LDA を用いた実生活 Tweet の二段階抽出法

山本　修平†　　佐藤　哲司††

† 筑波大学大学院図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学大学院図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †{yamahei,satoh}@ce.slis.tsukuba.ac.jp

あらまし　身近な出来事や感心事を投稿し共有する Twitter 上には，食事や交通，災害，気象など，様々な生活の局面で有益な Tweet が数多く投稿されている．本研究では，これらのような有益な Tweet を抽出するために，二段階抽出法を提案する．第一段階では，潜在的ディリクレ配分法（LDA）を用いて，大量の Tweet から複数のトピックを抽出する．第二段階では，局面がラベル付けされた少数の Tweet を用いて，トピックと局面の対応付けを構築する．精度を高めるため，特徴語の重みは情報利得によって計算する．プロトタイプシステムを実装し評価実験をした結果，未知の Tweet から局面を抽出できることを明らかにした．

キーワード　Twitter, 実生活, LDA, 二段階抽出法

Two Phase Extraction Method for Extracting Real Life Tweet using LDA

Shuhei YAMAMOTO† and Tetsuji SATOH††

† Graduate School of Library, Information and Media Studies, University of Tsukuba
1–2 Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
†† Faculty of Library, Information and Media Science, University of Tsukuba
1–2 Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
E-mail: †{yamahei,satoh}@ce.slis.tsukuba.ac.jp

Abstract　Recently, lots of users share their current events and opinions by using the Twitter. Hence, some of these tweets are beneficial in several aspects of user's real life, i.e. eating, appearance, living, disasters, and so on. In this paper, we propose a two phase extracting method for selecting beneficial tweets. In the first phase, many topics are extracted from a sea of tweets using Latent Dirichlet Allocation (LDA). In the second phase, associations between many topics and fewer aspects is built using a small set of labeled tweets. To enhance accuracy, the weight of feature words is calculated by information gain. Our prototype system demonstrates that the proposed method can extract the aspects of each unknown tweet.

Key words　Twitter, Real Life, LDA, Two Phase Extraction

1. はじめに

現在，知識共有コミュニティサイトやブログ，マイクロブログなど，多くの情報共有サービスが存在している．Twitter(注1)は，最も広く普及しているマイクロブログであり，最大 140 文字の短い文章からなる記事が投稿される．Twitter では，多くのユーザがリアルタイムに，自分の経験や意見，また日常生活の中のイベントなど，身近な「今」を投稿しているため，最新かつ有益な記事が多い．例えば，電車の遅延情報やスーパーの特売情報といった，地域性が高く新鮮な記事がある．本論文では，このような記事を「実生活 Tweet」と呼ぶ．大量の Tweet の中から実生活 Tweet を抽出することは，ユーザの生活を支援するために重要な研究課題であると考えられる．

実生活 Tweet が実際にユーザの生活を支援した例として，2011 年 3 月に起きた東日本大震災がある．地震が起きた直後，被災地では断水や食料供給の不足，電車の運行中止など，大きな混乱が生じた．その際，給水や食料配布が行われる場所，電車の運行情報などについて書かれた有益な Tweet が数多く投稿され，被災地のユーザを支援したとの報告がなされている [10].

このように，ユーザにとって有益な実生活 Tweet は，Twitter

(注1)：http://twitter.com

にますます投稿されるようになってきた．しかし，実生活 Tweet 以外の Tweet も数多く存在する．特に,「ありがとう」や「なるほど」のような，誰かの投稿に対する相槌や共感といった，ユーザの生活を直接支援しない Tweet が多い．このような Tweet は，実生活 Tweet の発見を妨げる原因となっている．

人々の生活は地域によって異なりがある．実生活 Tweet も，地域による異なりの影響を受ける可能性がある．例えば，電車の遅延情報を Twitter に投稿するとき，つくば市では「つくばエクスプレス」という単語が Tweet 内に出現することが考えられる．対して，那覇市では「モノレール」や「ゆいレール」という単語が Tweet 内に出現することが考えられる．以上のように，地域によって交通手段は異なるため，生活に関連する単語を人手で列挙して検索するだけでは，あらゆる地域の実生活 Tweet を抽出することは困難であると考えられる．本論文では，このような地域の異なりによって生じる問題を「地域依存性」と呼ぶ．

実生活 Tweet は，生活の様々な局面に対応できる．例えば,「電車が来ない」という Tweet は「交通」の局面に分類され，電車に乗ろうとしているユーザを支援できる.「今日は全商品半額です !!」という Tweet は「消費」の局面に分類され，買い物に行こうとしているユーザを支援できる．著者らは先行研究で，Wikipedia の「地域コミュニティ」(注2) と「生活」(注3) を参考に，実生活の局面を表 1 に示す 14 の局面に分類している [11]．分類した 14 の局面には，地域依存性が高い局面と低い局面とがあると思われる．

本論文では，地域に住むユーザを支援することを目的に，実生活 Tweet の二段階抽出法を提案する．第一段階では，大量の Tweet から潜在的ディリクレ配分法（LDA）を用いて，トピックを抽出する．第二段階では，少数のラベル付けされた Tweet を用いて，トピックと局面（表 1）の対応関係を構築する．精度を高めるため，情報利得を用いて特徴語の重みを計算する．

本論文の構成を以下に示す．第 2 章は関連研究について述べる．第 3 章は二段階抽出法について説明する．第 4 章は局面の抽出精度と地域依存性について評価し，第 5 章で考察を行う．最後に，第 6 でまとめと今後の課題を述べる．

