago	木 tree
total	611	292	291	277	262	260	183	174	163	156	146	142	141	136	133
Topic1	55	7	74	3	9	0	58	65	58	0	28	0	1	0	16
Topic2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Topic3	0	0	1	0	0	0	8	0	0	0	0	0	0	0	0
Topic4	0	2	3	0	0	1	0	0	0	0	0	0	0	0	0
Topic5	97	0	0	1	3	48	0	0	0	75	100	1	13	0	0
Topic6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Topic7	0	0	29	0	0	0	0	0	1	0	0	0	0	1	0
Topic8	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Topic9	8	0	7	0	0	0	0	1	1	0	0	0	0	0	0
Topic10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Topic11	0	19	0	38	0	0	70	71	0	0	0	97	0	135	66
Topic12	49	0	0	116	0	7	0	1	0	0	0	0	18	0	0
Topic13	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Topic14	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Topic15	0	1	0	0	0	4	0	0	0	0	1	4	0	0	0
Topic16	0	0	1	0	0	0	0	0	4	1	0	0	0	0	0
Topic17	17	39	0	0	0	1	0	0	0	0	0	0	7	0	39
Topic18	0	117	0	0	0	0	46	20	18	0	0	40	1	0	0
Topic19	0	0	0	0	0	1	0	0	0	1	0	0	0	0	2
Topic20	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Topic21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Topic22	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Topic23	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
Topic24	0	0	0	0	137	69	0	0	0	0	0	0	13	0	0
Topic25	73	0	103	0	0	103	0	0	72	0	0	0	22	0	0
Topic26	3	0	8	0	0	0	0	0	0	0	0	0	0	0	0
Topic27	234	106	34	111	67	20	0	0	2	78	1	0	64	0	10
Topic28	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
Topic29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Topic30	28	1	24	8	44	5	0	16	7	1	16	0	1	0	0

Figure 10: Topic spreading for term



Oct.19,'84

JakartaよりKotabumiへ行く。

[id:2]720km : (22)Jakarta出発
[id:3]830km : (12)Bakauhumi(*1)
① (12)マングローブ。(16)前面の(12)海には(12)バガン (魚取り用の(16)櫓)いくつもある。
② (27)ココヤシ(12)多い。この(5)下に少し(12)家ある。
③ (27)チョウジの(12)多い(5)斜面。
[id:4]853km : (25)稲。今若(25)実り。
[id:5]54km : このあたりより(4)チョウジ多くなる。その(5)下を時に耕している。(5)トウモロコシを植えるらしい。
[id:6]70km : (7)水田をよく見る。(7)東に(16)海見える。
[id:7]77-79km : (27)ココヤシが(27)多い。時に(27)水田あり、それ実っている。
[id:8]85km : (18)ココヤシ園(27)広い。時に(27)チョウジがある。
[id:9]90km : (3)西海岸に来る。(12)マングローブあるが、その(12)背後には(15)ココヤシ(27)多い。
[id:10]97km : (27)チョウジが(27)多い。この辺りは殆どが(27)ジャワ人だという。
[id:11]01km : (9)Sidomulyo。(27)周り、(9)シラス台地。
[id:12]11km : 5〜10年生の(27)ココヤシ(27)多い。(27)他に、(27)チョウジ、(27)バナナ、(27)ランブータン、(27)ドリアン。
[id:13]18km : (12)左の(12)海には(12)バガンが100基ほど見える。
[id:14]22km : (8)海岸は広く(27)ココヤシ。これ60年生。(1)高みは(27)チョウジ(27)多い。
[id:15]35km : (13)Panjang。ここの(23)ジャワ人の(14)店で(14)昼食。
[id:16]44km : (14)Tulukbetungの(15)町。
[id:17]50km : (7)Tanjungkaarang。これは(24)丘の(25)上の町。(27)ココヤシが(27)多い。(27)低い所は(25)水田。
[id:18]59km : (25)焼畑風の(25)水田。多くは(27)ジャワ人の入植者の(27)家だが(25)稀に(27)Abungの(27)立派な(27)家がある。

