第20回情報知識学会フォーラム予稿

フィールドノートに記述された場面を特徴付ける

—語彙による知識処理—

Characterizing Scenes in Field Note:

Knowledge Processing Using Vocabulary

山田太造^{1*}

Taizo YAMADA^{1*}

1 東京大学 史料編纂所

The University of Tokyo, Historiographical Institute

〒133-0033 東京都文京区本郷7-3-1

E-mail: t_yamada@hi.u-tokyo.ac.jp

*連絡先著者 Corresponding Author

フィールドノートは調査したフィールドの観察記録，観察したフィールドの場所・日時，その風景に関するスケッチ・写真などで構成されたものである．調査対象であるフィールドについて特定の日時での様子を詳細に理解することができるため，地域研究において非常に重要な研究資源の1つといえる．われわれはこれまでに，地域研究進展のためにフィールドノートを効率的かつ効果的に利用していく手法を模索しており，本研究ではテキストマイニングを用いてフィールドノートから記述されている場面を特徴づけ，かつセマンティックウェブを利用して表現する手法を提案する．

Field note consists of observation notes, locations, dates, drawings and images of fields which were observed by researchers of Area Studies. The note is one of important research resources for Area Studies, because it enables to understand details of the states of the field at the time. Due to progress of Area Studies we have explored some methods and/or techniques to utilize the note efficiently and effectively. In the paper we introduce a method which a scene described in the note can be characterized by text mining technique, and a method which the characterized scene can be represented by semantic web technique.

キーワード: 地域研究, トピックモデル, LDA, RDF, セマンティックウェブ
area studies, topic model, LDA, RDF, semantic web

1 はじめに

フィールドノートは調査したフィールドの観察記録、観察したフィールドの場所・日時、その風景に関するスケッチ・写真などで構成されたものである。記録されたフィールドについて特定の日時での様子を詳細に理解することができるため、地域研究において非常に重要な研究資源の1つといえる。記述されている内容は地域研究史資料に関する目録ベースの情報よりも詳細かつ膨大であり、地域研究推進において欠かせないエッセンスになりうる。

地域研究推進のために、ここ十年間で、さまざまなデータベースが構築され、公開されている。その多くは、研究資源の目録であり、さらには風景を撮影した写真・映像に関するデータベースの公開もされている。しかしながら、フィールドノートに関するデータベースとしては、公開されたとしても目録までであり、そのテキストに関するデータベースはあまり公開されていない。その理由としては、テキストの作成のコストの問題もあると考えられるが、テキストの地域研究等での利用方法が議論されていないためではないかと考えている。

著者らのプロジェクトでは、地域研究を展開していくために、フィールドノートを効率的かつ効果的に利用していく手法を模索している。そのため、フィールドノートから得られる情報がどのようなものがあるかを、情報学における手法を用いて提示し、提示された結果の地域研究における有益性を検証し、フィールドノート活用を地域研究における研究手法の1つとして確立することを目的としている。

本研究では、フィールドノートから記述されている場面の特徴を、テキストマイニングにより解析し、表現する手法を提案する。特に次の2点に着目する。

- 1) フィールドノートからの時空間的特徴の抽出と場面の構造化：テキストから地名・日付とともに各場面を特徴付ける用語を抽出する。
- 2) フィールドノート情報の利用：1) で表現した場面を検索できるフィールドノート検索システムのプロトタイピングを行う。また1) で表現した場面データの2次利用を図る仕組みを導入する。

一般的に、テキストからの情報抽出・話題検出などのテキストマイニングは、永らくデータ工学や自然言語処理などの分野で、提案され、改良されてきた。しかしながら、地域研究史資料に対する、地域研究への寄与を目的としたテキストマイニングは皆無に等しい。本研究が地域研究に与えるインパクト、特にフィールドノートやテキストデータの利活用の点で、地域研究の新たな研究パラダイムに繋がる可能性を秘めているというシーズとしての役割に展開していくと期待している。

本論文は以降、次のように構成している。2節で本研究にて用いたフィールドノートを示す。3章でフィールドノートのテキストから生成した場面データの構造を示す。4章で場面データを特徴づけていく手法を示す。5章で場面データの提示手法を示す。