2. 関 連 研 究

実生活 Tweet はユーザの経験や知識，地域の情報からなる．文書から経験情報を抽出するために,「経験マイニング」に関する研究がいくつか行われている．Kurashima ら [6] は，人間の経験を {状況，行動，主観} からなる情報と捉え，文章中から {時間，空間，動作，対象，感情} を自動抽出する手法を述べている．Inui ら [5] は，人間の経験を {時間，極性，話者態度} の観点から, {トピック，経験主，事態表現，事態タイプ，事実性} の各項目に索引付けする枠組みを提案している．これらの経験マイニングに関する研究は，ブログなどの長い文書に対して効果的であるが，Twitter に投稿される記事のような，非常に短い文書に対しては適切ではないと考えられる．加えて，Twitter に投稿される記事は，主語や目的語がよく省略されるため，経験マイニングをより難しくしている．

表 1　実生活の局面

局面	典型的な単語
服飾	衣服, 服装, 着る, 装飾, 化粧, 理髪, 衣装 ...
交流	約束, 出会い, 招待, 友人, 誘い, 勧誘, 飲み会 ...
災害	洪水, 竜巻, 地震, 火事, 津波, 二次災害 ...
食事	料理, 外食, 食べ物, レストラン, ジャンクフード ...
行事	祭り, 冠婚葬祭, 日程, 開催日, 学園祭, 文化祭 ...
消費	購入, 買う, 注文, 安売り, 特売, ショッピング ...
健康	風邪, 体調, 怪我, 痛み, 健康法, 病気予防 ...
趣味	余暇, 娯楽, おもちゃ, 音楽, テレビ, ゲーム ...
居住	掃除, 家具, 洗濯, 住まい, 隣人, アパート ...
地域	観光, 地域情報, 地理情報 ...
学校	勉強, 宿題, 課題, 試験, テスト, 資格, 研究 ...
交通	電車, バス, 飛行機, 時刻表, 渋滞, 混雑, 遅延 ...
気象	天気, 気温, 湿度, 風, 花粉, 雨量, 空模様 ...
労働	アルバイト, 研修, 就職活動, 営業, 仕事 ...

Twitter に関する研究は数多く行われている．Ramage ら [8] は，ハッシュタグなどのラベルを教師情報として利用できるように，LDA を拡張した Labeled LDA を用いることで，推薦の性能が向上することを示している．Bollen ら [2] は，Tweet の 6 次元の mood（tension，depression，anger，vigor，fatigue，confusion）について分析した結果，株価など実世界の出来事と相関があることを明らかにしている．Sakaki ら [9] は，Twitter ユーザをセンサーと想定し，地震などの現実世界でリアルタイムに起きるイベントを発見する手法を明らかにしている．Zhao ら [12] は，一つの Tweet は一つのトピックの内容を表すという仮説に基づいて，Twitter-LDA と呼ばれるモデルを提案し，Tweet 集合をトピック毎に分類した後で，トピックの内容を表すキーワードやフレーズを抽出している．Mathioudakis ら [7] は，収集した Tweet からバーストキーワードを見つけ出し，キーワードの共起を用いてグルーピングすることで，リアルタイムに変動するトレンドの発見をしている．

本論文では，人々の生活に有益である実生活 Tweet を抽出するための手法を提案する．実生活 Tweet は，ユーザの経験だけでなく，ユーザの知識に基づく情報も対象としているため，従来の研究とは大きく異なる．

3. 実生活 Tweet の二段階抽出法

実生活 Tweet は 1 章で述べたように，様々な局面を含んでいる．従って，全ての局面に関連するキーワードを列挙することは困難である．また，経験マイニングを用いたルールベースの解析手法は，Twitter に投稿される記事が短い文章で構成されているため，経験マイニングの精度を上げることは難しいと考えられる．

本論文では，二段階抽出法を提案する．二段階抽出法の概要を図 1 に示す．第一段階では，LDA を用いて大量の Tweet からトピックを抽出する．LDA は大量の文書集合をクラスタリン

(注2) : http://ja.wikipedia.org/wiki/地域コミュニティ
(注3) : http://ja.wikipedia.org/wiki/生活

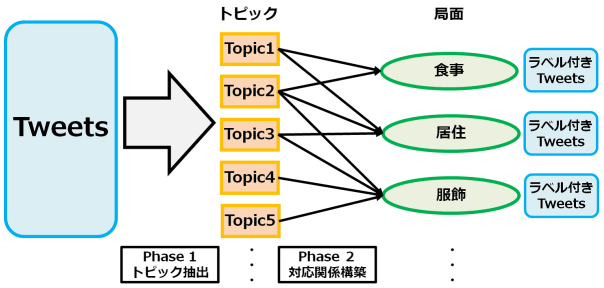

図 1　二段階抽出法

グするための，教師無し学習モデルである [1]．第二段階では，トピックと局面の対応関係を構築する．

この章の構成を以下に示す．第 3.1 節は LDA を用いたトピック抽出法について述べる．第 3.2 はトピックと局面の対応関係を構築する．第 3.3 は Tweet からの局面の抽出法について説明する．

3.1　LDA を用いたトピック抽出

第一段階のトピック抽出で使用する潜在的ディリクレ配分法（LDA）とは，Blei ら [1] によって考案された，確率トピックモデルを用いた潜在的トピック作成手法である．潜在的トピックモデルは，文書を複数のトピックからの混合分布であると仮定し，各トピックを単語の確率分布で表現する．

