Vol.2015-IFAT-117 No.2
2015/2/6

地域特有の単語共起にもとづく位置推定と地域トピックの考察

石田和成[†1]

マイクロブログにおける地域特有の単語および単語共起にもとづき，情報発信者の位置を推定する手法を提案し，地域トピックの考察を行う．スマートフォンの普及にともない，ソーシャルメディアのデータ量は増大を続けている．しかし，位置情報の付加されたデータが全体に占める割合は少ない．特に，大都市圏と比べ，地方都市の位置情報は非常に少ない．地方都市における地域トピックの考察には，情報発信者の大まかな位置推定にもとづく情報抽出が必要となる．そのため，位置情報付きデータに含まれる単語について，出現頻度，緯度経度の平均，標準偏差を求め，単語や単語共起の地域特定スコアを定義する．このスコアにもとづき，情報発信者の位置推定，および地域トピックの考察を行う．

Estimation of User Location and Discussion of Local Topics Based on Area Specific Term Co-occurrence

KAZUNARI ISHIDA[†1]

This paper proposes an estimation method of user location and discusses local topics, based on area specific term co-occurrence. The portion of geo-tagged information on social media has been quite small yet; however, the amount of information on social media has been increasing explosively due to widespread use of smartphone. Especially, The amount of information generated from regional cities is quite smaller than that of metropolitan cities. Hence, we have to estimate coerce user location to get enough estimated and geo-tagged information and to discuss local topics in a regional city. To get the information, we define area specific scores of terms and co-occurrences that are calculated with term frequency, average and standard deviation of longitude and latitude of raw geo-tagged information.

1. はじめに

マイクロブログにおける地域特有の単語共起にもとづき，情報発信者の位置推定を行い，地域トピックの考察を行う．スマートフォンの普及は，マイクロブログのデータ量急増をもたらした．しかし，ソーシャルメディアにおいて，位置情報の付加されたデータが占める割合は非常に少ない．地方都市は大都市と比較し，その傾向が顕著である．そのため，地方都市の地域トピックをソーシャルメディアから得られる情報から直接考察することは困難である．

各地域のトピックを抽出するために，本研究では情報発信者の大まかな位置推定手法を提案する．そのため，位置情報付きデータに含まれる単語について，出現頻度，緯度経度の平均，標準偏差を求め，単語の地域特定スコアを定義する．これら単語および単語共起にもとづき，情報発信者の位置推定を行う．単語共起は，多義的な単語の意味を表現するデータ構造として効果的であり，知識の体系化[1]やコミュニティの抽出[2], スパムの同定[3]などに用いられる．

位置推定された情報発信者の発信する情報にもとづき，各地域の単語頻度時系列を抽出し，地域の話題を考察する．地域区分は結果の解釈が容易な，都道府県，市町村を基準とし，話題考察時に粒度を選定する．ここで，マイクロブログのデータ量は膨大であるため，位置推定や時系列抽出に要する領域計算量は膨大となる傾向がある．そのため，大規模データセット処理に Hadoop を用いることにより，位置推定および単語時系列抽出を行い，地域ごとのトピックを考察する．

†1 広島工業大学
Hiroshima Institute of Technology

2. 関連研究

マイクロブログにおける位置推定手法として以下のような研究がある．Dalvi ら[4]は，空間的なモデルを用いたオブジェクトとツイートのマッチングを行うために，ユーザとオブジェクトの距離のモデル, 言語モデル(ユニグラム，バイグラム）を定義し，EM アルゴリズムによる学習を行った．位置の定まったオブジェクトとしてレストランを選び, Yahoo ローカルの 2009 年 12 月から 2011 年 1 月までのデータから，750,000 のレストランを抽出し，ツイートの位置推定を行った．これに対し，本研究では，オブジェクト（話題）の位置は地域に固定されないものとして取り扱う．

Bo ら[5]は，地域特定語を用いてテキスト分類問題にもとづくツイートの位置予測を行った．地域区分の方法として行政区分を用い，人口の少ない地域は，隣接する人口の多い地域と統合することにより，地域間の情報格差に対処する．この地域区分にもとづき，地域特定語の決定するため，語を，(1)ローカルワード（１地域に属する），(2)セミ

