Vol.2013-NL-214 No.17
2013/11/15

発言同期を用いたマイクロブログ著者の位置推定

高橋 哲朗[1,a)]

概要：twitter に代表されるマイクロブログは世の中の動向を知るためのセンサとしての活用が可能であるが，発言位置を特定できるユーザは一部でしかないため地理的な情報を用いた活用が難しいという課題がある．そこで本研究ではより網羅的にユーザの位置情報を推定するために，「雨」のような局所性のあるイベントに対する発言の同期を集計することにより，発言位置を特定できないユーザの位置情報を推定する手法を提案する．実験の結果，約 41% の精度で都道府県単位の位置を推定できることを確認した．また位置情報が既知であるユーザの発言を元に「雨」以外に局所的な発言の同期が起きる語を抽出する手法を提案し，実験の結果，雨などの自然状況や花火などのイベントを含む複数のカテゴリの語が得られることを確認した．

1. はじめに

近年 twitter や facebook などの SNS が盛んに使われており，多くのユーザの間で情報発信が身近なものとなってきている．それに伴いこれらのサービスに投稿されるデータの利用も進んできている．特に twitter はデータへのアクセスのための API も充実している．このデータを知識源として利用する研究も多数行なわれるようになっており，たとえば地震の検知 [1]，花粉の飛散 [2] インフルエンザの広がり [3] に関する研究が報告されている．

twitter への投稿はモバイル端末から行なわれることが多く，ユーザの位置情報を組み合わせることによって価値の高いサービスを構築できる．たとえば上述した応用以外にも，洪水やゲリラ豪雨などの自然現象や，渋滞や電車遅延などの交通関連の事象がどのような地域で起きているかをリアルタイムに把握することが可能となり twitter をより有効に活用できる．このような活用はすでに「ツイートフル[*1]」のような実サービスでも行なわれている．これらのサービスを実用的なものにするためには，それぞれの発言の位置情報が必要となるが，2 節で述べるように，twitter ユーザの位置情報を網羅的に得ることは難しい．そこで本研究では twitter ユーザの位置情報を推定するための一手法を提案する．

[1] 株式会社富士通研究所
[a)] takahashi.tet@jp.fujitsu.com

[*1] http://tweetflu.jp

2. 位置情報の推定方法

2.1 既存手法

twitter ユーザの位置情報を取得するためには，本節で後述するように複数の手法がある．これらの手法にはそれぞれ利用する情報や制限に違いがあり，また位置情報を取得できるユーザの範囲も異なっている．そしてこれらの手法を補完的に用いることにより，位置情報推定の網羅性を高めることができる．本研究で提案する位置推定も，一つの推定方法として補完的に用いられるものである．

2.1.1 GPS によって得られる情報

モバイル端末から tweet[*2]を投稿する場合，GPS によって測定された緯度・経度による位置情報を付与できる．しかし 2012 年 11 月時点で位置情報が付与された投稿の割り合いは全体の 1% 未満であり，災害検知などの応用に使うには不十分である．

2.1.2 プロフィール情報からの推定

twitter ではユーザのプロフィール情報として場所[*3]というフィールドがあり，一部のユーザはこのフィールドを用いて居住地に関連する情報を公開している．このプロフィール情報を用いることで，ユーザのおおまかな位置情報を知ることができる．調査の結果，日本語で tweet しているユーザの約 35% については都道府県単位の居住地を推定できることが分かったが，このフィールドに情報を書いていないユーザの数は多く，また書かれていても居住地を特定できないような文字列 (「夢の中」など) も多い．

[*2] ユーザの投稿した記述は「status」,「状態」,「つぶやき」,「tweet」などの名称で呼ばれるが本稿では「tweet」で統一する．
[*3] 英語版のサービスでは “location”

2.1.3 内容からの推定

twitter のユーザが過去に発信した tweet を分析することにより，ユーザの居住地を推定する技術が提案されている [4], [5], [6]．これらの手法は，ユーザが地名または特定の場所を示す手掛かりとなるランドマークなどの語句を tweet することを前提としている．そのためそのような語句を tweet したユーザの位置情報のみが推定の対象となる．