各文書はトピック分布 θ を持ち，単語はトピック z を選択した後，そのトピック z の単語分布 ϕ に従い生成される．ディリクレ事前分布を Dir()，多項分布を Multi() とすると LDA のモデル生成過程は以下で手順で与えられる．ここで，α はパラメータ θ が従うディリクレ事前分布のパラメータ，β はパラメータ ϕ が従うディリクレ事前分布のパラメータを示す．

（1）　文書 d について Dir(α) から多項分布パラメータ θ_d をサンプリング．

（2）　トピック t について Dir(β) から多項分布パラメータ ϕ_t をサンプリング．

（3）　文書 d に，N_d 個の単語があったとき，j 番目の単語 $w_{d,j}$ について

（a）　Multi(θ_d) から $z_{d,j}$ をサンプリング．

（b）　Multi($\phi_{t,j}$) から $w_{d,j}$ をサンプリング．

LDA では，教師無し学習によって，文書・単語空間からトピック集合 Z を推定する必要がある．推定方法には，差分ベイズ推定法，ギブスサンプリングなどがある．本論文では，LDA で一般的に使用されている崩壊型ギブスサンプリングを用いる．崩壊型ギブスサンプリングを用いたとき，文書 d の n 番目の単語 $w_{d,n} = v$ のトピック $z_i = k$ の更新式は以下の通りである．

$$P(z_i = k|Z_{-i}, W) = \frac{N^d_{k-i} + \alpha}{N^d_{-i} + T\alpha} \cdot \frac{N^d_{k-i} + \beta}{N^d_{k-i} + W\beta} \qquad (1)$$

ここで，-i はトピック集合全体から i（d 番目の文書の n 番目の単語）分を除くことを示す．N^d_k は，文書 d において，トピック k が割り当てられた回数，N^d は文書 d において単語が生成された回数，N^v_k はトピック k において単語 v が出現する回数，N_k は，トピック k に出現する単語の総数である．T はトピックの種類数，W は単語の語彙数である．

崩壊型ギブスサンプリングによって得られたサンプルから，文書 d において，トピック k が生成される確率 $\hat{\theta^k_d}$，トピック k から単語 w が生成される確率 $\hat{\phi^w_k}$ は以下の通りである．

$$\hat{\theta^k_d} = \frac{N^d_k + \alpha}{N^d + T\alpha} \qquad \hat{\phi^w_k} = \frac{N^v_k + \beta}{N_k + W\beta} \qquad (2)$$

本論文では文書集合を Tweet の集合と考え，潜在的トピックモデルを生成する．

3.2　トピックと局面の対応関係構築

トピックと局面の対応関係を構築するため，少数のラベル付けされた Tweet（正解データ）を用意する．局面 a としてラベル付けされた Tweet を形態素解析し，得られた単語の集合を W_a とする．ここで，各局面を特徴付ける単語の重みを計算するために，情報利得を用いる．これにより，特徴選択において良い特徴であるかどうかを数値で表現することができる．単語 w の情報利得 $IG(w)$ は，以下の式から計算される．

$$IG(w) = H(A) - (P(w)H(A|w) + P(\bar{w})H(A|\bar{w})) \qquad (3)$$

ここで，A は全ての局面を意味する．$P(w)$ は全ての Tweet の中で単語 w が出現する確率，$P(\bar{w})$ は全ての Tweet の中で単語 w が出現しない確率である．$H(A|w)$ は単語 w が出現する時の，全局面 A における条件付きエントロピー，$H(A|\bar{w})$ は単語 w が出現しない時の，全局面 A における条件付きエントロピーである．$IG(w)$ の値が高い時，単語 w は良い特徴であることを意味する．

トピック t と局面 a の関連度 $R(a,k)$ は，以下の式から計算される．

$$R(a,t) = \frac{1}{length(W_a)} \sum_{w \in W_a} IG(w) * p_{w,t} \qquad (4)$$

$p_{w,t}$ は，LDA を用いて抽出したトピック t における単語 w の生起確率である．$length(W_a)$ は単語集合 W_a の大きさを表す．この式は，単語の生起確率と情報利得を用いて関連度を算出する．

値を 0 から 1 の範囲にするため，各局面で関連度 $R(a,t)$ を正規化する．ここで，各局面で正規化した関連度 $\hat{Ra}(a,t)$ と，各トピックで正規化した関連度 $\hat{Rt}(a,t)$ を計算する．二つの関連度は，以下の式から計算される．

$$\hat{Ra}(a,t) = \frac{R(a,t)}{\sum_{t \in T} R(a,t)} \qquad \hat{Rt}(a,t) = \frac{R(a,t)}{\sum_{a \in A} R(a,t)} \qquad (5)$$

T は LDA を用いて抽出した全てのトピック，A は全ての局面である．$\hat{Ra}$ は，局面がどのトピックによって表現されるかを示す指標であり，$\hat{Rt}$ は，トピックがどの局面を支持しているかを示す指標である．

正規化された関連度 $\hat{Ra}(a,t)$ が，各局面 a における閾値 $\gamma(a)$ を超えた時，トピックと局面の対応関係を構築する．各局面 a の閾値は，以下の式で求められる．

$$\gamma(a) = \max_{t \in T} \hat{Ra}(a,t) - std(\hat{Ra}(a,T)) * d \quad (6)$$

$std(\hat{Ra}(a,T))$ は，局面 a と全トピック T の関連度 $\hat{Ra}(a,T)$ の標準偏差である．パラメータ d を大きくするとより多くのトピックが局面に関連付けらる．第 4 章では d を変化させて，抽出精度を評価する．