ローカルワード（n 地域に属する），(3)コモンワードの 3 種類に分類した．語の特徴量として，単語頻度と地域頻度に加え，情報利得を用いた．この研究では，地域の範囲について，モデル構築時に地域統合の制約を加えているが，本研究では，分析時に地域の粒度を選択できる手法を提案する．

Cheng ら[6]は，地域特定のキーワードにもとづくユーザ位置の推定アルゴリズムを提案した．地域特定キーワードを選定するために，Backstrom ら[7]が提案した，語の地理的な集中と散らばりのモデルを用い，ユーザ位置推定を行った．この研究では，単語の地域性を用いているが，本研究では, 地域特有の単語共起も合わせて用いることにより，位置推定精度を改善する．

Ishida[8]は総務省統計局の定める地域メッシュにもとづき位置推定を行った．Roller ら[9]は，言語モデルと地域区分として適応的グリッドを用いた位置推定を行った．それに対し，本研究では，得られた結果の解釈が容易な，行政地域ごとの地域区分を用いる．

3. 位置推定

位置情報付きマイクロブログから，単語毎に緯度経度を集計し，地域特有の単語を特定する．そのため，位置情報付きツイートから，ツイートにおける名詞を単語として抽出する．また，単語毎に緯度経度の平均，標準偏差を求める．これら統計量にもとづき，単語の地域固有スコアを定義する．以下の手順で位置推定と推定精度の評価を行う．

1. 位置情報付きツイートをデータセットから抽出
2. 位置情報付きツイートを発信したユーザ（ジオユーザ）を抽出
3. ツイートから名詞を単語として抽出
4. 単語の緯度経度の平均，標準偏差を求める
5. 単語（単語共起）の地域固有得点データベースを構築
6. ジオユーザの全ツイートをデータセットから抽出
7. ジオユーザの全ツイートと，単語（単語共起）の地域固有得点データベースにもとづき，ジオユーザの位置推定を行う
8. ジオユーザの推定位置と実際の存在位置を比較し，推定精度の評価を行う

用いるデータセットは，Twitter Streaming API を用いて，2011 年 3 月から 2014 年 5 月までに収集したツイートである．データセットにおけるツイート数は 347,742,872，単語数は 4,124,568,983，単語の種類は 58,994,705，ユーザ数は 17,251,905 である. また, 位置情報付きツイートは 1,132,580 と全ツイートの 0.33%，位置情報付きツイートを発信したユーザ数は 311,812 と全ユーザの 1.8%である．

3.1 手法 1：単語を用いた位置推定

る．ここで，予備的な実験にもとづき，扱う単語は名詞のみとした．表 1 に抽出された位置情報付き単語の統計量を示す．ここで得られた単語の平均緯度経度について，国土地理院のデータにもとづき作成した，緯度経度と住所のデータベースを用いて, 単語と住所の対応関係を抽出する．さらに，単語の地域特定スコアを定義する（式 1）．

表 1　単語頻度と出現緯度経度

単語	頻度	平均経度	平均緯度	経度標準偏差	緯度標準偏差
東京	34875	139.52	35.67	1.12	0.66
京都	8257	135.86	35.03	0.87	0.44
新宿	6951	139.67	35.69	0.48	0.24
地震	7997	138.74	36.51	3.63	2.75
津波	230	138.64	36.35	3.90	3.32
. . .	. . .	. . .	. . .	. . .	. . .

$$Score = tf \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots (1)$$

ここで，各単語についての位置情報付き単語の頻度(***tf***)，経度の標準偏差(***sx***)，緯度の標準偏差(***sy***)を用いている．この定義により，地理的分散が小さく出現頻度の単語は，地域を特定する単語として高いスコアを得る．このスコアにもとづき，全ツイートに含まれる単語を用いて，ジオユーザの位置推定を行う．各ジオユーザについて，ツイートから抽出した単語に対応する住所のスコアを加算する．これをこのユーザの全単語について行うことにより，ユーザの推定住所のランキングが得られる．このランキングでトップの地域をユーザの推定位置とする．

3.2 手法 2：地理的散らばり，頻度を制限した位置推定

手法 1 では，地域特定スコアを定義し，単語と地域の関連の強さを計算することにより，単語を用いたユーザの位置推定手法を定義した．ただし，この手法では，出現頻度が非常に高い単語の場合，緯度経度の散らばりが大きい場合でも，比較的高いスコアが得られる可能性がある．
そのため第 2 の方法では，地理的散らばりと単語頻度に閾値を設け，位置推定に用いる単語を制限し，地域特定スコアを用いる．これらの閾値の設定により，出現頻度が高く緯度経度の散らばりの大きい単語による，位置推定精度の低下を防ぐ．予備的実験にもとづき，単語出現頻度の上限については 50000，緯度経度の標準偏差の上限については 2.0 を用いることとした．