2 対象のフィールドノート

京都大学地域研究統合情報センターは、東南アジア地域研究において著名な高谷好一氏によるフィールドノートを書籍と

図1 フィールドノートの例。左図はOct. 19 '84の、右図はOct. 28 '84の記述。

して2012年3月に5冊、2013年3月に3冊合計で4,405ページを出版した。このシリーズは『地域研究アーカイブズ フィールドノート集成』と冠しており、調査対象地域は、東南アジア、インド、ヨーロッパ、アフリカなど様々である。[2]によれば、このフィールドノート集成は、記録者である高谷氏の全面的な協力の下、現場で観察した記録（調査した当日の夜にその日の観察記録をカード化し整理したもの）に対する文字起こしを行い、さらに写真やイラストの整理も行うことで、出版に至ったようだ。

本研究ではこのうち、「スマトラ1984.10.19 - 1985.1.18 スマトほぼ全域」[1]を利用した。該当のフィールドノートは、分量としてはA4サイズ198ページ、165,757文字程度であるが、スマトラ島全

域をカバーしている。図1は実際のフィールドノート集成のページを示す。前述のとおり、この観察記録は日を単位に整理されており、その日の観察した風景、現地の人へのインタビュー、土地利用などに関する記述、現場でスケッチしたイラスト、撮影した写真などで構成されている。図1左はこの調査全体の初日の冒頭を示す。書き出しは“Oct. 19 '84”であり、これは日付を示す。日付に続く“JakartaよりKotabumiへ行く。”はこの後に続く観察した記録の主題に相当する。主題は1日に複数回登場することもあるが、全く登場しない場合もある。移動しながら風景を記録している場合、端的に1行から数行程度の短い記述が続く。ここで、行が変われば記録している場面が変わることを意味する。図

図2 場面データの作成フロー

1右に“B) Agrarian に連れられて周辺を歩く”という主題があり、その後に記述が続く。このような記述はその場所の風景、土地利用または統計的情報を集中的に記録している場合に見られる。現地の人へのインタビューの際も同様の記述形態である。本研究では、記録の最小単位であるこの記録を“場面”と呼ぶことにする。

3 場면을構造化する

本節では場面の構造化について述べる。前節で述べたとおり、本フィールドノートは日単位にまとめられており、記述内容は場面の景観、土地利用、現地の人へのインタビューであるため、データとしての場面は時間および場所に関する情報を持つ。そこで、本研究では、これらの要素を用いて各場면을構造化することにした。

図2は場面データの作成フローを示す。

本フィールドノートは書き起こしたテキストがあるため[2]、テキスト化は不必要だった。まずフィールドノートのテキストを場面ごとに区切った。その後、場面データを作成した。前述のとおり、場面データは場面の記述内容、時間と場所を表すデータを用いて表現した。

次に、時間軸・空間軸で識別される場面を特徴付けた。マングローブ、チョウジ、ココヤシ、赤土、丘陵のようにその場面で観察された“もの”・“こと”が記述されており、場面の特徴を端的に表現している。本研究では、これらを抽出し、分類することでその場面を特徴付けていくことにした。まず、場面を特徴付けるこれらを用語としてテキストから抽出した。このとき形態素解析などを行うことで抽出した。場面の分類を行うため、抽出した用語とその出現頻度をもとに、トピックモデルを用いて場面を特徴付けた。このように特徴付けら

れた場面データを多目的に利活用可能な形式で表現する．その方式としてRDFを用いて表現した．これらは次節以降に述べる．

4 場面を特徴付ける

4.1 用語を抽出する

場面データを特徴付けるため，まずテキストから用語を抽出する．用語抽出ではMeCab[3]を用いて形態素解析を行い，そのうち名詞および形容詞を用語として抽出する．ただし名詞のうち，代名詞，数，接尾，副詞可能，形容動詞語幹，ナイ形容詞語幹，接続詞的，非自立は対象外とした．形態素解析辞書としてIPADICを用いた．また，頻出するがIPADICには登録されていない用語（ランブータン，ジャックフルーツ，キャッサバ，サゴヤシ）をユーザ辞書に登録した．