2.1.4 ネットワークを使った推定

twitter などの SNS では ユーザ間のつながりに関する情報があり，このつながりによって形成されるソーシャルグラフを用いて居住地を推定する方法が提案されている [7], [8]．これらの手法は，ソーシャルメディア上でのつながりがあるユーザは居住地においても近くに住んでいる，という仮定に基いている．一部のつながりにおいてはこの仮定が成り立つと思われるが，すべてのつながりでも成り立つ保証はないため，誤った推定を引き起こす可能性がある．

2.2 発言同期による位置推定の提案

多くの twitter ユーザは，雨や花火などのイベントを観たときにそれに関した投稿を行なっている[*4]．そして発言の対象となったイベントが，雨や花火のように局所性のあるイベントであった場合その発言をした人々は同一の地域にいる可能性が高い．そこで本稿では，雨や花火などの局所性のある特定のイベントに対する同じ時刻帯での投稿を手掛かりにし，位置情報が既知のユーザの位置情報を元に位置情報の分からないユーザの位置情報を推定する手法を提案する．本稿では，同じ時間帯に特定の共通したイベントに対する発言が行なわれることを「発言同期」と呼ぶ．

異なる地域で雨などの同じイベントが同時に起き，異なる地域間で偶然発言同期が起きる可能性はある．しかし複数回の発言同期から得られる場所の候補が蓄積されれば，ユーザの生活圏外に比べ生活圏内の位置で発言同期が起きる確率が高まることが見込まれ，したがって推定位置は生活圏内に集まることが期待できる．

本稿が提案する手法は，tweet の発言位置ではなくユーザの生活圏を推定する．また 2.1 節で述べた手法のうち GPS によって得られる情報以外の手法は，発言をした位置を推定しているのではなく提案手法と同様に生活圏を推定している．そのため，これらの手法によって得られた位置情報を使って厳密な位置の推定はできないが，たとえば都道府県や市町村といったある程度の範囲を持つエリアの単位で活用する場合には有用と言える．

3. 発言同期による位置推定実験

発言同期が起きるイベントとしては「雨」や「花火」，「雷」，「流れ星」などいくつかの候補が考えられたが，本研究では特に記述頻度が高い「雨」を用いて発言同期によるユーザの位置推定実験を行なった．なお評価のためには対象ユーザの位置情報が既知でなければならないため，まず 2.1.2 節で述べたプロフィールを用いた位置情報抽出を行ない位置情報の特定できるユーザ集合を作成した．位置情報の単位は都道府県とした．

3.1 プロフィールによる位置推定対象データ

twitter 社は全 tweet からランダムにサンプリングされたデータを公開している．これらのデータは StreamingAPI という API により取得できる．今回の実験ではユーザの位置情報推定のために，2010 年 4 月から 2012 年 2 月までに収集した約 3 年分のサンプリングデータを用いた．この期間に抽出された tweet 数とユーザの数は 表 1 の通りである．これらのユーザに対して，プロフィールからの位置推定を行なった．

表 1 StreamingAPI により得られたサンプリング tweet

tweet 数	216,332,178
user 数	11,015,728

プロフィールの場所情報は個々のユーザが自然言語で書いたものであるため，その記述から都道府県名を推定する必要がある．この推定のために，国土地理院が公開している日本の地名[*5] のうち，市区町村，主要な地域，山，丘，川，湖を用い，地名と都道府県名との対応表を作成した．twitter の場所情報には「琵琶湖の近く」のような表現が用いられることもしばしばあるため行政区画以外の名前も用いた．これらの地名の延べ数は 3,196 であり，さらにこれをひらがなとローマ字に展開すると，延べ数は 11,584 になる．場所情報は「さっぽろ」や「Kanagawa, Japan」のようにひらがなやアルファベットで書かれることが少なくないためこの展開は必要であった．この 11,584 表現の中には異なり数で 664 の重複表現があったため，それらの重複した地名については 「都道府県名 & 地名」というクエリによる Web 検索のヒット件数を用い，もっとも多くのヒット件数を得た都道府県名をその地名に対応付けた．

プロフィールの位置情報には「現在は東京に住んでいます」のように地名以外も含まれることがあるため，プロフィールの位置情報と対応表とのマッチングには文字列長で正規化した bigram の類似度をスコアとして用い，対応表と一定の閾値以上のスコアでマッチングできた位置情報を採用した．閾値はサンプリングにより得られた 1,000 件の開発データを用いて求めた．

プロフィールからの位置情報の推定結果を 表 2 に示す．タイプの詳細はそれぞれ以下である．

[*4] 我々の予備調査では，「雨」という表現を含む tweet を行なったことがあるユーザは約 40% に上る (詳細は 3.2 節).