3.3 実生活 Tweet の抽出

実生活 Tweet を抽出するため，トピックと局面の対応関係を用いる．未知の Tweet と各局面のスコアは，以下の式で算出する．

$$Score(a) = \sum_{t \in T_a} \sum_{w \in W} p_{w,t} * \hat{Ra}(a,t) * std(\hat{Rt}(A,t)) \quad (7)$$

ここで，W は未知の Tweet から抽出した単語の集合である．T_a はトピックと局面の対応関係構築において，関連度 $\hat{Ra}(a,t)$ が閾値 $\gamma(a)$ を超えたトピック t の集合である．$std(\hat{Rt}(A,t))$ は，トピック t と全局面 A の関連度 $\hat{Rt}(A,t)$ の標準偏差を意味する．$std(\hat{Rt}(A,t))$ が高いとき，そのトピックは特定の局面を強く支持しており，局面にとって有用なトピックであると言える．

未知の Tweet から，以下の式を満たす局面を抽出する．

$$aspect = \arg\max_{a \in A} Score(a) \quad (8)$$

4. 評 価 実 験

第 3 章で提案した二段階抽出法による，局面の抽出精度を評価する．実生活 Tweet の地域依存性について評価するため，つくば，札幌，京都，沖縄の 4 地域で集めた Tweet 集合から，LDA を用いてトピックを抽出する．また，高い精度を示す局面と，低い精度を示す局面の原因を明確にするため，局面に高い関連度で繋がるトピックにも着目する．

以下．第 4.1 節は，評価実験に用いたデータセットとパラメータについて述べる．第 4.2 節は評価尺度について説明し，第 4.3 節は実験結果について議論する．

4.1 データセットとパラメータ設定

4.1.1 データセット：トピック抽出のための Tweet

LDA を用いてトピックを抽出するため，大量の Tweet 集合を用意する．実験は，2012 年 4 月 15 日から 2012 年 8 月 14 日の間に，日本語で Twitter に投稿された Tweet を用いる．その中から，つくば，札幌，京都，沖縄の 4 地域の Tweet を抽出する．抽出条件は，各 Tweet のロケーション情報に日本語表記の地名（例：「つくば」），あるいはアルファベット表記の地名（例：「Tsukuba」）が入力されている Tweet とした．各地域で，抽出した Tweet の数を表 2 に示す．

表 2　地域別 Tweet 数

地域	つくば	札幌	京都	沖縄
Tweet 数	1,966,746	2,491,168	2,390,553	2,097,016

表 3　人手判定の分類一致 Tweet 数

局面	二人以上一致	三人一致
服飾	97	73
交流	94	37
災害	97	69
食事	120	82
行事	78	11
消費	74	14
健康	98	68
趣味	105	64
居住	96	66
地域	56	20
学校	107	80
交通	104	72
気象	105	71
労働	79	45
非実	72	31
合計	1,382	803

4.1.2 データセット：実生活 Tweet

LDA で抽出したトピックと，生活の局面の対応関係を構築するため，局面がラベル付けされた Tweet を用意する．1,500 件の Tweet に対して，第一著者と他 2 名の合計 3 名の実験者で人手判定を行った．実験者にはガイドラインとして，表 1 に示す局面に含まれる典型的な単語と，その局面に分類される例文（各局面 1 文ずつ）と，それが分類された理由を提示した．人手判定では，各 Tweet に対して最も適切な局面を一つだけ付与することとした．いずれの局面にも適さないと判断した場合，「非実生活」を付与することとした．なお，1,500 件の Tweet はいずれもロケーション情報に「つくば」あるいは「Tsukuba」と表記されたものであり，実験者 3 名ははいずれも「つくば市」在住の大学生である．

人手判定によって分類が一致した Tweet 数を表 3 に示す．実験者間の κ 値 [3] は，実験者 A と実験者 B の κ 値が 0.609，実験者 A と実験者 C の κ 値が 0.645，実験者 B と実験者 C の κ 値が 0.664 となった．κ 値の平均は 0.639 であり，高い一致（substantial）が得られた．

トピックと局面の対応関係の構築，及び抽出精度の評価では，被験者 3 名のうち，2 名以上の判定が一致した Tweet を用いる．

4.1.3 パラメータ設定

LDA は，いくつかのパラメータを設定する必要がある．関連研究 [4] を参考にハイパーパラメータである α は $50/T$，また β は 0.1 とした．LDA のイテレーション回数は常に 100 とし，トピック数は 10, 20, 50, 100, 200, 500, 1,000 と変化させた．

4.2 評 価 尺 度

4.2.1 トピック数

最適なトピック数を決定するため，各局面間の JS Divergence を用いて，ある一つの局面と他の局面との類似度を計算する．JS Divergence 値が 0 の時，二つの局面の確率分布は同じであることを意味する．本論文の場合は，各局面間の JS Divergence の合計値が最大であるとき，最適なトピック数であるとした．

JS Divergence の合計値 JS_{sum} は，以下の式で求められる．

$$JS_{sum} = \sum_{(\forall P, \forall Q) \in A} JS(P||Q) \tag{9}$$

$$JS(P||Q) = \frac{1}{2}\left(\sum_x P(x)\log\frac{P(x)}{R(x)} + \sum_x Q(x)\log\frac{Q(x)}{R(x)}\right)$$

ここで，P と Q は確率分布を，A は全ての局面である．R は確率分布 P と Q の平均であり，$R = \frac{P+Q}{2}$ である．

4.2.2 抽出精度

局面の抽出精度は，10 分割交差検定によって評価する．10 分割されたうちの 9 割の Tweet で局面とトピックの対応関係を構築し，残りの 1 割の Tweet で抽出精度を評価する．以上の操作を 10 回繰り返し，抽出精度の平均を算出する．