3.3 手法 3：単語共起を用いた位置推定

手法 2 では，地理的散らばりや出現頻度の閾値にもとづく地域特定スコアを用いた，ユーザの位置推定手法を定義した．しかし，通常，単語は多義的で，複数の意味を持つものが多いため，異なる意味で用いられている同一表記の単語が，位置推定精度を悪化させる可能性がある．そのため，第 3 の手法では，単語共起を用いた位置推定手法を定義する．単語共起における 2 つの単語のうち，一方の単語のみ，地理的散らばりや出現頻度の閾値を用い，位置推定に用いる単語共起を制限する．

双方の単語に制限を課す場合，有効な単語共起が得られる確率が非常に低く，位置推定に利用できる十分な，単語共起と住所の対応関係が得られない．また，双方の単語とも制限無しとした場合，位置推定にとって有用な情報を持たない単語共起が多数含まれるため，位置推定精度の低下や，計算量の爆発といった問題が生じる．

そこで一方の単語のみに閾値を設定した単語共起について，方法 1 でのべた単語のスコアと同様に，単語共起にもとづく地域特定スコアを定義する．この単語共起は，両方の単語が閾値の制約を満たす場合もある．ここで 1 つのツイートにおける単語共起は，含まれる単語すべてのペアである．ツイートの文字列の最大は１４０文字と非常に短いため，同一ツイート内にある単語共起には有意な意味があると考える．表 2 に単語共起の統計量を示す．

表 2　単語共起頻度と出現緯度経度

単語 1	単語 2	住所 1	緯度平均	経度平均	緯度標準偏差	経度標準偏差
津波	茨城	茨城県	36.50	140.62	0.48	0.30
千葉	津波	千葉県	35.65	140.31	0.24	0.43
岩手	津波	岩手県	39.32	141.76	0.43	0.16
宮城	津波	宮城県	38.43	141.31	0.99	0.62
津波	高萩	茨城県	36.72	140.71	0.00	0.00
一関	津波	岩手県	38.98	141.64	0.17	0.12
津波	若葉	千葉県	35.62	140.18	0.00	0.00

3.4 位置推定精度の評価

単語や単語共起にもとづく位置推定結果について，ツイートに付与された実際の位置情報にもとづき評価を行う．ユーザの位置推定結果は，地域特定単語や単語共起による，位置スコアの合計にもとづき，推定された地域が順位付けされる．そのうち，一番得点の高い地域をユーザの推定地域とする．この推定地域と，実際にユーザが滞在した地域との距離にもとづき，位置推定結果を評価する．この評価方法にもとづき，3.1，3.2，3.3 でそれぞれ定義した，単語による位置推定（手法 1），制限付き単語による位置推定（手法 2），制限付き単語共起による位置推定（手法 3）を比較する．

図 1 は，推定された位置と実際の位置との誤差について，ユーザの度数分布の推移を表す．手法 1 （Method 1，単語を用いた位置推定）では，誤差 250～300km のユーザ頻度が高い．手法 2（Method 2，単語の地理的分散，頻度を制限した位置推定）では，誤差 50~100km のユーザ頻度が高い．位置推定に用いる単語に制限を加えることにより，誤った位置推定が低減していることがわかる．手法 3（Method 3，単語共起を用いた位置推定）では，誤差 50km 以下のユーザ頻度が高い．地域特定単語の代わりに，地域特定単語共起を用いることにより，位置推定の精度が向上することがわかる．

図 2 は，誤差距離についての累積ユーザの割合を示す．8 割のユーザが含まれる誤差の許容範囲について，手法 1 は 350km，手法 2 は 300km，手法 3 は 100km である．