[*5] http://www.gsi.go.jp/kihonjohochousa/gazetteer.html

表 2 プロフィールによる位置情報推定結果

タイプ	数	割合 (%)
1) 空欄	4,772,071	43.3
2) 推定不可	2,106,881	19.1
3) 推定可	3,256,788	29.6
4) 複数/同一	588,562	5.3
5) 複数/変更	291,426	2.7
計	11,015,728	100.0

表 3 プロフィールによる位置推定結果例

項番	正誤	場所情報	推定地
1	正	Tokyo	東京都
2	正	横須賀の浦賀のあたり	神奈川県
3	正	日本のダウンタウン、尼崎市	兵庫県
4	正	23 区の多摩川沿い	東京都
5	正	奈良和歌山 県境をうろうろ	奈良県
6	正	長野 東京	東京都
7	誤	カーテンからの日光が優しい季節	山口県
8	誤	@ Hiroshima	香川県

1) 空欄　プロフィールの位置情報が空欄であった

2) 推定不可　期間中に書かれたプロフィール情報からは位置情報を閾値以上のスコアで推定できなかった

3) 推定可　期間中に位置情報の変更が無く，位置情報を閾値以上のスコアで推定できた

4) 複数/同一　期間中に位置情報の変更があったが，そこから推定される都道府県名は同一であった

5) 複数/変更　期間中に位置情報の変更があり，そこから推定される都道府県名が異なっていた

約 3 年間の収集期間の中では多くのユーザにおいて複数の tweet が得られたが，各 tweet から得られる場所情報が途中で変更されていた場合はそれぞれの場所情報毎に都道府県の推定を行なった．その結果，複数の場所情報があり，いずれの場所情報からも都道府県を推定できなかったユーザは **2) 推定不可**に分類した．またすべての場所情報から同一の都道府県を推定できた場合は **4) 複数/同一**に分類した．たとえば「神奈川県横浜市」から「神奈川県川崎」への変更や，「横浜」から「Yokohama」への変更などがその例である．推定した都道府県が異なる場合は **5) 複数/変更**に分類した．

以降の実験では，場所情報から都道府県を一意に推定できた，表 2 中の **3) 推定可**と **4) 複数/同一**のユーザ計 3,845,350 人を対象とする．

この都道府県名の推定は自動で行なっているため，その結果には誤りが含まれている可能性がある．そこで対象ユーザ 3,845,350 人の中から 1,000 人をサンプリングし都道府県推定の精度を人手で評価した．結果を 表 3 に示す．項番 2, 3 の事例からは bigram によるマッチングが，また項番 4 の事例からは行政区画以外の地名がうまく働いていることを示している．項番 5,6 のように，複数の都道府県が書かれている場合もあったが，今回の評価ではいずれかの都道府県を推定できていれば正解とした．項番 7,8 は誤りの例である．項番 7 は bigram マッチングの副作用が原因であり，項番 8 は「香川県丸亀市広島町」とマッチしたことが原因である．行政区画に優先度を持たせるなどの拡張の余地が残っている．

総合的に見ると，1,000 事例のうち推定を誤った事例数は 23 事例のみであり，97.7% の精度で正しく推定できることを確認できたため，以降の実験ではこの結果をこのまま用いた．

本研究で提案する雨による位置推定だけが目的であれば雨について発言したユーザのみの位置を推定をすれば良いが，プロフィールの場所情報から推定できる割り合いを確認するために，また多数の開発データを用いた推定精度の向上のために，サンプリングによって得られたすべてのユーザの都道府県推定を行なった．

3.2 雨に関する tweet の収集

雨に関する tweet は 2012 年 8 月から 2013 年 2 月までの範囲で収集した．収集には 2 つの方法を使った．一つは 3.1 節で述べた StreamingAPI により得られるサンプリングされた tweet のうち，2012 年 8 月から 2013 年 2 月の範囲で「雨」という表現が含まれる tweet を抽出した．もう一つは SearchAPI を使って「雨」という表現が含まれる tweet を収集した．SearchAPI も twitter 社が提供している API であり，検索クエリなどの検索条件に該当する tweet を取得できる．これらをマージした結果，tweet 数およびユーザ数は 表 4 のようになった．以降の処理ではこのマージ後のデータを用いる．

