Vol.2015-CH-105 No.4
2015/1/31

地域研究資料と対象とした時空間情報に着目したデータの構造化

山田 太造[1,a)]

概要：本研究は，地域研究史資料を対象として，記述内容における時空間変化を追跡可能なデータ構造の確立を目指す．本報告では特に，本研究における地域研究史資料からのテキスト作成，テキストからの時空間情報の抽出・トピックの検出の各手法について述べる．

1. はじめに

地域研究では，各地域における政治，経済，産業（産業構造），法律，文化，民俗，歴史など多岐にわたる事象・情報を加味しつつ，対象とする地域の特色を比較しながら考察していく．文学や史学などでは，当然のことではあるが，主に文献資料を対象として研究を進めていくが，地域研究では文献資料だけでなく，映画・音楽のような動画・音声についても研究対象に含まれ，研究素材も多様である．対象とする地域で得られる観察記録，メモ，スケッチ，写真などのように観察対象であった現地のことを知る上で重要であるフィールドノートは地域研究において重要な “地域研究史資料” として位置づけられる．またフィールドノートの作成においては，研究者個人が対象地域をその研究的視点により多様な情報の管理方法を体現した結果として見ることもできる．

しかしながら，フィールドノートはそれを記した研究者個人のみが利用していくことを前提に作成されていることが多いため，他者が利用するなどそれを前提とした地域研究の推進は行われていない．フィールドノートの全体を包含するような論文・公な報告はないため，膨大な研究資料が死蔵されているのが現状である．

本研究は，地域研究の推進のためには欠かせないはずであるフィールドノートを，記した研究者だけでなく，他者も利用できるように共有し，利用していくことを目的とする．本稿では，この目的を達成のためのプロジェクトである京都大学地域研究統合情報センターにおける地域情報学プロジェクト「「地域の知」の情報学―時間・空間・語彙に注目した地域情報学の展開―」[*1] における研究課題（複合共同研究ユニット）の 1 つである「地域研究データにおけるトピックの検出と時空間変化に関する研究」[*2]（平成 25 年 4 月～平成 27 年 3 月（2 年間））での研究報告を行うものである．ここでは，どの地域でいつ何がどうなったか，という情報が地域研究において重要であると考えている．そこで，フィールドノートを対象として，そこからトピック（ここでは共起する語彙により特徴付けられるカテゴリを指す）を検出し，時空間的特色を明らかにするための手法の確立を目指している．特に以下の 2 点に着目している．

(1) フィールドノートからのトピック検出および時空間変化追跡手法の確立：テキストから特徴語を抽出し，それをもとにトピックを検出する．特に，時空間語彙に着目した時間的・空間的特徴を検出する．
(2) トピック追跡をもとにした地域に関するデータの構造化：データ間において潜在的に形成されるリンクを構造化する．機械可読可能な形式で表現することにより，フィールドノートの特徴を再利用や他の目的で使用可能かつ意味構造を表現しうる形式で表現する．

2. 対象とするフィールドノート

高谷好一氏は京都大学東南アジア研究センター（現在の東南アジア研究所）に 1967 年～1995 年に在籍していた．主に東南アジアを対象に，地形・植生などの自然環境条件と地域社会の生業体系との関係をテーマとして詳細にフィールド調査を記録していた．本研究では，高谷氏によるフィールドノートの 1 つである『地域研究アーカイブズ フィールドノート集成 2 スマトラ 1984.10.19-1985.1.18 スマトラほぼ全域』[1]（以下高谷ノート）を対象としてい

[1] 東京大学史料編纂所
Historiographical Institute The University of Tokyo
[a)] t_yamada@hi.u-tokyo.ac.jp

[*1] http://www.cias.kyoto-u.ac.jp/project/kyodo25-3-1.html
[*2] http://www.cias.kyoto-u.ac.jp/project/kyodo25-3-1-2.html