Figure 11: An example of topics for each scene

4.4 Scene Comparison

LDA can reflect the analysis result (i.e., the detection result) in the analyzed text; thus, we can confirm the result, unlike with principal component analysis (PCA) (Jolliffe, 2002). Figure 11 shows a part of the results for the topics assigned to terms in each scene, where an underlined string indicates an extracted term and number in parentheses indicate the assigned topic number. For example, the term and detected topic of the scene beginning with "830km: Bakauhumi" are "Bakauhumi (a place name)": 12, "ココヤシ (coco palm)": 27, "チョウジ (clove)": 27, "バガン (scaffold for fishing)": 12, "マングローブ (mangrove)": 12, "下 (bottom)": 5, "前面 (front)": 16, "多い (many)": 12, "家 (house)": 12, "斜面 (slope)": 5, "櫓 (scaffold)": 16 and "海 (ocean)": 12. From this result, if the scene can be characterized by the detected topics and their frequencies, we can obtain {Topic 5: 2, Topic 12: 7, Topic 16: 2, Topic 27: 2} as a bag-of-words. Using the bag-of-words, the similarity between scenes d_1 and d_2 can be expressed as follows:

$$sim(d_1, d_2) = \frac{\sum_k weight(z_{1,k}) \cdot weight(z_{2,k})}{\sqrt{\sum_k weight(z_{1,k})^2} \cdot \sqrt{\sum_k weight(z_{2,k})^2}}$$

Equation 7

Here, $weight(z_{i,k})$ indicates the weight of topic k in scene d_i, and we used the frequency of the term in the scene as the weight value. For example, by calculating the similarity of the above scene to other scenes using formula (7), we obtained the following scenes:

- sim=0.950654, date: Oct 26, id=383, "32.5km: ココヤシの多い集落。"
- sim=0.950656, date: Oct 19, id=9, "90km: 西海岸に来る。マングローブあるが、その背後にはココヤシ多い。"
- sim=0.949866, date: Jan 6, id=1185, "68.2km: 山の上の村。クミリが大変多い。対岸にはオカボの実ったものが見えている。チョウジもある。"
- sim=0.870228, date: Oct 19, id=5, "54km: このあたりよりチョウジ多くなる。その下を時に耕している。トウモロコシを植えるらしい。"

According to the field notes, Takaya visited Pekanbaru four times. Figure 12 shows the trip routes, and an overview is given as follows:

- (October 23 to 26) Takaya went from Solok to Pekanbaru, where he stayed three nights. He then went to Rengat.
- (November 1 to 2) Takaya went from Taluk to Pekanbaru. The next day he went to Bangkinan.
- (November 2) Takaya returned to Pekanbaru and the surrounding cities and villages. (November 5) Takaya went to Ujang batu.
- (November 23 to 26) Takaya went from Selat panjang to Pekanbaru by ship. The next day he went to Tembilahan by plane.

Unfortunately, there are few descriptions about the town of Pekanbaru; thus, we collected topic results for scenes that could be judged as being within a 40-km radius from the center of Pekanbaru. The aggregation results are shown in Figure 13. As shown, the scenes around Pekanbaru comprise various topics, which are summarized as follows:

- Topic 27 is the most frequent. In the scenes, the following terms are assigned to Topic 27: "多い (many)," "家 (house)," "ゴム (rubber)," "ランブータン (Rambutan)," "村 (village)," "ココヤシ (coco palm)," "コーヒー (coffee)," "チョウジ (clove)," and "バナナ (banana)."
- Topic 5 has many terms that are also assigned to Topic 27. This result is consistent with the

result discussed in Section 4.2.

- The terms of Topics 1, 30, and 12 are very similar to those shown in Figure 8.