雨について発言したユーザの数は 4,410,011 であり，表 1 に示したサンプリングにより得られたユーザ数 11,015,728 のうちの 40.0% となっている．約 3 年間のサンプリングで全ユーザアカウントを収集できていたと仮定すると，日本語で tweet する全ユーザのうち 40.0% が「雨」について発言したことがあると言える．

表 4 雨発言の tweet(2012 年 8〜2013 年 2 月)

	SearchAPI	StreamingAPI	マージ後
tweet 数	28,933,737	356,776	29,021,033
user 数	4,402,056	294,040	4,410,011

各ユーザ毎の発言数のばらつきは図 1 に示すように非常に大きかった．発言回数が 10 回未満のユーザ数は 3,680,654 人であり全体の 83.5% になる．多くのユーザが少ない数の雨発言しかしていない中，非常に多くの雨発言をしているユーザアカウントもあった．図 1 は 1,200 回までの発言についてプロットしているが，雨発言の回数の多い上位 5 位までのアカウントの発言数は表 5 に示す値で

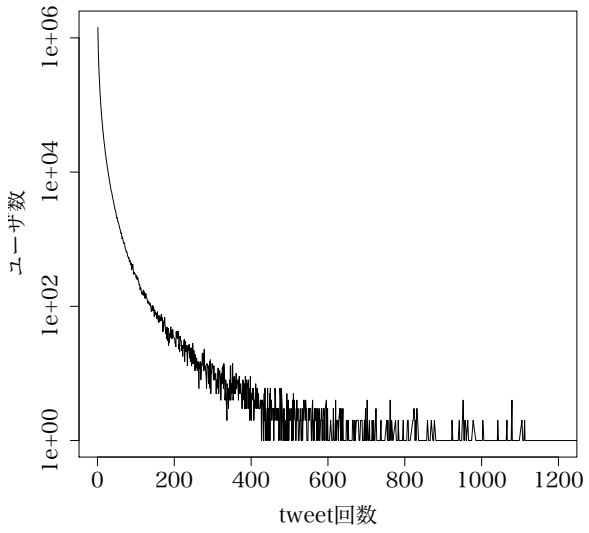

図 1 発言回数ごとのユーザ分布

あった．この数は 212 日間の発言数としては非常に多いが，tweet の内容を確認したところ，これらは機械的に天気予報などの情報を投稿する bot と呼ばれるアカウントによるものだった．このようなアカウントの発言についての対応は 3.4.3 節で後述する．

表 5 雨発言の回数の上位 5 位

順位	期間中の tweet 数
1	51,627
2	11,901
3	10,969
4	9,295
5	8,746

今回の発言同期による位置推定の実験には，位置情報が分かっており，かつ雨について 1 回以上発言したことがあるユーザが必要となる．すなわち，表 2 中の **3) 推定可** と **4) 複数/同一**のユーザ 3,845,350 人と雨についての発言のあった 4,410,011 人の積集合が対象となるが，その数は 1,607,257 人だった．この人数のユーザが以降で報告する実験の対象である．

3.3 位置情報推定アルゴリズム

発言同期を用いたユーザの位置情報推定アルゴリズムを以下に示す．

位置情報が未知である `userA` の各雨発言において，その雨発言と同じ時間帯における位置情報が既知であるすべてのユーザの位置情報を `userA` の位置候補として投票し，最も多くの投票を受けた位置を `userA` の位置と推定する．疑似コードによるアルゴリズムを 図 2 に示す．

なお 図 2 中の各変数および関数は以下とする．

`userA.rainTweet(timestamp)`
: `timestamp` において `userA` に雨発言があったときに `true` を返す関数

`knownUserList`
: 位置情報が既知であるユーザのリスト変数

`pref[userB]`
: 位置情報が既知であるユーザ (e.g. `userB`) の位置 (都道府県) を持つハッシュ変数

`locationCand[prefCandidate]`
: 各位置 (都道府県) の推定値を記録するハッシュ変数

`weight[pref]`
: 各位置 (都道府県) のスコアに対する重みを保持したハッシュ変数

```
Beginnig of code
function getLocation(userA) {
  timestamp = beginDate;
  locationCand = Hash::new()
  while timestamp <= endDate {
    if userA.rainTweet(timestamp) == true {
      foreach userB in knownUserList {
        if userB.rainTweet(timestamp) == true {
          prefCandidate = pref[userB];
          locationCand[prefCandidate] += weight[prefCandidate];
        }
      }
    }
    timestamp += timeRange;
  }

  return locationCand.keys.max;
}
End of code
```