ⓒ 2015 Information Processing Society of Japan

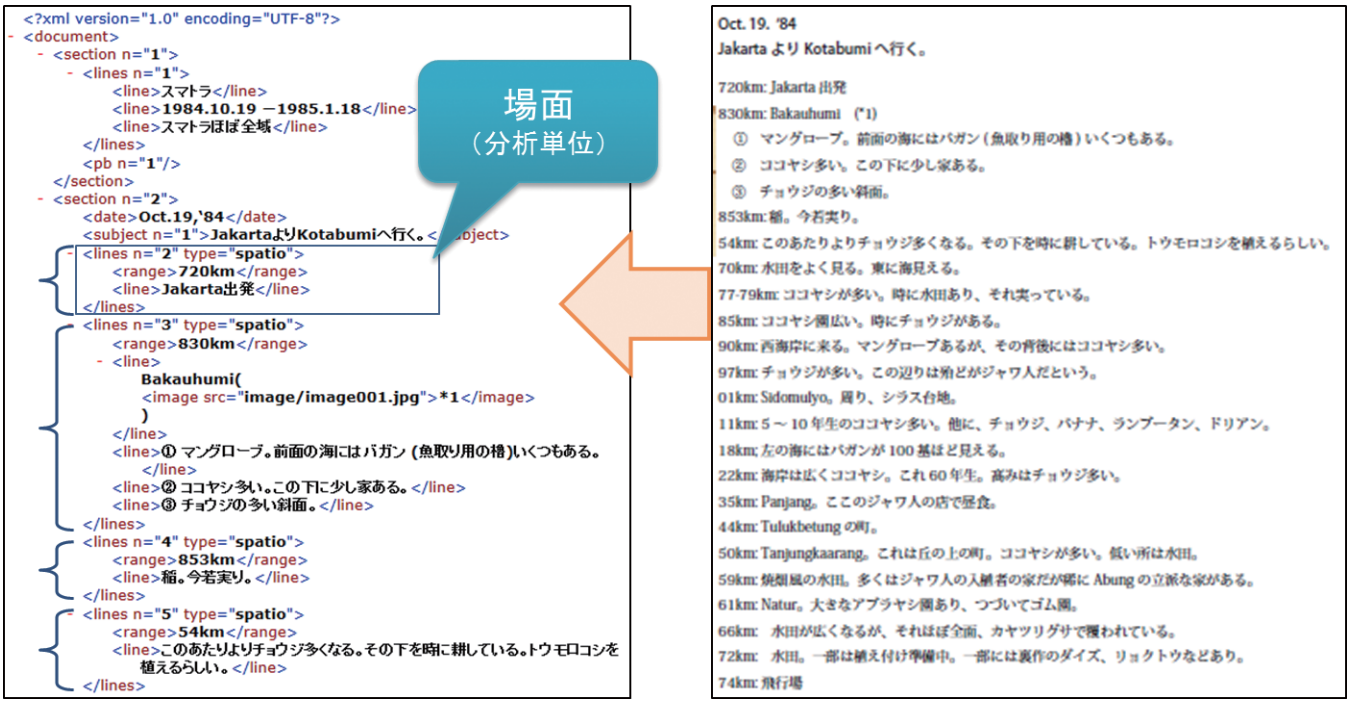


図 1 フィールドノートとテキスト
Fig. 1 Fieldnote and the text

る．高谷ノートはスマトラ島全域をカバーしており，このテキストは 197 ページ，延べ 165,757 文字である．

高谷ノートからのテキスト化は [2] にその作成について言及があり，高谷好一氏本人により，手書き資料のテキスト化を行い，情報の確認，写真・イラストの整理を行ったとある．

図 1(右) に高谷ノートの例を示す．この図のように，高谷ノートには，その日の記録の始めに日付（Oct.19.'84）があり，該当日の概要（Jakarta より Kotabumi へ行く。）が記されており，その後に，場面ごとの記録，スケッチ画像，写真画像が掲載されている．記録としては場面ごとの観察記録だけでなく，現地の人から聞いたことや調べたことなどもある．

3. 場面の特徴付け

整理された高谷ノートはテキスト付き PDF の形式であるため，全文検索程度であれば，直接 PDF 内のテキストを検索することで検索要求を満たすことができる．しかしながら，フィールドノート全体の把握や各場面の特徴は，フィールドノートを読んだ人にしか理解できないであろう．そこで，本研究では，フィールドノートの利活用のため，ある場面と “似ている場面” を検索することができ，フィールドノート全体を抽象化するための手法を導入する．これを**図 2** に示す．ここでは，1) テキストの構造化，2) 場面を特徴づける用語の抽出，3) 場面の特徴付けを行う．