Pekanbaru (located in central Sumatra) is the capital of Riau province. The city name is derived from the Indonesian words for “new market.” In the late 19th century, the city was developed to serve the coffee and coal industries, and the Dutch built roads to help ship goods to Singapore and Malacca. Figure 13 shows the results of the scenes of suburban Pekanbaru. However, topics related to coffee and waterside were detected, and, for some reason, the results appear to be related to Pekanbaru.

5.1 Text Analysis by LDA

Because LDA is an unsupervised learning method, preparing learning data is not required. Supervised learning, which is a machine learning method (like support vector machines (Crammer and Singer, 2001), naive Bayes classifiers (Mozina et al., 2004), and random forests (Breiman, 2001)), outputs analysis results according to prepared learning data. Therefore, the classes to be output are determined prior to analysis. On the other hand, unsupervised learning, such as LDA, can determine the number of classes to classify prior to analysis but cannot decide the output class. Thus, with LDA, it is difficult to achieve highly accurate classification as with supervised learning; however, LDA is very useful when a user does not understand the content of the data. In this study, one purpose was to share the analysis results of a field note that only the investigator(s) would have used; thus, we consider that the extraction of features from data with content that is not understood is important. Therefore, LDA is an optimal analysis method. In addition, like support vector machines, LDA can support tracing the factors of the analysis results to the original resource, which means that LDA can realize detailed analysis of field notes.

5. Discussion

In this section, we discuss the text database we constructed, the LDA-based text analysis method and applicable scope of our method. In an experiment, we set the number of topics to 30. Note that this value was determined heuristically. We confirmed each experimental result (corresponding to Figure 8) when the number of topics was set to 20, 30, and 50 for area studies.

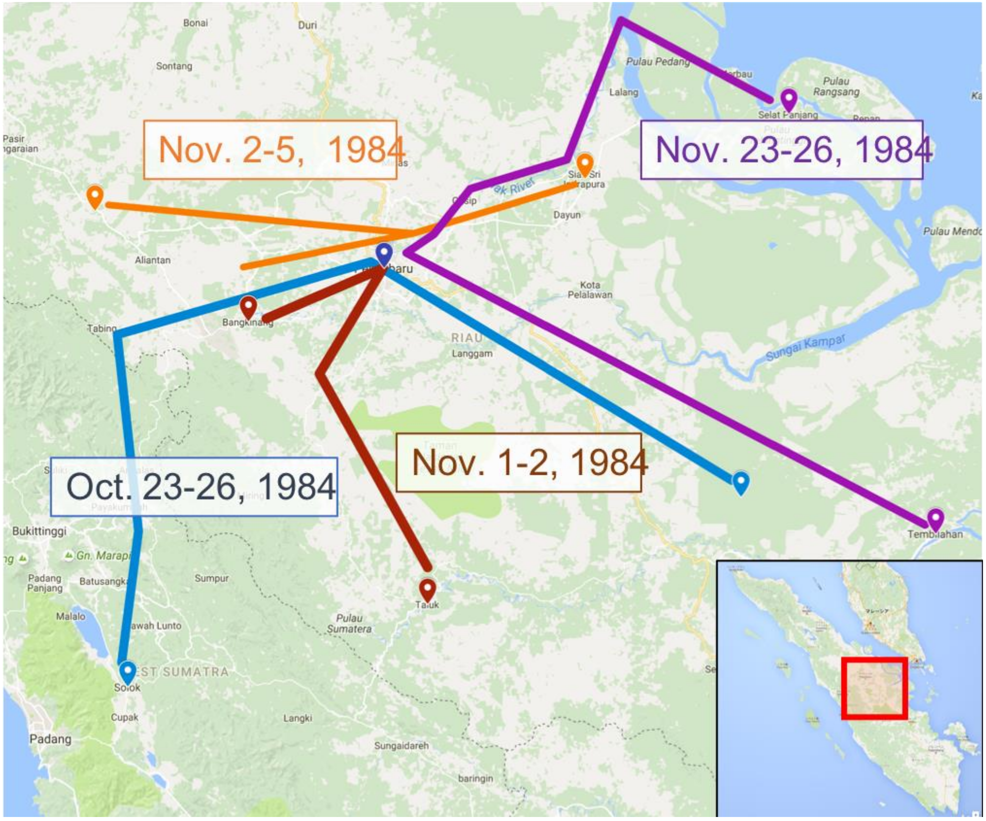
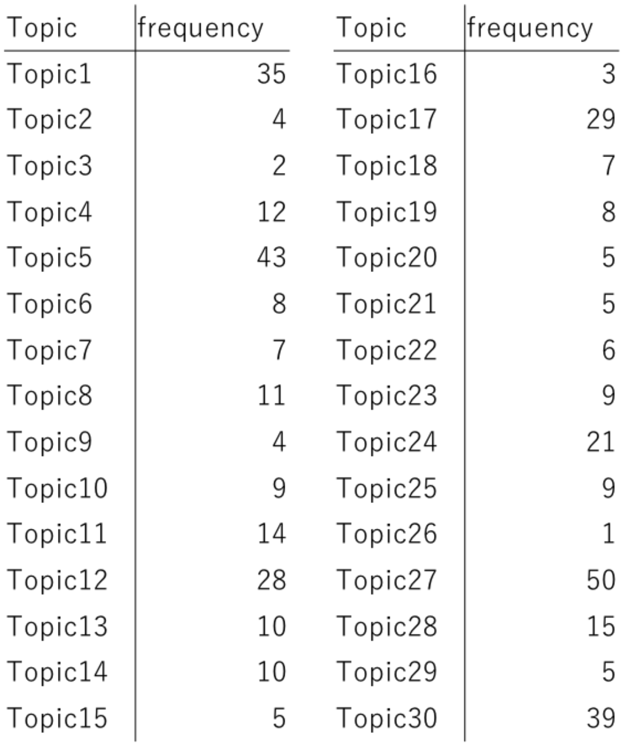