図 2 疑似コードによるアルゴリズム

本研究では 2012/08/01 から 2013/02/28 の 212 日間の tweet を用いたので, 図 2 中の `beginDate` には 2012/08/01, `endDate` には 2013/02/28 を用いた．

発言同期の起きる時間範囲を指定する `timeRange` は 1 時間とした．2 時間以上の時間範囲でも実験を行なったが，1 時間のときに最も良い結果が得られたためである．

人口密度の違いなどにより，都道府県毎に雨発言の行なわれやすさに差がある．この差を吸収するために，212 日間における都道府県毎の雨発言の頻度をあらかじめ算出しておき，その逆数を各都道府県に対する重みとして推定候補地 (都道府県) のスコアに用いることで正規化を行なった．これにより雨発言の多い地域の発言数は抑えられ，雨発言の少ない地域と同等に比較できるようになる．

3.4 実験結果

本実験では，実験対象としたユーザのうち 90% のユーザの位置情報を既知であるとし，その位置情報を用いて残りの 10% のユーザの位置推定を行ない，推定精度を確認した．

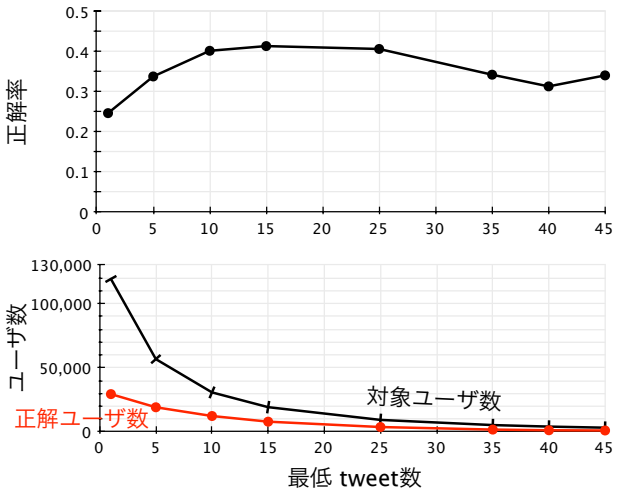

図 3 最低 tweet 数に対する位置推定精度

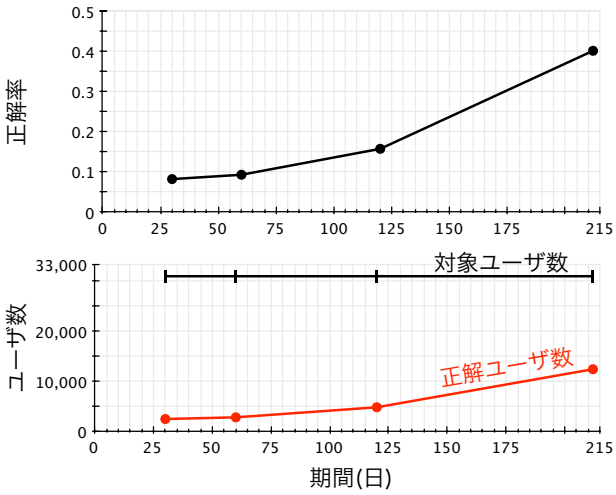

図 4 期間に対する位置推定精度

なお，以降の節では雨発言数，期間，類似度の閾値のそれぞれのパラメータを変化させて精度の変化を見ているが，それ以外のパラメータには 表 6 の値を用いた．

表 6 パラメータの既定値

雨発言数	10 回
期間	212 日
類似度の閾値	0.7

3.4.1 雨発言数による推定精度の変化

本研究の提案手法では，雨発言の数が多くなればなるほど発言同期が起きる回数も増え，推定精度が高くなることが期待できる．そこで雨発言の回数に閾値を設け，その閾値以上の雨発言があったユーザのみを対象として発言同期による位置推定を行なった．