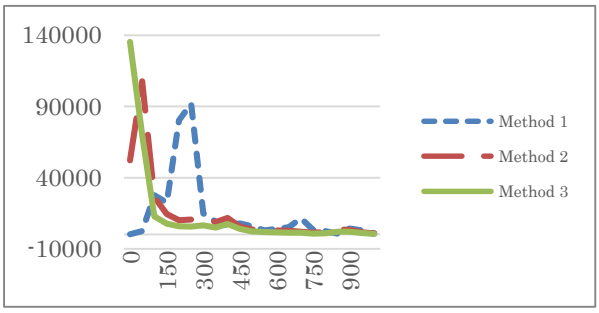

図 1　推定誤差とユーザ数

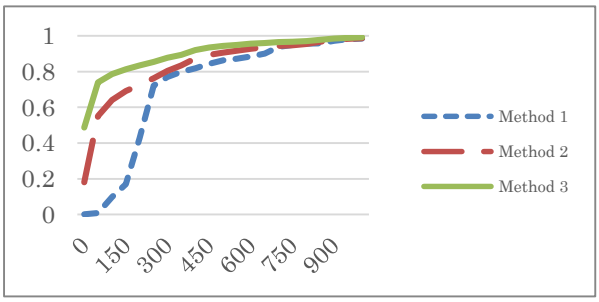

図 2　推定誤差と累積ユーザ割合

4. ユーザ位置推定にもとづく地域別単語時系列抽出手法

手法 3 で構築した単語共起と位置の対応データベースにもとづき，データセット内の全ユーザの位置を推定し，ユーザのつぶやきを地域毎に集計することにより，地域別話題の推定を行う．手法 3 を全データに適用して位置推定できたユーザ数は，約 8 千万(8,575,766)であった．これは全ユーザ数の約 50%に達する．データセットにおいて位置情報付きツイートを発信したユーザ数は，約 30 万(311,812)であるため，約 28 倍の地域特定ユーザが得られたこととなる．

また，データ処理手順を以下に示すとおり，膨大なデー

タから単語共起を抽出するため，データ処理の一部に Hadoop を用いた．処理手順を以下に示す．

1. 各ユーザのツイートのグループ化 (Hadoop 利用)
2. 各ユーザの単語共起抽出
3. 単語共起による各ユーザの位置推定
4. 位置推定ユーザによる各地域の単語時系列の抽出
5. 各地域，各時間帯における単語の集計 (Hadoop 利用)

Twitter Streaming API を用いて収集したデータセットにおいては，ツイートは発信時刻順に並んでいる．このデータを直接ステップ 2（各ユーザの単語共起抽出）で処理すると，全データを主記憶上に読み込んだ後，各ユーザのツイートから単語共起を抽出する必要がある．あるいは，全データについて，各ツイートを読み込みながら単語共起を抽出し，全データ読み込み完了まで，各ユーザの単語共起を主記憶上に保持する必要がある．先に述べたとおり，全ユーザ数は 17,251,905(約 1 千 7 百万)，ツイート数は 347,742,872(約 3 億 4 千 8 百万), 単語の種類は 58,994,705(約 5 千 9 百万)である．そのため，主記憶上に全ユーザの全ツイートを保持する，または全ユーザの単語共起を保持することはできない．

しかし，データセットのツイートを，ユーザ毎に連続した行にグループ化することにより, ステップ 2 の処理では，連続する個別ユーザのツイートの処理完了時に，各ユーザの単語共起を補助記憶上のファイルに出力できるため，主記憶上に保持する必要が無い．

データセットにおけるツイートを，ユーザ毎のグループ化するには，Hadoop の Key-Value ペアにおいて，Key をユーザ ID，Value をツイートとした出力により，ユーザ ID でソート，グループ化されたファイルを得ることができる．このファイルをステップ 2 で処理することにより，処理結果を主記憶上に保持し続けることなく，補助記憶上に膨大な各ユーザの単語共起を出力することができる．

ただし，Hadoop のクラスター規模や，各ノードの主記憶量によっては，データセット一度に処理することができない．そのような場合は，データセットを複数のサブセットに分割し，ステップ 1 の処理を行う必要がある．今回は主記憶装置が 32G バイトの計算機 1 台において，疑似分散モードの Hadoop を用いた．そのため，今回は 6 か月毎に分けたサブセットをそれぞれ Hadoop でソート，グループ化し，それぞれの出力に対し，ステップ 2 を実行し，その結果をユーザ毎に統合した．