1) については前節で述べた高谷ノートからテキストや画像を抽出し，XML で構造化していく．これを図 1 に示す．ここでは < 節 >–< 場面 > の階層構造を表現するためのテキスト構造化を行っている．節は section タグ，場面は lines タグ，場面の日付は date タグ，場面の概要は subject タグで示している．

2) は 1) で作成した XML から場面ごとの記録に関する記述を抽出し，それから名詞を抽出し，それを場面を特徴づける用語とした．高谷ノートは現代日本語での記述であるため，形態素解析器として mecab [*3]，形態素解析用辞書として IPADic [*4] を用いた．**図 3** は用語抽出の例を示す．用語抽出の対象は名詞であるが，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．また，連続する名詞はチャンクした．抽出対象の名詞の直後の接尾，[a-zA-Z]+な連続はチャンクした．また，出力は Bag-of-Words で表現した．これにより，抽出した用語とその出現頻度を表現することができる．

3) では，時間語彙，空間語彙，および場面のトピックにより表現する．時間語彙は前節の date タグの値をそのまま利用する．空間語彙は，場面の概要，もしくはテキスト中に出現する地名を用いて表現する．場面のトピックはテキスト中に明示されておらず，隠れていると仮定する．そこで，本研究では，場面のトピックを，トピックモデル LDA（Latent Dirichlet Allocation）[3] を用いて抽出することにした．LDA により 1 場面から複数の潜在トピックが生成され，さらに 1 トピックから複数の用語を生成する文書生成モデルとみなすことができる．これにより，潜在ト