結果を 図 3 に示す．雨発言の範囲 1 から 45 の範囲においては，閾値が 15 の時に最も精度が高く 0.414 であり，また閾値が 1 の時に最も精度が低く 0.247 であった．雨発言の回数の閾値が 15 回までは推定精度が上がっているが，それ以降は落ちている．これは，閾値が大きくなると共に対象ユーザ数が減るため，推定に用いている位置情報が既知のユーザの数も減ってしまうことが原因であると考えられる．

47 都道府県をランダムに回答した場合 1/47 の確率で正解することになるので精度は 0.021 となる．また，雨発言を行なうユーザ数が最も多かった都道府県は東京であり，全体の雨発言に占める東京ユーザの雨発言の割り合いは 0.189 であった．そのためすべてのユーザの位置 (都道府県) を「東京」と推定した場合の精度も同様に 0.189 となる．これらのベースラインと比較したとき，閾値 15 の時の最高精度および閾値 1 の時の最低精度のどちらも上回っており，この結果から提案手法の発言同期による位置推定が機能していると言える．

3.4.2 期間による推定精度の変化

本実験では最大で 212 日間のデータを使用できたが，この日数を少なくしたときの推定精度の変化を見た結果を図 4 に示す．

このグラフから明らかなように，期間が増えるにつれて正解率が上がってきており，212 日の時点でもまだ精度は飽和していない．したがってさらに長い期間のデータを使うことでより正確に推定できることが期待できる．

3.4.3 類似度の閾値による推定精度の変化

今回対象とした twitter アカウントの中には各地の天気情報を機械的に tweet する bot アカウントもあった．そのようなアカウントの tweet は，位置推定を行なう際にはノイズとなることが予想されたので，これらの発言を除く処理を行なった．予備調査を行なったところ，bot アカウントから生成される tweet のうち，特に天気に関する tweet は tweet 間の類似度が高いという特徴を持っていた．そこでアカウント毎に tweet 間の類似度を測定し，類似度が指定した閾値よりも低いユーザのみを用いる実験を行なった．

ユーザ毎の記事の類似度の計算には，編集距離を用いた類似度を定義し用いた．$tweetA$ と $tweetB$ の間の類似度 $sim(tweetA, tweetB)$ の式は以下である．

$$sim(tweetA, tweetB) = \frac{length - ed(tweetA, tweetB)}{length}$$

ここで $ed(tweetA, tweetB)$ は 2 つの tweet の編集距離であり，$length$ は 2 つの tweet の文字数の平均値である．

アカウント毎に最大 100 件の記事をサンプリングし，それらのうちのすべてのペアにおいて類似度を計算し，この類似度が高い順に半数を選択し，それらの類似度の平均値を求め，各アカウントの類似度とした．

類似度の閾値を変化させながら位置情報の推定を行なった結果を図 5 に示す．類似度の閾値が低いほど，つまり類似した tweet ばかりを発言するアカウントを排除するほど精度が高いという結果が得られた．この結果は予想してい

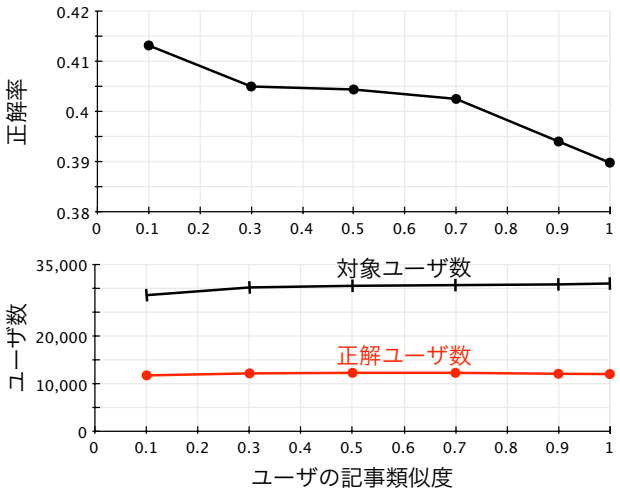

図 5 ユーザの記事類似度に対する位置推定精度

た通りではあるが，図 5 が示すようにその差は大きくはなかった．

4. 発言同期の起きる表現の獲得

3 節の実験では「雨」という表現のみを用いて位置の推定実験を行なった．しかし 3.2 節でも示したように，twitter ユーザの中で「雨」を発言したことのあるユーザは 40.0% であり，残りの 60.0% のユーザについては雨発言による位置推定が行なえない．そのため，より網羅的に位置を推定するためには，「雨」以外の表現も使って発言同期をとらえ，位置推定に用いなければならない．