4.3 実験結果

4.3.1 トピック数の決定

最適なトピック数を決定するため，式 (9) に示す JS_{sum} を算出した．トピック数を変化させたときの JS_{sum} の値を，表 4 に示す．この表から，最大の JS_{sum} となるトピック数 500 を用いて，以降の評価を行うこととした．

表 4 各トピック数のときの JS_{sum}

トピック数	10	20	50	100	200	**500**	1,000
JS_{sum}	6.81	11.74	16.67	20.10	21.63	**22.91**	22.31

4.3.2 抽出精度

トピックとの対応関係を決定する閾値 d にともなう抽出精度の変化について，地域 [つくば] トピックの時を図 2 と図 3，地域 [沖縄] トピックの時を図 4 と図 5 に示す．いずれの図も，横軸はトピックと局面の対応関係を決定する閾値 d，縦軸は精度である．両地域とも，d の増加に従って精度が大きく変わっているが，$d \geqq 15$ で安定した精度が得られている．学校の局面は，地域 [つくば] トピックでは d の増加に従って精度が高くなっていくが，地域 [沖縄] トピックでは $d = 8$ で精度が高くなり，更に d が増加するにつれて精度が低くなっていることが分かる．

$d = 24$ のとき各地域のトピックで抽出精度を算出した結果を，表 5 に示す．全ての地域において，災害，食事，居住，交通，気象の局面の精度が高いことが分かる．対して，全ての地域において，交流，健康，地域，非実生活の抽出精度は低かった．服飾の局面は，他の地域に比べて地域 [沖縄] だけ精度が低く，地域と学校の局面は，地域 [つくば] のみ精度が高い結果となった．行事の局面は，地域 [つくば] 以外の地域の精度が高いことも分かる．全ての局面の抽出精度の平均は，地域 [つくば] が最も高く，地域 [沖縄] が最も低い結果となった．

局面に高い関連度 $\hat{Ra}$ で関連付けられたトピックを，表 6 に示す．表中の括弧内の数字は $\hat{Ra}$ である．順位は，$std(\hat{Rt}(A,t))$ でトピックを降順に並べた時の順位を示している．災害と最も強く関連付けられたトピックは topic144 であり，関連度 $\hat{Ra}$ は 0.1502，$std(\hat{Rt}(A,t))$ の順位は 1 位である．topic465 と topic320 は災害，行事，地域，交通，気象の局面と，関連度 $\hat{Ra}$ が高いものから 5 位以内に関連付けられていることが分かる．

表 5 各地域における，$d = 24$ のときの抽出精度

局面	つくば	札幌	京都	沖縄	平均
服飾	0.344	0.356	0.378	0.156	0.309
交流	0.156	0.100	0.144	0.056	0.114
災害	0.656	0.711	0.556	0.611	0.634
食事	0.633	0.625	0.650	0.633	0.635
行事	0.471	0.657	0.586	0.657	0.593
消費	0.286	0.270	0.300	0.357	0.303
健康	0.256	0.278	0.233	0.267	0.259
趣味	0.410	0.490	0.410	0.400	0.428
居住	0.556	0.511	0.533	0.567	0.542
地域	0.320	0.080	0.100	0.040	0.135
学校	0.570	0.300	0.470	0.310	0.412
交通	0.620	0.620	0.680	0.580	0.625
気象	0.670	0.700	0.800	0.820	0.748
労働	0.429	0.500	0.457	0.486	0.468
非実	0.114	0.114	0.100	0.057	0.096
平均	0.433	0.421	0.426	0.400	0.420

5. 考　　察

5.1 抽出精度に関する考察

図 2 と図 3 より，地域 [つくば] トピックで災害，食事，居住，学校，交通，気象の局面において高い精度を示していることが分かる．表 5 から，災害と交通は，$\hat{Ra}1$ 位のトピックの $std(\hat{Rt}(A,t))$ が，$\hat{Ra}2$ 位のトピックに比べてかなり高い順位を示していることが分かる．$std(\hat{Rt}(A,t))$ が高いトピックは，特定の局面に対して強く支持していることを示す．$\hat{Ra}$ と $std(\hat{Rt}(A,t))$ が共に高いとき，そのトピックは特定の局面にのみ，強い関連度を持っていると言える．以上のことから災害，交通は，その局面を表す典型的なトピックが高い関連度で関連付けられていると考えられる．

食事，居住，学校，気象は，$\hat{Ra}1$ 位のトピックにおける $std(\hat{Rt}(A,t))$ の順位が，$\hat{Ra}2$ 位以下のトピックに比べて大きく離れていない．しかし，$\hat{Ra}2$ 位以下のトピックもその他のトピックに比べて，高い順位を示していることが分かる．以上のことから，食事，居住，学校，気象は，$\hat{Ra}1$ 位のトピックだけではその局面を表現することができていないが，$\hat{Ra}2$ 位以下のトピック数個によって，局面を適切に表現することができていると考えられる．

図 4 と図 5 より，地域 [沖縄] トピックにおいては災害，行事，食事，居住，交通，気象の局面において高い精度を示していることが分かる．表 5 から，災害，食事，居住，交通，気象は 4 つの地域いずれも高い精度を示しているが，行事は地域 [つくば] のみ精度が低いことが分かる．表 6 より，行事は topic320 が $\hat{Ra}1$ 位，topic465 が $\hat{Ra}2$ 位の関連度を持っていることが分かる．topic320 と topic465 について見ると，この二つのトピックは，災害，地域，交通，気象とも $\hat{Ra}5$ 位以内の関連度で関連付けられている．また，topic320 と topic465 の $std(\hat{Rt}(A,t))$ の順位は，他のトピックに比べて低いことが分かる．以上のことから，topic320 と topic465 は他の局面にも出現しやすいト