ステップ 4 における，各地域の単語時系列抽出においては, ステップ 3 で得られた各ユーザの推定位置にもとづき，ステップ 1 で得られたユーザごとにグループ化されたツイートを，地域ごとに分類し，単語とその発信時刻を抽出する．ステップ 5 の単語集計においても Hadoop を用いる．ステップ 4 で得られた地域ごとの発信時刻付単語データを，関係データベースで集計することもできるが，レコード数が莫大なため，処理時間が非常に長くなるため，Hadoop を用いることとした．

4.1 位置推定による地域情報の増大と地理的区分選定

位置情報付きツイートを発信しているジオユーザと，データセットから位置推定されたユーザについて，発信された単語の種類，頻度について比較する．データセットにおいて位置情報付きツイートを発信したユーザのツイートのみを用いた単語の集計では，約 5 百万(5,032,683)種類の単語が抽出された．各単語の平均頻度は 30.45359197 であった．これに対して，位置推定ユーザのツイートを用いた場合, 平均頻度は 613.5067108 と, 20 倍以上の頻度であった．さらに，全位置推定ユーザのツイートのみに現れた単語数は 49,309,065 で，単語の種類は 10 倍程度拡大した．このように，位置推定ユーザの利用により，得られる地域情報は飛躍的に増大したことが分かる．

ここで，位置推定により得られた位置情報にもとづき地域トピックを考察するために, 地域分割の粒度を決定する．そのため，位置推定ユーザを実際の人口の分布と比較し，位置推定の粒度を選定する．都道府県ごとの位置推定ユーザ数と，総務省の市区町村別の人口及び世帯数[a]について，都道府県ごとに集計した値との相関を求めたところ，男性 0.807，女性 0.799，全員 0.803，世帯 0.823 が得られた．同様に，市区町村ごとの位置推定ユーザ数と，総務省の市区町村別の人口及び世帯数との相関を求めたところ，男性 0.350，女性 0.345，全員 0.348，世帯 0.381 が得られた．都道府県ごとのユーザ数と人口との相関関係と比べ，市町村ごとのユーザ数と人口との相関関係はかなり低いことが分かる．そのため，今回得られた位置情報にもとづく地域別話題の地理的区分として，都道府県ごとの区分を用いることとする．

4.2 位置推定ユーザを用いた重み付き単語抽出

ユーザの位置は推定値であるため，完全に正確であるとは限らない. そのため, ユーザの推定位置の精度を定義し，その精度を単語頻度の重みづけとして用いることにより，精度の低いユーザの影響を低減する．

具体的には，各ユーザについて住所ランキングにある地域のスコアを重みとして緯度経度の重み付き平均と標準偏差を求める．ここで得られた平均緯度経度の地点を推定位置として用いる. また, 緯度経度の標準偏差を $\boldsymbol{sx}$, $\boldsymbol{sy}$ とし，住所ランキングの平均スコアを $\boldsymbol{scave}$ としたとき，ユーザ推定位置での重み(***Weight***)を式(2)で定義する．この重みにもとづき，ユーザ推定位置での重み付き単語頻度(***WTF***)を

a http://www.soumu.go.jp/menu_news/s-news/17216_1.html

式(3)で定義する．

$$Weight = scave \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots (2)$$

$$\mathrm{W}TF = \left(1 - \exp\left(-\left(1 + Weight\right)\right)\right) \quad \cdots (3)$$

これらの定式化により，位置推定の平均スコアが高く，標準偏差が小さい場合，単語頻度の重みは重くなり，逆に，スコアが低く標準偏差が大きい場合は軽くなる．つまり，位置推定の精度の低いユーザについては，単語頻度を割り引いて集計することとする．***WTF*** を用いて地域ごとに抽出した単語の時系列にもとづき話題の考察を行う．

具体的な地域としてここでは，広島県について，単語の集計を行い，頻度順にランキングしたところ，上位には 2 位「広島」(2,034,632)，27 位「呉」(173,125)といった地名が多く得られた．その他キーワードとして，51 位「刺身」(80,569)，53 位「お好み焼き」(79,038)が得られた．ここでは，地域の名物である「お好み焼き」の地域ランキングや単語の時系列推移を 4.3 で，他の地域の名物である「たこ焼き」について 4.4 で示す．さらに，近年各地域で頻発に発生している自然災害に関連した「雪」「津波」について，4.5，4.6 でそれぞれ考察する．