発言同期が起きる表現を獲得するためには，すでに位置が分かっているユーザの発言を用いる方法が考えられる．ある地域では多く発言されるが他の地域ではあまり発言されない表現があれば，それは地域的局所性を表わす表現であり，そのため発言同期による位置推定に有効であると言える．

この仮説のもと，発言同期の起きる表現 (以降，同期単語と呼ぶ) の抽出を試みた．本節ではこの結果を報告する．

4.1 発言同期の起こりやすさの指標化

位置情報が既知であるユーザ群が発言した tweet を用いる．これらの tweet の本文から形態素解析により形態素列を抽出し，品詞により選択した内容語のみを候補単語とした．そして各地域毎および各時間帯毎に，その地域および時間帯に tweet された候補単語のそれぞれにスコアを付与した．地域 a において 時刻 t から $t+r$ の間での単語 w のスコア $S(a,t,r,w)$ は 式 (1) で求める．

$$S(a,t,r,w) = \frac{R(a,t,r,w)}{R(!a,t,r,w)} \quad (1)$$

ここで $R(a,t,r,w)$，$R(!a,t,r,w)$ はそれぞれ次の式で求める．

$$R(a,t,r,w) = \frac{F(a,t,r,w)}{N(a,t,r)} \quad (2)$$

$$R(!a,t,r,w) = \frac{F(!a,t,r,w)}{N(!a,t,r)} \quad (3)$$

ここで用いている各関数および変数は以下の通り．

$F(a,t,r,w)$	地域 a において 時刻 t から $t+r$ の間に単語 w を発言した人数
$F(!a,t,r,w)$	地域 a 以外において 時刻 t から $t+r$ の間に単語 w を発言した人数
$N(a,t,r)$	地域 a において 時刻 t から $t+r$ の間に発言した人数
$N(!a,t,r)$	地域 a 以外において 時刻 t から $t+r$ の間に発言した人数

式 (1) は，ある時間帯において，他の地域に対して対象の地域での出現確率の比を表わしている．たとえば以下の式は神奈川県において 2012 年 8 月 1 日 12 時から 13 時の間に「雨」が tweet される割り合いは，他の 46 都道府県と比較して 5 倍だったということを意味する．

$$S(\text{“神奈川”}, 2012080112, 1, \text{“雨”}) = 5$$

4.2 抽出結果

3.1 節で説明した，約 3 年分の tweet データを用いて，同期単語の抽出を行なった．地域 a の区分には都道府県を用い，時間間隔 r は 1 時間とした．また，不適切な語が抽出されるのを防ぐために $F(a,t,r,w)$ が 10 以上となる単語だけを対象とした．つまり今回の実験の設定では 1 つの地域で 1 時間のうちに，10 人以上が発言した単語のみを用いたことになる．

この設定で 式 (1) で定義したスコアが 5 以上の単語を抽出した結果，表 7 に示す 177 の単語が得られた．

表 7 同期単語の抽出結果

カテゴリ	数	例
自然	23	雷，雪，地震，雨，虹，吹雪，台風，初雪
イベント	12	花火，魂祭，開港，七夕，祇園祭
災害	10	停電，人身，警報，冠水，勧告，断水
場所	49	福岡，札幌，名古屋，横浜，御堂筋，淀川
鉄道会社	5	相鉄，京王，西鉄，名鉄，京浜東北
その他	78	位，送料，無料，レポ，秒，容量，電気
計	177	

自然に関する単語や，イベント，災害に関する単語は，同期単語として想定していたものであり，本研究で用いた「雨」も抽出されていた．

場所や鉄道会社などの単語も多く出現した．これらの単語も発言同期には有効であるとは考えられるが，これらには単語自体に局所性があるため，2.1.3 節で紹介した tweet 中のランドマークを手掛かりにユーザの位置情報を推定す

る手法がより直接的である．

その他の単語について誤り分析を行なった結果，複数のアカウントを使って，同じ時間帯にまったく同一の tweet が投稿された場合に，そこに含まれる単語を同期単語として抽出してしまうケースが多かった．この問題に対しては，同一時間帯に投稿された tweet における相互の類似度が高いユーザアカウントをあらかじめ見付けておき，それらのユーザアカウントの tweet は処理に用いないようにするなどの対策が考えられる．