Figure 12: Investigation routes around Pekanbaru



Topic	frequency	Topic	frequency
Topic1	35	Topic16	3
Topic2	4	Topic17	29
Topic3	2	Topic18	7
Topic4	12	Topic19	8
Topic5	43	Topic20	5
Topic6	8	Topic21	5
Topic7	7	Topic22	6
Topic8	11	Topic23	9
Topic9	4	Topic24	21
Topic10	9	Topic25	9
Topic11	14	Topic26	1
Topic12	28	Topic27	50
Topic13	10	Topic28	15
Topic14	10	Topic29	5
Topic15	5	Topic30	39

Figure 13: Topics from scenes around Pekanbaru

The optimum number of topics was found to be 30. To estimate the number of topics in LDA, a method that uses a hierarchical Dirichlet process (HDP) has been proposed (The et al., 2006a). In future, we plan to analyze field notes using this HDP and compare the current results.

5.2 Textual Database for a Field Note

An overview of the field note was given in Section 2, and our text modeling method was described in Section 3.1. We were able to extract dates easily from the field notes; however, it was difficult to extract detailed spatial information from the text because there was a lack of decisive data, such as latitude and longitude information. In addition, because IPADic lacked data about specific Sumatran place names, the results of the word segmentation and the POS tagging were only classified as nouns despite being place names. Therefore, a place name extraction method that uses place name data in a place name data service (e.g., Geonames.org[6]) and latitude and longitude determination method are required. We assume that the text data and analytical results will be shared on the web; thus, our text database was designed based on the RDF model, which makes it easy to link to external LOD. By actively linking to external data, we believe that the usefulness of the text data will increase. Note that, as linking with external data advances, vocabulary correction and addition become increasingly important. Therefore, we believe that the completeness of the text data and text database will improve.