形態素解析の結果からチャンキングを行った．固有名詞や専門用語の出現をカバーするためである．日本語係り受け解析器であるCaboCha[4]は固有表現解析の機能を有しており，これを用いてチャンキングを行った．CaboChaはIREX（Information Retrieval and Extraction Exercis）の定義にもとづき固有表現タグをIOB2形式で出力することができる．本フィールドノートでは“Banda Aceh”や“Sungai Tumbesi”のようにローマライズされた表記が出現するが，残念ながらこれらに対して有効に機能しない．そこで，CaboChaでの固有表現解析の結果に対して，さらに下記のルールを適用した．これは汎用的ではなく，本フィールドノートにおいてのみ有効なヒューリスティックな手法である．

- ・抽出対象が連続する場合

- ・名詞の直後に名詞-接尾が出現する場合
- ・[a-zA-Z]+にマッチする形態素が連続する場合

4.2 場面からトピックを検出する

次に，場面を特徴付ける．情報検索等ではしばしばベクトル空間モデルを用いて特徴付けることがある．文書を対象に特徴づける場合，ある文書での各用語の出現頻度等で算出される重みで文書ベクトルを作成する．本研究では文書は場面に相当する．場面における記述は1行程度である場合や数ページに渡る場合もあり，長さが不均一である．そこで本研究では，各場面のトピックにもとづいてその場面を分類していくことで特徴づけていくことにした．

予めどのようなトピックがあるかが判断することができなかった．そこで，教師なし学習 (unsupervised Learning) の手法であるLDA (Latent Dirichlet Allocation) [5]を用いて潜在トピックの検出を行った．LDAを用いると，場面と潜在トピックの関係，および潜在トピックと用語の関係を検出することができる．LDAは（1）式により表現することができる．

$$p(d|\alpha, \beta) = \int \text{Dir}(\theta|\alpha) \cdot \left(\prod_{n=1}^{|d|} \sum_{k=1}^C p(w_n|z_{k,\beta}) p(z_k|\theta) \right) d\theta \quad (1)$$

ここで α, β はパラメータ， $z = z_1, z_2, \dots, z_C$ は潜在トピック， $\theta = \theta_1, \theta_2, \dots, \theta_C$ は潜在トピックの生成確率， $\text{Dir}(\theta|\alpha)$ はディリクレ分布， $d = (w_1, w_2, \dots, w_{|d|})$ は場面， w_n は用語， $|d|$ は場面 d の総用語数を示す．LDA は潜在トピックの生成確率がディリクレ分布に従うと仮定した文書生成モデルといえる．