5. まとめ

本稿では「雨」などのイベントに関する発言の同期を使って位置の推定を行なう手法を提案し，実験によりその推定精度を評価した．その結果，約 41% の精度で雨発言を行なったユーザの都道府県単位の位置を推定できることを確認した．また，推定においてパラメータとなった雨発言の回数の最低値や対象データの期間，類似度の閾値についても，推定精度に対するそれぞれの影響を整理した．

2 節でも述べたようにユーザの位置を推定するためのさまざまな手法が提案されており，これらを組み合わせることによってより網羅性の高い twitter の活用が可能になる．本研究もその手法の一つを提案したという位置付けになる．

今回用いた推定手法では雨について発言したユーザの位置情報しか推定できず，その割り合いは 40.0% であるので，雨について発言をしていない 60.0% のユーザの位置を推定することはできない．しかし推定した位置を活用するアプリケーションがたとえば雨発言を用いた洪水の検知であれば，雨を発言するユーザだけが位置推定の対象となる．網羅的なユーザの位置情報の推定は難しいが，この例のように，アプリケーションを限定することで現実的な網羅性を獲得できる可能性はある．

4 節では発言同期の起きる表現の獲得も行なった．ここで獲得された「雨」以外の表現を用いることでさらに新しいユーザの位置推定を行なえる．すなわち位置推定と，同期単語の抽出を相互に繰り返すことにより，両方の知識をある程度増やせることが考えられ，今後の課題の一つととらえている．

本稿で提案した位置推定アルゴリズムでは，発言同期により求めた位置の候補の中から，重みを考慮した上でもっとも値の大きい候補地を推定値として採用した．より正確に推定するためには，候補地がどれだけ一箇所に集中しているかという局所性の考慮が考えられる．今回の実験では位置の単位を都道府県としたため局所性の判定が難しかったが，推定位置の単位としてメッシュを用いることで局所性を容易に判定できるようになる．これも今後の課題としたい．

本研究では雨に対する tweet を用いたが，雨自体センサによる観測も行なわれているので，センサで得られる情報と発言を組み合わせてユーザの位置を推定する方法も考えられる．たとえば XRAIN[*6] では時間・空間・雨量それぞれおいて詳細なデータを観測している．このようなデータを用いることにより，上述したメッシュ状の位置推定も可能になる．

参考文献

[1] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, *WWW 2010: Proceedings of the 19th international conference on World wide web*, ACM, pp. 851–860 (online), DOI: http://doi.acm.org/10.1145/1772690.1772777 (2010).
[2] Takahashi, T. and Noda, Y.: Can Twitter be an alternative of Real-World Sensors?, *IEICE technical report. Natural language understanding and models of communication*, Vol. 110, No. 400, pp. 43–48 (online), available from ⟨http://ci.nii.ac.jp/naid/110008676785/en/⟩ (2011).
[3] Achrekar, H., Gandhe, A., Lazarus, R., Ssu-Hsin Yu and Liu, B.: Predicting Flu Trends using Twitter data, *The First International Workshop on Cyber-Physical Networking Systems* (2011).
[4] Cheng, Z., J.Caverlee and Lee, K.: You are where you tweet: A content-based approach to geo-locationg Twitter users, *CIKM* (2010).
[5] Kinsella, S., Murdock, V., N.O'Hare：I'm Eating a Sandwich in Glasgow: Modeling Locations with Tweets, *SMUC* (2011).
[6] Ikawa, Y., Miki, M. and Tatsubori, M.: Location Inference using Microblogging Messages, *WWW* (2012).
[7] Backstrom, L., Sun, E. and Marlow, C.: Find me if you can: Improving geographical prediction with social and spatial proximity, *WWW 2010: Proceeding of the 19th international conference on World Wide Web*, New York, NY, USA, ACM, ACM (2010).
[8] Davis Jr., C. A., Pappa, G. L., de Oliveira, D. R. R. and de L. Arcanjo, F.: Inferring the Location of Twitter Messages Based on User Relationships, *Transactions in GIS*, Vol. 15, No. 6, pp. 735–751 (2011).

[*6] 国土交通省 X バンド MP レーダネットワーク