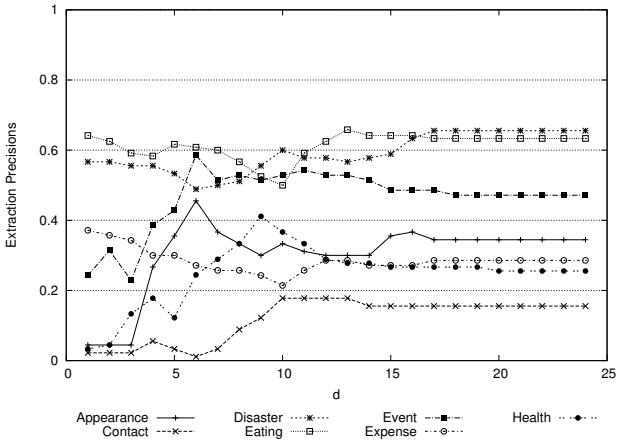

図 2　地域 [つくば] トピックの時の抽出精度：服飾ー健康

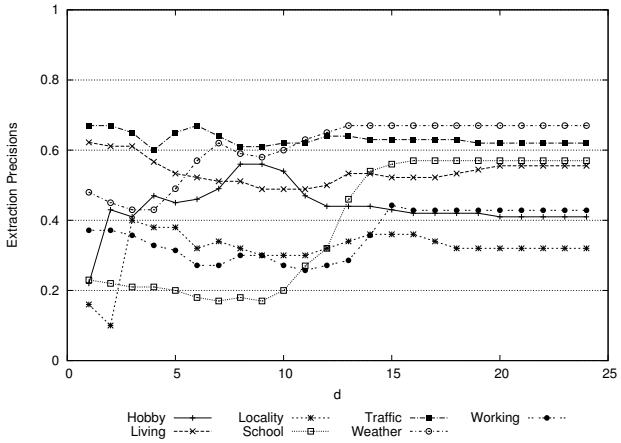

図 3　地域 [つくば] トピックの時の抽出精度：趣味ー労働

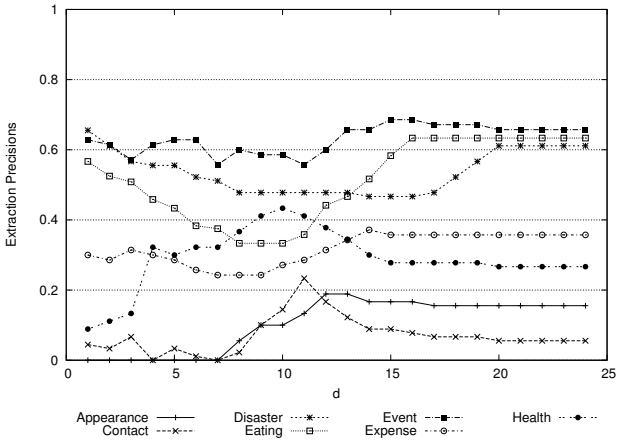

図 4　地域 [沖縄] トピックの時の抽出精度：服飾ー健康

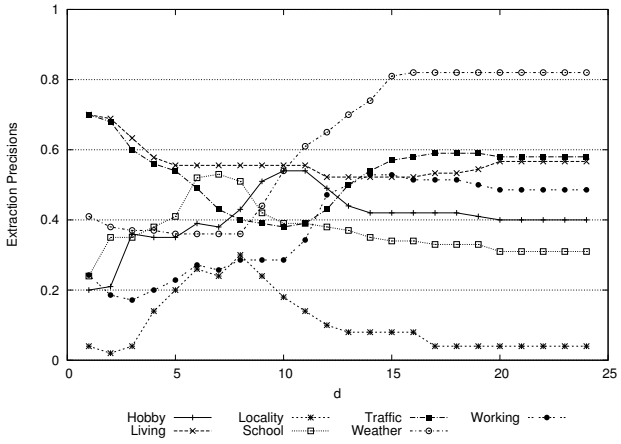

図 5　地域 [沖縄] トピックの時の抽出精度：趣味ー労働

表 6　$\hat{Ra}$ の高いトピックの関連度と，$std(\hat{Rt}(A,t))$ の順位（地域 [つくば] トピックのとき）

	$\hat{Ra}1$ 位		$\hat{Ra}2$ 位		$\hat{Ra}3$ 位		$\hat{Ra}4$ 位		$\hat{Ra}5$ 位	
局面	トピック	順位	トピック	順位	トピック	順位	トピック	順位	トピック	順位
服飾	topic290(0.0668)	91	topic134(0.0936)	8	topic382(0.0350)	96	topic55(0.0347)	97	topic136(0.0312)	93
交流	topic23(0.0404)	102	topic302(0.0270)	95	topic384(0.0254)	79	topic484(0.0198)	110	topic342(0.0178)	104
災害	topic144(0.1502)	1	topic465(0.0884)	198	topic280(0.0825)	135	topic320(0.0717)	207	topic311(0.0510)	196
食事	topic297(0.1014)	42	topic383(0.0970)	40	topic479(0.0760)	45	topic388(0.0628)	34	topic290(0.0374)	91
行事	topic320(0.0533)	207	topic465(0.0525)	198	topic490(0.0413)	73	topic89(0.0346)	60	topic467(0.0313)	63
消費	topic290(0.1416)	91	topic382(0.0528)	96	topic55(0.0521)	97	topic136(0.0468)	93	topic313(0.0391)	100
健康	topic23(0.0410)	102	topic302(0.0269)	95	topic297(0.0245)	42	topic214(0.0229)	313	topic497(0.0219)	217
趣味	topic315(0.0521)	92	topic150(0.0513)	89	topic407(0.0449)	47	topic290(0.0447)	91	topic305(0.0422)	90
居住	topic418(0.1571)	11	topic173(0.0593)	10	topic290(0.0251)	91	topic97(0.0159)	123	topic411(0.0135)	20
地域	topic465(0.1025)	198	topic320(0.0867)	207	topic311(0.0560)	196	topic280(0.0415)	135	topic354(0.0195)	180
学校	topic198(0.0648)	24	topic216(0.0210)	13	topic318(0.0205)	26	topic477(0.0186)	87	topic405(0.0184)	2
交通	topic220(0.0781)	3	topic465(0.0770)	198	topic320(0.0637)	207	topic316(0.0476)	16	topic311(0.0428)	196
気象	topic411(0.0842)	20	topic162(0.0556)	80	topic175(0.0506)	15	topic465(0.0495)	198	topic320(0.0405)	207
労働	topic341(0.0892)	61	topic484(0.0467)	110	topic165(0.0444)	267	topic465(0.0284)	198	topic320(0.0254)	207