*3 https://code.google.com/p/mecab/
*4 http://sourceforge.jp/projects/ipadic/

ⓒ 2015 Information Processing Society of Japan

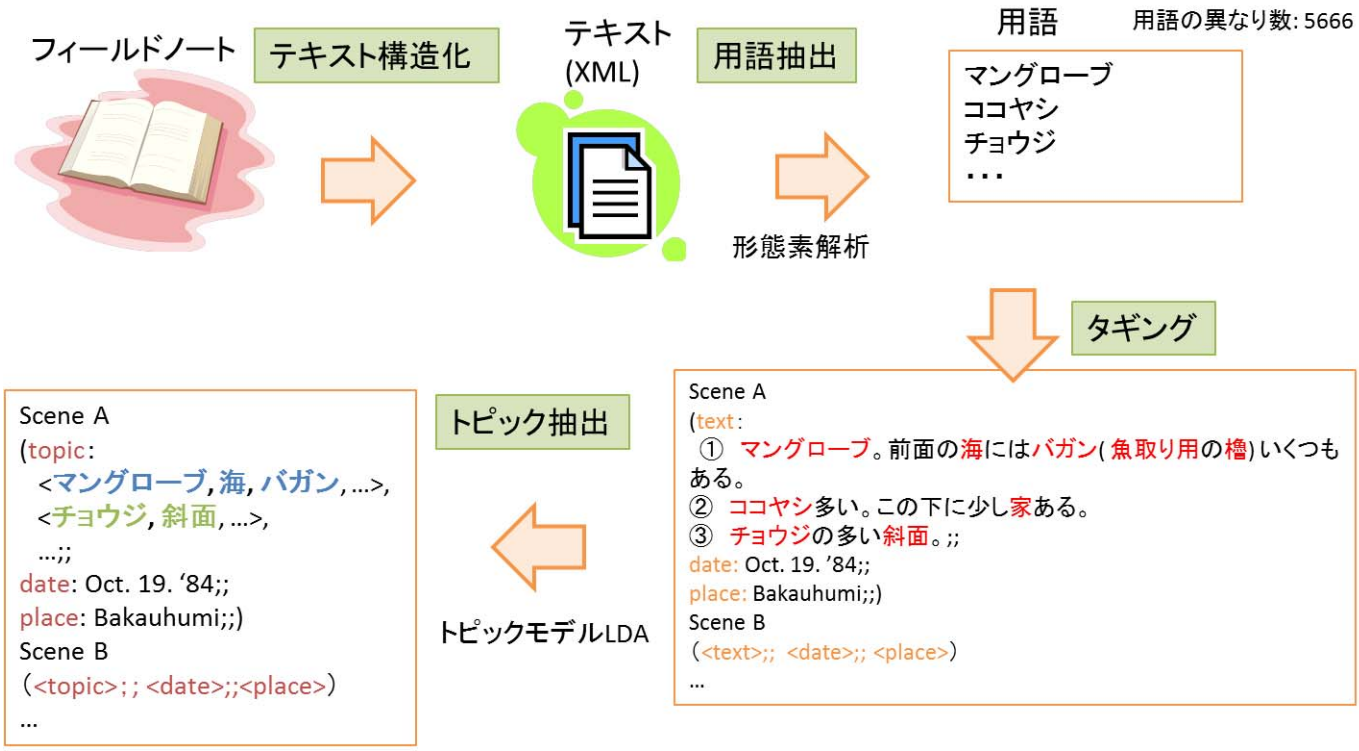


図 2 フィールドノートからのトピック検出
Fig. 2 Topic detection from a fieldnote

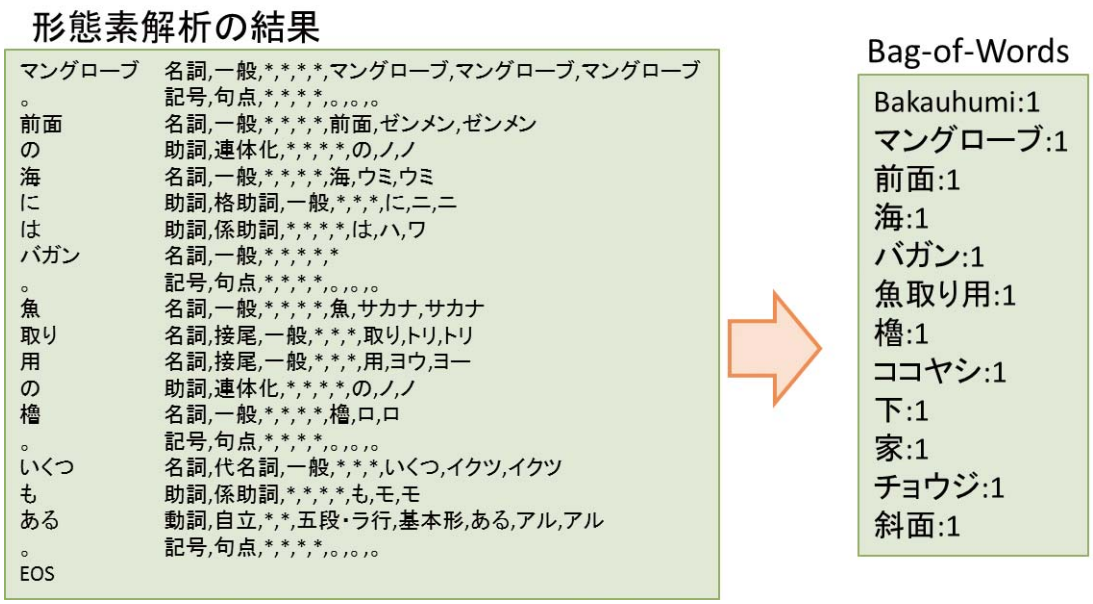


図 3 用語抽出の例
Fig. 3 Example of term extraction

ピックの検出，および，潜在トピックと用語の関係，潜在トピックと場面の関係を確率的に明示することができる．LDA は次式で表現することができる．

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

ここで，ここで α,β は LDA のパラメータ，$z = z_1, z_2, \ldots, z_C$ は潜在トピック，$\theta = \theta_1, \theta_2, \ldots, \theta_C$ は潜在トピックの生成確率，$Dir(\theta|\alpha)$ はディリクレ分布，$d = (w_1, w_2, \ldots, w_{|d|})$ は場面，w_n は用語，$|d|$ は場面 d の総単語数を示す．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．(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ラを用い