5.3 Applicable Scope of Our Method

We describe the applicability of our method to materials other than Takaya' field note explained in Section 2. We believe that our method could be applied to a material in which an analysis unit and a feature vector (such as bag-of-words) are determined. Our method may be applied to other field notes and other resources related to area studies (like newspaper) if the resources meet the conditions. Also, term co-occurrence is one of the conditions for using LDA. The same applies to the applicability to resources other than text materials. The method of extracting features from images and movies is different from the method of extracting from texts. We think that establishing the method (like the method of bag-of-visual-words (Csurka et al., 2004)) will be important if dealing with them.

6. Conclusion and Future Work

In this paper, we have introduced a method to construct a text database of area studies resources (specifically field notes) and an analysis method for the resources that uses LDA. In text analysis experiments, the amount of text data was too small to use a machine learning method. We initially thought that it may be more effective for users to read and understand the field notes manually. However, a researcher who was familiar with the content of the field notes is reported that LDA results can give an attention point that he has not noticed before (Yanagisawa et al., 2016). We believe that the impact of topic model analysis has been confirmed in areal studies, and we would like to position topic analysis as a practical service, e.g., as a common database search service, which we expect to be of value to researchers in various fields.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Numbers 16H01897, 18H03576, and 18H03588 and the Japan-ASEAN Platform for Transdisciplinary Studies project (Center for Southeast Asian Studies, Kyoto University).

References

Amorim, R. C., 2015, Feature Relevance in Ward's Hierarchical Clustering Using the L_p Norm. *Journal of Classification*, Vol. 32(1), 46–62.

Asuncion, A., Welling, M., Smyth, P. and Teh, Y. W., 2009, On Smoothing and Inference for Topic Models. *Proceedings of the 25th Conference Conference on Uncertainty in Artificial Intelligence*, 27-34.

Berners-Lee, T., 2006, Linked Data, https://www.w3.org/DesignIssues/LinkedData.html.

Blei, D. M., Ng, Andrew, Y. and Jordan, M. I., 2003, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, 993-1022.

Breiman, L., 2001, Random Forests, *Machine Learning*, 45, 5–32.

Crammer, K. and Singer, Y., 2001, On the Algorithmic Implementation of Multi-Class Kernel-Based Vector Machines. *Machine Learning Research*, Vol. 2, 265-292.

Csurka, G., Bray, C., Dance, C. and Fan, L., 2004, Visual Categorization with Bags of Keypoints, *Proceedings of the 8th European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision*, 59-74.

DCMI, 2012, DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/.

Griffiths, T. L. and Steyvers, M., 2004, Finding scientific topics. *Proceedings of the National Academy of Sciences*. Vol. 101–1, 5228–5235.

Jolliffe, I. T., 2002, *Principal Component Analysis*. Springer.

Mozina, M., Demsar, J., Kattan, M. and Zupan, B., 2004, Nomograms for Visualization of Naive Bayesian Classifier. *Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 337-348.

Takaya, Y., 2012, The Field Note Collection 2 (地域研究アーカイブズ フィールドノート集成 2). *CIAS Discussion Paper Series*, Vol. 22, 1-199 (in Japanese).

Tanigawa, R. and Yamamoto, H., 2013, *Area Studies' Database*. Center for Integrated Area Studies, Kyoto University.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M., 2006a, Hierarchical Dirichlet Process. *Journal of the American Statistical. Association*, Vol. 101, 1566-1581.

Teh, Y. W., Newman, D. and Welling, M., 2006b, A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 1353-1360.

W3C, 1999, Resource Description Framework (RDF) Model and Syntax Specification, https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

W3C, 2015, Semantic Web, https://www.w3.org/standards/semanticweb/.

Wallach, H. M., Mimno, D. M. and McCallum, A., 2009, Rethinking LDA: Why priors matter. *Proceedings of the 22th Neural Information Processing Systems Conference*, 1973–1981.

Yanagisawa, M., Takata, Y. and Yamada, T., 2016, Interpretation of Regional Informatics– Spatiotemporal Indication and Text Analysis as a Discovery Tool. *JCAS Review*, Vol. 16–2, 267–291. (in Japanese).