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1集落 8	島 10	Loc 8	Bengkalis 8	チガヤ 4	多い 94	松 12	きれいい 6	池 28	町 30	Sultan 5	ゴム 59	地区 11	中国人 103	木 16
2乳産 6	多い 8	北道 4	土手 6	Tembilahan 3	オカボ 92	悪い 7	広大 5	水田 25	店 17	pres 5	水田 44	九太 7	人 103	炭 10
3女 4	周辺 6	島 3	登 5	急 3	トウモロコシ 86	溝 7	ゴム園 4	魚池 16	市場 9	森 5	ゴム園 21	会社 6	自分 82	室 10
4新しい 4	ton 5	苗木 3	オランブニヤ 4	松林 3	広い 80	平地 6	広い 4	魚 13	北 7	簡単 4	Minangkabau 19	レジ 5	無い 79	炭焼き小屋 8
5苗木 4	松林 5	Pekanbaru 2	乾季 4	煙瓦 3	コーヒー 52	所々 6	ヨシ原 3	小池 6	Ansad 5	Ungku Tugut 3	suka 16	分かれ 5	家 78	直径 2
6箱 4	美 4	Rokan 2	井戸 4	草地 3	焼酎 37	産 5	Sunga Lela 2	稚魚 6	Kapas 5	村 3	焼酎 16	山 5	ココヤシ 76	幅 6
7盆地 3	Talawa 3	Tembilahan 2	作期 3	Batak 2	斜面 33	高い 5	cair 2	Loc 5	丘 5	班 3	人口 14	uma 4	Rp 74	壁 5
8Kampar 2	シラス 3	baris 2	分かれ 3	Makanan Padang 2	周 31	急 4	cungkilan 2	hedong 5	植 5	Blasting 2	オカボ 13	Transmigrasi 3	Melayu 68	灌漑水路 4
9Transmigrasi 2	丘地帯 3	karet 2	川 3	Medan 2	シナモン 30	村 4	gulungan 2	上流 5	全 4	Tengku Syarif 2	トウモロコシ 10	集落 12	ft 3	土地 51
10Balai adat 2	川沿い 3	ホテル 2	満期時 3	ドラム缶 2	クミリ 29	村長 4	ばら 2	圃 5	Duma 3		星夜水 3	多い 49	クーボン 3	泰 3
V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
1オカボ 5	サゴ 91	オランダ 33	ゴム 142	バガン 19	水田 108	魚 43	家 77	多い 227	左 4	Banjor 49	草 108	牛 16	コショウ 19	ココヤシ 103
2Bukit 4	工場 59	下 16	広い 90	Tebing Tinggi 6	広い 84	網 28	多い 59	家 118	下り 3	Rugs 37	排 98	長い 16	トリアン 12	木 75
3松 4	tuat 44	ムラユ 14	ゴム園 41	Tanjung Dauk 5	稲 65	長い 26	右 40	コーヒー 95	川口 3	Sapa 37	田 95	クビキ 8	根元 7	自分 58
4Amuntai 3	サゴヤシ 35	Raja Kecil 12	タビダ 27	Tanjung Pinang 5	多い 61	inch 23	マンダローブ 29	ココヤシ 91	屋敷 3	Tembilahan 29	多い 86	土 8	中心 5	良い 42
5Loc 3	水 34	間 9	丘 24	核 5	幅 61	エビ 18	集落 24	村 61	湿地 3	稲 27	水田 22	祖先 7	成木 5	泥炭 41
6マラヤ 3	Rp 33	kata 8	植 12	内皮 4	継目 60	depa 14	左 23	ゴム 69	長穴 3	Jawa 21	無い 72	草原 6	コーヒー 4	サゴヤシ 36
7鳥 3	濡れサゴ 23	人達 8	悪い 8	原 4	川 43	目 14	ニッパヤシ 15	周 54	雑魚 3	Pari 19	苗代 72	鶏卵 5	地主 4	水路 34
8小村 3	ton 20	王 8	サゴヤシ 6	Huk Teicu 3	谷地田 35	深い 13	大変 14	ラングータン 39	Tungku Saleh 2	稲作 19	月 68	翼 5	木の 下 4	下 32
9Tanjung Pinang 2	穀粉 20	役人 7	低み 6	Punggur Jawa 3	周 28	gambang 12	町 14	マンゴー 35	karet 2	Melayu 18	稲 60	Sukauk 4	Palem bang 3	粘土 32
10Jama 2	洗い場 18	王宮 7	高木 6	日本 3	冊 28	坪 9	昔後 14	チウジ 31	paang 2	ココヤシ園 13	水牛 57	Bucun 3	Takungon 3	川 30

図 3 LDA による潜在トピック検出の結果

本研究に適用した場合，LDA は，1 場面におけるトピックは複数あり，トピックはそれぞれに複数の用語を生成することをモデル化している．(1) 式をそのまま計算することはかなり困難であるが，崩壊形ギブスサンプリングを用いた解法が知られており [6]，本研究ではこれを用いて潜在トピックを算出する．

トピックモデルとして LDA 以外にも LSI (Latent Semantic Indexing) [7]や pLSI (probabilistic LSI) [8]がある．LSI は 1 場面につき 1 トピックを仮定するため，多角的な関連性を考慮できない．pLSI は LDA と同様に LSI を拡張し 1 場面につき複数トピックを仮定する．しかしながら，潜在トピックの生成確率，つまり式 (1) における $p(z|\theta)$ を最尤推定するなどして事前に算出する必要がある．そのため，学習データにはない場面への