ⓒ 2015 Information Processing Society of Japan

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	kota	家	水田	Minangkabau	ゴム	幅	家	木	サゴ	人	炭	ゴム林	町	地区	オカボ
2	王宮	マングローブ	棚田	周辺	ゴム園	Kalua	ココヤシ	収穫	工場	父	窯	井戸	店	Transmigrasi	トウモロコシ
3	幅	中国人	稲	Bilis	タッピング	更新	コーヒー	実	Rp	Singapore	炭焼き小屋	内皮	中心	transmigrasi	コーヒー
4	枠	集落	周り	チガヤ原	手入れ	Encik Ibrahim	ゴム	油	サゴヤシ	サゴ	直径	土手	北	計画	焼畑
5	目	バガン	川	下り	植	Penyengat	周り	自分	中国人	土地	鶏	食堂	川沿い	高木	水田
6	Kapsa	右	幅	煉瓦	周り	Riau	村	幹	木	オランダ	松林	ココヤシ	肉	Bangkinang	斜面
7	樹脂	船	谷地田	裏作	アブラヤシ	kincil	ドリアン	周り	自分	人達	火事	大抵	Empang	ft	ゴム
8	水槽	Bekawan	村	Kampar川	Java	家	ランブータン	ココヤシ	水	中国人	入り口	播	区画	ニッパヤシ	畑
9	王	背後	山	Payakumbuh	desa	立方	バナナ	葉	tual	町	雨季	西進	湖	丸太	チョウジ
10	Batak	Melayu	柵	Tengku Bujang	丘	Payakumbu	マンゴー	大木	人	自分	中央	Raja	高み	禁止	バナナ

	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
1	M	魚	草	Empang	家族	牛	ton	ココヤシ	A	Minangkabau	漁業	水田	ゴム	Loc	泥炭
2	J	エビ	鍬	Jambi	Haji Ruppek	右	ゴム林	自分	C	オランダ	Rangsan	池	suku	Raja Kecil	粘土
3	村	網	田	下	Ugum	水牛	壁	人	中央	Melayu	サトウヤシ	魚池	Banjar語	Mandah	下
4	墓	Rp	苗代	島	Ungku Tugut	左	鳥	水路	B	人口	Desa	砂	丘陵	造船所	ココヤシ
5	水	inch	稲	左右	村	山	墓	ココヤシ園	土手	ゴム園	昼食	雑魚	村	Sultan	湿地
6	Kuala Enok	月	水田	Air Molek	ゴム	empang	松	Banjar	海岸	集落	森林	便所	Java人	島	所々
7	上流	坪	除草	Piring	サゴヤシ	クビキ	Muchtar氏	Bugis	Medan	Bengkalis	rumbia	区画水田	Kuantan	Kampong Ibrahim	水
8	中国	塩	月	Sago lemak	実り中	犁	Y	Sapat	Padang	野菜	人口	氾濫原	商業	Penyagat	マングローブ
9	家番号	目	水	kepuk	幹線道路沿い	犁先	オカボ	森	囲い	土地	作物	Batang Sumpur	所々	息子	ココヤシ園
10	Mageri	m inch	人	ジャワ人	Kuala Gete	翼	乾季	最初	大池	高地	名	Solok	Barabai	Haji Muhamad	土