非実	topic79(0.0440)	129	topic290(0.0329)	91	topic102(0.0251)	146	topic419(0.0179)	139	topic334(0.0160)	260

ピックで，これらのトピックが上位に関連付けられていることが原因で，行事の抽出精度が他の地域に比べて低くなっていると考えられる．地域の抽出精度が低い原因も，同様の理由であると考えられる．

服飾で $\hat{Ra}1$ 位の topic290 は，消費の局面にも 1 位で関連付けられている．しかし，服飾は $\hat{Ra}2$ 位に $std(\hat{Rt}(A,t))$ の順位が高いトピックがある．災害や交通では，$std(\hat{Rt}(A,t))$ が高い順位のトピックが $\hat{Ra}1$ 位に関連付けられていたため，精度が高かったと考えられる．以上のことから，服飾は局面を表す典型的なトピックに高い関連度を持っているが，topic290 のような複数の局面と高い関連度を持つトピックが原因で，精度が高くなっていないと考えられる．改善手法として，複数の局面に高い関連度を持つトピックは，局面との対応関係構築の際に除くことが考えられる．

5.2 地域依存性に関する考察

表 5 より，学校と地域の局面で，地域 [つくば] の精度が高いことが分かる．このように，精度が大きく異なる原因を分析するため，局面に関連付けられている各地域のトピック内の単語について評価する．各局面に高い関連度を持つ上位 3 トピックについて，生起確率が高い上位 20 単語の集合から，Jaccard 係数を算出する．Jaccard 係数は二つの集合の積を，二つの集合の和で除すことによって求められる．Jaccard 係数が 1 のとき，二つの集合の要素は完全に一致しており，Jaccard 係数が 0 のとき，二つの集合の要素は一つも一致していないことを意味する．各地域間の Jaccard 係数を算出した結果を，図 6 に示す．横軸は各局面，縦軸は Jaccard 係数である．地域 [つくば] と他の地域では，地域の局面で Jaccard 係数が 0 になっていることが分かる．災害，趣味，学校，交通も Jaccard 係数が低くなっていることが分かる．服飾，食事，気象の局面では Jaccard 係数が高くなっていることが分かる．

表 5 と図 6 より，災害，交通は抽出精度は高いが，Jaccard 係数が低いという特徴がある．この二つの局面について，最も関連度の高いトピック内の生起確率が高い単語を，表 7 に示す．災害は，4 つの地域で地震に関連する単語が出現している．特に地域 [札幌] では,「青森」や「群馬」など地名も出現していることが分かる．交通は 4 つの地域で「乗る」「降りる」などの動詞が出現している．地域 [つくば] では「秋葉原」，札幌では「新琴似」等の駅名．地域 [沖縄] では「バス」や「モノレール」など，沖縄でよく使われる交通手段が単語として出現している．交通と災害は，地域特有の単語が出現しており，地域依存性は存在するものの，関連度が最も高いトピックによって適切に吸収できていると言える．以上のことから，全ての地域で抽出精度が高かったと考えられる．

表 5 と図 6 より，学校と地域の局面は，地域 [つくば] の精度が他の地域に比べて高いが，Jaccard 係数が低いという特徴がある．この二つの局面について，最も関連度の高いトピック内の生起確率が高い単語を，表 8 に示す．地域の局面は，地域 [つくば] に「茨城」や「つくば」といった地域特有の単語が出現しているが，他の地域では地域特有の単語は出現していない．

学校の局面は，地域 [つくば] と地域 [京都] では「研究」や

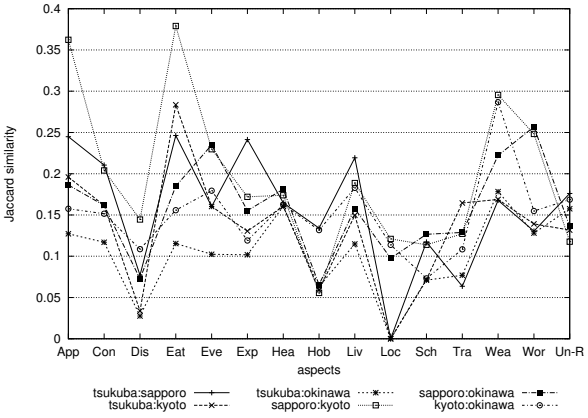

図 6 関連度が高いトピックにおける内容語の Jaccard 係数

「学会」といった研究関連の単語，地域 [札幌] では「レポート」や「課題」など講義に関連する単語が出現している．地域 [沖縄] は，学校に関連する単語がトピック内の生起確率上位に含まれていなかった．学校の局面が付与された Tweet に，研究関連の内容が言及された Tweet が多かったことが原因で，地域 [札幌] では学校関連の単語がトピック内に出現するにも関わらず，精度が低くなったと考えられる．また，地域 [沖縄] は学校に関連する単語が生起確率上位に含まれるようなトピックが存在しないことが，精度を低下させる原因となったと考えられる．