4.3 地域別話題：お好み焼き

地域別話題として「お好み焼き」の地域ランキングトップ 5 を表 3 に示す．総務省統計局の平成 18 年度事業所・企業統計調査bによると，お好み焼き屋店舗数第一位は広島県である．また，ランキングトップの「広島県」における単語頻度の推移を図 3 に示す．ジオタグデータの場合(g)は，散発的に単語が出現しているのに対し，推定データの場合(a)は，継続的に単語が出現していることが分かる．

表 3：地域ランキング「お好み焼き」

順位	地域	頻度
1	広島県	79038
2	岡山県	47387
3	兵庫県	24474
4	島根県	22697
5	大阪府	21812

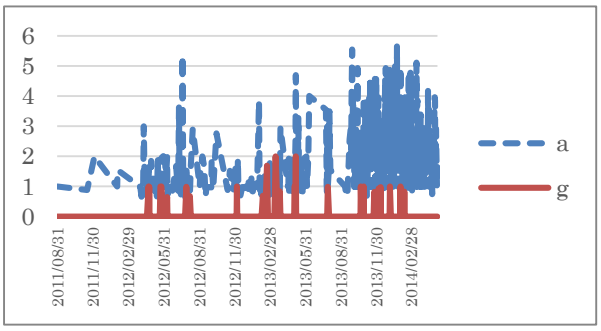

図 3：単語頻度の推移：「広島県」の「お好み焼き」

4.4 地域別話題：たこ焼き

地域別話題として「たこ焼き」の地域ランキングトップ 5 を表 4 に示す．NTT タウンページによる「たこ焼き店/お好み焼き店」に関する調査結果cによると，たこ焼き屋店舗数第一位は大阪府である．また，ランキングトップの「大阪府」における単語頻度の推移を図 4 に示す．ジオタグデータの場合(g)は，散発的に単語が出現しているのに対し，推定データの場合(a)は，継続的に単語が出現していることが分かる．

表 4：地域ランキング「たこ焼き」

順位	地域	頻度
1	大阪府	32216
2	京都府	28439
3	徳島県	23925
4	滋賀県	22526
5	兵庫県	21302

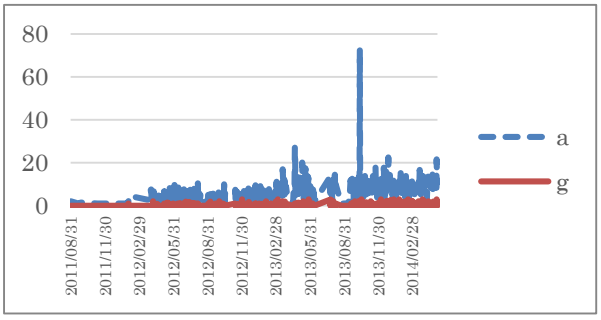

図 4：単語頻度の推移：「大阪府」の「たこ焼き」

4.5 地域別話題：雪

地域別話題として「雪」の地域ランキングトップ 5 を表 5 に示す．また，ランキングトップの「北海道」における単語頻度の推移を図 5 に示す．気象庁による直近の 24 時間降雪量dによると，北海道地域がランキング上位にあること

b http://www.data.jma.go.jp/obd/stats/data/mdrr/rank_daily/data00.html#snf24h, http://todo-ran.com/t/kiji/13448

c http://tpdb.jp/townpage/order?nid=TP01&gid=TP01&scrid=TPDB_GY01, http://news.mynavi.jp/news/2014/01/29/328/

d http://www.data.jma.go.jp/obd/stats/data/mdrr/rank_daily/data00.html#snf24h

がわかる．ジオタグデータのみ(g)と比較し，推定データ(a)では単語頻度が多く，推移の変化をより詳細に把握できる．

表 5：地域ランキング「雪」

順位	地域	頻度
1	北海道	429000
2	青森県	413661
3	秋田県	407565
4	山形県	317954
5	岩手県	313065

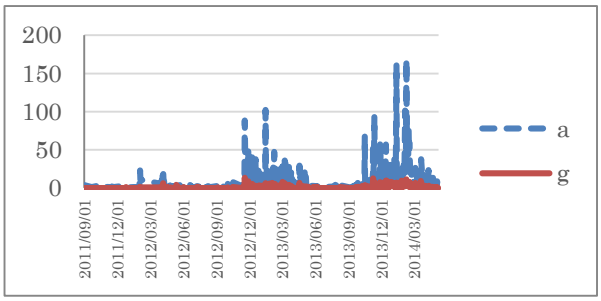

図 5：単語頻度の推移：「北海道」の「雪」

4.6 地域別話題：津波

地域別話題として「津波」の地域ランキングトップ 5 を表 6 に示す．また，ランキングトップの「宮城県」における単語頻度の推移を図 6 に示す．推定データ(a)において，2012 年 12 月 7 日に突出した頻度増大があることが分かる．