対応は高コストになってしまう．また $p(z|\theta)$ は学習データの量に応じて計算コストが増大してしまうためアドホックな手法で求めることが多い．これに対し LDA は $p(z_k|\theta)$ を確率的に算出する生成モデルである．

4.3 フィールドノートへ適用する

本フィールドノートに対し，LDA を適用することで潜在トピックの検出を行った．LDA のパラメータとしてはトピック数を 30，ギブスサンプリングの回数を 2,000 とした．

この結果を図 3 に示す．この図では 30 のトピックと各トピックを構成する用語のうち出現頻度順に上位 10 件を示している．各トピックの概要を端的に示すことは難しい．そこで，水田とココヤシが出現するトピックに着目してみる．水田

図4 類似する場面の検索例

が出現するトピックは9, 12, 21, 27である。トピック9では、水田以外に、池、魚池、小池、稚魚などがある。同様に、トピック12では、ゴム・焼畑・オカボ・集落・トウモロコシなど、トピック21では、稲・幅・棚田・川・谷地田など、トピック27では、草・鋤・苗代・水牛などがあつた。これにより、トピック9は池や水田と魚関係、トピック12はゴムや水田などの作目がある風景、トピック21は水田のある風景、トピック27は田畑・耕作等関係を示すと考えられる。ココヤシが出現するトピックは、14, 24, 30だった。トピック14では中国人・家・Melayu・土地など、トピック24ではコーヒー・村・ゴム・ランブータン・マンゴー・チョウジなど、トピック30では木・泥炭・サゴヤシ・水路・粘土・川などがあつた。これらの検出された用語より、トピック14は土地利用関係、トピック24はコーヒー・ココヤシなどがある風景、トピック

30は水辺・川辺での風景を示すと考えられる。また、文字列もしくは指し示す“もの”・“こと”は同じであっても、トピックが異なればその用語の意味合いが異なることがわかった。

5 場面を提示する

5.1 類似する場面を提示する

前節で示したLDAの結果を用いて、類似する場面を提示する方法について考えてみる。各場面から抽出した用語に対し、LDAを適用することにより、それぞれがどのトピックとして検出されたかが結果として得られる。検出された用語の頻度を場面ごとに集計し、これをベースに場面ごとに各トピックの重み付けを行えば場面ベクトルを作成することが可能である。ある場面 x におけるトピック i の重み $weight(x_i)$ を次式により算出する。

図5 RDFを用いた場面データ

$$\text{weight}(x_i) = tf(x_i) \cdot \left(\log \frac{N}{df(i)+1} \right) \quad (2)$$

これはtf-idf重み付けに相当する。ここで、 $tf(x_i)$ は場面 x におけるトピック i の出現頻度、 $df(i)$ はトピック i を含む場面数、 N は場面の総数を示す。次に場面の類似度次式で算出する。

$$\text{sim}(x, y) = \frac{\sum_i \text{weight}(x_i) \cdot \text{weight}(y_i)}{\sqrt{\sum_i \text{weight}(x_i)^2 \cdot \sum_i \text{weight}(y_i)^2}} \quad (3)$$

(3) 式は場面ベクトル x と y のコサイン類似度を示している。図4はOct. 19 '84の2つ目の場面（場面ID：3、青の矩形で囲んだ部分）と類似する場面を検索したときの検索結果（赤の矩形で囲んだ部分）の一部を示す。この場面における各トピックの出現頻度を図4の上部に示している。これは場面ベクトルの一例であり、これを用いて

(2) 式および (3) 式により、場面ベクトル間の類似度が算出され、類似性の高い場

面を検索することができる。

5.2 場面をデータとして提示する

図2に示したフローにより場面データを作成していく。このフローにおける各処理については前節までに示したとおりである。作成された場面データを単一の目的、例えば5.1に示した類似する場面データを提示するための利用目的、もしくは単一のシステムでの利用だけではせっかく作成した場面データを広く共有・利活用するのは限界があると考えている。場面データをシステムとは切り離し、汎用的な形式で記述することができれば、既存のシステム・ツールで利用するなど多様な目的で利活用することが可能だと考えられる。