図 4 LDA の結果
Fig. 4 The result of LDA

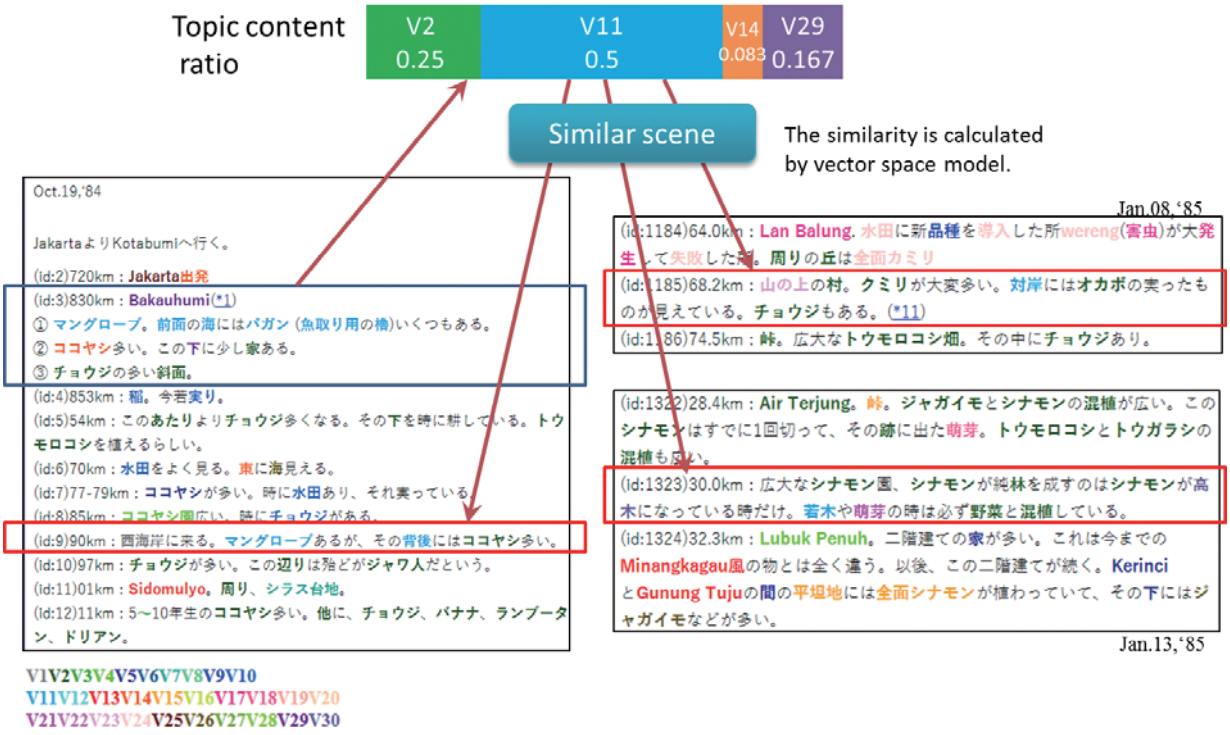


図 5 類似した場面の検索
Fig. 5 Search for similar scene

た解法が知られており [4]，本研究ではこれを用いて潜在トピック z を算出する．潜在トピックの算出により，場面とトピックの関係，およびトピックと用語の関係を定量的に明示することが可能となる．

4. 高谷ノートの潜在トピック

高谷ノートに対し，LDA による潜在トピックの検出を試みた．LDA のパラメータとして潜在トピック数を 30，ギ

ⓒ 2015 Information Processing Society of Japan

ブスサンプリングの回数を 2000 とした．

図 4 は LDA による潜在トピックの検出結果のうち，30 のトピックとそのトピックを構成する上位 10 件の用語を示す．それぞれのトピックが何を示すかを端的に表現することは難しいが，例えば，V2 は水辺での土地利用関係，V5 はゴム関係，V9 はサゴヤシ関係，V18 および V21 は農耕技術関係を示すと考えられる．また，“水田” は V3，V15，V18，V27 に出現するが，トピックを構成する他の用語から，V3 は風景，V15 は作物，V27 は池・水田利用の関するトピックと考えることができるため，同じ用語でもトピックが異なれば，テキスト中での意味合いが異なると考えられる．

LDA の結果により，各トピックに属する用語の出現頻度をもとに，場面ごとに各トピックの出現頻度を算出することができる．これに対して，ベクトル空間モデル [5] を適用することで，定量的に場面間の類似度を算出することができる．

ベクトルの各要素はトピックを構成する各用語の出現頻度の総和となる．これをここではトピックの出現頻度として扱う．ある場面 x におけるトピック i の重み $weight(x)_i$ を次式に示す tf-idf 重み付けにより計算する．

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

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 を含む場面の個数，N は場面数を示す．tf-idf 重み付けは i が少数の場面で多く出現するとその重みを大きくし，反対に，多くの場面で出現すると重みを小さくする．重みの大きなトピックはその場面のテキストの性質を特徴付けると考えられる．