表 6：地域ランキング「津波」

順位	地域	頻度
1	宮城県	155011
2	福島県	97773
3	茨城県	66514
4	栃木県	63260
5	和歌山県	54506

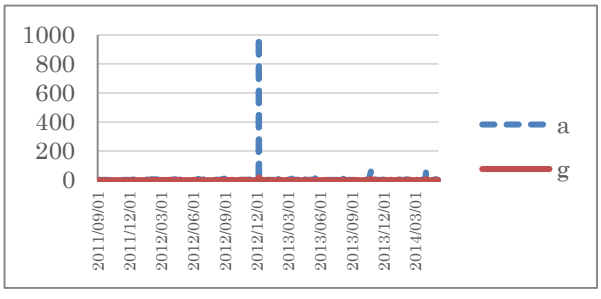

図 6：単語頻度の推移：「宮城県」の「津波」

消防庁災害対策本部により発表された三陸沖を震源とする地震（確定報）eによると，地震の発生日時は，平成 24 年 12 月 7 日 17 時 18 分頃，震源は三陸沖，規模はマグニチュード 7.4 であった．宮城県は津波警報が 7 日 17 時 22 分に発令された．また，気象庁の発表fによると石巻市鮎川で 7 日 18:03 に最大 98cm の津波が観測された．

5. おわりに

マイクロブログにおける地域固有表現にもとづき，情報発信者の位置を推定する手法を提案した．そのため，位置情報付きデータに含まれる単語について，出現頻度，緯度経度の平均，標準偏差を求め，単語の地域特定スコアを定義した．これら単語や単語共起にもとづき，情報発信者の位置推定および話題の考察を行った．

参考文献

1) Ishida, K. and Ohta T., “An approach for organizing knowledge according to terminology and representing it visually,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. 32, No. 4, pp. 366-373, 2002.
2) Ishida, K., “Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graphs,” Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem - Aggregation, Analysis and Dynamics in the 14th International World Wide Web Conference (WWW2005), Makuhari Messe, Chiba, Japan, May 10 - 14, 2005.
3) Ishida, K, "Extracting Spam Blogs with Co-citation Clusters," Proc. Of the 17th International World Wide Web Conference (WWW2008), April 21 - 25, 2008.
4) Dalvi N., Kumar R., and Pang B., "Object Matching in Tweets with Spatial Models," WSDM’12, February 8-12, 2012, Seattle, Washington, USA.
5) Bo H., Cook P., and Baldwin T., "Geolocation Prediction in Social Media Data by Finding Location Indicative Words," Proceedings of COLING 2012: Technical Papers, pages 1045-1062, COLING 2012, Mumbai, December 2012
6) Cheng Z., Caverlee J., and Lee K., "A Content-Driven Framework for Geolocating Microblog Users," ACM Transactions on Intelligent Systems and Technology, Vol. 4, No. 1, Article 2, Publication date: January 2013.
7) Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. (2008). Spatial variation in search engine queries. In Proceeding of the 17th international conference on World Wide Web, WWW ’08, pages 357-366, Beijing, China. ACM.
8) Ishida K., "Extracting Geo-Social Information based on Geo-Tagged Social Media," 4th World Congress on Social Simulation (WCSS 2012), National Chengchi University, Taipei, Taiwan, September 4-7, 2012 .
9) Roller S., Speriosu M., Rallapalli S., and Wing R., Jason Baldridge, "Supervised Text-based Geolocation Using Language Models on an Adaptive Grid," Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1500-1510, Jeju Island, Korea, 12-14 July 2012.

e http://www.fdma.go.jp/bn/%E4%B8%89%E9%99%B8%E6%B2%96%E3%82%92%E9%9C%87%E6%BA%90%E3%81%A8%E3%81%99%E3%82%8B%E5%9C%B0%E9%9C%87%28%E7%A2%BA%E5%AE%9A%E5%A0%B1%EF%BC%89.pdf

fhttp://www.jma.go.jp/jma/press/1301/10a/1212tohoku.pdf