地域の局面が付与された Tweet は，つくばや茨城特有の単語が多かったため，地域 [つくば] トピックでは地域特有の単語を含むトピックに高い関連度を持つことができたと考えられる．他の地域のトピックでは，つくばや茨城に関するトピックが存在しなかったため，地域の局面の精度が低くなったと考えられる．以上のことから，学校と地域の局面は地域依存性が高い局面であるといえる．

6. 結　　論

本論文では，実生活 Tweet を抽出するための二段階抽出法を提案した．第一段階では，LDA を用いて大量の Tweet から多数のトピックを抽出する．第二段階では，少数のラベル付き Tweet を用いて．トピックと局面の対応関係を構築する．精度を向上させるため，特徴語の重みは情報利得によって計算する．

評価実験の結果，未知の Tweet から局面を抽出できることが分かった．抽出精度が高い局面は，その局面を表す典型的な一つのトピックと高い関連度を持っているか，少数のトピックによってその局面を適切に表していることが分かった．抽出精度が低い局面は，複数の局面と高い関連度を持っているトピックが，その局面の関連度の高いトピックとして関連付けられている特徴があることが分かった．Jaccard 係数を用いた分析によって，災害や交通は地域依存性が存在するが，トピックによって吸収できていることが分かった，地域や学校の局面は地域依存性が高いことが明らかになった．

今後の課題は，局面とトピックの関連度を算出する手法を改善し，複数の局面に強い関連度を持つトピックへの対処があげられる．また，実生活 Tweet の季節依存性についても評価す

表 7　災害，交通と最も高い関連度で繋がるトピックの単語

	つくば	札幌	京都	沖縄
災害	最大，震源，予想	支庁，地方，群馬	揺れる，風邪，震度	パンダ，予防，子供
	速報，震度，発生	揺れる，南部，震度	引く，地震，治る	感染，揺れる，震度
	茨城，地震，揺れる	地震，青森，北部	大事，大丈夫，ひく	地震，仙台，原因
交通	新幹線，秋葉原，通勤	終着，北海道，下り	遅れる，列車，乗る	乗れる，バス，飛行機
	乗る，列車，降りる	あいの里，石狩，新琴似	特急，乗車，降りる	乗る，着く，モノレール
	電車，常磐，快速	上り，札幌，電車	阪急，電車，快速	降りる，空港，電車

表 8　地域，学校と最も高い関連度で繋がるトピックの単語

	つくば	札幌	京都	沖縄
地域	美術館，都内，秋葉原	行く，参加，久々	リアルタイム，闘士，見る	お出かけ，つくる，抜ける
	展示，エクスプレス，茨城	買い物，募金，中山	神話，見れる，久しぶり	充実，明日，用事
	市民，つくば，博物館	クリック，今日，昨日	瞬間，思い出す，黄金	夕飯，久しぶり，帰宅
学校	計画，学園，発表	早寝，提出，レポート	閉まる，論文，研究	お出かけ，つくる，抜ける
	論文，研究，学会	早起き，実験，学校	学会，講演，開く	充実，明日，用事
	成果，テーマ，分野	終わる，ゼミ，課題	技術，科学，会議	夕飯，久しぶり，帰宅

ることを考えている．

謝　　辞

本研究の一部は，筑波大学図書館情報メディア系プロジェクト研究による助成を受けたものである．ここに記して謝意を示す．

文　　献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Proceedings of WWW 2010*, pp. 450–453, 2010.

[3] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.

[4] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Science*, Vol. 101, pp. 5228–5235, 2004.

[5] Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. Experience mining: Building a large-scale database of personal experiences and opinions from web documents. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[6] Takeshi Kurashima, Taro Tezuka, and Katumi Tanaka. Extracting and geographically mapping visitor experiences from urban blogs. *WISE 2005*, pp. 496–503, 2005.

[7] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. *In Proceedings of the 2010 International Conference on Management of Data*, pp. 1155–1158, 2010.

[8] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. *Proceedings of ICWSM 2010*, pp. 130–137, 2010.

[9] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. *In Proceedings of 18th International World Wide Web Conference (WWW2010)*, pp. 851–860, 2010.

[10] Masahito Yamamoto, Hiroya Ogasawara, Ikuo Suzuki, and Masashi Furukawa. Tourism informatics:9. information propagation network for 2012 tohoku earthquake and tsunami on twitter. *IPSJ Magazine*, Vol. 53(11), pp. 1184–1191, 2012 (in Japanese).

[11] Shuhei Yamamoto and Tetsuji Satoh. Real life information extraction method from twitter. *The 4th Forum on Data Engineering and Information Management(DEIM2012) F3-4*, 2012 (in Japanese).

[12] Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee Peng LIM, and Xiaoming Li. Topical key phrase extraction from twitter. *The 49th Annual Meeting of the Association for Computational Linguistics*, pp. 379–388, 2011.