本研究では、特にweb上でのデータ取得・共有、あるいはwebアプリケーション・システムでの利用を考慮し、場面データを

RDF (Resource Description Framework) を用いて表現することにした。図5はRDFモデル化した場面データのグラフを示す。RDFはWeb上で扱う情報 (リソース) を広く共有・利用するための共通化した標準的な枠組みである。RDFではリソースを主語・述語・目的語の3つ組 (トリプル) で表現する。主語は表現対象である。場面データを対象としているため、場面 (図5はScene) 自体が主語である。目的語も主語と同じくリソースである。述語は主語と目的語の関係を明示した情報と見なすことができる。述語はプロパティと呼ばれることもある。図5では楕円は主語、矢印は述語、矩形は目的語を示す。RDFではリソースの識別のためURI (Uniform Resource Identifier) が標準化されている。本研究でも同様に、場面の識別子としてURIを用いる。目的語はURIもしくはIRI (Internationalized Resource Identifier) で表現する。述語はDublin Coreなど既知・既存のメタデータ語彙を用いた。しかしながら潜在トピックの表現では、同様の語彙が見当たらなかったため独自語彙として定義した。この独自語彙は、図5において、接頭辞がfnである語彙がこれに相当する。図5は、場面Sceneは、その場面の識別子 (fn:descId)、タイトル (dc:title)、主題 (dc:subject)、日時 (dcterms:temporal)、場所 (dcterms:spatial)、記述内容 (dc:description) および潜在トピック (fn:topic) で構成されていることを示している。ここで、fn:termは用語、fn:topicClassは潜在トピックのクラスを示す。

6 おわりに

本研究では、高谷好一氏著『地域研究アーカイブズ フィールドノート集成』の一部を用いて、フィールドノートに記録された場面の記述をもとに、フィールドの各場面を自動的に特徴付けていく手法について述べた。また、その成果を特定のシステム・アプリケーションだけではなく、広く利活用するため、RDFによるデータ表現についても述べた。

地域研究におけるテキストデータの利活用、特に地域研究における研究的側面に耐えうる利用方法を目指していくことを想定し本プロジェクトを開始した。しかしながら現在では、地域研究としてのニーズに耐えうるだけのアプローチを提案することは大変困難であるかもしれないと考えるに至った。テキストの利用方法についてはシーズ的に研究者に提供するほうがもしかしたらニーズに寄り添ったものになるかもしれないと考えるに至り、現在の研究過程を経た。これをより、研究者サイドで利用しやすいものへと転換する必要がある。これはテキストおよびテキスト分析結果等の提示もしくは新たな加工・分析を可能とするユーザインターフェースも必要かもしれない。もしくは、さらに掘り下げたデータ分析結果かもしれない。それについて、地域研究者との対話を重ねながら本研究を進めていく予定である。

謝辞

本研究の成果の一部は、日本学術振興会科学研究費若手研究 (B) (26730167)、基盤研究 (B) (26280122)、および京都

大学地域研究統合情報センター共同利用・共同研究拠点個別共同研究ユニット「フィールドノートを対象としたテキストマイニングに関する研究」の助成を受けたものによる。

参考文献

- [1] 高谷好一：「スマトラ」，地域研究アーカイブズ フィールドノート集成2，京都大学地域研究統合情報センター，Vol. 2，No. 22，pp. 1-198，2012.
- [2] 柳澤雅之：「フィールドノート・プロジェクト」，Seeder，昭和堂，No. 11，pp. 14-22，2014.
- [3] MeCab: Yet Another Part-of-Speech and Morphological Analyzer,
<http://taku910.github.io/mecab/>.
- [4] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer,
<https://taku910.github.io/cabocha/>.
- [5] D. M. Blei; A. Y. Ng; M. I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.
- [6] T. L. Griffiths; M. Steyvers: Finding scientific topics, Proc. of the National Academy of Sciences of the United States of America, vol. 101, pp. 5228-5235, 2004.
- [7] S. Deerwester; S. T. Dumais; G. W. Furnas; T. K. Landauer; R. Harshman: Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, Vol. 41, No. 6, pp. 391-407, 1990.
- [8] T. Hofmann: Probabilistic Latent Semantic Indexing, Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57, 1999.