場面間の類似性を量化するため，次式で求める 2 つのベクトル x と y のコサイン類似度（cosine similarity）を計算する．

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

本研究では，この値を場面間の類似性を示すスコア（類似度）として用いることにした．

$$score(x,y) = sim(x,y) \quad (4)$$

また，検索結果として得られる類似した場面は，この $score(x,y)$ を用いることでランキングすることができる．**図 5** は類似した場面のうち，上位 3 件の検索結果を示している．

5. 考察

得られれた潜在トピックは何を意味するのか．地域研究者のような専門家により，もしくは専門家でなくても，“見れば” わかるトピックもあれば，専門家であっても表現しにくい，もしくは不可思議なトピックが検出されることも多々ある．記述した本人にしかわからないことや，記述した本人以外にはわかること，があるかもしれない．また，まだ誰も気づいていない地域研究には重要な事柄を示唆しているかもしれない．フィールドノートは主観的に記述されたものであり，その分析はようやく始まった段階である．いわば研究としての宝庫であるとも考えられる．

トピックを検出していくうえで，用語の抽出は重要なタスクである．図 4 に示したように，分析を進めていくうえで不必要な用語が出現している．これらの除去を推し進めて行く必要がある．反対に，“足りない” 用語がある．例えば，“オカボ” が出現したとき，それは何を意味しているか，という情報が不足している．その場面ではオカボが多いのか，少ないのか，（そこにあっても不思議ではないのに，もしくは以前はあったが）全く無いのか，のように用語を修飾することでその意味合いが変わってくる．これに応じて，おそらくは検出されるトピックも変化すると考えられる．

現時点では空間語彙は，場面の概要から抽出した地名を用いている．実際には場面のテキストに地名が出現している．それらの地名の殆どは，今回の形態素解析では未知語として解析される．そのため，地名の抽出を進めていく必要があると考えている．地名の抽出では，手作業で抽出していくことも方法の 1 つではあるが，フィールドノートはこの高谷ノートだけではない．汎用的に，継続的にフィールドノートを利用していくのであれば，自動的に抽出できる方法を模索すべきである．

6. おわりに

本研究では，地域研究を推進するため，これまで利用されることが殆ど無かったフィールドノートに対し，時間語彙・空間語彙とともにトピックモデルを用いてテキストを特徴づけることでその利活用を行うための方法について述べた．

今後は，単なるテキスト分析だけでなく，さらに検索機能を充実させることで，フィールドノートという資料を地域研究を進める上で重要な素材としていく．そこで，これまでの取り組みに関して，いっそう評価を進め，より精度の高い分析方法を目指す．テキスト分析結果の深化させるため，用語抽出の精度向上，トピックとして検出されたものが何であるかを分析していく方法，フィールドノート内の時空間変化追跡を可能とするための可視化などに取り組んでいく必要があると考えている．また，フィールドノートを利活用していくうえで，単なる全文検索可能な公開方法ではなく，テキスト分析の成果を広く利活用可能な形式でのデータ構造を探求していくことも重要であろうと考えている．

一方，共有するための機能・分析性能を高めるだけ考え

ⓒ 2015 Information Processing Society of Japan

るだけでは新たな進展はないと考えている．もっとも重要なことは地域研究者がフィールドノートをどう利用したいのか，であろう．これが明確化されていくことで，検索シナリオをうまく設定していくことが可能となり，新たな課題が見出されていくことになると考えている．

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「地域研究データにおけるトピックの検出と時空間変化に関する研究」の助成を受けたものによる．

参考文献

[1] 高谷好一：地域研究アーカイブズ フィールドノート集成 1〜5, CIAS Discussion Paper Series, 京都大学地域研究統合情報センター, No.22(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.14–21(2014).
[3] D.M.Blei,A.Y.Ng, and M.I.Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3, pp.993-1022(2003).
[4] T.L.Griffiths and M.Steyvers: Finding scientific topics, Proc. of the National Academy of Sciences of the United States of America, vol.101, pp.5228-5235(2004).
[5] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information Processing and Management: an International Journal*, Vol. 24, No. 5, pp. 513–523 (1988).

ⓒ 2015 Information Processing Society of Japan